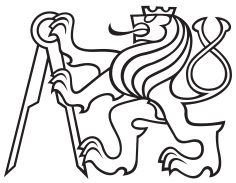
Master's thesis



Czech Technical University in Prague

F3

Faculty of Electrical Engineering
Department of Circuit Theory

Evaluating Digital Profiles of Patients in the Field of Addiction

Bc. Evgenii Grigorev

Supervisor: doc. Ing. Daniel Novák, Ph.D., ing. David Kolečkář
May 2024

MASTER‘S THESIS ASSIGNMENT

I. Personal and study details

Student's name: **Grigorev Evgenii** Personal ID number: **492106**

Faculty / Institute: **Faculty of Electrical Engineering**

Department / Institute: **Department of Circuit Theory**

Study program: **Medical Electronics and Bioinformatics**

Specialisation: **Signal processing**

II. Master’s thesis details

Master’s thesis title in English:

Evaluating Digital Profiles of Patients in the Field of Addiction

Master’s thesis title in Czech:

Analýza digitálních profilů pacientů v adiktologické doméně

Guidelines:

Given the widespread use of mobile technology, which mirrors people's real-life experiences, it can be leveraged to develop accurate and individualized disease phenotypes and markers for diagnosis, monitoring, and treatment.
1. Develop a comprehensive understanding of phenotype analysis and individualized healthcare, with a focus on advanced techniques such as topological data analysis, generalized linear models, temporal multivariate longitudinal data clustering, latent variable models, and semi-supervised deep clustering.
2. Conduct exploratory analysis and implement data visualization methods.
3. Evaluate the effectiveness of the therapy using the methodology of henotype analysis.
4. Carry out statistical validation of the results using augmentation techniques such as bootstrap or permutation tests.

Bibliography / sources:

[1]Torous, Kiang MV, Lorme J, Onnela JP. JMIR Ment Health. 2016;3(2)
[2] Onnela JP, Rauch SL. H Neuropsychopharmacology, 41(7):1691-6. 2016
[3] Benoit J, Onyeaka H, Keshavan M, Torous J. Harv Rev Psychiatry. 2020 4. Bilgen Esmer, Tijen Sengezer, Funda Aksu, Adem Özkara, Kurtulus Aksu, European Respiratory 2016

Name and workplace of master’s thesis supervisor:

doc. Ing. Daniel Novák, Ph.D. Analysis and Interpretation of Biomedical Data FEE

Name and workplace of second master’s thesis supervisor or consultant:

Ing. David Kolečkář The MAMA AI, Praha

Date of master’s thesis assignment: **31.01.2024** Deadline for master's thesis submission: **24.05.2024**

Assignment valid until: **21.09.2025**

doc. Ing. Daniel Novák, Ph.D. Supervisor’s signature	doc. Ing. Radoslav Bortel, Ph.D. Head of department’s signature	prof. Mgr. Petr Páta, Ph.D. Dean’s signature

III. Assignment receipt

The student acknowledges that the master’s thesis is an individual work. The student must produce his thesis without the assistance of others, with the exception of provided consultations. Within the master’s thesis, the author must state the names of consultants and include a list of references.

. | Date of assignment receipt

Student’s signature

Acknowledgements

I want to thank my supervisors, doc. Ing. Daniel Novák, Ph.D. and ing. David Kolečkář, who helped me develop this thesis, provided ideas on improving it and guided me through this process.
Also, I want to thank Ing. Václav Burda, who helped me understand the structure of the application I was analysing, and Ing. Jiří Anýž, Ph.D. for helping me with distributed computations.
Next, I want to thank my good friend Denis Stashkevich, who helped and supported me during our bachelor's and master's studies, and our Czech teacher, Mgr. Jana Eichlerova, thanks to her lectures, I can study in this university in Czech.
Last but not least, I want to thank my mother for her moral and financial support during all these years.

Declaration

I declare that the presented work was developed independently and that I have listed all sources of information used within it in accordance with the methodical instructions for observing the ethical principles in the preparation of university thesis.

Prague, May 23, 2024

Prohlašuji, že jsem předloženou práci vypracoval samostatně, a že jsem uvedl veškeré použité informační zdroje v souladu s Metodickým pokynem o dodržování etických principu při přípravě vysokoškolských závěrečných prací.

V Praze, 23. května 2024

Abstract

This master thesis is dedicated to the analysis of data from users of an application designed to treat nicotine addiction. We have compiled several machine-learning models for solving classification and regression problems.

First, we constructed a series of classification models to predict the likelihood of successful completion of therapy in individual patients. By examining in detail many of the relevant functions, we tried to identify the most influential parameters that significantly affect the results of treatment.

Second, Our investigation extended to regression models focused on adherence parameters. These models were tuned to predict the degree of adherence of patients to the prescribed treatment. By analyzing adherence models, we sought to uncover valuable insights about the effectiveness of therapy. As part of our research, we also measured similarities between different adherence variables, which shed light on potential correlations and revealed new avenues for personalized treatment approaches.

Keywords: smoking cessation app, machine learning, SVM, Support Vector Machines, Logistical Regression, Random Forest

Supervisor: doc. Ing. Daniel Novák, Ph.D., ing. David Kolečkář

Abstrakt

Tato diplomová práce je věnována analýze dat od uživatelů aplikace navržené pro léčbu nikotinové závislosti. Postavili jsme několik machine-learning modelů pro řešení úloh klasifikace a regrese.

Za prvé jsme sestavili sérii klasifikačních modelů, které mají za cíl předpovědět pravděpodobnost úspěšného dokončení terapie u jednotlivých pacientů. Prozkoumáním mnoha relevantních faktorů jsme se snažili identifikovat nejvlivnější parametry, které významně ovlivňují výsledky léčby.

Za druhé se naše zkoumání rozšířilo na regresní modely zaměřené na parametry dodržování terapie. Tyto modely byly laděny tak, aby předpovídaly míru dodržování předepsané terapie u pacientů. Analýzou vzorů dodržování jsme se snažili odhalit cenné poznatky o účinnosti terapie. Naš výzkum také zahrnoval měření podobnosti mezi různými proměnnými dodržování, což osvětlilo potenciální korelace a odhalilo nové přístupy k personalizované léčbě.

Klíčová slova: smoking cessation app, strojové učení, SVM, Support Vector Machines, Logistická Regrese, Náhodný les

Překlad názvu: Analýza digitálních profilů pacientů v adiktologické doméně

Contents

1 Introduction 3
Motivation . . . 3
Goal . . . 3

2 State of the Art 5
Factors Affecting Smoking Cessation 5
Conclusion . . . 6

3 Application 7
Phases . . . 7
Data . . . 7

4 Methods and Algorithms 9
Train Test Split . . . 9
Metrics . . . 9
Classification . . . 9
Regression . . . 10
Random Forest . . . 10
Classification . . . 10
Regression . . . 10
Logistical Regression . . . 11
SVM . . . 11
Classification . . . 11
Regression . . . 11
Statistical testing . . . 11
ANOVA . . . 11
Kruskal-Wallis . . . 12
Pearson Correlation . . . 12
Spearman Correlation . . . 13

5 Exploratory Analysis 15
Completed therapy . . . 17
Multivariate . . . 18
Lapses . . . 23
Conclusion . . . 24

6 Analysis of the logs 25
Analysis method . . . 25
Screen time . . . 26
Mean screen time . . . 30
Number of sessions . . . 34
Regularity . . . 36
Correlation between the adherence variables . . . 39
Adherence at the first 10 days . . . 40
Conclusion . . . 42

7 Classification 45
Random Forest . . . 45
Logistical Regression . . . 47
SVM . . . 48
Comparison . . . 48
Conclusion . . . 49

8 Regression 51
Definition . . . 51
Random Forest . . . 51
Screen time . . . 51
Mean screen time . . . 52
Number of sessions . . . 52
Regularity . . . 52
SVM . . . 53
Screen time . . . 53
Mean screen time . . . 53
Number of sessions . . . 53
Regularity . . . 54
Comparison of the two methods . . . 54
Screen time . . . 54
Mean screen time . . . 55
Number of sessions . . . 56
Regularity . . . 57
Conclusion . . . 58

9 Conclusion 59

A Bibliography 61

B Unused graphs 63

Figures

5.1 Distribution of age 15
5.2 Distribution of cigarettes per day now 16
5.3 Distribution of quitting attempts count 16
5.4 Distribution of smoking since age regularly 16
5.5 Distribution of smoking time . . . 17
5.6 Distribution of lapse count 17
5.7 Age vs. Smoking age since regularly 18
5.8 Quitting attempts count vs. Smoking age since regularly 19
5.9 Lapse vs. Quitting attempts count 19
5.10 Quitting attempts count vs. Last non-smoking period duration 20
5.11 Lapse count vs. Last non-smoking period duration 20
5.12 Lapse count vs. Quitting attempts count 20
5.13 Lapse count vs. Cigarettes per day before therapy 21
5.14 Quitting attempts count vs. Sex 21
5.15 Cigarettes per day before therapy vs. Sex 22
5.16 Smoking since age regularly vs. Sex 22
5.17 Distribution of lapses by phase 23
5.18 Distribution of lapses by state . 24

6.1 Total time spent 26
6.2 Total time spent histograms 26
6.3 Screen time vs age 27
6.4 Smoking time vs Screen time . . . 27
6.5 Cigarettes per day before therapy vs Screen time 28
6.6 Last non-smoking period duration vs Screen time 28
6.7 Town size vs Screen time 28
6.8 Educational background vs Screen time 29
6.9 Screen time vs completed therapy 29
6.10 Screen time vs Lapse count . . . 30
6.11 Mean time spent per day 31
6.12 Mean time spent per day histograms 31
6.13 Cigarettes per day before therapy vs Mean screen time 32
6.14 Last non-smoking period duration vs Mean screen time 32
6.15 Educational background vs Mean screen time 33
6.16 Mean screen time vs completed therapy 33
6.17 Number of sessions vs Date joined 34
6.18 Number of sessions histograms 34
6.19 Last non-smoking period duration vs Number of sessions 35
6.20 Educational background vs Number of sessions 35
6.21 Number of sessions vs completed therapy 36
6.22 Regularity 36
6.23 Regularity histograms 37
6.24 Cigarettes per day before therapy vs Regularity 37
6.25 Last non-smoking period duration vs Regularity 38
6.26 Educational background vs Regularity 38
6.27 Regularity vs completed therapy 39
6.28 Correlation between the Screen time and the Number of sessions . . 39
6.29 Correlation between the Regularity and the Screen time . . . 40
6.30 Correlation between the Regularity and the Number of sessions 40
6.31 Screen time vs Screen time for the first 10 days 41
6.32 Number of sessions vs Number of sessions for the first 10 days 41
6.33 Regularity vs Regularity for the first 10 days 42

7.1 Feature importances 46
7.2 Feature importance with only 10 days of data 47

8.1 Support Vector regression for the screen time for all data 55

Tables

5.1 Correlation with completed therapy 18
5.2 Correlation with completed therapy for men 22
5.3 Correlation with completed therapy for women 23

6.1 Basic statistics of the Screen time 30
6.2 Basic statistics of the Mean screen time 33
6.3 Basic statistics of the Number of sessions 36
6.4 Basic statistics of the Regularity 39
6.5 Basic statistics of the Screen time in 10 days 42
6.6 Basic statistics of the Number of sessions in 10 days 42
6.7 Basic statistics of the Regularity in 10 days 42

7.1 Classification metrics for the Random Forest 46
7.2 Classification metrics for the trimmed Random Forest 46
7.3 Classification metrics for the trimmed Random Forest for 10 days 47
7.4 Classification metrics for the Logistical Regression 47
7.5 Classification metrics for the Logistical Regression for 10 days . . 48
7.6 Classification metrics for the SVM 48
7.7 Classification metrics for the SVM for 10 days 48
7.8 Classification metrics for all used methods 48
7.9 Classification metrics for all used methods for 10 days 48

8.1 Regression metrics for the RF for the screen time 51
8.2 Regression metrics for the RF for the mean screen time 52
8.3 Regression metrics for the RF for the number of sessions 52
8.4 Regression metrics for the RF for the regularity 52

8.2 Support Vector regression for the mean screen time for all data 56
8.3 Support Vector regression for the number of sessions for all data 57
8.4 Support Vector regression for the regularity for all data 58

B.1 Mean screen time vs age 63
B.2 Smoking time vs Mean screen time 63
B.3 Town size vs Mean screen time 64
B.4 Mean screen time vs Lapse count 64
B.5 Number of sessions vs age 64
B.6 Smoking time vs Number of sessions 65
B.7 Cigarettes per day before therapy vs Number of sessions 65
B.8 Town size vs Number of sessions 66
B.9 Number of sessions vs Lapse count 66
B.10 Regularity vs age 66
B.11 Smoking time vs Regularity . . 67
B.12 Town size vs Regularity 67
B.13 Regularity vs Lapse count 68

8.5 Regression metrics for the SVM for the screen time 53
8.6 Regression metrics for the SVM for the mean screen time 53
8.7 Regression metrics for the SVM for the number of sessions 53
8.8 Regression metrics for the SVM for the regularity 54
8.9 Regression metrics for the RF and SVM for the screen time 54
8.10 Regression metrics for the RF and SVM for the mean screen time . . . 55
8.11 Regression metrics for the RF and SVM for the number of sessions . . 56
8.12 Regression metrics for the RF and SVM for the regularity 57

Chapter 1

Introduction

This work continues David Kolečkář's Master Thesis [1] on the visualisation and analysis of patients' digital phenotypes.

Motivation

Addiction remains a pressing global health concern, affecting individuals, families, and communities worldwide. Despite significant advancements in addiction treatment modalities, the complex nature of addiction demands innovative approaches for effective intervention and support. In this digital age, where technology permeates nearly every aspect of daily life, leveraging digital tools and platforms has emerged as a promising avenue for enhancing addiction treatment outcomes.
Digital technologies offer unprecedented opportunities to revolutionize the delivery of addiction treatment services. Digital profiles, comprising comprehensive collections of patient data gathered from various sources such as wearable devices, mobile applications, electronic health records, and social media platforms, hold immense potential in providing personalized, data-driven interventions tailored to individual needs and circumstances. By harnessing the power of data analytics and machine learning algorithms, healthcare providers can gain deeper insights into patient behaviours, preferences, and treatment responses, thereby optimizing treatment strategies and improving patient outcomes.
In our case, we use the mobile application, the structure of which will be discussed later.

Goal

This work aims to analyse factors that can influence therapy's success, like sex, education level, and previous experience with nicotine cessation, building classification models that will help predict if the patient will finish the therapy and building regression models that will predict users' adherence to the treatment.

Chapter 2

State of the Art

The cessation of smoking poses a complex challenge, influenced by a myriad of factors. This research explores various determinants affecting successful smoking cessation, including demographic, psychological, and behavioural aspects. Understanding these factors is crucial for developing effective smoking cessation interventions.

Factors Affecting Smoking Cessation

There are different kinds of factors that may affect if the patient will successfully quit smoking.
Younger age, female gender, and lower socioeconomic status are identified as potential obstacles to quitting. Heavy smoking, lack of social support, and the external health locus of control contribute to the difficulty in cessation. [2]
Approximately half of the respondents reported success in their first attempt, with 17.9% requiring more than six attempts.
Successful first attempts are associated with factors such as marital status, abrupt cessation, personal beliefs, willpower, and effective thought diversion. Conversely, the use of cessation aids and family promptings are inversely related to success on the first attempt. While initial smoking cessation rates are lower in men, women exhibit challenges in maintaining long-term cessation. Conflicting studies suggest varying impacts of age and gender on smoking cessation rates. Educational levels show inconsistent effects on cessation success. Socioeconomic factors, including education, employment, and socioeconomic status, exhibit conflicting associations with smoking cessation. [3]
Factors influencing smoking cessation include pack consumption and adherence to treatment. Cessation efforts should prioritise reducing packs smoked per year for improved success; these findings may inform future programs, though further studies in diverse populations are needed. [4]
Marital status, childbearing status, household smoking, psychiatric history, and age of starting regular smoking influence cessation periods. Educational level and depression impact cessation duration, with higher education associated with longer periods. [5]
Nicotine Replacement Therapies (NRTs) are identified as the first treatment choice in the absence of contraindications. Behavioural therapy tailored to

individual needs is crucial for addressing relapses. [6]
NRT, bupropion, and varenicline exhibit different mechanisms of action in aiding smoking cessation. The effectiveness of these pharmacological agents is discussed, emphasising the need for tailored interventions. [7]
The USPSTF (United States Preventive Services Task Force) suggests insufficient evidence to assess the balance of benefits and harms of pharmacotherapy interventions in pregnant persons. E-cigarettes are not recommended for tobacco cessation due to concerns about nicotine addiction. [8]
Limited evidence supports the use of smartphone apps as monotherapy for smoking cessation. Combinations of FDA-approved medications and behavioural counselling are cost-effective strategies, enhancing the likelihood of success. [9]
Lifetime tobacco exposure, educational attainment, alcohol drinking status, and birth cohort are identified as potent determinants for smoking cessation success. Additional factors, including marriage, occupational classification, disease morbidity, and secondhand smoke exposure, play roles in the initiation and termination of smoking.[10]
One of the main subjects in our study was adherence; in different studies, it is defined differently: for example, it may be defined as meeting the minimal predefined activity time per week [11], which was not suitable for us because we have no guidelines on it, but it is still fascinating statistics we can look at. Another option is to measure the number of sessions the user has completed of their total engagement time [12]. From [13], we can see the number of logins to the application is the most commonly reported measure of adherence, followed by the number of sessions completed.

Conclusion

Smoking cessation involves numerous factors like demographics, beliefs, behaviours, and medical support. Challenges include age, gender, and socioeconomic status, while factors like willpower and support can aid initial success. Long-term quitting faces obstacles like pack consumption and social influences. Tailored interventions combining therapy and counselling show promise, but ongoing research is crucial, especially in diverse populations like pregnant women. A comprehensive approach, considering individual needs and evidence-based strategies, is essential for lasting success in quitting smoking.
The main definitions of adherence used are the number of completed sessions and total screen time, which we will explore in the next chapters.

Chapter 3

Application

We are using the data we got from the adiquit app. The users manually enter the data we use, and the app generates the data. The user enters their initial information, like how many times they've tried to stop smoking, their sex, how many cigarettes they smoke every day, and others, described below, during the first EE phase. Another part of the data is the data we are getting from the application's internal logs.

Phases

The EE phase is introductory, which gets information from the user and provides them with information on how to work with the application. It consists of 10 sessions that usually happen once per day. The user may also start completing all these sessions in one day, which is, in our terms, called "bujon". This is the only phase of the application that is available for non-paid users.
The EQ phase is the phase when the user has to stop smoking. It usually takes only one day and session, but if they don't stop smoking, they'll have 3 more unique sessions that will try to help them. The 4th session will repeat until the user manages to stop smoking.
The FU phase is the follow-up phase, which helps the user not give up on the cessation. It takes 21 sessions, which happen once daily, to complete.
The WR phase is the weekly rotation phase and maintenance phase; it was introduced later than the other phases, on 28.04.2021. It consists of 70 sessions, which happen once in 3 days.
The FIN phase is our application's last phase, indicating that the user has finished their therapy.
After the user reports they've stopped smoking, after the EQ phase, the application will ask the user once per day, usually in the morning when it sends a notification about the new session or in the evening if the user has not responded in the morning, if they've smoked.

Data

We are using the dataset created in [1], and the procedure will be described below. For more information on its creation and preprocessing, see [1]. Later, we added the new features computed from the application log. Our dataset consists of 1212 patients, or 661 after the filtration.

