

**Data Driven Decision Making**

*Addiction Treatment*

*Student Name(s):*

*Alex Brulliea, Tyler Staffin, Labib Nazer, Alvin Mathew, James Grissel*

**School of Graduate Professional Studies**

MPS/MS in Data Analytics

DAAN 881 – Data Driven Decision Making

Spring 2022

**Document Control**

**Work carried out by:**

|  |  |  |
| --- | --- | --- |
| **Name** | **Email Address** | **Task description** |
| Labib Nazer | lfn5029@psu.edu |  |
| Alex Brulliea | ajb6058@psu.edu |  |
| Alvin Mathew | afm6089@psu.edu |  |
| Tyler Staffin | tms5882@psu.edu |  |
| James Grissel | jmg7446@psu.edu |  |

**Revision Sheet:**

|  |  |  |
| --- | --- | --- |
| **Release No.** | **Date** | **Revision Description** |
| Deliverable 1 | 01/23/2022 | Defined the research goal of “Can data-driven techniques be leveraged to increase patient engagement on the SPARx platform?” Identified the queries that should be addressed, and identified the sources to be used for the analysis: IGNTD SPARx platform data (page 2) |
| Deliverable 2 | 2/6/2022 | Created table of attribute names, data types, number of instances, sample values, number of missing values, and measurement scale (see excel document attached). Created a [private GitLab Repo](https://github.com/labibnazer/DAAN-881-Project) and imported all raw data files into an Azure MS SQL database so that it would be accessible by all team members. Updated the queries (page 2 paragraph 2) to re-imagine our approach based on the reading. |
| Deliverable 3 | 2/20/2022 | Updated Page 3 Data Description section to move the original graphic (summarizing data types) to the appendix and add statistical summary and NA outputs. Created Exploratory Analysis section to call out our features of importance for the proposed research goal and provided supporting graphics. Created a Missing Values section to present the analysis and visualization of missing values in the data for our selected variables. Created a Data Issues section on to call current known issues and a plan for fixing these issues in the next deliverable. |
| Deliverable 4 | 3/6/2022 | Performed data cleansing activities addressed in Deliverable 3 to increase the quality of the data prior to modeling using both SQL and R. Created tables to capture conversion to Boolean variable types in the intake\_assessment and sds\_user\_activations tables. Provided the correlation matrix between intake assessment values which will be utilized for determining engagement. |
| Deliverable 5 | 3/27/2022 | Performed various feature selection techniques to determine which features to use in modeling efforts. Narrowed down the potential modeling approaches for the model generation phase. |
|  |  |  |
|  |  |  |

**Table of Contents**

[Introduction 2](#_Toc99309032)

[Data Description 3](#_Toc99309033)

[Exploratory Analysis 4](#_Toc99309034)

[Missing Values 10](#_Toc99309035)

[Data Issues 12](#_Toc99309036)

[Data Cleaning 13](#_Toc99309037)

[Feature Selection 16](#_Toc99309038)

[Data Modelling Plan 25](#_Toc99309039)

[Deliverable 6 Assignment 26](#_Toc99309040)

[Deliverable 7 Assignment 27](#_Toc99309041)

[Appendix A 28](#_Toc99309042)

[Appendix B 31](#_Toc99309043)

[Appendix C 34](#_Toc99309044)

[References 35](#_Toc99309045)

# Introduction

The global opioid epidemic is challenging our current understanding of managing and treating addiction, posing a significant financial burden and strain on our already over-utilized medical systems. As is the case with other modern challenges, the introduction of data-driven technology could address this issue and improve the likelihood of long-term patient success. SPARx, a Smart and Adaptive Recovery System developed by IGNTD, utilizes machine learning and customized programming to help individuals throughout the addiction recovery process. SPARx aims “to transform the current addiction industry [by promoting] empowerment, solution-oriented and compassionate recovery.” (*IGNTD Announces The Launch of SPARx, A Smart Personalized Adaptive Recovery System That Finally Fixes The Failing One-Size-Fits-All System That is Killing More People Every Year 2*021). Long-term platform usage is critical to the success of all addiction patients; thus, it is essential that IGNTD continues to identify and implement methods for improving user engagement. This research will serve to address the viability of the SPARx platform through the scope of the following question: Can data-driven techniques be leveraged to increase patient engagement on the SPARx platform?

|  |
| --- |
| **Business Question** |
| What content should we be offering users of the platform? |
|  |
| **Descriptive Question** |
| What are the platform usage patterns? How do different types of users engage? |
|  |
| **Predictive Question** |
| Given each user's history, how are they more likely to engage with the platform next? |
|  |
| **Prescriptive Question** |
| What are the best types of content to present to ensure patients are engaging with the platform and progressing in their road to recovery? |

These questions were formed utilizing the outline from *Analytical AI and Data Science* (Vaughn, 2020). They will help to support the overall research goal more effectively.The source which has been identified to answer the above business questions is the IGNTD SPARx platform data.

# Data Description

The Data Dictionary for the database and underlying columns can be found in [Table 1.1](#_Appendix) in the Appendix.

In this worksheet, there are 2 tabs. The first is all the column-level names, definitions, measurement scales, data types, null counts, relevance flag (for the analysis), and any associated notes with the column. The second is all the table-level names, definitions, and row counts.

We loaded all data files sent over by the IGNTD team into an Azure MS SQL relational database to provide accessibility to all team members and establish an environment for query execution. We have also set up a [private GitHub repository](https://github.com/labibnazer/DAAN-881-Project) to manage version control for all code employed for the remainder of this project.

The data provided by IGNTD consists of 10 delimited text files. These files fit the profile for data extracted via IGNTD’s web application relational database in that they are normalized with primary and foreign key references. There are also numerous metadata fields including record creation and update date/time stamps.

The tables contain logs of the users’ assessments and engagements while using the SPARx platform. Of the available data fields, 29 were initially have been identified as key fields, which are integer values used to uniquely define records and make references to one or more records in other tables. Although they are essential to the database model, these fields will not be used in analysis.

Please find a table with the volume of empty strings of each dataset as presented in Table 1 below the (note that all empty strings and NULL values were converted to NA). The statistical summary of all the fields by table in the dataset are shown in the Figure 6 in the Appendix.