During the EE phase, the user enters these variables:

- region
- sex
- age
- income
- cigarettes per day before therapy
- list of the nicotine products they've tried
- list of the nicotine products they use at least once per week
- reason for quitting smoking
- quitting attempts count
- town size
- employment type
- educational background
- list of their health conditions
- if they are taking medications regularly
- if they've suffered the COVID-19
- last withdrawal method
- reason for quitting smoking
- last non-smoking period duration

The [1] dataset also uses these variables:

- date joined
- app purchased
- lapse count
- bujon

These data are stored in the PosgreSQL database in multiple tables, which is queried once weekly, and then stored on the local PC as the 'pickled' object. The data we use are stored in the users_user and payments_payment tables. The logs processed are stored in the events_event table.

Chapter 4

Methods and Algorithms

A systematic and multifaceted research methodology was employed to achieve this study's objectives and rigorously evaluate patients' digital profiles. This section outlines the key components of the methodology, including data collection methods, sample selection criteria, data analysis techniques and machine learning methods.

Train Test Split

We are splitting our data into 3 sets:

- Testing set, which contains 20% of all data
- The remaining 80% are split for training and validation set during the parameter search phase using the k-fold method; in our case, k equals 5.
- The training set is used to train the chosen model.
- The validation set is used to measure the performance of the chosen method on data it has not seen.

We chose the best models after validation based on their roc_auc score, or the Area under the ROC (receiver operating characteristic) Curve

Metrics

This work contains two parts that may require some metrics: the classification part and the regression part, which use different kinds of metrics.

Classification

For the classification part, we are using these metrics:

- Accuracy $= (TP + TN)/(TP + TN + FP + FN)$ answer the question: how often the model is correct?

- Precision= $TP/(TP + FP)$ answers the question: how often are the positive predictions correct?
- Recall= $TP/(TP + FN)$ answers the question: can an ML model find all instances of the positive class?

Where TP means the number of True Positives, TN means the number of True Negatives, FP means the number of False Positives, and FN means the number of False Negatives in our classified data.

Regression

For the regression part, we are using these metrics:

- Mean Squared Error (MSE)= $1/N \sum (Y_{itrue} - Y_{ipred})^2$
- Determination coefficient R^2

MSE metrics can not be interpreted so easily; their values depend on our data's magnitudes.
R^2 metrics may have values from -inf to 1, where $R^2 = 1$ means that our model explains all variance in the data. $R^2 = 0$ means that our model explains variance in the data as well as the mean value constant model does. $R^2 < 0$ means our model explains variance worse than the mean value constant model.

Random Forest

Classification

A decision forest classifier works by constructing multiple decision trees during training and combining their predictions through a voting mechanism to classify new data points, with each tree considering different subsets of features and instances to improve overall accuracy and robustness.
We used the random forest classifier to choose the features that may be useful for predicting whether the patient will finish their therapy. We used feature importances metrics based on the Mean Decrease in Impurity in the sklearn package. Feature importances are provided by the fitted attribute feature_importances_ and they are computed as the mean and standard deviation of accumulation of the impurity decrease within each tree.
Sklearn DecisionTreeClassifier takes diffenet parameters as input. We changed the number of trees in any forest (n_estimators parameter) from 50 to 500, and the maximum depth of a tree in the classifier (max_depth parameter) from 1 to 20.

Regression

Decision tree regression works by recursively partitioning the feature space into smaller regions and assigning a constant value (usually the mean or

median) to each region, with splits based on feature values that minimize the variance of the target variable within each partition, thus creating a piece-wise linear approximation of the underlying relationship between the features and the target variable.
We used DecisionTreeRegression from the sklearn package to predict patients' adherence.

Logistical Regression

Logistic regression works by modelling the probability that a given instance belongs to a particular class using a logistic function, which transforms the output of a linear combination of features into a value between 0 and 1, with parameters (coefficients) learned through optimization to minimize the discrepancy between predicted probabilities and actual class labels, making it suitable for binary classification tasks.

SVM

Classification

Support Vector Machine (SVM) for classification works by finding the hyperplane that best separates the data points of different classes in the feature space while maximizing the margin between the nearest data points (support vectors) of each class, with the hyperplane chosen to minimize classification errors and generalize well to unseen data through the use of a kernel function to map the data into a higher-dimensional space, allowing for nonlinear decision boundaries.

Regression

Support Vector Machine (SVM) for regression, often referred to as Support Vector Regression (SVR), works by fitting a hyperplane in the feature space that captures as many data points as possible within a specified margin while also minimizing the deviations (errors) of the data points from the hyperplane, with the width of the margin and the amount of deviation allowed controlled by hyperparameters, thus effectively modelling the relationships between features and target variables continuously, allowing for both linear and nonlinear regression tasks.

Statistical testing

ANOVA

ANOVA helps us dissect the observed variance in a dataset into two distinct components:

Systematic Factors: These factors have a statistical influence on the data. They are the variables we intentionally manipulate or study.
Random Factors: These factors do not significantly impact the data. They represent the inherent variability or noise. ANOVA allows us to investigate how different groups or treatments contribute to the overall variability.
The F-ratio (also known as the F-statistic) is the key output of ANOVA. It quantifies the variability between and within groups. If there's no true variance between the groups (i.e., null hypothesis holds), the F-ratio will be close to 1. A significant F-ratio indicates that at least one group differs from the others.
The main assumptions of the ANOVA are:

- Independence: Observations within each group (or treatment) are independent of each other
- Normality: The residuals (differences between observed values and predicted values) within each group follow a normal distribution.
- Homoscedasticity: The variance of residuals is approximately equal across all groups.

If these assumptions are not satisfied, we will use the non-parametric Kruskal-Wallis test.

Kruskal-Wallis

The Kruskal-Wallis test assesses whether there is a statistically significant difference between the medians of multiple independent groups. It serves as the nonparametric counterpart to the ANOVA.
The main assumptions of the Kruskal-Wallis are:

- Independence: Observations within each group (or treatment) are independent of each other
- Similar Distributions: The distributions in each group should have similar shapes

The test statistic, denoted as (H), is calculated based on the ranks of the data.

Pearson Correlation

The Pearson correlation coefficient is a widely used measure of linear correlation between two data sets.

The main assumptions of the Pearson Correlation are:

- Independence: Observations within each group (or treatment) are independent of each other
- Normality: The residuals (differences between observed values and predicted values) within each group follow a normal distribution.

- Homoscedasticity: The variance of residuals is approximately equal across all groups.

If there are no repeated data values, a perfect Pearson correlation of +1 or -1 occurs when each variable is a perfect monotone function of the other.

To calculate Pearson's r, we have to compute it like this:

$$r = (cov_x y)/(\sigma_x * \sigma_y)$$

Where r is the correlation coefficient, $cov_x y$ is covariance of the variables and σ_x and σ_y are standard deviations of the variables.

Spearman Correlation

A nonparametric analogue of the Pearson correlation, the Spearman correlation assesses how well the relationship between two variables can be described using a monotonic function. Unlike Pearson's correlation, which assesses linear relationships, Spearman's correlation focuses on monotonic relationships (whether linear or not).

If there are no repeated data values, a perfect Spearman correlation of +1 or -1 occurs when each variable is a perfect monotone function of the other.

To calculate Spearman's ρ, we have to convert the raw scores of both variables to ranks and then compute it like this:

$$\rho = (cov_x y)/(\sigma_x * \sigma_y)$$

Where ρ is the correlation coefficient, $cov_x y$ is covariance of the rank variables and σ_x and σ_y are standard deviations of the rank variables.

Chapter 5

Exploratory Analysis

Exploratory data analysis was conducted in [1]. Since my work continues this analysis, I must include the original graphs from [1] here.

The exploratory analysis shows that the average patient is 31 years old, with a standard deviation of 10.

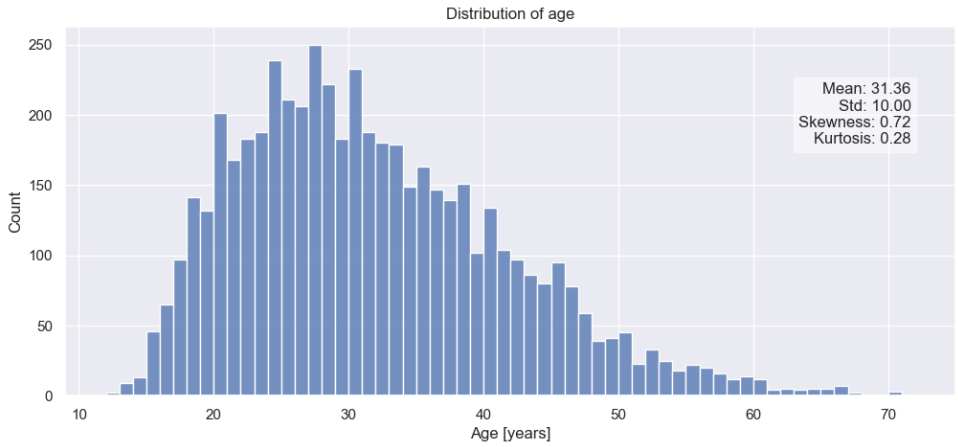


Figure 5.1: Distribution of age

They smoke about 20 cigarettes (=1 pack) per day, but that may be caused by the fact that most of the patients do not count the exact number of cigarettes they smoke and, therefore, use some fractions of one pack of cigarettes to describe it, like 1 pack, ½ of the pack (= 15 cigarettes) etc.

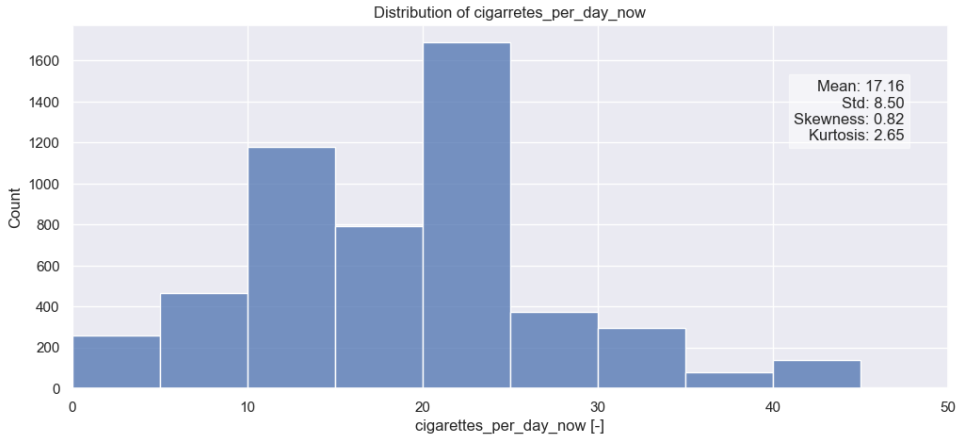


Figure 5.2: Distribution of cigarettes per day now

Most of the patients tried to quit smoking a few times, with a peak at 2, and some of them tried 10 or more times, which caused the second peak.

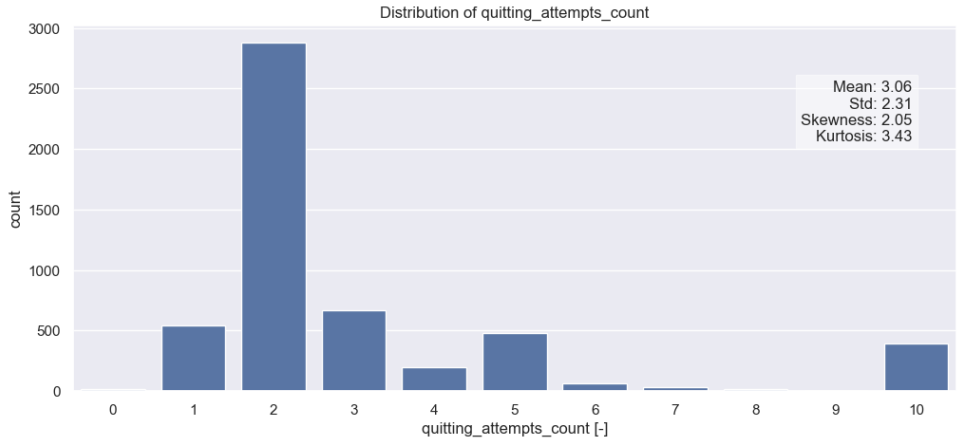


Figure 5.3: Distribution of quitting attempts count

The majority of patients started smoking around 15 to 17 years of age.

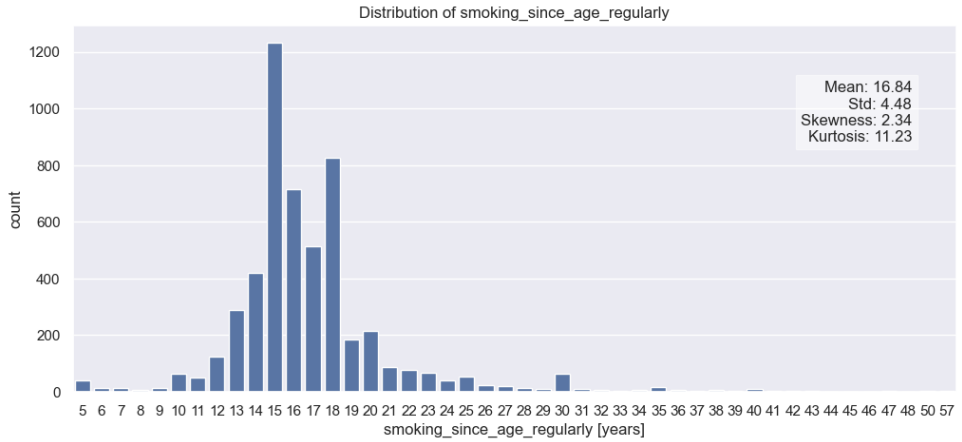


Figure 5.4: Distribution of smoking since age regularly

The majority of patients smoke for around 15 years. About 60% of them smoke from 3 to 20 years.

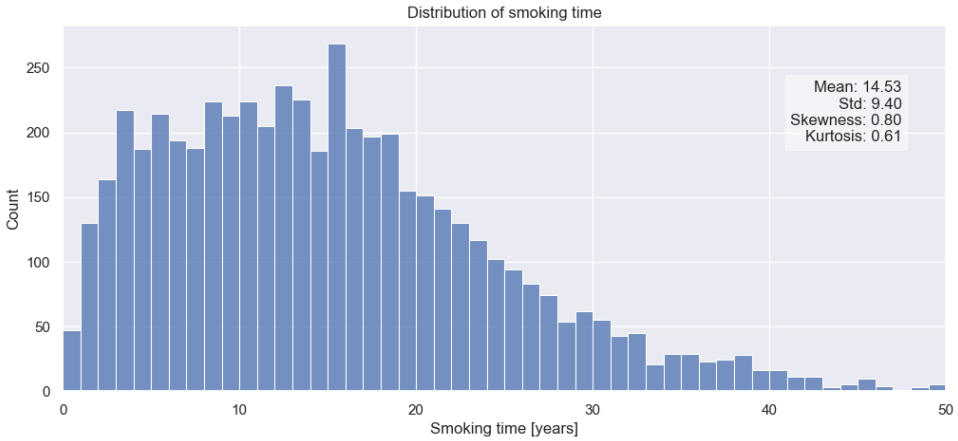


Figure 5.5: Distribution of smoking time

The majority of patients had no relapses.

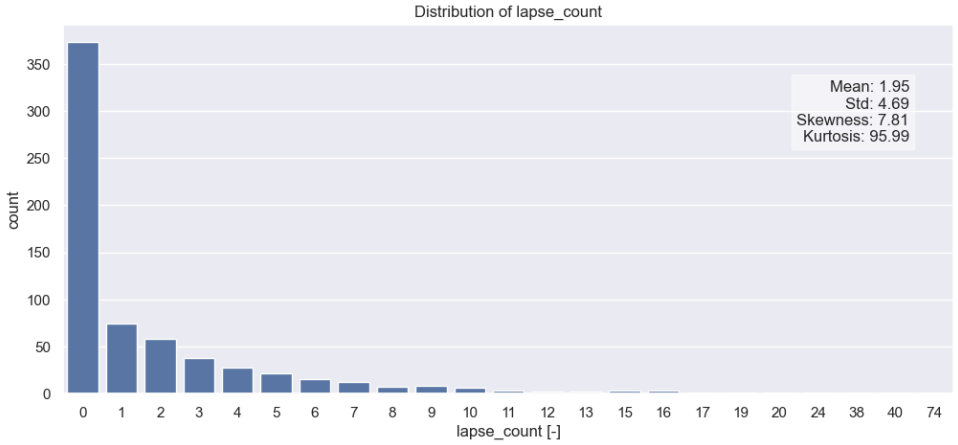


Figure 5.6: Distribution of lapse count

Completed therapy

One of the problems was that we were unsure how to define if the user had finished their therapy. We came to this definition:

1. The user has passed the EQ phase and, therefore, started the therapy
2. User had at most 1 lapse
3. The user used the application one week after their last lapse - they have not given up after their first fail, OR they have reached the WR phase

By our definition, 16% or 105 of all our patients have completed the therapy.

Let's test the correlation between our parameters and the completed therapy variable with Spearman Correlation:

Variable	r	p
Age	-0.02	0.62
Smoking since age regularly	0.02	0.55
Lapse count	-0.3	< 0.001
Quitting attempts count	0.06	0.08
Cigarettes per day before therapy	-0.02	0.82
Smoking time	-0.04	0.28
Sex	0.09	0.017
Last non-smoking period duration	0.01	0.82

Table 5.1: Correlation with completed therapy

As we can see, the only two statistically significant correlations we have found are the weak correlation with the sex variable and the medium correlation with lapse count. It shows that women have a slightly higher chance of finishing the therapy; in the case of lapse count, it was probably caused by our definition of the completed therapy, which depends on the lapse count. All other studied variables are statistically insignificant and weak, with almost no correlation; surprisingly, even correlation with lapse count was neither significant nor strong, even though our definition builds upon it.

Multivariate

In this section, we've completed a multivariate exploratory analysis.

In this graph, we can see no clear dependency between the age at which patient started smoking, their age and if they have completed the therapy.

Figure 5.7: Age vs. Smoking age since regularly

In this graph, we can see that patients who have not tried or tried quit only once but completed the therapy started smoking earlier than their counterparts who have not completed the therapy.

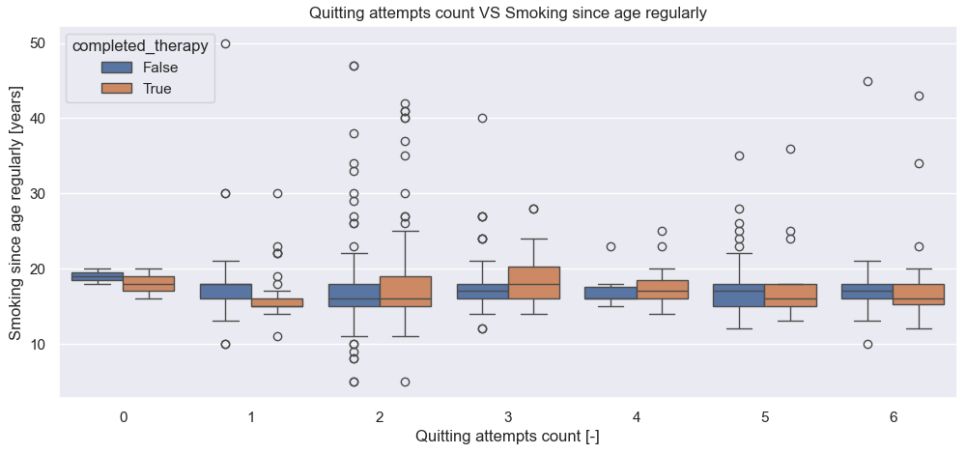


Figure 5.8: Quitting attempts count vs. Smoking age since regularly

In this graph, we can see that patients who have finished the therapy had fewer lapses, which comes from our definition.

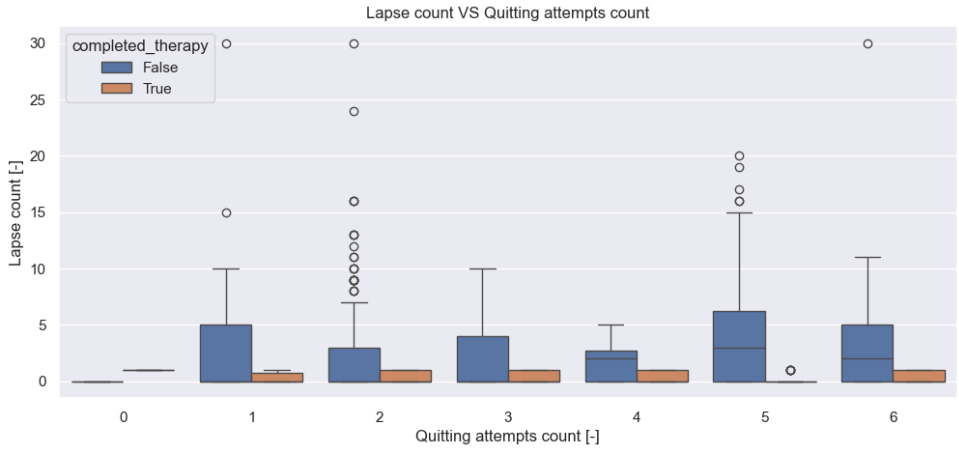


Figure 5.9: Lapse vs. Quitting attempts count

In this graph, we can see no clear dependency between the quitting attempts count, the last non-smoking period, and whether they have completed the therapy, except that successful patients show greater variability in the quitting attempts count.

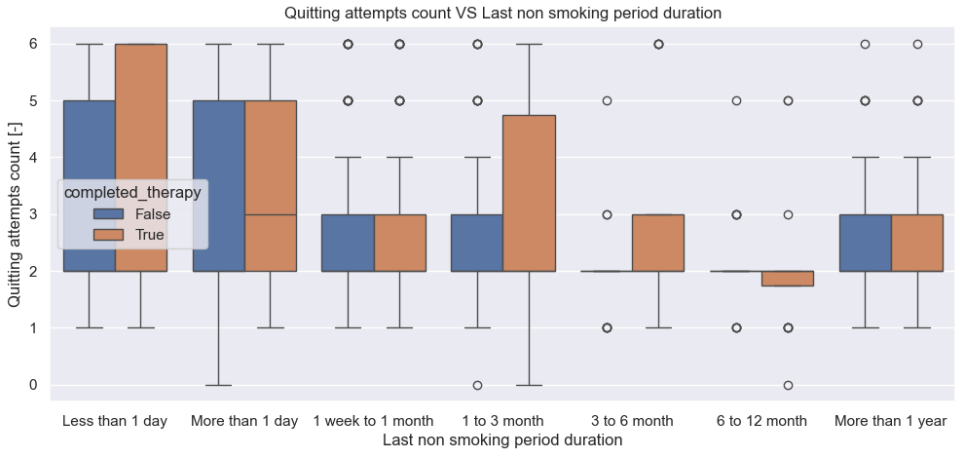


Figure 5.10: Quitting attempts count vs. Last non-smoking period duration

In this graph, the lapse count won't differ with the duration of the last non-smoking period.

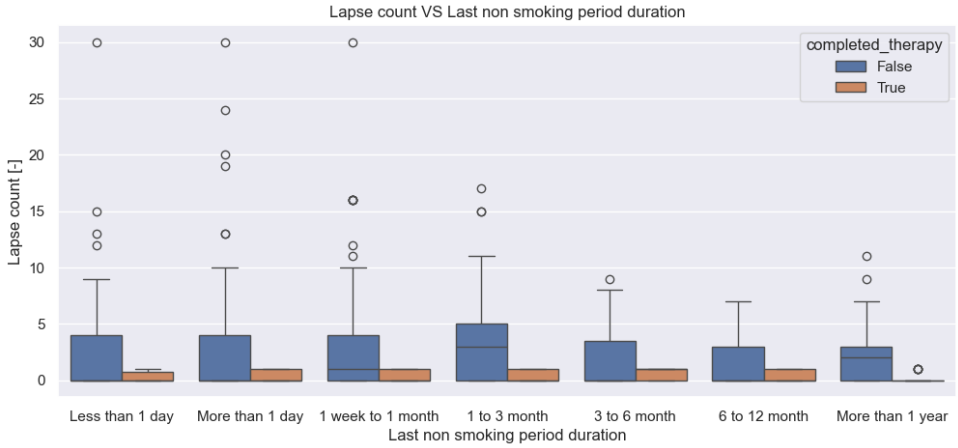


Figure 5.11: Lapse count vs. Last non-smoking period duration

In this graph, we can see that the lapse count won't differ with different attempt counts.

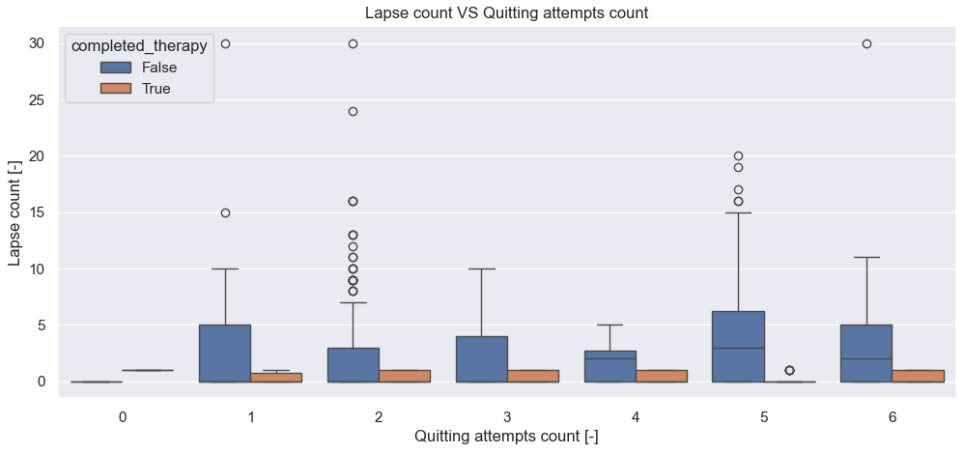


Figure 5.12: Lapse count vs. Quitting attempts count

In this graph, we can see that patients who smoke an average amount of cigarettes have lower variability in lapses.

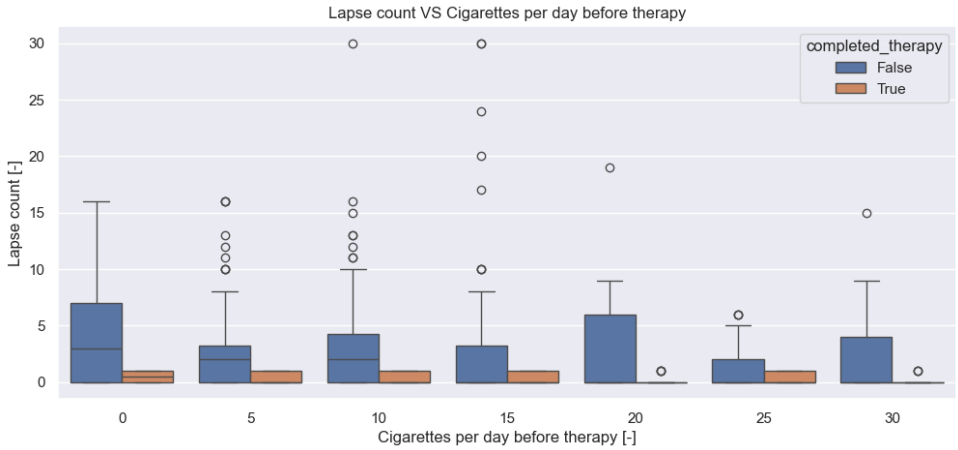


Figure 5.13: Lapse count vs. Cigarettes per day before therapy

In this graph, we can see that successful men have higher variability in their quitting attempts count than unsuccessful men. Conversely, successful women have lower variability in their quitting attempts count than successful women.

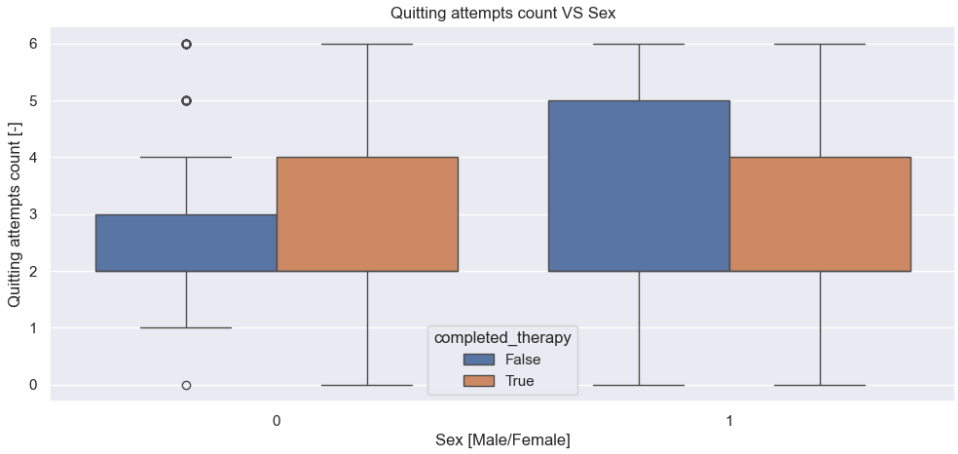


Figure 5.14: Quitting attempts count vs. Sex

In this graph, we can see that successful men have a slightly higher mean in the number of cigarettes they smoke than unsuccessful men. Women generally smoke more than men, but their means do not differ between the two groups.

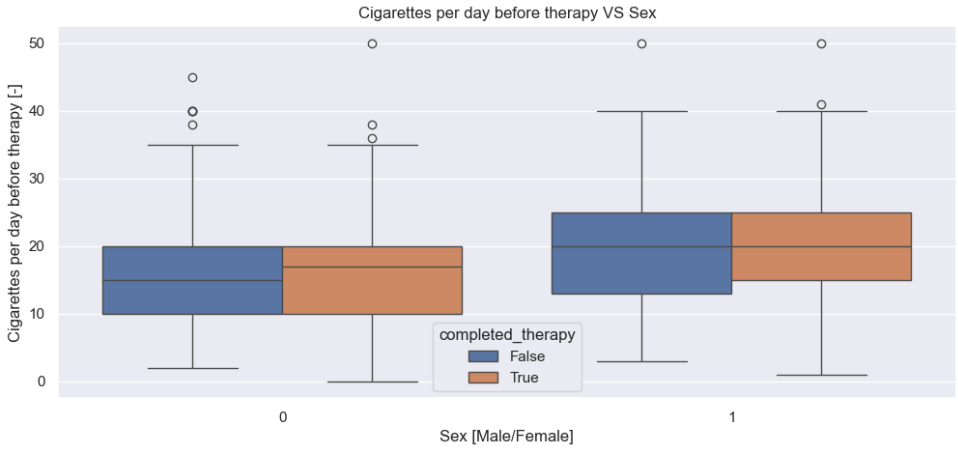


Figure 5.15: Cigarettes per day before therapy vs. Sex

In this graph, we can see that successful men started smoking later than unsuccessful men. Conversely, successful women started smoking earlier than unsuccessful women

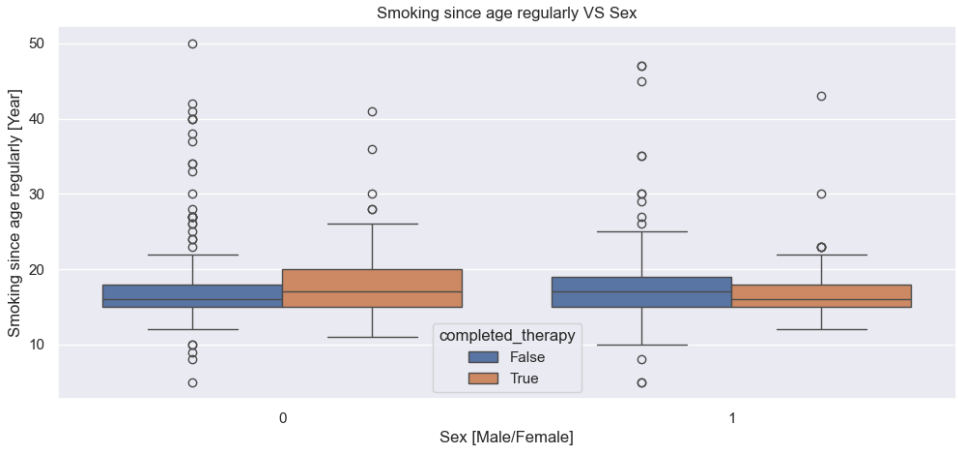


Figure 5.16: Smoking since age regularly vs. Sex

Because we had the opposite results for men and women in a few graphs, we measured correlations for them separately for both sexes.

These are the results for men:

Variable	r	p
Age	-0.05	0.29
Smoking since age regularly	0.10	0.04
Lapse count	-0.32	< 0.001
Quitting attempts count	0.06	0.21
Cigarettes per day before therapy	0.02	0.73
Smoking time	-0.08	0.10
Last non-smoking period duration	0.05	0.3

Table 5.2: Correlation with completed therapy for men

These are the results for women:

Variable	r	p
Age	0.01	0.81
Smoking since age regularly	-0.12	0.02
Lapse count	-0.26	< 0.001
Quitting attempts count	-0.04	0.38
Cigarettes per day before therapy	0.03	0.52
Smoking time	0.05	0.33
Last non-smoking period duration	0.06	0.28

Table 5.3: Correlation with completed therapy for women

As we can see, the Spearman correlation confirms our hypothesis that the later the man started smoking, the higher the probability that he will successfully finish the therapy, and on the contrary, the earlier the woman started smoking, the higher the probability that she will successfully finish the therapy. That correlation is weak and may be caused by the type I error because the p-value is close to 0.05. If we will stricken it, or use Bonferonni correction, it will become insignificant.

Lapses

We have analysed the application's logs to get information about patients' relapses. The log may contain these values: 'Lapse', which means that the patient has smoked; 'Abstinent', which means that the patient has not smoked; 'No reply', which means that the patient has not replied on notification.
We have plotted distributions of the patients' responses:

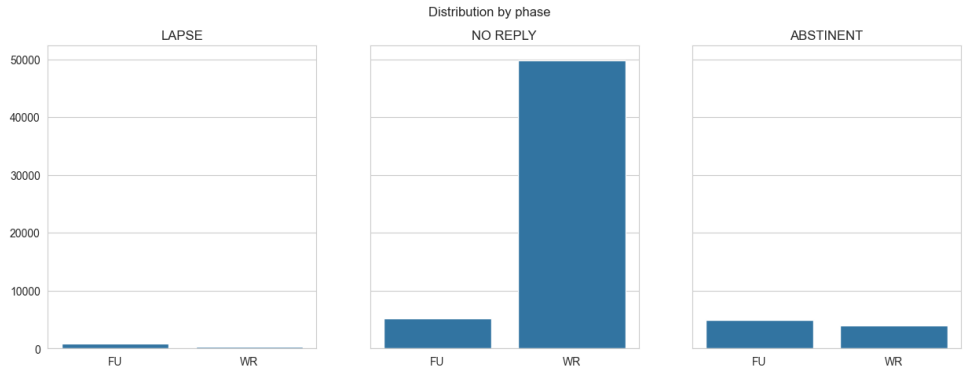


Figure 5.17: Distribution of lapses by phase

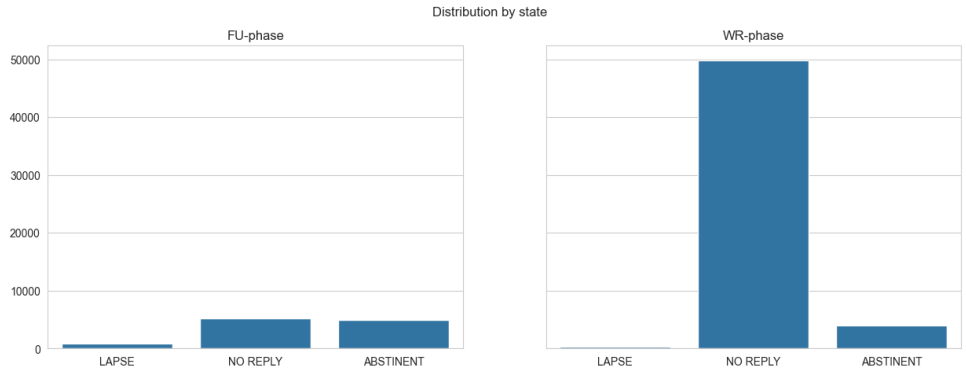


Figure 5.18: Distribution of lapses by state

As we can see from these graphs, the number of relapses the patients have is pretty low. Users mostly stop responding to notifications during the long WR phase. Lapses may only appear in the FU and WR phases and can be counted only for users with the paid version.

Conclusion

In this chapter, we can see that almost all of our variables do not correlate with the success of the therapy. Lapse count has a medium correlation, which is caused by our definition of the completion of therapy and the weak correlation with age when the patient started smoking, the effect of which surprisingly differs depending on the sex of the patient.

Chapter 6

Analysis of the logs

This chapter's main idea is to quantify users' adherence to the therapy based on their behaviour in the application, which we can get from the activity logs. The main problem was to define adherence correctly, as we can see in our State of the Art section, so we have chosen to use metrics like number of sessions completed, regularity, total screen time and mean screen time per week.
The user also reports to the application daily after they finish the EQ phase, if they have a relapse, we can measure the number of lapses the user has. Still, it is unreliable because the user can provide incorrect information.

Analysis method

The pipeline is simple: we query the data from the events_event table and merge it with the users_user table by actor_id and id keys, taking only paid users with the same criterion as in [1], non-staff and non-superusers so that they won't influence our results. By doing that, we got a list of all users' IDs sql_query_ids that we can use. Then we take data from the events_event table again, taking only rows that belong to the users included in the sql_query_ids list and only rows in which variable column equals 'timer'. Then we grouped the rows for every user in sql_query_ids and ran the functions we used for computing the parameters we wanted to know. These functions go through every user's log and compute the chosen parameters.
The rows have this structure: id, time, post_value, actor_id, where id is the id of the log, time is the time when that event happened in the format 'YYYY-MM-DD HH:MM:SS', post_value contains information about the screen of the event and how much time it took for the user to complete the whole session, and id is the id of the user.
We should look more at the structure of the post_value column. For example, it may look like this: 'EE02.1, 6864 ms'. EE is the phase code, which may be B0 (bujon), EQ, FU, WR, LM (lapse management) and other purely technical phases that are discarded from the computation. 02 is the session number. 1 is the screen number; each session may contain a few screens. 6864 is the time the user spent in this **session** in ms; this value is capped at 300000 ms, so we have to compute time for every screen from the time column.