Table 1: Selected Predictor Variables

|  |  |
| --- | --- |
| **Table** | **Number of empty strings** |
| completed\_lessons | 2280 |
| daily\_assessment | 2214 |
| intake\_assessment | 6208 |
| lesson\_ratings | 506 |
| sds\_intake\_question\_options | 304 |
| sds\_user\_activations | 3078 |
| site\_visits | 9066 |
| user\_points | 0 |
| user\_rewards | 0 |
| wol\_items | 0 |

# Exploratory Analysis

When performing any statistical analysis or developing models, a crucial preliminary step involves thorough quantitative and qualitative analysis of all available data. The Cross-Industry Standard Process for Data Mining (CRISP-DM) advocates for the importance of proper data understanding prior to pursuing any data preparation or model development steps. The initial dataset consists of 13 tables, each of which contains a ranging number of variables and records. After further assessment of each variable and its relevance to the research question, the team decided to focus on the following 18 features as presented in Table 1:

Table 2: Selected Predictor Variables

|  |  |  |
| --- | --- | --- |
| **Table** | **Variable** | |
| **Name** | **Description** |
| intake\_assessment | age | User age |
| sds\_user\_activations | gender | User gender |
| intake\_assessment | days\_drink\_per\_week | Days per week user drinks |
| intake\_assessment | days\_cannabis\_per\_week | Days per week user uses cannabis |
| intake\_assessment | other\_drugs\_per\_week | Days per week user uses other drugs |
| intake\_assessment | gambling\_per\_week | Days per week user gambles |
| intake\_assessment | nicotine\_per\_week | Days per week user uses nicotine |
| intake\_assessment | bothered\_by\_mental\_health | Scale of how bothered user is by mental health habits from 0 to 10 |
| sds\_user\_activations | welcome\_video | If user watched welcome video |
| sds\_user\_activations | bought\_hero | If user bought Hero package |
| sds\_user\_activations | intake | If user has finished an intake assessment |
| sds\_user\_activations | daily\_assessment | If user has completed a daily assessment |
| sds\_user\_activations | started\_lesson | If user has started a lesson |
| sds\_user\_activations | upload\_worksheet | If user has uploaded their worksheet |
| sds\_user\_activations | rated\_lesson\_1 | If user has rated the first lesson |
| sds\_user\_activations | finished\_module\_1 | If user has finished the first module |
| sds\_user\_activations | created\_community\_acct | If user has created a community account |
| sds\_user\_activations | completed\_lesson | If user has completed a lesson |

As described in the prior section, the team developed an Azure MS SQL relational database to properly store and maintain all provided data. The R statistical computing language was utilized in this phase and for the remaining steps of the CRISP-DM process. The team prepared histograms for each continuous predictor using ggplot2 to derive an improved quantitative perspective of their underlying distributions, as displayed in Figure 2. Since most of the instances for variables capturing “number of days” were zeros, the distributions were significantly skewed. To address this (and for visual clarity), all zero values were removed from the histogram plots, but not removed from the underlying data.

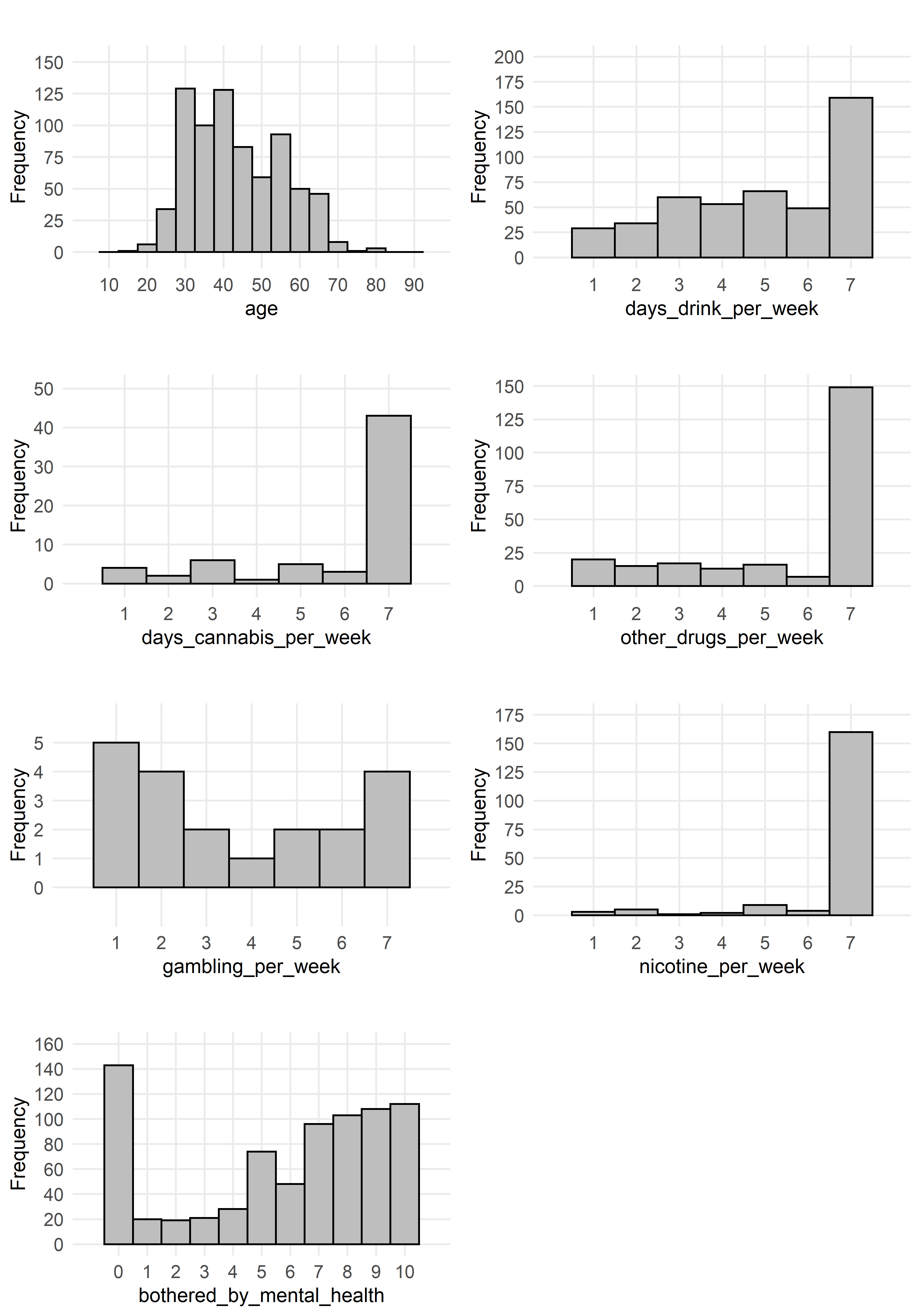


Figure 1: Histograms of Continuous Predictors

As mentioned above, most of the variables expressed frequency values of 0. This is likely due to a combination of both true and inaccurate responses, as individuals may be reluctant to share truthful data about their drug, alcohol, and gambling habits. Although the data for these self-reported variables is contingent upon the validity of the user responses, it is critical that all subsequent analysis and model evaluation processes operate under the assumption that the information recorded by users of the SPARx platform is accurate.