Screen time

The screen time is computed like this: the main pipeline runs through every log for the selected user and computes the time the user spends in the application every day. The first screen time per day is taken from the post_value column; others are taken from the difference in time between the last log and this one. They are summed. This is the total screen time
We got these results:

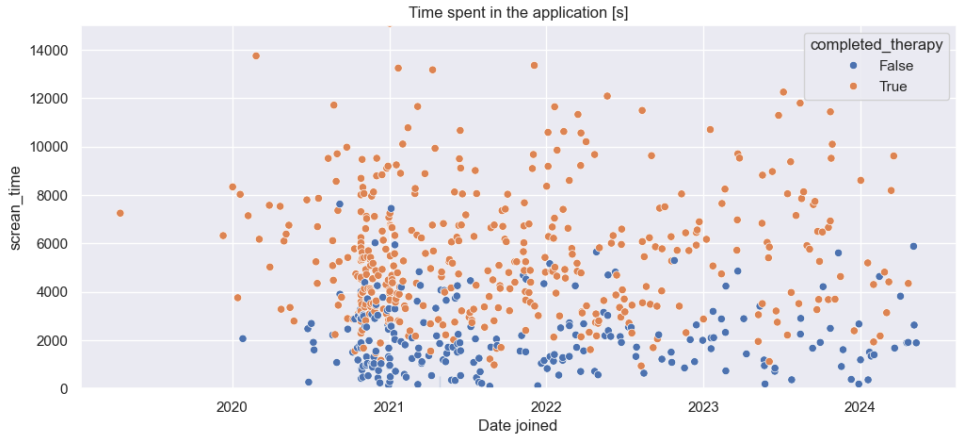


Figure 6.1: Total time spent

There does not seem to be a clear trend or pattern between the date joined and screen time. The data points are scattered throughout the graph. Users who completed therapy have higher total screen time than users who have not completed it.

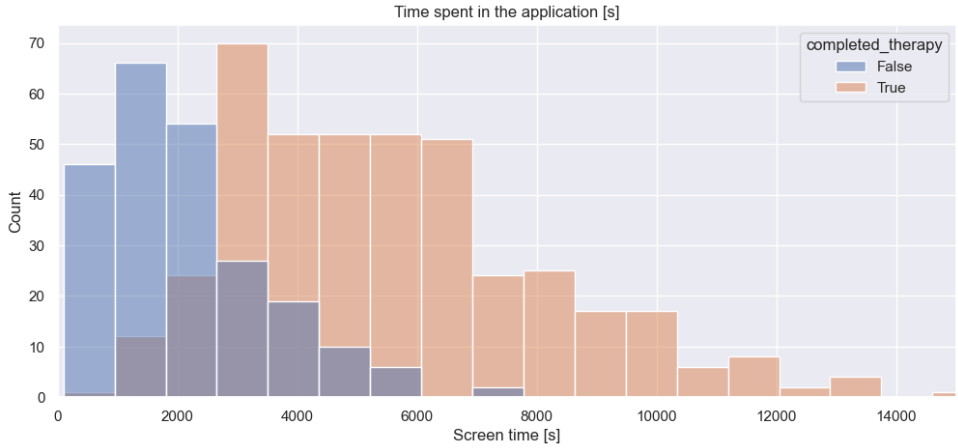


Figure 6.2: Total time spent histograms

From these histograms, we can see that two groups of patients have different behaviour.

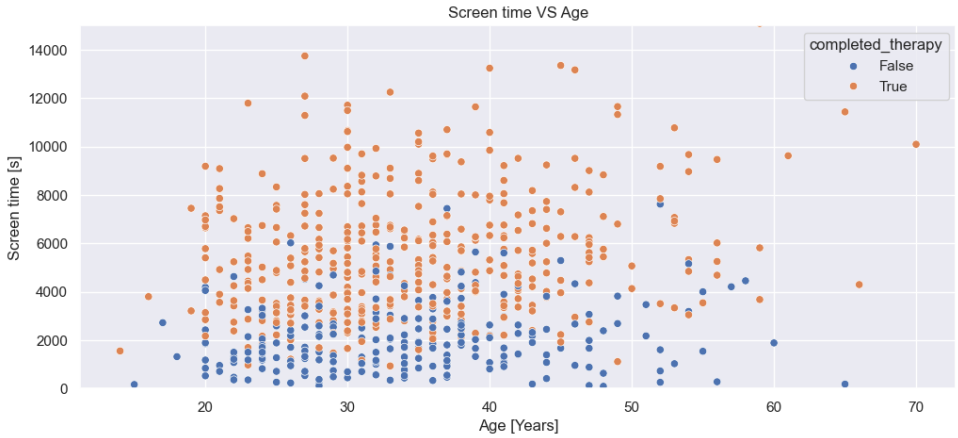


Figure 6.3: Screen time vs age

Smoking time is defined as the difference between the user's age and the age at which they've started to smoke.

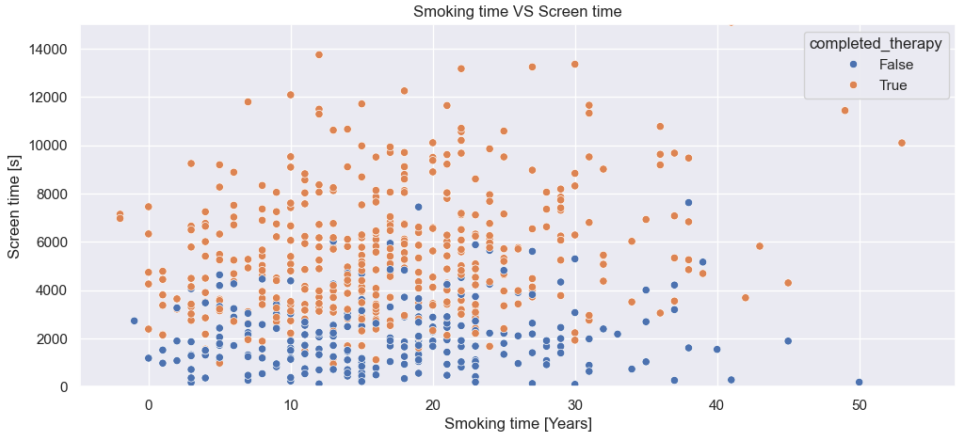


Figure 6.4: Smoking time vs Screen time

There is no clear pattern or trend between age and screen time. Data points are scattered throughout the graph. Both completed and non-completed therapy groups exhibit varying screen time durations across different age groups.

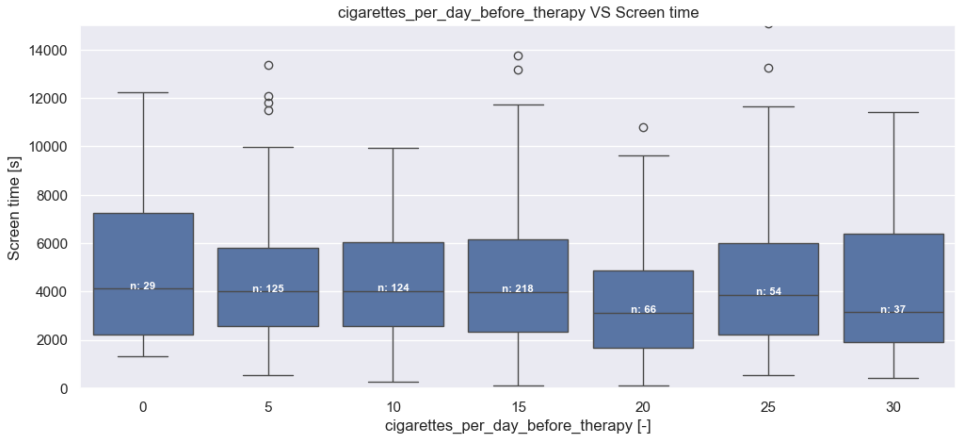


Figure 6.5: Cigarettes per day before therapy vs Screen time

There is no clear connection between the number of cigarettes patient smoked and their screen time.

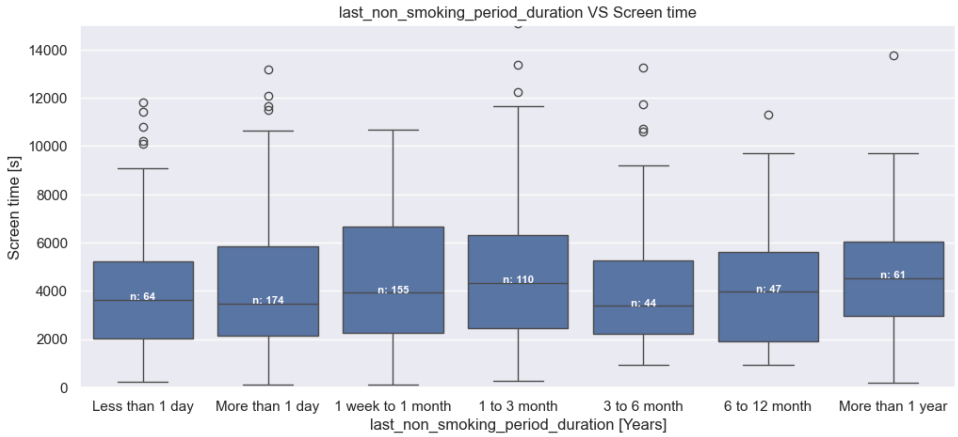


Figure 6.6: Last non-smoking period duration vs Screen time

It seems that patients who managed not to smoke for a longer period of time have slightly higher screen time, but the Kruskall-Wallis test tells us that with p-value=0.4, these groups are not different.

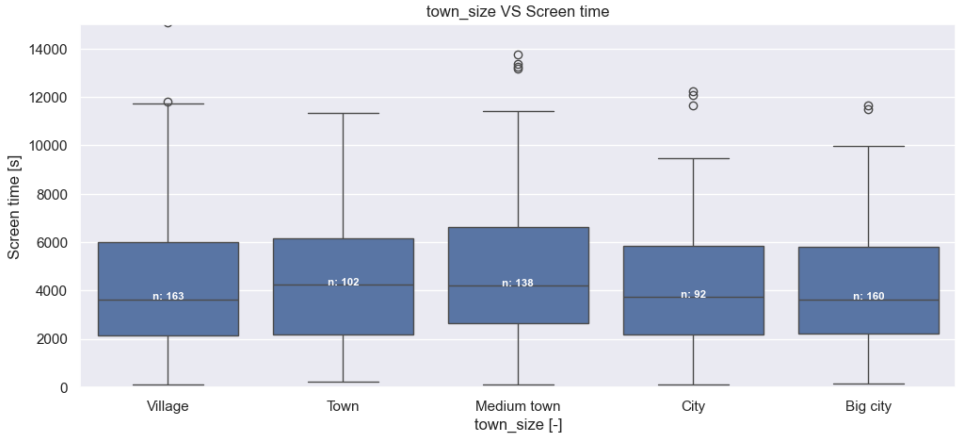


Figure 6.7: Town size vs Screen time

There is no connection between town size and screen time.

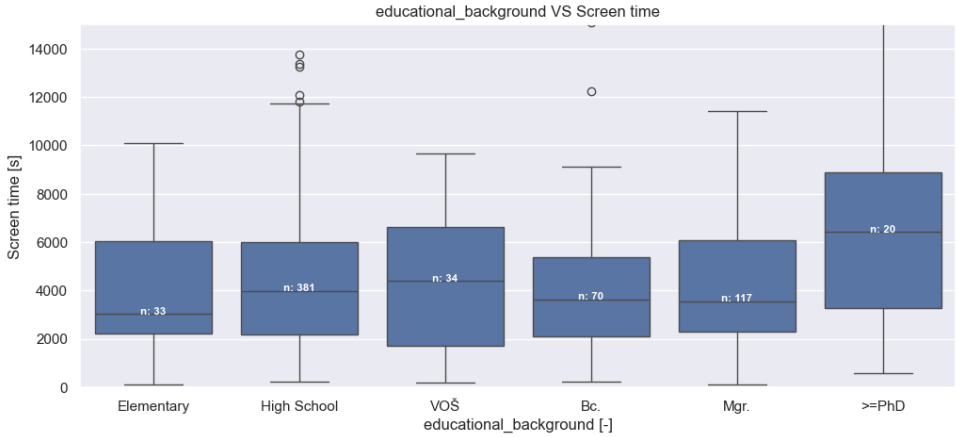


Figure 6.8: Educational background vs Screen time

Patients who have completed PhD have higher screen time, but the Kruskall-Wallis test tells us that the differences between the groups are insignificant, with p-value=0.24. The Mann-Whitney U test finds statistically significant differences between the PhD group and other groups, but it becomes insignificant after the Bonferroni correction.

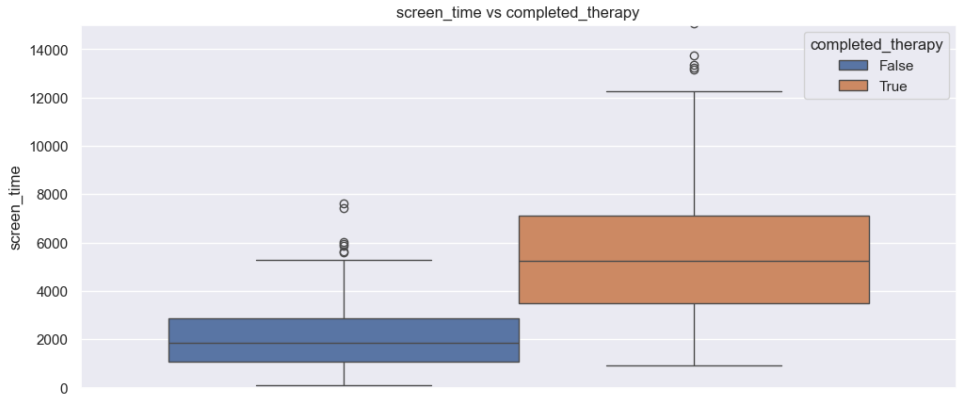


Figure 6.9: Screen time vs completed therapy

Once again, we can see that these groups have different means. Mann-Whitney U test agrees with p-value<0.01.

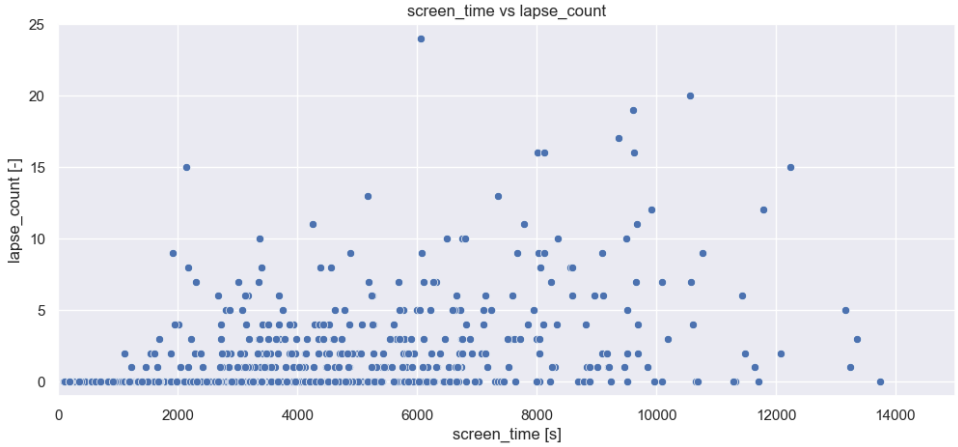


Figure 6.10: Screen time vs Lapse count

Patients with higher screen time have more lapses, which makes sense because they must spend time on the lapse management sessions.

Let's look closer to their basic statistics:

Mean	4432 s / 73.9 min
Median	3810 s / 63.5 min
Std	3149 s / 52.5 min
90% quantile	8470 s / 141.2 min

Table 6.1: Basic statistics of the Screen time

Mean screen time

The screen time is computed like this: the main pipeline runs through every log for the selected user and computes the time the user spends in the application every day. The first screen time per day is taken from the post_value column; others are taken from the difference in time between the last log and this one. They are summed for every day. If the user has not used the application that day, the value for that day is set to 0. Later, we smoothed the array we got with the moving average filter with the window size of one week and took the mean value from that array.
Some graphs that do not contain useful information are located in Appendix B. We got these results:

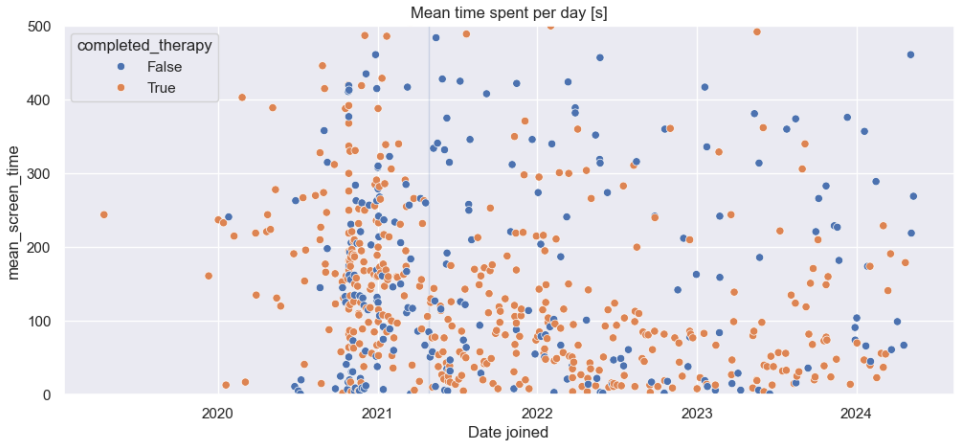


Figure 6.11: Mean time spent per day

There is no clear trend or pattern between the date joined and mean screen time. The data points are scattered throughout the graph. Users who completed therapy have lower mean screen time than users who have not completed it.

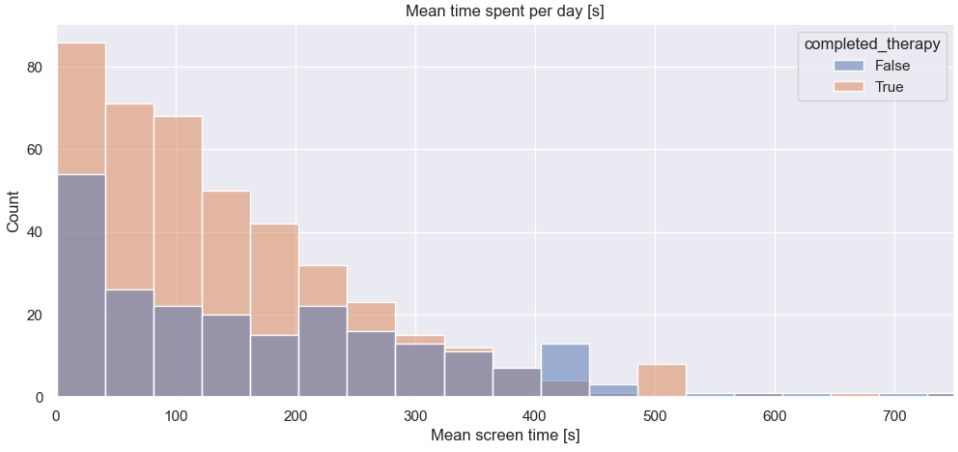


Figure 6.12: Mean time spent per day histograms

Both groups have the highest count at the lower end of the mean screen time spectrum. As screen time increases, the counts decrease for both categories.

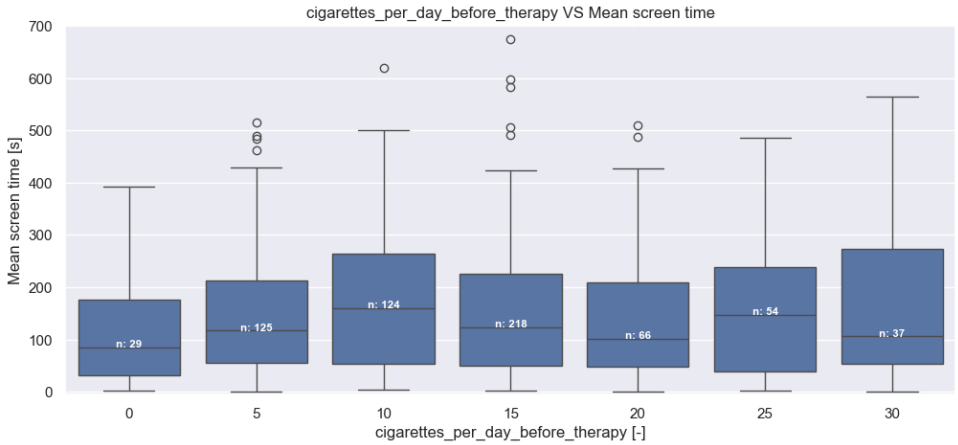


Figure 6.13: Cigarettes per day before therapy vs Mean screen time

Individuals who smoked more cigarettes per day generally exhibit higher mean screen times, though differences are not statistically significant with p-value=0.47.

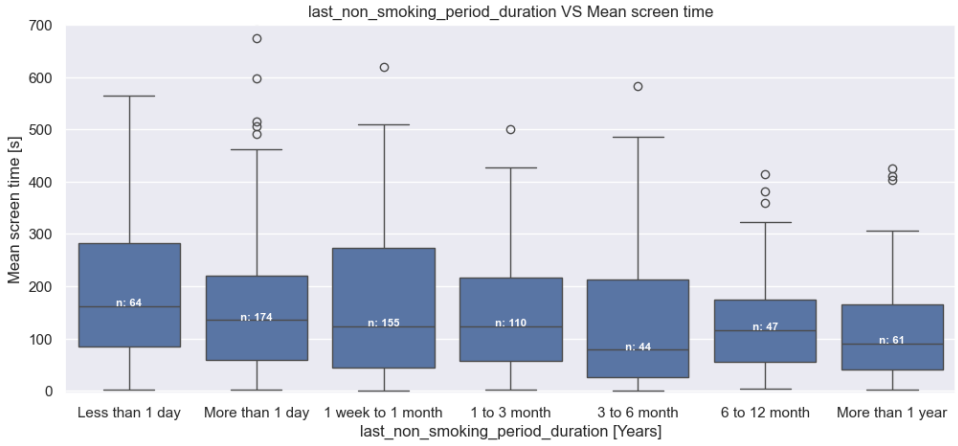


Figure 6.14: Last non-smoking period duration vs Mean screen time

As the duration of the non-smoking period increases, the mean screen time tends to decrease, though it is not statistically significant, with p-value=0.07. Mann-Whitney U tells us that there are differences between the shortest and the longest groups, but after Bonferonni correction, they become insignificant.

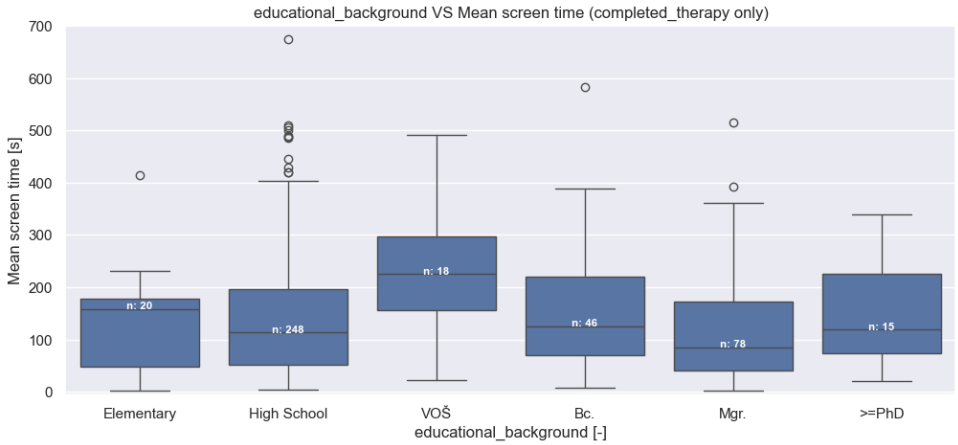


Figure 6.15: Educational background vs Mean screen time

We can see that VOŠ group differs from others. Kruskall-Wallis test agrees with p-value=0.034, which shows that there are differences between the groups. Mann-Whitney U test shows us that groups VOŠ and Mgr. are different even after Bonferonni correction with p-value=0.002.

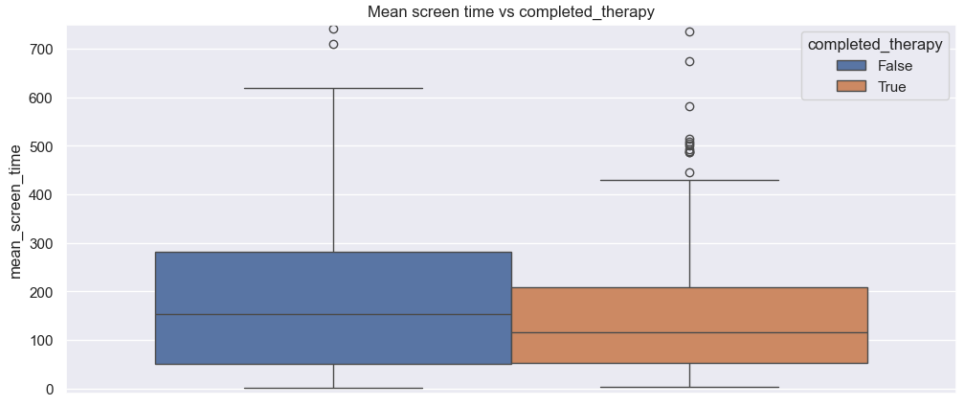


Figure 6.16: Mean screen time vs completed therapy

We can see that patients who have completed therapy have lower mean screen time. Mann Whitney U test agrees with p-value=0.03 that that difference is significant.

Let's look closer to their basic statistics:

Mean	186 s / 3.1 min
Median	144 s / 2.4 min
Std	223 s / 3.7 min
90% quantile	366 s / 6.1 min

Table 6.2: Basic statistics of the Mean screen time

Number of sessions

The number of sessions is computed like this: the main pipeline runs through every log for the selected user and counts every new phase-session number combination in the logs, which are not purely technical.
We got these results:

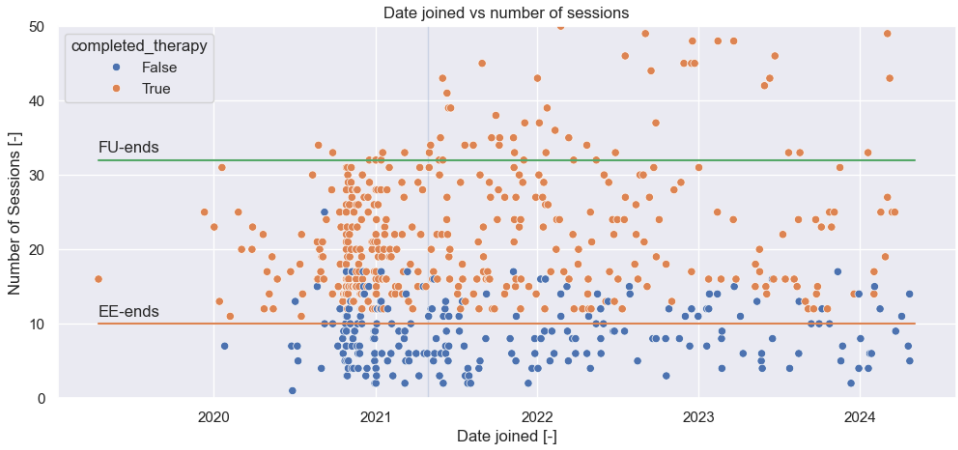


Figure 6.17: Number of sessions vs Date joined

Patients who have completed the therapy usually have more completed sessions.

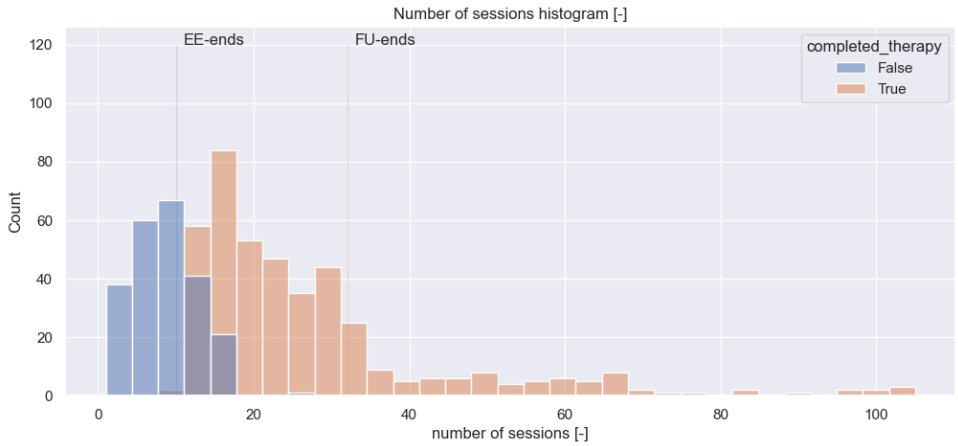


Figure 6.18: Number of sessions histograms

We can once again see that patients who have completed the therapy usually have more completed sessions.

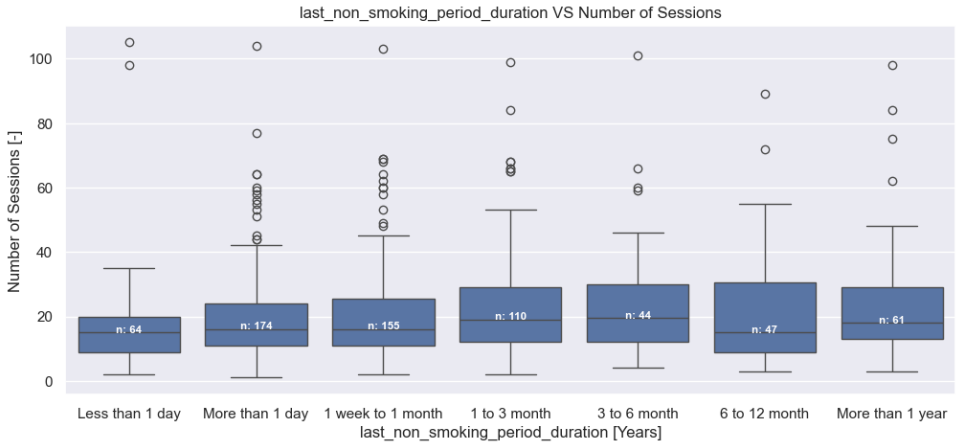


Figure 6.19: Last non-smoking period duration vs Number of sessions

We can see some rising trend in this graph, and the Kruskall-Wallis test with p-value=0.032 shows that differences are statistically significant; the Mann-Whitney U test finds differences between the first group and the four last, but after the Bonferonni correction, they come insignificant.

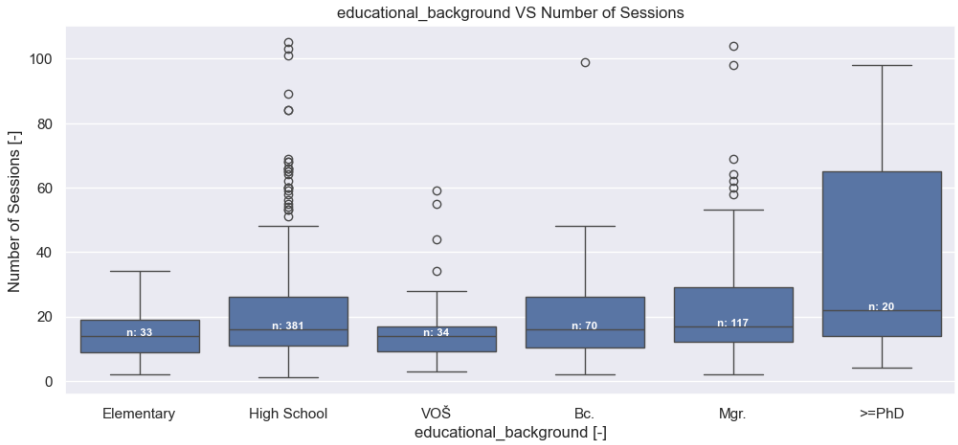


Figure 6.20: Educational background vs Number of sessions

We can see that the PhD. group differs from the others, and the Kruskall-Wallis test agrees that there are differences with p-value=0.022. Mann-Whitney U test agrees that the PhD. group is different from others, but after the Bonferonni correction, it becomes insignificant.

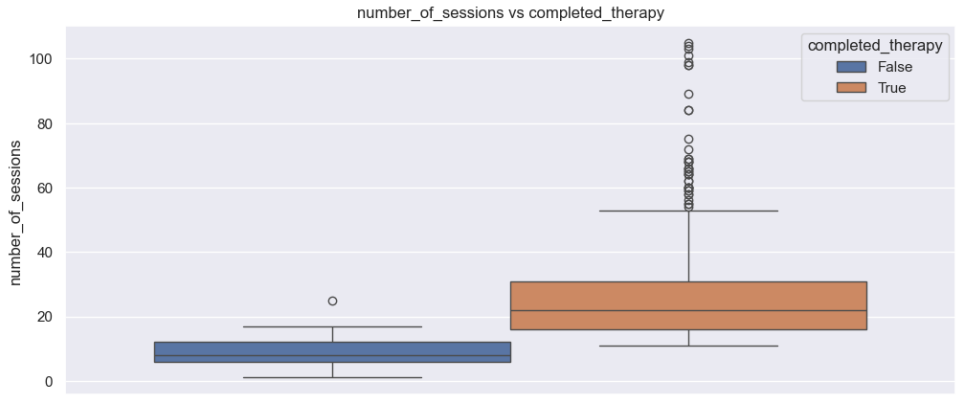


Figure 6.21: Number of sessions vs completed therapy

Patients who have completed the therapy have a higher number of sessions completed. Mann-Whitney U test with p-value<0.001 tells us that this difference is statistically significant.

Let's look closer to their basic statistics:

Mean	21
Median	16
Std	16.7
90% quantile	38.9

Table 6.3: Basic statistics of the Number of sessions

Regularity

The regularity is computed like this: the main pipeline runs through every log for the selected user and marks in the array if the user has used the application that day. Later, we smoothed the array we got with the moving average filter with the window size of one week and took the mean value from that array.

We got these results:

Figure 6.22: Regularity

We can see that patients who have completed the therapy have higher regularity.

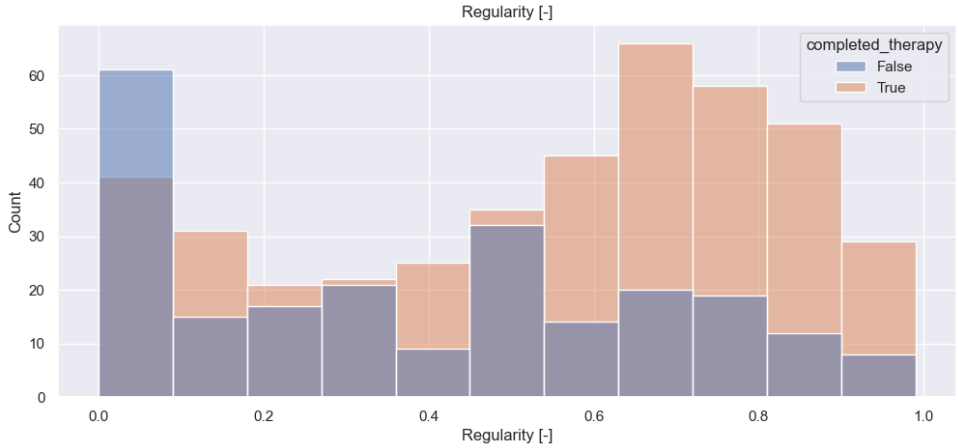


Figure 6.23: Regularity histograms

Once again, we can see that patients who have completed the therapy have higher regularity.

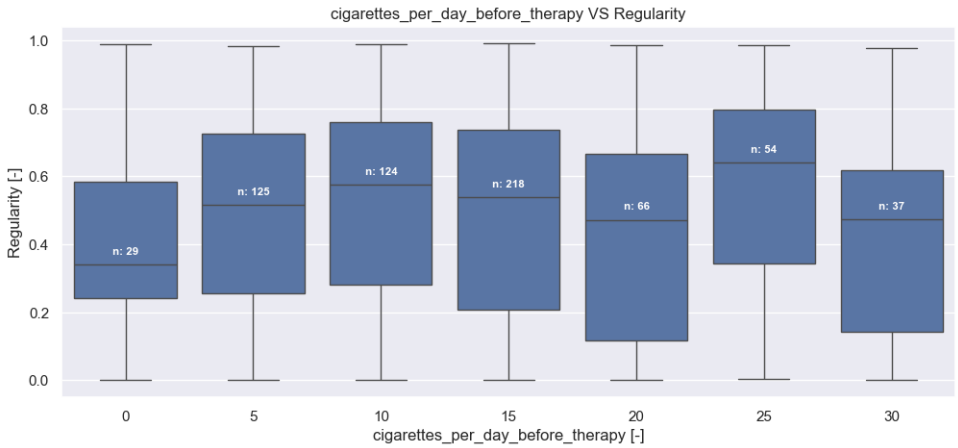


Figure 6.24: Cigarettes per day before therapy vs Regularity

There seem to be differences between the groups, but the Kruskall-Wallis test thinks it is insignificant with p-value=0.67.

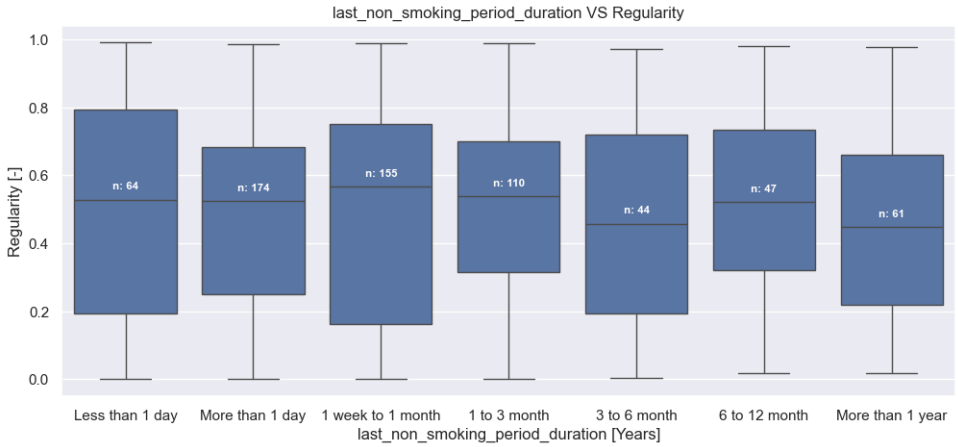


Figure 6.25: Last non-smoking period duration vs Regularity

There is no clear connection between regularity and the duration of the last non-smoking period. Kruskall-Wallis test agrees with p-value=0.9.

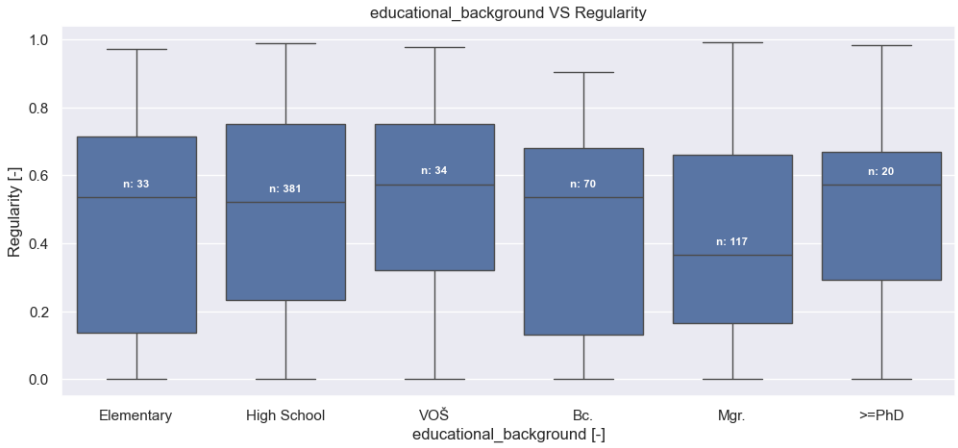


Figure 6.26: Educational background vs Regularity

There is no clear connection between regularity and the duration of the last non-smoking period. Kruskall-Wallis test agrees with p-value=0.2. Mann-Whitney U test thinks that differences between High School and Mgr. and VOŠ and Mgr. are statistically significant, but after the Bonferonni correction, they become insignificant.