Evidently, the age of SPARx platform users varies widely, with the lowest and highest recorded ages (excluding zeros) are 16 and 79, respectively. The average and median ages are currently 41.91 and 41 years, respectively, with a standard deviation of approximately 14.1 years. It appears that most platform users fall between 30 and 40 years of age, but there is a large volume of users with ages on the higher and lower ends of the distribution. Through the model development and evaluation process, the team will certainly explore the impact of user age on their continued engagement with the SPARx platform. The summary statistics for the remaining continuous variables were computed and displayed as follows in Table 2:

Table 2: Summary Statistics for Continuous Predictors

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **Min.** | **Q1** | **Med.** | **Mean** | **Q3** | **Max.** | **SD** |
| age | 0 | 32 | 41 | 41.91 | 53 | 79 | 25.89 |
| days\_drink\_per\_week | 0 | 0 | 2 | 2.84 | 6 | 7 | 2.97 |
| days\_cannabis\_per\_week | 0 | 0 | 0 | 0.49 | 0 | 7 | 2.82 |
| other\_drugs\_per\_week | 0 | 0 | 0 | 1.73 | 3 | 7 | 2.76 |
| gambling\_per\_week | 0 | 0 | 0 | 0.1 | 0 | 7 | 2.85 |
| nicotine\_per\_week | 0 | 0 | 0 | 1.58 | 0 | 7 | 2.8 |
| bothered\_by\_mental\_health | 0 | 3 | 7 | 5.8 | 9 | 10 | 3.76 |

Additionally, there are several Boolean variables the team identified as being potentially relevant for answering the business question. Although the original dataset did not characterize these variables as Boolean, the team modified their variable classes to accurately reflect their types. While the first factor is gender (descriptive of the user), the remaining variables relate to the engagement with and usage of the SPARx platform. The bar plot in Figure 3 captures these factors in terms of their relative percentage of the total number of instances:

Timeline

Description automatically generated

Figure 2: Bar Plot of Boolean Variables and % of "True" Values

The above bar plot provides some interesting initial observations into the refined data subset. It appears that roughly two-thirds of the platform users are female, with a population consisting of 292 females and 140 males. 68 instances of gender were removed due to NULL or “other” values (specifically, 67 NULL and 1 “other”). Most users completed an intake assessment (~88%), and just over two-thirds of users reported having completed at least one daily assessment. While 65% of users started the first lesson, a much smaller number of participants completed lesson in its entirety. Just about 15% of platform users successfully completed of the first module. Throughout this research, the team will look to identify which factors contributed to longer platform engagement to ensure that users continue to monitor their treatment process with the SPARx platform.

For correlations between values, we decided to focus on the numeric values of the intake assessment. Based on the values, there were no outstanding unexpected correlations negatively or positively. 2 correlations of interest, however, were nicotine\_per\_week and cigarettes\_per\_day both positively correlated with sex\_issues\_per\_week and other\_drugs\_per\_week. Based on that correlation, we could make the assumption that those who have sex issues and other drug use tend to also have a reliance on nicotine. The full correlation matrix can be seen in figure 4 below, however, a more easily readable document called CorrelationMatrix.xlsx has been attached for reference.

Graphical user interface

Description automatically generated

Figure 3: Correlation Matrix of intake\_assessment

# Missing Values

To visualize missing values in the data, any NULL instances were converted to NA values and plotted. The tables sds\_user\_activations and intake\_assessment contain the variables selected for analysis, and the number of present and missing values for all variables in these tables is captured in Figure 4 below. We can see that many of these fields contain missing data and that some fields are missing data in the majority of instances.

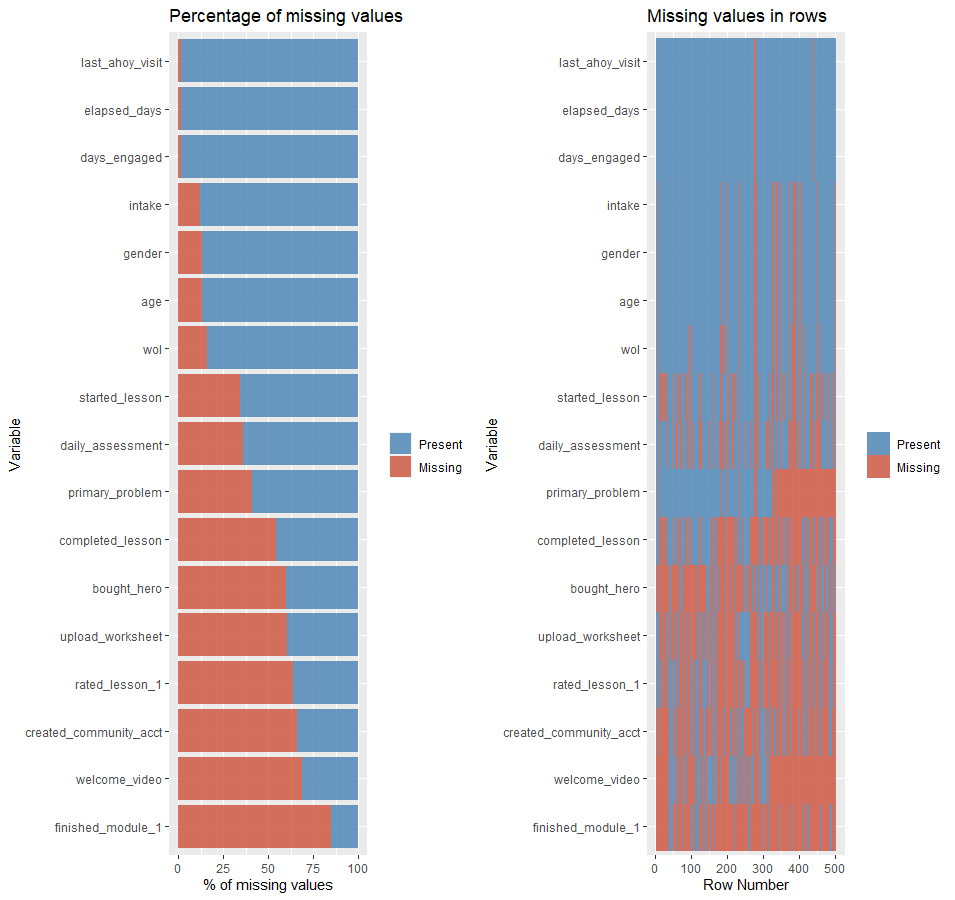


Figure 4: Bar Plot of Missing Values in sds\_user\_activations by Variable and Row

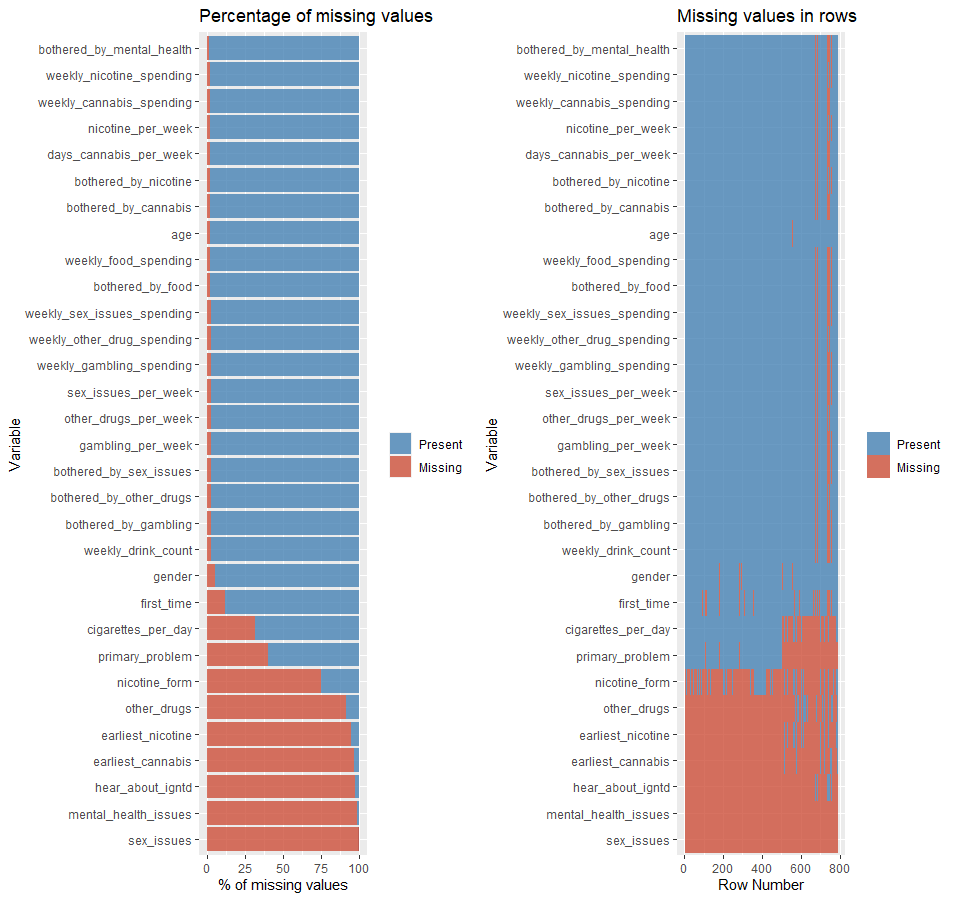


Figure 5: Bar Plot of Missing Values in intake\_assessment by Variable and Row

# Data Issues

The following list captures the data issues identified throughout the initial phases of this research. Additionally, potential solutions are listed below each issue and will be modified accordingly as solutions are implemented in future deliverables:

1. Boolean variables listed as “NA” instead of 0
   * To address this issue, the team with utilize R’s is.na() function on all boolean variables and set all instances where is.na() is TRUE to 0. This should resolve this data issue.
2. Boolean data is inconsistently recorded
   * For example, the first\_time field has “TRUE” and “FALSE” values in the intake\_assessment table, whereas the bought\_hero attribute in the sds\_user\_activations tables contains either a “1” or “NULL”. The team will need to standardize these values in the data.
3. NULL values prevalent in data
   * NULL values are represented in inconsistent ways in the character variables. For example, the mental\_health\_issues field has four different representations of NULL, including “NULL”, “{NULL}”, “{}”, and “”. The team will need to implement a consistent pattern for representing NULL values across all variables.
   * NULL values are represented as “NULL” character values in columns that should only contain integers. For example, the fields bought\_hero and upload\_worksheet demonstrate this issue, while others including wol and daily\_assessment do not. The team will need to implement a consistent solution for representing NULLs across all integer fields.
4. Outlier values in data (i.e., age has values of 0, etc.)
   * Outliers will be addressed on a case-by-case basis, with potential systematic solutions resulting from outlier detection methods.
5. Some of the tables have entries for user\_id (1629, 1630, and 1631) that do not exist in sds\_user\_activations, which should contain all users. The team will need to decide whether to omit this data entirely or create dummy user\_id values.
6. The intake\_assessment table contains more than one response in some cases for specific users. After clarification from our collaborators, we now understand that this is intended within the data, as intake assessments are supposed to occur monthly for each user.
7. sds\_intake\_question\_options was provided as a table with enum descriptions that give information about what the data in the tables refer to. These will need to be mapped to the corresponding data fields in the intake\_assessment table.

# Data Cleaning

To address the previously mentioned data issues in Deliverable 3, the team has taken the following actions to ensure the issues are mitigated using data cleansing techniques in both SQL and R.

1. NA values in Boolean variables were set to zeroes.
2. TRUE/FALSE values were converted to dummy/binary values for consistency with Boolean variables within the IGTND platform.
3. NULL values across the tables were inconsistent based on the IGTND platform’s front-end usage, therefore the team normalized the NULL values irrespective of the data type.

|  |  |  |
| --- | --- | --- |
| **SDS\_USER\_ACTIVATION**  **(Pre- Boolean conversion)** | | |
| Column | Column Value | Count |
| bought\_hero | 1 | 201 |
| bought\_hero | NULL | 299 |
| completed\_lesson | NULL | 273 |
| completed\_lesson | 1 | 227 |
| created\_community\_acct | 1 | 170 |
| created\_community\_acct | NULL | 330 |
| daily\_assessment | NULL | 179 |
| daily\_assessment | 1 | 321 |
| finished\_module\_1 | 1 | 75 |
| finished\_module\_1 | NULL | 425 |
| intake | NULL | 61 |
| intake | 1 | 439 |
| rated\_lesson\_1 | 1 | 182 |
| rated\_lesson\_1 | NULL | 318 |
| started\_lesson | NULL | 173 |
| started\_lesson | 1 | 327 |
| upload\_worksheet | 1 | 197 |
| upload\_worksheet | NULL | 303 |
| welcome\_video | 1 | 157 |
| welcome\_video | NULL | 343 |