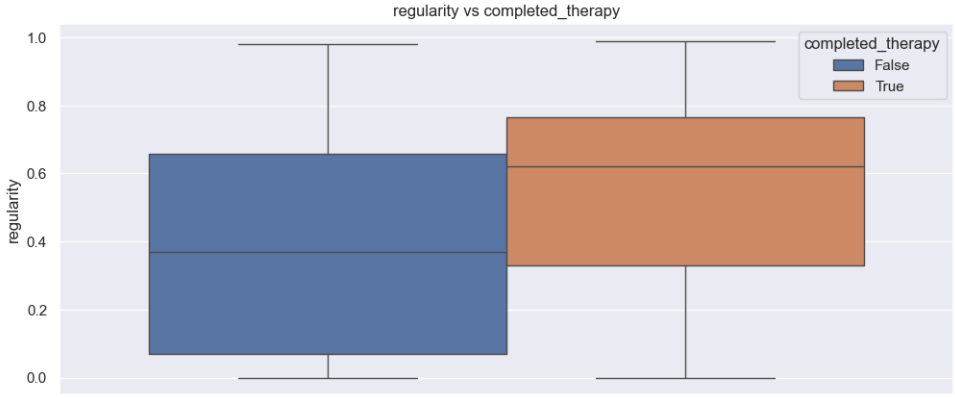


Figure 6.27: Regularity vs completed therapy

Patients who completed the therapy had higher regularity. Mann-Whitney U test agrees with p-value<0.001.

Let's look closer to their basic statistics:

Mean	0.51
Median	0.57
Std	0.30
90% quantile	0.87

Table 6.4: Basic statistics of the Regularity

Correlation between the adherence variables

We have decided to measure the correlation between every pair of our chosen measurements and show their respective graphs.

For the Screen time and the Number of session, Spearman's rank correlation coefficient equals -0.22 with the p-value < 0.001. Surprisingly, the screen time declines with the number of sessions increase.

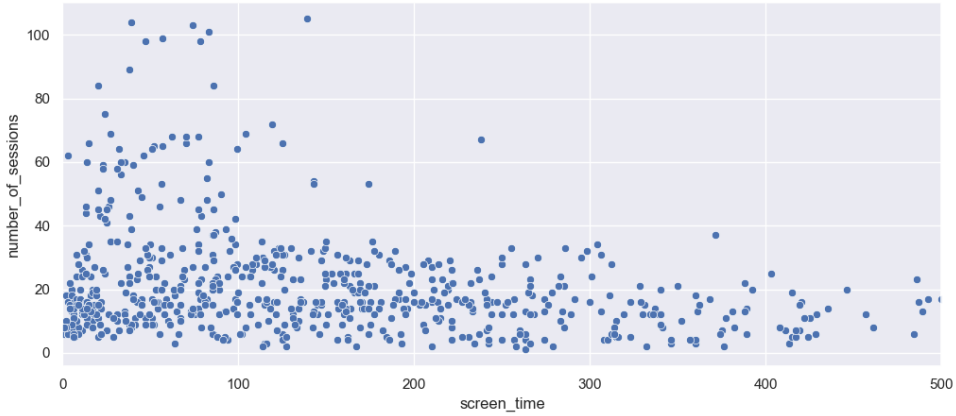


Figure 6.28: Correlation between the Screen time and the Number of sessions

For the Screen time and the Regularity, Spearman's rank correlation coefficient equals 0.23 with the p-value < 0.001. The correlation in medium.

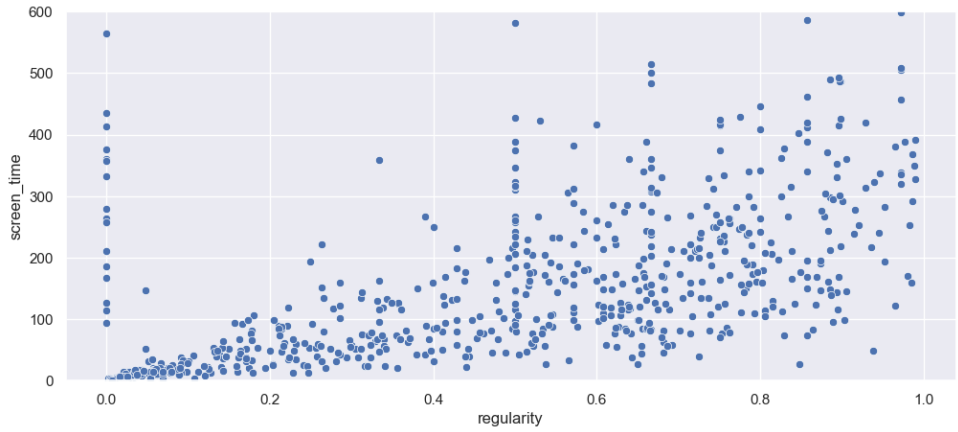


Figure 6.29: Correlation between the Regularity and the Screen time

For the Regularity and the Number of session, Spearman's rank correlation coefficient equals 0.64 with the p-value < 0.001. The correlation is very strong, which makes sense because if users don't use the application one day, they'll have 0 screen time that day.

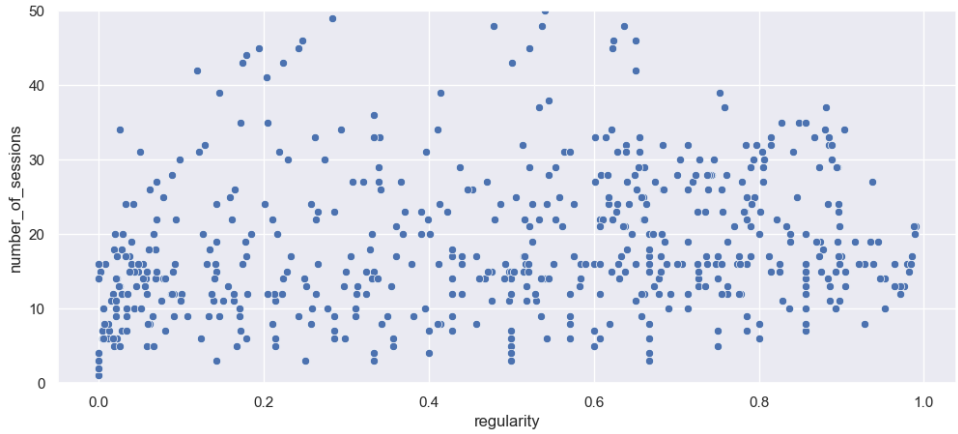


Figure 6.30: Correlation between the Regularity and the Number of sessions

Adherence at the first 10 days

For the classification task, we have decided to use other data - screen time, regularity and number of sessions for users only during their first 10 days after they've started using the application.
The reason for it is simple - we want to use data we got from the first stages of the therapy to understand if the person will finish it. Therefore, we can not use data from all their therapy.

For the Screen time variables, Spearman's rank correlation coefficient equals 0.70 with the p-value < 0.001. This graph shows that for almost every user, their screen time at the beginning of the therapy is higher than their screen time during the whole therapy, which means that screen time declines at the latter phases.

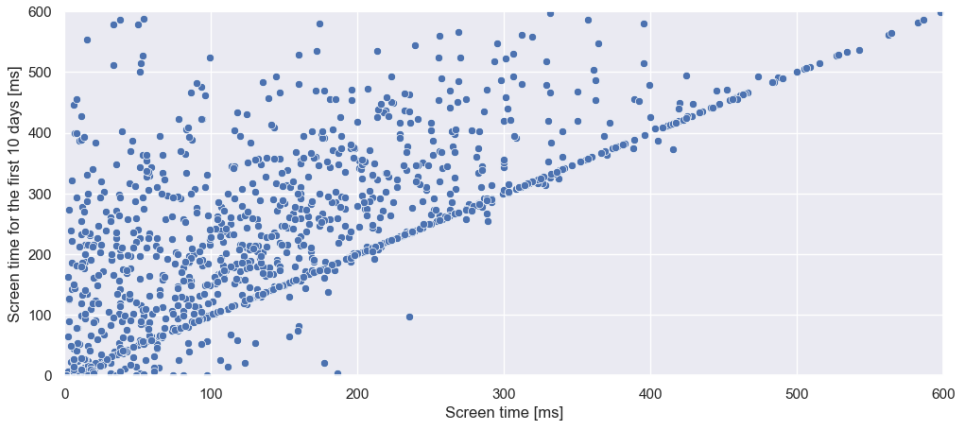


Figure 6.31: Screen time vs Screen time for the first 10 days

For the Number of sessions variables, Spearman's rank correlation coefficient equals 0.65 with the p-value < 0.001.

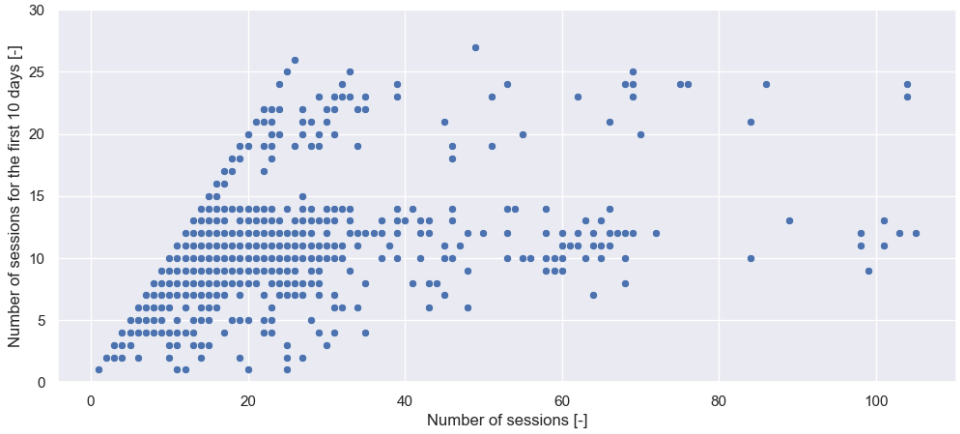


Figure 6.32: Number of sessions vs Number of sessions for the first 10 days

For the Regularity variables, Spearman's rank correlation coefficient equals 0.73 with the p-value < 0.001.

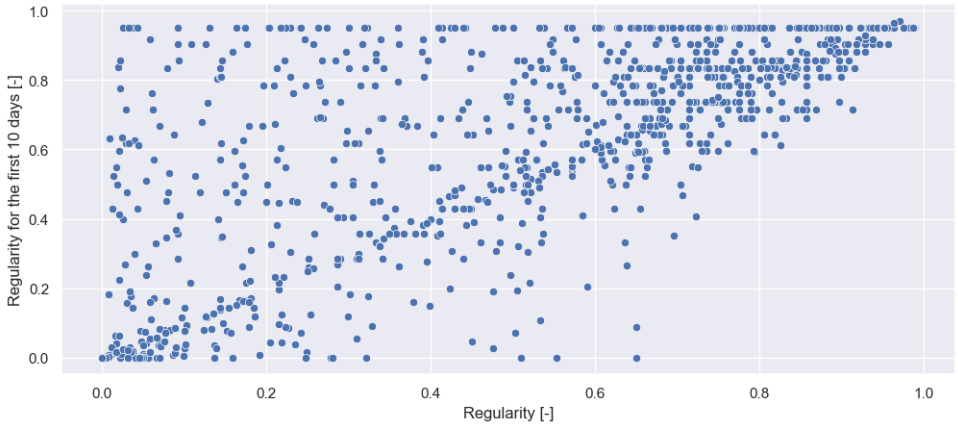


Figure 6.33: Regularity vs Regularity for the first 10 days

Those variables are highly correlated, which is great because we will not lose much information if we use it instead of the full data.

Let's look closer to their basic statistics:

Mean	273 s / 4.5 min
Median	246 s / 4.1 min
Std	236 s / 3.9 min
90% quantile	470 s / 7.8 min

Table 6.5: Basic statistics of the Screen time in 10 days

Mean	9.9
Median	10
Std	4.9
90% quantile	14

Table 6.6: Basic statistics of the Number of sessions in 10 days

Mean	0.6
Median	0.68
Std	0.31
90% quantile	0.95

Table 6.7: Basic statistics of the Regularity in 10 days

Conclusion

From this chapter, we can see that our chosen adherence variables may influence the success of the therapy. The patients behave similarly during their first 10 days of therapy and during the whole therapy, which means we can use only the first 10 days for prediction and regression tasks.

Our adherence variables are correlated, which is a good sign because it may mean that they may be used to describe one thing: adherence.

On the other hand, our newly introduced adherence variables do not have

much connection with the data we got previously. When we can see some differences between the groups, it usually comes as a type I error after the Bonferonni correction.

Chapter 7
Classification

We have tried to build classifiers that will try to predict if the user will finish the therapy using the parameters we have measured in our dataset. We used all adherence metrics from the whole time and only from the first 10 days of the therapy.

Random Forest

We decided to start with the Random Forest classifier because it is the easiest one to interpret by looking at the structure of the trees it contains. We have used it to select the best features to predict if the user will finish the therapy. We are intentionally not using the lapse count variable to avoid data leakage because our finishing condition depends on it.

Let's start with all the data we had:

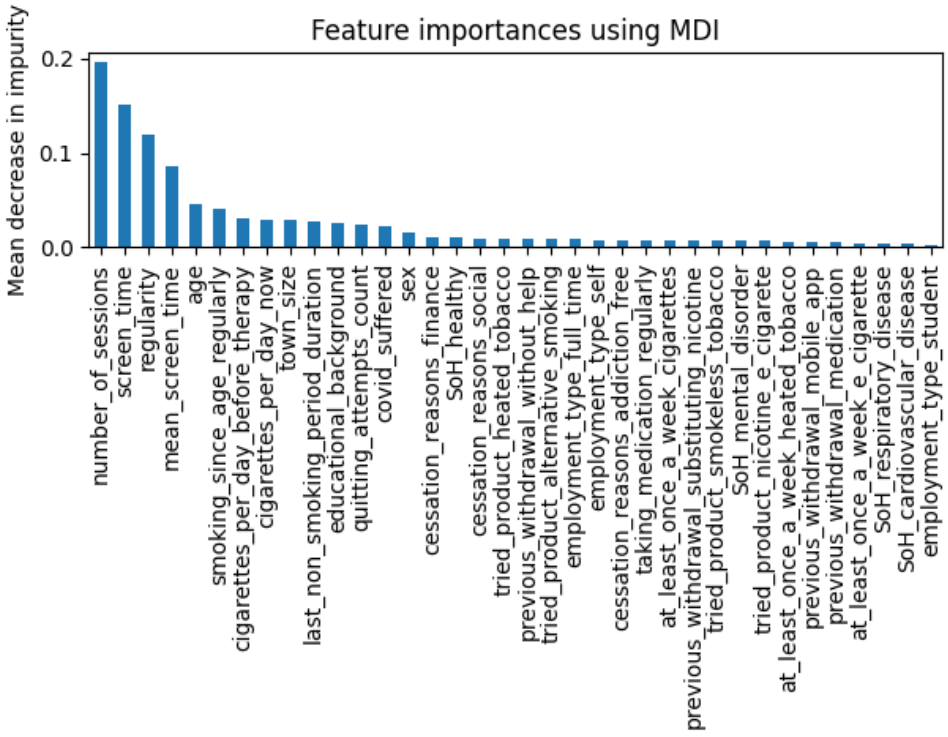


Figure 7.1: Feature importances

The four best features are 'number of sessions', 'regularity', 'screen time' and 'mean screen time'. They may explain 55% of all the variability in the data. Our adherence metrics are in the first 4 places, which is a good sanity check.

After the cross-validation by choosing the best classifier maximising the roc-auc score, we got these metrics:

Accuracy	0.76
Precision	0.75
Recall	0.69

Table 7.1: Classification metrics for the Random Forest

As we can see, these classifications are pretty good.

Later, we decided to use only the four best parameters to see how well they could help us classify our data and get these results:

Accuracy	0.76
Precision	0.71
Recall	0.88

Table 7.2: Classification metrics for the trimmed Random Forest

Surprisingly, our accuracy scores had increased, probably because our trees won't overfit as much as before; because of their random nature, the result is pretty unstable and may change.

Now let's see how our classifier will handle the classification with only adherence data available from the first 10 days:

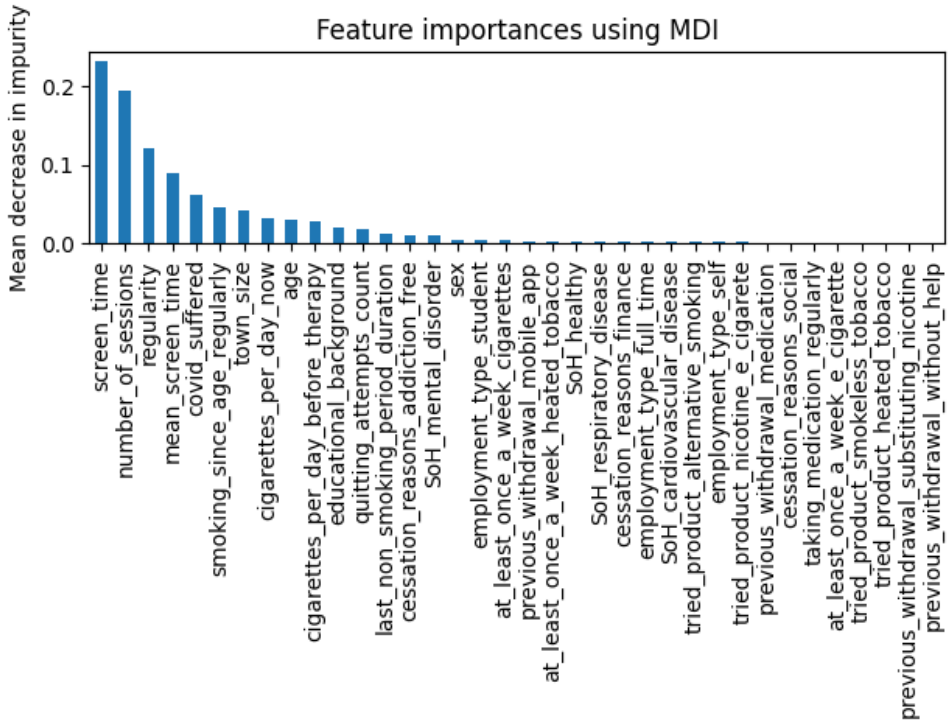


Figure 7.2: Feature importance with only 10 days of data

The 4 best features may now explain 64% of the variance in the data. The order of the features has slightly changed.

Let's run our classification with the 4 features we have chosen before:

Accuracy	0.63
Precision	0.63
Recall	0.56

Table 7.3: Classification metrics for the trimmed Random Forest for 10 days

Our precision metrics have declined, which makes sense because we lost some information about our users.

Logistical Regression

We have also tried to classify our data with the 100 logistical regressions using the best 4 parameters. Here are their mean statistics:

Accuracy	0.74
Precision	0.78
Recall	0.73

Table 7.4: Classification metrics for the Logistical Regression

And using only 10 days of data:

Accuracy	0.72
Precision	0.74
Recall	0.77

Table 7.5: Classification metrics for the Logistical Regression for 10 days

The accuracy and precision are slightly lower than before, but recall is higher.