|  |  |  |
| --- | --- | --- |
| **INTAKE\_ASSESSMENT**  **(Pre-Boolean conversion)** | | |
| Column | Column Value | Count |
| First\_Time | TRUE | 551 |
| First\_Time | NULL | 93 |
| First\_Time | FALSE | 141 |

|  |  |  |
| --- | --- | --- |
| **SDS\_USER\_ACTIVATION**  **(Post Boolean conversion)** | | |
| Column | Column Value | Count |
| bought\_hero | 0 | 299 |
| bought\_hero | 1 | 201 |
| completed\_lesson | 0 | 273 |
| completed\_lesson | 1 | 227 |
| created\_community\_acct | 0 | 330 |
| created\_community\_acct | 1 | 170 |
| daily\_assessment | 0 | 179 |
| daily\_assessment | 1 | 321 |
| finished\_module\_1 | 0 | 425 |
| finished\_module\_1 | 1 | 75 |
| intake | 0 | 61 |
| intake | 1 | 439 |
| rated\_lesson\_1 | 0 | 318 |
| rated\_lesson\_1 | 1 | 182 |
| started\_lesson | 0 | 173 |
| started\_lesson | 1 | 327 |
| upload\_worksheet | 0 | 303 |
| upload\_worksheet | 1 | 197 |
| welcome\_video | 0 | 343 |
| welcome\_video | 1 | 157 |

|  |  |  |
| --- | --- | --- |
| **INTAKE\_ASSESSMENT**  **(Post-Boolean conversion)** | | |
| Column | Column Value | Count |
| First\_Time | 1 | 551 |
| First\_Time | 0 | 234 |

The table sds\_intake\_question\_options contains text enum mappings that describe the data in the table intake\_assessment for gender, primary\_problem, and nicotine\_form. However, there are many observations in intake\_assessment with NULL values. To handle this problem, we created a SQL view titled intake\_assessment\_question\_texts and will leverage it as a replacement for the intake\_assessment raw data table. The SQL statement for this view is in Appendix B.

The team developed a script in R (Appendix C) to handle the remaining issues regarding NULL and missing (NA) values. The first step was to identify the columns in the intake\_assessment\_question\_texts view and sds\_user\_activations that contained NULLs, and then treat them as missing values. A function titled get\_null\_cols() was developed to search for NULL values (i.e., NULL, {NULL}, and {}) in these tables to determine where to resolve this problem. NULL values were only present in character variables, so the first step was to convert all NULL values to NAs. The presence of NULLs in numeric variables caused the values to be imported as characters, thus the classes of these variables were converted back to numeric. The class of the first\_time variable in the intake\_assessment\_question\_texts view was converted to logical to enable the proper conversion to the numeric variable type.

The next step involved the conversion of missing values in Boolean variables with zeroes, as the unique values for these variables only consisted of 1s and NAs. Lastly, zeroes were regarded as outliers in the age column of the intake\_assessment\_question\_texts view and replaced with NAs. All steps were necessary to ensure adequate preparation of the data prior to performing any model development-related steps.

# Feature Selection

One of the major steps in preparation for model development is the identification and selection of variables. Not only does reducing the number of variables increase the speed of the modelling process, it can also reduce noisiness while improving both model performance and accuracy. The team employed various feature selection and engineering approaches in this step, including Principal Component Analysis (PCA), Exploratory Factor Analysis (EFA), Confirmatory Factor Analysis (CFA), and the Random Forest Model (via the Boruta R package). The following sections outline these preparation procedures in further detail.

## Principal Component Analysis (PCA)

As discussed in Lecture 5, the use of PCA helps to reduce the amount of correlation in the underlying dataset through the elimination of redundancy. The team decided to leverage the Eigenvalue-One Criterion, in which principal components with eigenvalues greater than one are maintained, and the other principal components are discarded. The Proportion of Variance approach was also implemented, but required that a larger number of principal components were maintained (using a 90% variance threshold).

These computed principal component enable significant dimension reduction for this data subset. As a result, we can decrease the number of variables in this subset from 23 to 8, without drastically impacting the amount of explained variance. These resulting principal components will be combined with the remaining variables in the prepared data set and then used in the subsequent model development phase. Figure 6 graphically depicts the eigenvalues of each principal component, and distinguishes signficiant eigenvalues by color.