SVM

We have also tried classifying our data with the 10 linear-kernel SVMs using the best 4 parameters. Here are their mean statistics:

Accuracy	0.74
Precision	0.69
Recall	0.79

Table 7.6: Classification metrics for the SVM

And using only 10 days of data:

Accuracy	0.69
Precision	0.6
Recall	0.84

Table 7.7: Classification metrics for the SVM for 10 days

The accuracy and precision are slightly lower than before, but recall is again higher.

Comparison

Metric	RF	LR	SVM
Accuracy	0.76	0.74	0.74
Precision	0.71	0.78	0.69
Recall	0.88	0.73	0.79

Table 7.8: Classification metrics for all used methods

If we are using all the data available, the Random Forest has the highest Accuracy and Recall, and the Logistical Regression has the highest Precision.

Metric	RF	LR	SVM
Accuracy	0.63	0.72	0.69
Precision	0.63	0.74	0.6
Recall	0.56	0.77	0.84

Table 7.9: Classification metrics for all used methods for 10 days

If we are using data from the first 10 days, the Random Forest's Accuracy and Recall drops significantly, probably because of its previous overfitting. Now, Logistical Regression has the highest Accuracy and Precision, and SVM has the highest Recall, which not only has not decreased but has increased.

We already know that according to Random Forest, the most important variable for the classification is screen time or number of sessions, depending on the task. Let's look at what other methods can say:

- Mean coefficients of the Logistical Regression are [9.400e-02 1.438e+00 0.000e+00 -2.000e-03 -2.081e+00] for the whole data and [6.000e-02 1.791e+00 0.000e+00 -1.000e-03 -2.261e+00] for the first 10 days, where the first four are the best features in order 'number_of_sessions', 'regularity', 'screen_time', 'mean_screen_time' and the last one is the bias.
- Mean coefficients of the SVM are [0.037 2.33 0.008 -0.041 -2.281] for the whole data and [0.079 1.847 0.022 -0.173 -2.142] for the first 10 days, where the first four are the best features in order 'number_of_sessions', 'regularity', 'screen_time', 'mean_screen_time' and the last one is the bias.

As we can see, the highest coefficient is the coefficient for the second variable for all cases. This means that both Logistical Regression and SVM agree that Regularity is the most important parameter for classifying our data.

Conclusion

Logistical Regression and Linear kernel SVM are both well-suited for our classification problem with their metrics above 70%. According to them, the most important parameter determining whether the patient will complete the therapy is how regularly they use the application, which is described by our regularity variable.

Chapter 8

Regression

We have tried to estimate our adherence variables with Random Forest regression and Support Vector regression using the variables we got during our EE phase together with our adherence variables. We have used a linear SVM kernel.

Definition

We have made 3 different kinds of experiments with our data:

1. We are trying to estimate one of our adherence variables based on the data gathered during the whole therapy
2. We are trying to estimate one of our adherence variables based on the data gathered during the first 10 days of the therapy without using values for this specific variable, e.g. if we are estimating the screen time, we can not use the screen time for the first 10 days
3. We are trying to estimate one of our adherence variables based on the data gathered during the first 10 days of the therapy

Mean and total screen time were converted to minutes from seconds.

Random Forest

Screen time

Our results for the regression of the screen time look like this:

Experiment	1	2	3
MSE	992	2659	1612
MAE	19	35	26
R2	0.63	-0.04	0.45

Table 8.1: Regression metrics for the RF for the screen time

As we can see, as expected, in our third experiment, we had worse results than in the first one because we lost some data. Our model can not work without knowing the screen time for the first 10 days; its predictive ability is no better than a regular mean value regression.

Mean screen time

Our results for the regression of the mean screen time look like this:

Experiment	1	2	3
MSE	2.3	13	3.7
MAE	0.76	2	1.3
R2	0.75	-0.93	0.6

Table 8.2: Regression metrics for the RF for the mean screen time

As we can see, as expected, in our third experiment, we had worse results than in the first one because we lost some data. Our model can not work without knowing the screen time for the first 10 days; its predictive ability is much worse than a regular mean value regression.

Number of sessions

Our results for the regression of the number of sessions look like this:

Experiment	1	2	3
MSE	152	485	438
MAE	40	14	13
R2	0.48	-0.87	-0.55

Table 8.3: Regression metrics for the RF for the number of sessions

Surprisingly, our model can not predict the number of sessions based only on data we collected during the first 10 days of the therapy. The results of the second and third experiments are much worse than a regular mean value regression.

Regularity

Our results for the regression of the regularity look like this:

Experiment	1	2	3
MSE	0.02	0.11	0.09
MAE	0.1	0.25	0.21
R2	0.75	-0.29	-0.02

Table 8.4: Regression metrics for the RF for the regularity

Surprisingly, our model can not predict the regularity based only on data we collected during the first 10 days of the therapy. The results of the second are worse than a regular mean value regression, and the results of the third are no better than that.

SVM

Screen time

Our results for the regression of the screen time look like this:

Experiment	1	2	3
MSE	374	1988	752
MAE	12	31	19
R2	0.86	0.25	0.68

Table 8.5: Regression metrics for the SVM for the screen time

As we can see, as expected, in our third experiment, we had worse results than in the first one because we lost some data. Our model can work without knowing the screen time for the first 10 days; its predictive ability is better than a regular mean value regression.

Mean screen time

Our results for the regression of the mean screen time look like this:

Experiment	1	2	3
MSE	0.76	8.6	2.2
MAE	0.63	1.6	1.1
R2	0.91	-0.01	0.75

Table 8.6: Regression metrics for the SVM for the mean screen time

As we can see, as expected, in our third experiment, we had worse results than in the first one because we lost some data. Our model can not work without knowing the screen time for the first 10 days; its predictive ability is no better than a regular mean value regression.

Number of sessions

Our results for the regression of the number of sessions look like this:

Experiment	1	2	3
MSE	87	207	236
MAE	5	9	8.5
R2	0.67	0.21	0.25

Table 8.7: Regression metrics for the SVM for the number of sessions

As we can see, as expected, in our third experiment, we had worse results than in the first one because we lost some data. Our model can work without knowing the screen time for the first 10 days; its predictive ability is better than a regular mean value regression.

Regularity

Our results for the regression of the regularity look like this:

Experiment	1	2	3
MSE	0.02	0.09	0.05
MAE	0.1	0.23	0.16
R2	0.76	-0.03	0.43

Table 8.8: Regression metrics for the SVM for the regularity

As we can see, as expected, in our third experiment, we had worse results than in the first one because we lost some data. Our model can not work without knowing the screen time for the first 10 days; its predictive ability is no better than a regular mean value regression.

Comparison of the two methods

Let's look at the results of both methods for every parameter we have tried to estimate:

Screen time

Experiment	RF 1	RF 2	RF 3	SVM 1	SVM 2	SVM 3
MSE	992	2659	1612	374	1988	752
MAE	19	35	26	12	31	19
R2	0.63	-0.04	0.45	0.86	0.25	0.68

Table 8.9: Regression metrics for the RF and SVM for the screen time

SVM had better results in all experiments, especially in the second and third ones; unlike RF, it became at least usable in the second experiment.

Let's visualise the results:

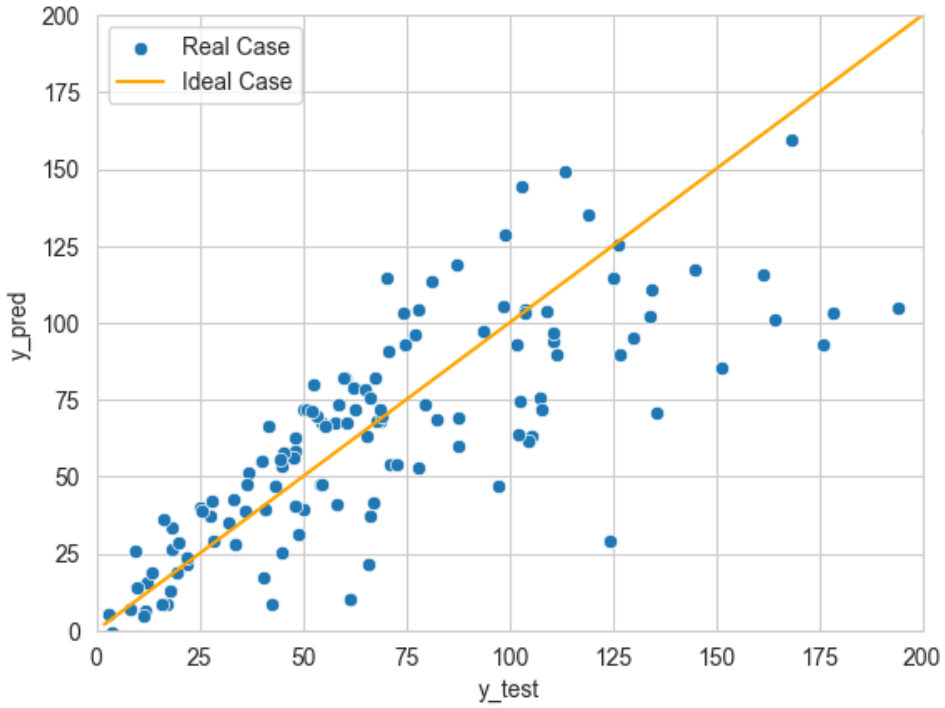


Figure 8.1: Support Vector regression for the screen time for all data

On the x-axis, we can see the values in the test set we have tried to estimate; on the y-axis, we can see the values estimated by our model. In the ideal case, we want to see a straight line.

Our estimated data are initially close to the ideal line but start to deviate with higher values.

Mean screen time

Experiment	RF 1	RF 2	RF 3	SVM 1	SVM 2	SVM 3
MSE	2.3	13	3.7	0.76	8.6	2.2
MAE	0.76	2	1.3	0.63	1.6	1.1
R2	0.75	-0.93	0.6	0.91	-0.01	0.75

Table 8.10: Regression metrics for the RF and SVM for the mean screen time

SVM had better results in all experiments, especially in the second and third ones.

Let's visualise the results:

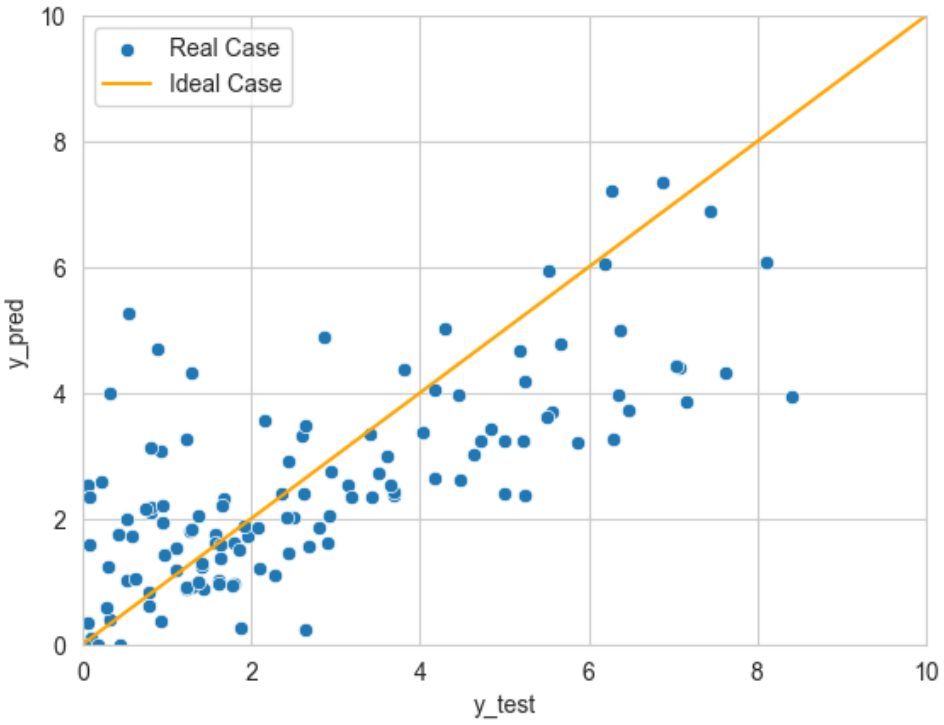


Figure 8.2: Support Vector regression for the mean screen time for all data

On the x-axis, we can see the values in the test set we have tried to estimate; on the y-axis, we can see the values estimated by our model. In the ideal case, we want to see a straight line.

Our estimated data are initially close to the ideal line but start to deviate with higher values.

Number of sessions

Experiment	RF 1	RF 2	RF 3	SVM 1	SVM 2	SVM 3
MSE	152	485	438	87	207	236
MAE	40	14	13	5	9	8.5
R2	0.48	-0.87	-0.55	0.67	0.21	0.25

Table 8.11: Regression metrics for the RF and SVM for the number of sessions

SVM had better results in all experiments, especially in the second and third ones; unlike RF, it became at least usable.

Let's visualise the results:

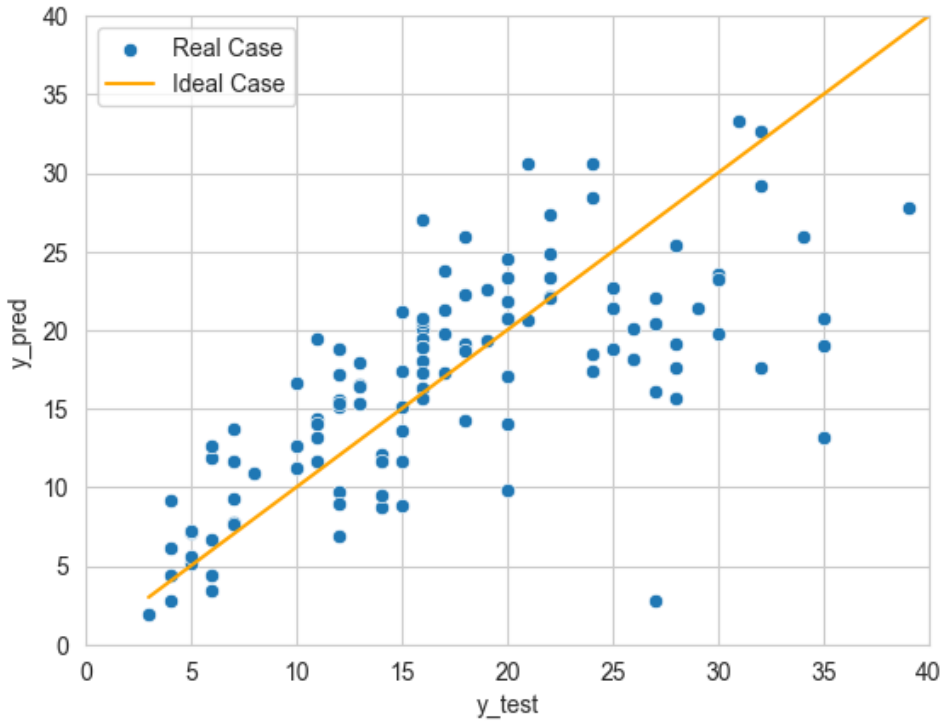


Figure 8.3: Support Vector regression for the number of sessions for all data

On the x-axis, we can see the values in the test set we have tried to estimate; on the y-axis, we can see the values estimated by our model. In the ideal case, we want to see a straight line.

Our estimated data are initially close to the ideal line but start to deviate with higher values.

Regularity

Experiment	RF 1	RF 2	RF 3	SVM 1	SVM 2	SVM 3
MSE	0.02	0.11	0.09	0.02	0.09	0.05
MAE	0.1	0.25	0.21	0.1	0.23	0.16
R2	0.75	-0.29	-0.02	0.76	-0.03	0.43

Table 8.12: Regression metrics for the RF and SVM for the regularity

SVM had better results in all experiments, especially in the second and third ones.

Let's visualise the results:

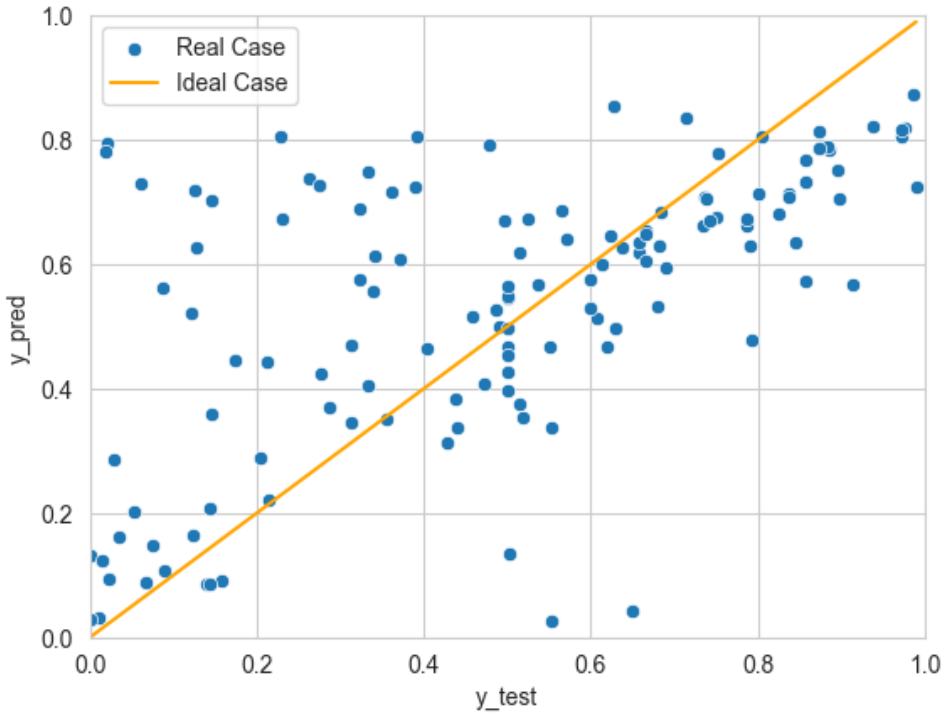


Figure 8.4: Support Vector regression for the regularity for all data

On the x-axis, we can see the values in the test set we have tried to estimate; on the y-axis, we can see the values estimated by our model. In the ideal case, we want to see a straight line.

Our estimated data are close to the ideal line for the higher values but start to deviate with lower values.

Conclusion

Support Vector regression had overall better results than a Random Forest regression. Mean screen time is the easiest variable to predict. Number of sessions is the hardest variable to predict. Even though our four variables are highly correlated, it is very hard to predict the value of the chosen variable without knowing its values during the first 10 days; it is impossible for the Random Forest.

Chapter 9

Conclusion

We have introduced 4 new adherence variables - number of sessions, total screen time, mean screen time and regularity. These variables are correlated with each other but not with other variables from our dataset. We can use that for our classification and regression tasks.
Our exploratory part tells us that there is almost all correlation between chosen demographic factors. The only weak correlations we found, which weren't caused by the data leakage, were correlations with sex and the age at which the patient started smoking. Our analysis shows that women are more likely to complete the therapy and that the age at which a patient started smoking influences men and women differently - if the woman started smoking earlier, it is more likely she will complete the therapy, and the opposite applies to men. Previous studies show that one variable may positively and negatively influence success or have no effect at all, depending on different studies.
In our classification task, we tried to predict if the patient would complete the therapy based on the chosen parameters. We have used the Random Forest classifier to understand which variables are the most important for the successful prediction. It happened that the 4 most important features are our adherence variables, which is great because it means that our work makes sense. We have used these variables for classification with SVM and Logistical Regression. They have better results in classification than Random Forest - higher accuracy, precision and recall. We have also tried to predict the success of the therapy based on only the patient's behaviour during the first 10 days of the therapy - because we have lost some data, we got worse results, even though it still can be successfully used for the classification. Both Logistical Regression and SVM agree that regularity is the most important factor in completing the therapy.
In our regression task, we tried to predict different adherence variables based on the values of other adherence variables using Random Forest Regression and SVM Regression. SVM had better results than the RF. It is hard to predict one variable's value if we do not know it during the first 10 days of therapy, even though our adherence variables are correlated.

Appendix A

Bibliography

[1] David Kolečkář, Master's thesis "Visualization and analysis of patients digital phenotypes", 2023

[2] Esen et al. "Factors affecting success and abstinence within a smoking cessation clinic: A one-year follow-up study in Turkey", Tobacco Prevention and Cessation, 2020, doi: https://doi.org/10.18332/tpc/130471

[3] Martins et al., "Factors motivating smoking cessation: a cross-sectional study in a lower-middle-income country", BMC Public Health, 2021, doi: https://doi.org/10.1186/s12889-021-11477-2

[4] Bacha et al., "Factors associated with smoking cessation success in Lebanon", Pharmacy Practice, 2018, doi: https://dx.doi.org/10.18549/pharmpract.2018.01.1111

[5] Esmer et al., "Clinical and demographical factors influencing smoking cessation rates", European Respiratory Journal, 2016, doi: https://doi.org/10.1183/13993003.congress-2016.PA1181

[6] Ucar et al., "Effectiveness of pharmacologic therapies on smoking cessation success: three years results of a smoking cessation clinic", Multidisciplinary Respiratory Medicine, 2014, https://doi.org/10.1186/2049-6958-9-9

[7] Wu, P., Wilson, K., Dimoulas, P. et al. "Effectiveness of smoking cessation therapies: a systematic review and meta-analysis." BMC Public Health 6, 300 (2006). https://doi.org/10.1186/1471-2458-6-300

[8] Interventions for Tobacco Smoking Cessation in Adults, Including Pregnant Persons: Recommendation Statement. Am Fam Physician. 2021 Jun 15;103(12):Online. PMID: 34128620.

[9] Choi et al., "The current state of tobacco cessation treatment", Cleveland Journal of Medicine, 2021, https://doi.org/10.3949/ccjm.88a.20099

[10] Yang et al., "What Are the Major Determinants in the Success of Smoking Cessation: Results from the Health Examinees Study", PLOS One, 2015, https://doi.org/10.1371/journal.pone.0143303

[11] Bullard, T., Ji, M., An, R. et al. A systematic review and meta-analysis of adherence to physical activity interventions among three chronic conditions: cancer, cardiovascular disease, and diabetes. BMC Public Health 19, 636 (2019). https://doi.org/10.1186/s12889-019-6877-z

[12] Baumel, A., & Yom-Tov, E. (2018). Predicting user adherence to behavioral eHealth interventions in the real world: examining which aspects of intervention design matter most. Translational Behavioral Medicine. doi:10.1093/tbm/ibx037

[13] Donkin L, Christensen H, Naismith SL, Neal B, Hickie IB, Glozier N A Systematic Review of the Impact of Adherence on the Effectiveness of e-Therapies J Med Internet Res 2011;13(3):e52 doi:10.2196/jmir.1772

Appendix B

Unused graphs

This Appendix contains unused graphs that show no dependency between variables and, therefore, were moved there so they won't take up space in the main part of the thesis.

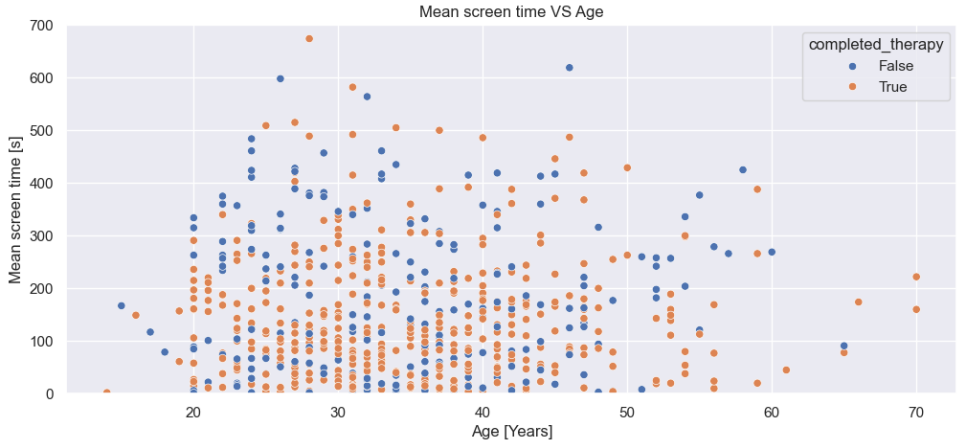


Figure B.1: Mean screen time vs age

There is no clear linear pattern or correlation between age and mean screen time.

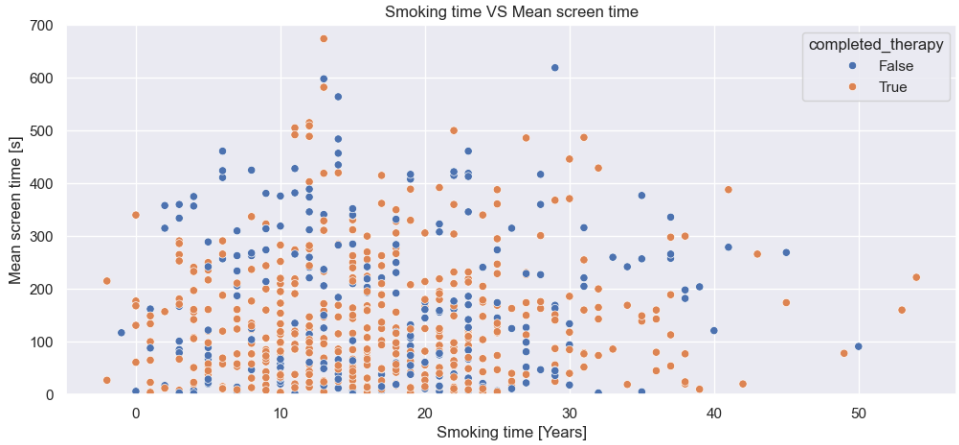


Figure B.2: Smoking time vs Mean screen time

There is no clear linear relationship between smoking time and mean screen time.

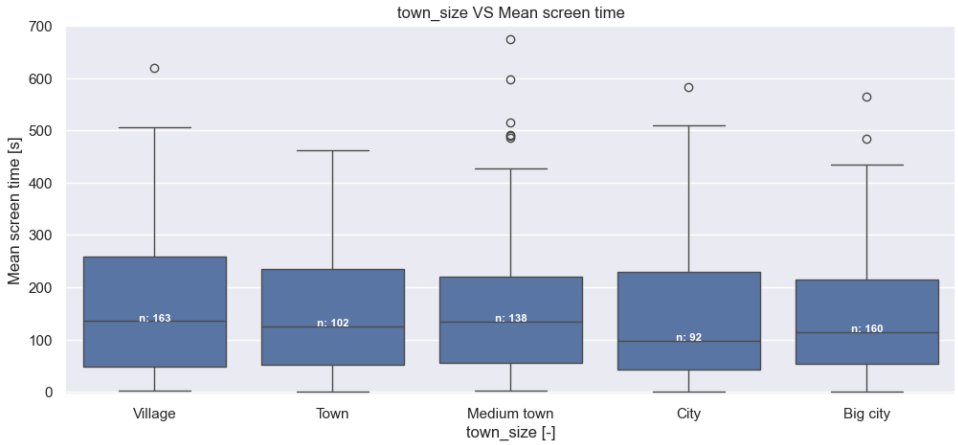


Figure B.3: Town size vs Mean screen time

There is no connection between the town size and mean screen time.

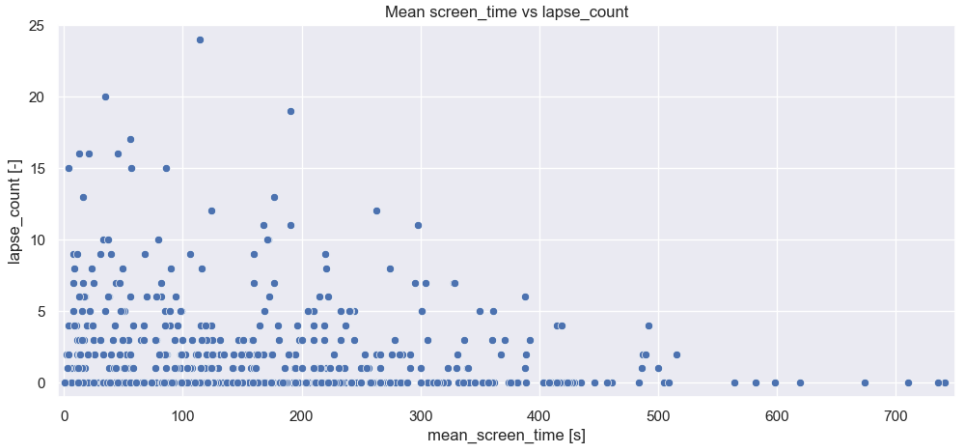


Figure B.4: Mean screen time vs Lapse count

There is no connection between mean screen time and lapse count.

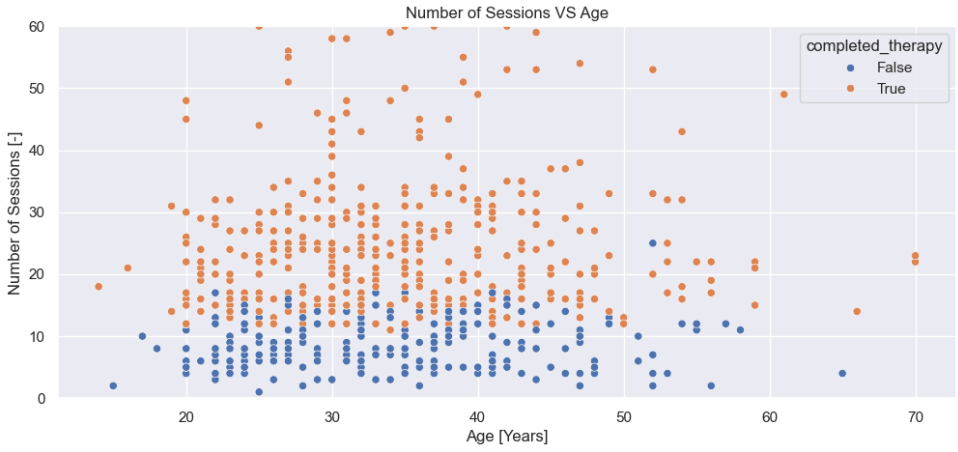


Figure B.5: Number of sessions vs age

There is no connection between the number of sessions completed and age.

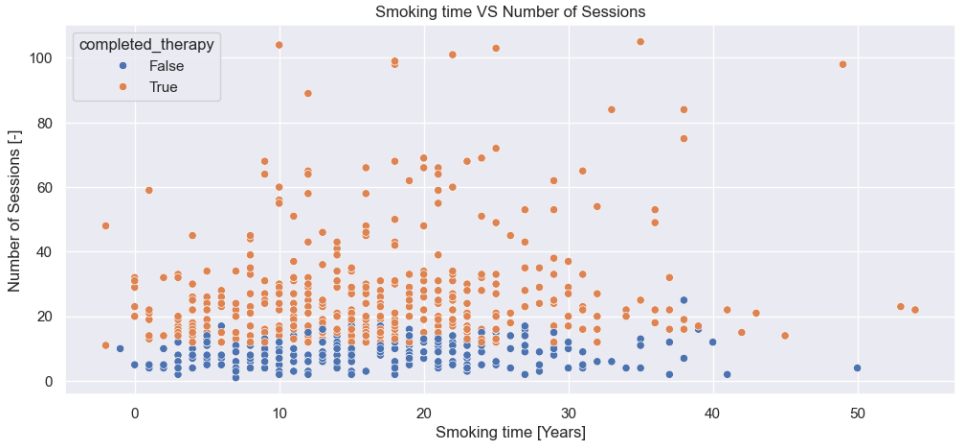


Figure B.6: Smoking time vs Number of sessions

Despite the attempt to show a relationship between smoking time and the number of sessions, there is no clear trend visible.

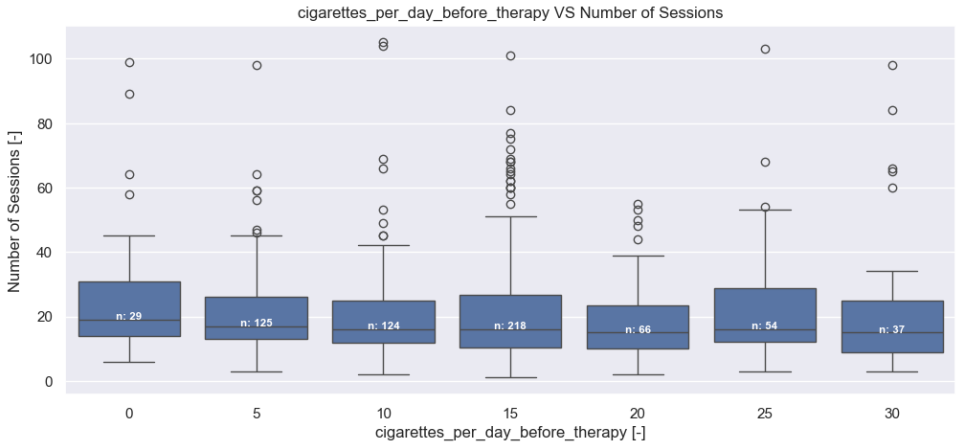


Figure B.7: Cigarettes per day before therapy vs Number of sessions

There is no connection between the number of sessions completed and how many cigarettes the patient has smoked before.

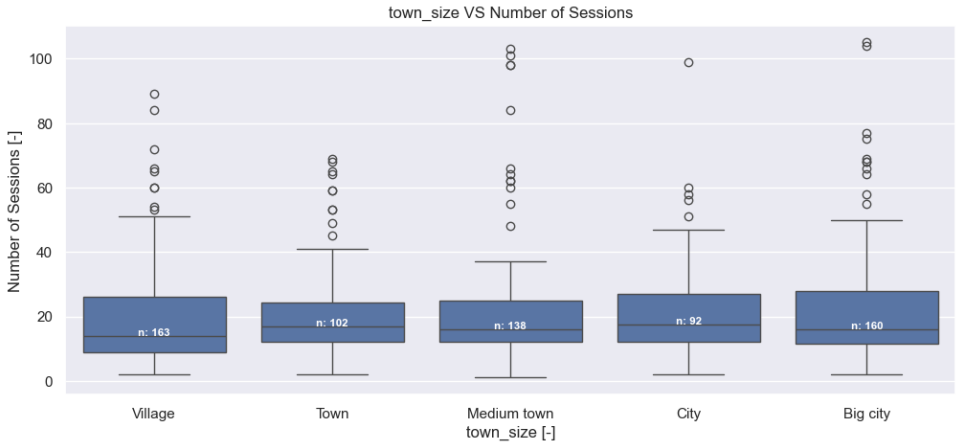


Figure B.8: Town size vs Number of sessions

There is no connection between the number of sessions completed and town size.

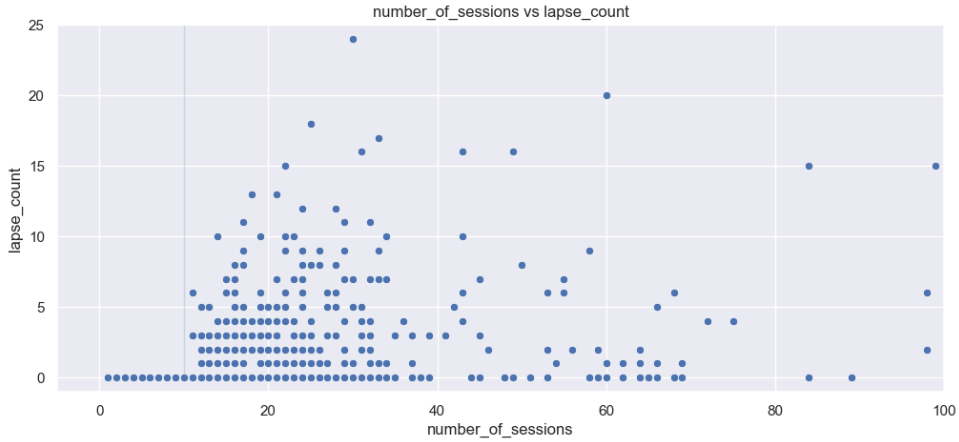


Figure B.9: Number of sessions vs Lapse count

There is no clear connection between the number of sessions and the lapse count.

Figure B.10: Regularity vs age

There is no connection between the regularity and age.

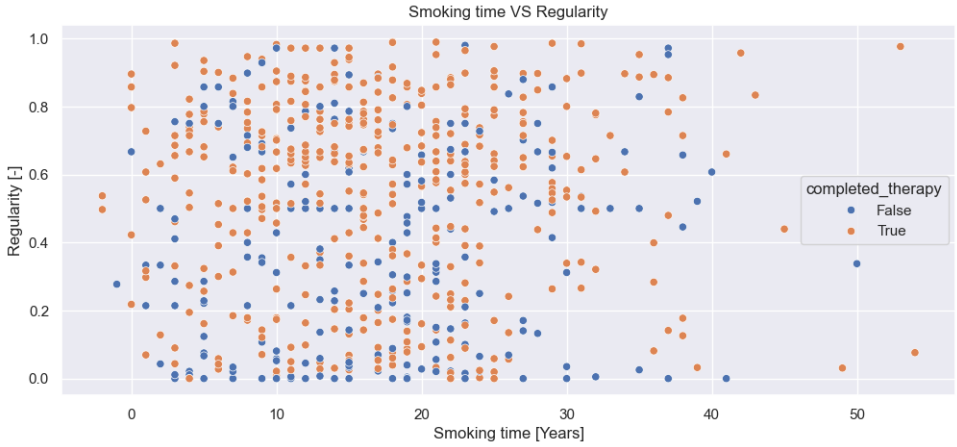


Figure B.11: Smoking time vs Regularity

There is no connection between the regularity and smoking time.

Figure B.12: Town size vs Regularity

There is no connection between regularity and town size.

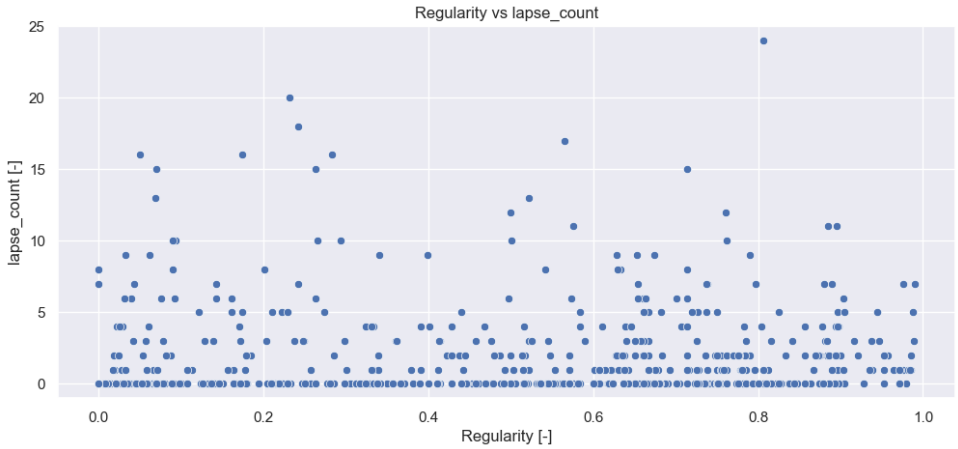


Figure B.13: Regularity vs Lapse count

There is no clear connection between regularity and lapse count.