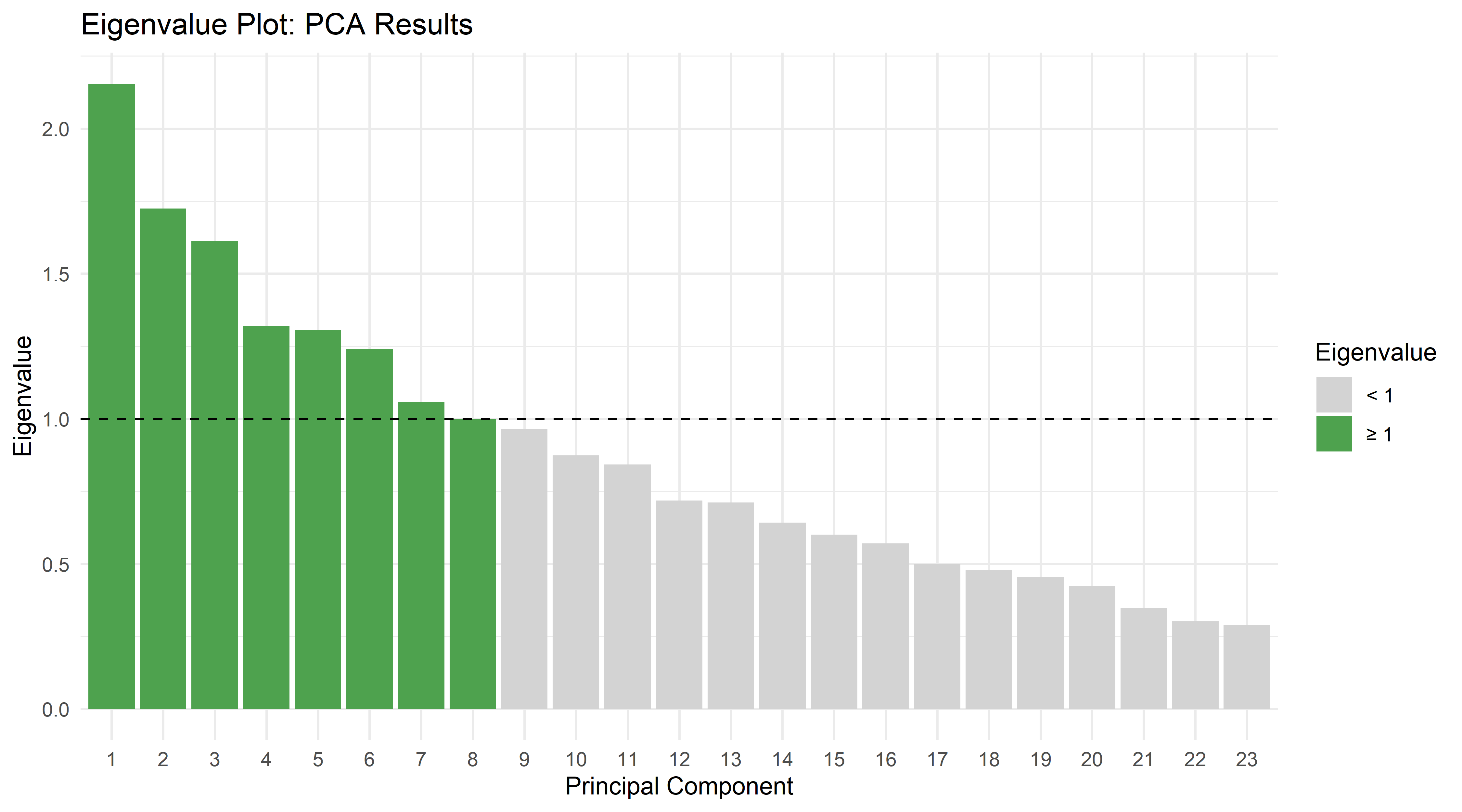


Figure 6: PCA Eigenvalue Plot

## Random Forest Model (Boruta Algorithm)

Random Forest models can be applied to the variable reduction process. The model identifies the importance of each variable in the decision-making process of the algorithm, outputting a numerical value of importance to how each variable interacted with the output. For this, we utilized the Boruta algorithm, which is a wrapper around Random Forest. We had identified factors of potential relevance and chose to use these across all our explored variable reduction methods. When plugged in to the Boruta training model and searching for predictors of completed\_lesson, the following was found:

Chart

Description automatically generated

Figure 8: Boruta Feature Importance Boxplots

The most important factor was weekly\_drink\_count, followed by bothered\_by\_mental\_health, weekly\_alcohol\_spending, other\_drugs\_per\_week, and bothered\_by\_gambling. Days\_drink\_per\_week and weekly\_other\_drug spending were just on the threshold and were left out of the final feature selection by Boruta.

Table

Description automatically generated

Figure 9: Boruta Feature Importance and decision

## Factor Analysis

Factor Analysis (FA) can be applied to selected variables in the data to determine the underlying factors from which the observed variables were generated. Both Exploratory Factor Analysis (EFA) and Confirmatory Factor Analysis (CFA) will be used for the variable assessments.

### Exploratory Factor Analysis

Exploratory Factor Analysis (EFA) can be used to identify constructs that help explain covariation amongst a set of obversved variables. In our case, we identified common factors derived from variables in the intake assesment survey. The team determined that the optimal number of factors was 8 after further analysis:

#### Varimax Rotation

Under a varimax rotation, the factors nearly perfectly lined up with a number of central themes. The output generated in R is presented in Figure 10. These factors and themes included:

**Factor 1:** Days\_drink\_per\_week, weekly\_drink\_count and bothered\_by\_drinking; aligned with “drinking” theme

**Factor 2:** Nicotine\_per\_week, cigarettes\_per\_day, bothered\_by\_nicotines; aligned with “smoking” theme

**Factor 3:** Gambling\_per\_week, weekly\_gambling\_spending, bothered\_by\_gambling; aligned with “gambling” theme

**Factor 4:** other\_drugs\_per\_week, weekly\_other\_drug\_spending, bothered\_by\_other\_drugs; aligned with “other drugs” theme

**Factor 5:** weekly\_food\_spending, bothered\_by\_food; aligned with the “food” theme

**Factor 6:** days\_canabis\_per\_week, weekly\_canabis\_spending, bothered\_by\_canabis; aligned with the “cannabis” theme

**Factor 7:** no apparent central theme for this factor, but consisted of variables pertaining to sex, drugs and smoking

**Factor 8:** weekly\_sex\_issues, sex\_issues\_per\_week, bothered\_by\_sex\_issues; aligned with the “sex” theme

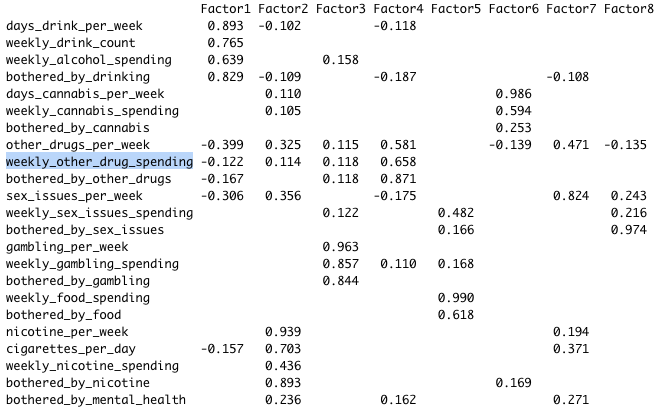


Figure 10: EFA Results Under Varimax Rotation

As captured in Figure 11, 66% of the variance could be explained under these 8 factors under the varimax rotation approach:

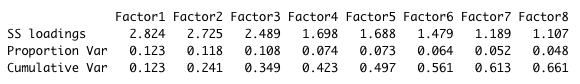


Figure 11: Variance Explained by Factors Under Varimax Rotation

#### Oblique Rotation

The factor alignment under the oblique rotation approach was similar to that of the varimax rotation approach, suggesting that 63% of the variance can be explained under the 8 factors. This information is presented in the following output:

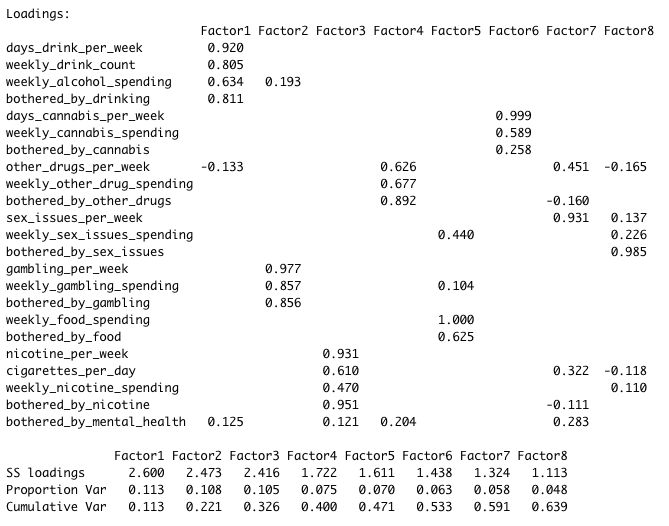


Figure 12: EFA Results Under Oblique Rotation

Clear boundaries are visible for each of the factor loadings, as demonstrated in Figure 13:

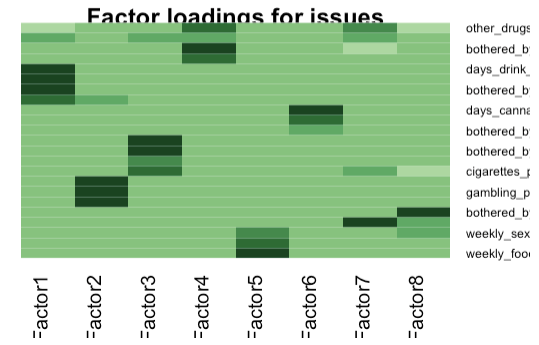


Figure 13: EFA Factor Loadings

The following path diagram shows the clear distinctions for each factor and their respective loadings:

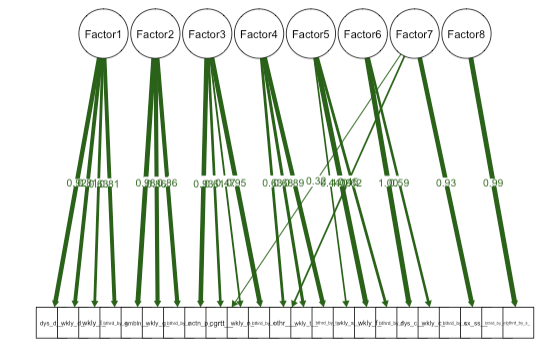


Figure 14: EFA Path Diagram

### Confirmatory Factor Analysis

Confirmatory Factor Analysis (CFA) can be used to evaluate factor significance using a hypothesized model. In this analysis, we construct a model that groups factors as:

* “Alcohol” – any factors that have to do with alcohol or drinking
* “Substance” – any factors that are substance-related (excluding alcohol), including smoking, cannabis, or other drugs
* “NonSubstance” – any factors related to addiction that are not alcohol, smoking, or drugs, such as eating, gambling, sex, or mental health

Our initial model included all factors in our intake assessment data as grouped into the above categories. The data was also scaled for normalization purposes. As presented in Figure 15, the model did not fit well, producing a CFI of 0.635 and RMSEA of 0.143.

A screenshot of a computer

Description automatically generated with medium confidence

Figure 15: First CFA Model Summary

A second attempt was constructed by removing several factors that had weak estimates or high *p*-values. [Some analysts](https://benwhalley.github.io/just-enough-r/gof.html) suggest that simplifying a CFA model, especially in the presence of a low number of observations, can be helpful for improving the model fitness.

Evidently, this second model exhibited a greatly improved fit, producing a CFI of 0.933. However, the RMSEA was still relatively high at value of 0.121. Additionally, the estimates of the modeled latent variables were not strong, and deemed to be insignificant with *p*-values greater than 0.05 in each case. This information is presented in Figure 16 and Figure 17.

A screenshot of a computer

Description automatically generated with low confidence

Figure 16: Second CFA Model Summary

A picture containing text, receipt

Description automatically generated

Figure 17: Second CFA Model Latent Variable Estimates and Significance

This updated CFA model produced the path diagram illustrated in Figure 18.

Diagram

Description automatically generated

Figure 18: CFA Path Diagram

Although interesting to perform, the hypothesized CFA models are not strongly supported by the available data. This could possibly be improved with larger data samples or by developing different hypotheses to analyze.

# Data Modelling Plan

Once the data reduction procedure is completed and resulting preparation steps are employed, consideration of the data model to be built should begin. There are several factors that must be taken into consideration when selecting model candidates:

* How large is our data set?
* How many variables will be used as predictor variables?
* What is our response variable? What data type is it (i.e., continuous, binary, categorical)?
* How much time does the team have to run the model?

Taking these factors into consideration, the team will initially aim to build a logistic regression model. This model was selected due to its reliability, the small underlying data size, and its inherent ability to perform efficient binary classification. Due to these factors, models such as the Neural Network or the Multiple Linear Regression model were not considered. The team will also be investigating CART and *k*-Nearest Neighbors (KNN) models for comparisons in terms of performance and prediction accuracy.

# Deliverable 6 Assignment

This week the focus will be on data modeling. You will update the template document with the following:

* Describe all the steps performed for data modelling.
* Discuss the performance measure(s) used to determine the goodness of fit for the proposed model.
* Do you suspect your initial effort (model) suffers overfitting? If so, discuss what steps you took to overcome overfitting.
* Discuss your findings or inferences for each of the research/ business queries you identified in Deliverable 1.

Submit your completed **Deliverable 6** to the Canvas assignment.

# Deliverable 7 Assignment

**Purpose:**

To provide a demonstration of your project in this course. In addition to submitting the final document, you will prepare a video presentation to share with the rest of the class.

**Tasks:**

**Part 1**

1. Before making the final submission, please ensure that you have implemented all instructor feedback from the previous Deliverables (1-6).
2. Please ensure that the title page, document control section, table of contents and the reference section is updated. Any appendix section should also be updated.
3. Submit the final document to the Deliverable 7 Canvas assignment in Lesson 14.
4. Submit your dataset(s), code/script, and PowerPoint to the Box account provided by your instructor.

**Part 2**

1. You will prepare a video presentation to demonstrate the designed data-driven Analytics System to the rest of the class.
2. The presentation should include PowerPoint slides and video demonstration.
3. Submit the link for your video presentation to the Project Presentation Discussion in Lesson 14.

Make all submissions by the due date noted in the Course Schedule.

More instructions (if needed) will be provided by the instructor via e-mail or Canvas Announcements.

# Appendix A

Table 3: Data Dictionary



Figure : Count of Data Fields by Data Type

Chart

Description automatically generated

Figure : Statistical Summaries and NULL Counts

Calendar

Description automatically generated with medium confidence Calendar

Description automatically generated with low confidence Diagram

Description automatically generated with medium confidence A screenshot of a computer

Description automatically generated with medium confidence Scatter chart

Description automatically generated A screenshot of a computer

Description automatically generated with medium confidence A picture containing text

Description automatically generated A picture containing text

Description automatically generated A picture containing scatter chart

Description automatically generated A picture containing scatter chart

Description automatically generated

# Appendix B

SQL for view intake\_assessment\_question\_texts:

**SELECT**

a**.**id**,**

a**.user\_id,**

a**.**age**,**

a**.**gender**,**

gender\_enum**.**title **AS** gender\_text**,**

a**.**how\_are\_you\_feeling\_today**,**

a**.**primary\_problem**,**

pp\_enum**.**title **AS** primary\_problem\_text**,**

a**.**days\_drink\_per\_week**,**

a**.**weekly\_drink\_count**,**

a**.**weekly\_alcohol\_spending**,**

a**.**bothered\_by\_drinking**,**

a**.**days\_cannabis\_per\_week**,**

a**.**earliest\_cannabis**,**

a**.**weekly\_cannabis\_spending**,**

a**.**bothered\_by\_cannabis**,**

a**.**other\_drugs**,**

a**.**other\_drugs\_per\_week**,**

a**.**weekly\_other\_drug\_spending**,**

a**.**bothered\_by\_other\_drugs**,**

a**.**sex\_issues**,**

a**.**sex\_issues\_per\_week**,**

a**.**weekly\_sex\_issues\_spending**,**

a**.**bothered\_by\_sex\_issues**,**

a**.**gambling\_per\_week**,**

a**.**weekly\_gambling\_spending**,**

a**.**bothered\_by\_gambling**,**

a**.**weekly\_food\_spending**,**

a**.**bothered\_by\_food**,**

a**.**nicotine\_per\_week**,**

a**.**cigarettes\_per\_day**,**

a**.**earliest\_nicotine**,**

a**.**nicotine\_form**,**

nic\_form\_enum**.**title **AS** nicotine\_form\_text**,**

a**.**weekly\_nicotine\_spending**,**

a**.**bothered\_by\_nicotine**,**

a**.**mental\_health\_issues**,**

a**.**bothered\_by\_mental\_health**,**

a**.**first\_time**,**

a**.**hear\_about\_igntd**,**

a**.**created\_at

**FROM**

dbo**.**intake\_assessment **AS** a

**LEFT** **OUTER** **JOIN** **(**

**SELECT**

**DISTINCT** enum\_value**,**

title

**FROM**

dbo**.**sds\_intake\_question\_options

**WHERE**

**(**intake\_question\_id **=** 817**)**

**)** **AS** gender\_enum **ON** **NULLIF** **(**a**.**gender**,** 'NULL'**)** **=** gender\_enum**.**enum\_value

**LEFT** **OUTER** **JOIN** **(**

**SELECT**

**DISTINCT** enum\_value**,**

title

**FROM**

dbo**.**sds\_intake\_question\_options **AS** sds\_intake\_question\_options\_2

**WHERE**

**(**intake\_question\_id **=** 820**)**

**)** **AS** pp\_enum **ON** **NULLIF** **(**a**.**primary\_problem**,** 'NULL'**)** **=** pp\_enum**.**enum\_value

**LEFT** **OUTER** **JOIN** **(**

**SELECT**

**DISTINCT** enum\_value**,**

title

**FROM**

dbo**.**sds\_intake\_question\_options **AS** sds\_intake\_question\_options\_1

**WHERE**

**(**intake\_question\_id **=** 871**)**

**)** **AS** nic\_form\_enum **ON** **NULLIF** **(**a**.**nicotine\_form**,** 'NULL'**)** **=** nic\_form\_enum**.**enum\_value

# Appendix C

Addressing NULL and NA valus in sds\_user\_activations and intake\_assessment\_question\_texts”

library(dplyr)

library(purrr)

library(stringr)

library(magrittr)

# ====== NULL Values Resolution ======

# Defines NULL values in data

values\_null <- c("NULL", "\\{\\}")

# Function to identify columns with null values in selected table

get\_null\_cols <- function(tbl) {

tbl %>%

map\_df(~ any(str\_detect(.x, paste(values\_null, collapse = "|")))) %>%

t() %>%

data.frame() %>%

select(has\_nulls = 1) %>%

filter(has\_nulls) %>%

rownames()

}

# Determines columns in database tables containing null values

null\_cols <- list\_data %>%

.[c("sds\_user\_activations", "intake\_assessment\_question\_texts")] %>%

map(get\_null\_cols)

# Resolves NULL values in sds\_user\_activations data

list\_data$sds\_user\_activations %<>%

mutate(across(null\_cols$sds\_user\_activations, ~ str\_replace(.x, paste(values\_null, collapse = "|"), NA\_character\_)),

across(c("bought\_hero", "welcome\_video", "upload\_worksheet", "rated\_lesson\_1",

"finished\_module\_1", "created\_community\_acct"), as.numeric),

across(c("bought\_hero", "welcome\_video", "upload\_worksheet", "rated\_lesson\_1",

"finished\_module\_1", "created\_community\_acct", "intake", "welcome\_video", "intake", "wol", "daily\_assessment", "started\_lesson", "completed\_lesson"), ~ replace(.x, is.na(.x), 0)))

# Resolves NULL values in intake\_assessment data

list\_data$intake\_assessment\_question\_texts %<>%

mutate(across(c(null\_cols$intake\_assessment\_question\_texts, "first\_time"), ~ str\_replace(.x, paste(values\_null, collapse = "|"), NA\_character\_)),

across(first\_time, as.logical),

across(c("age", "how\_are\_you\_feeling\_today", "weekly\_drink\_count",

"days\_cannabis\_per\_week", "weekly\_cannabis\_spending", "bothered\_by\_cannabis",

"other\_drugs\_per\_week", "weekly\_other\_drug\_spending", "bothered\_by\_other\_drugs",

"sex\_issues\_per\_week", "weekly\_sex\_issues\_spending", "bothered\_by\_sex\_issues",

"gambling\_per\_week", "weekly\_gambling\_spending", "bothered\_by\_gambling",

"weekly\_food\_spending", "bothered\_by\_food", "nicotine\_per\_week",

"cigarettes\_per\_day", "weekly\_nicotine\_spending",

"bothered\_by\_nicotine", "bothered\_by\_mental\_health", "first\_time"), as.numeric),

across(age, ~ replace(.x, .x == 0, NA\_integer\_)))

# References

*IGNTD Announces The Launch of SPARx, A Smart Personalized Adaptive Recovery System That Finally Fixes The Failing One-Size-Fits-All System That is Killing More People Every Year*. 24-7pressreleases.com. (2021, February 8). Retrieved January 23, 2022, from <https://www.24-7pressrelease.com/press-release/479198/igntd-announces-the-launch-of-sparx-a-smart-personalized-adaptive-recovery-system-that-finally-fixes-the-failing-one-size-fits-all-system-that-is-killing-more-people-every-year>#

Vaughan, Daniel. “3. Learning to Ask Good Business Questions.” *Analytical Skills for AI and Data Science: Building Skills for an AI-Driven Enterprise*, O'reilly, Beijing, 2020.