Data622 - Group2 - Homework4

Zachary Palmore, Kevin Potter, Amit Kapoor, Adam Gersowitz, Paul Perez

10/21/2021

Table of Contents

# Overview

In this project, we will analyze a real-life mental health dataset to provide context around suicide prediction given the variety of unidentifiable demographic data.

# Approach

We will first perform exploratory data analysis on dataset to later inform our modeling approaches for Clustering, Principal Compnent Analysis, Gradient Boosting, and Support Vector Machines.

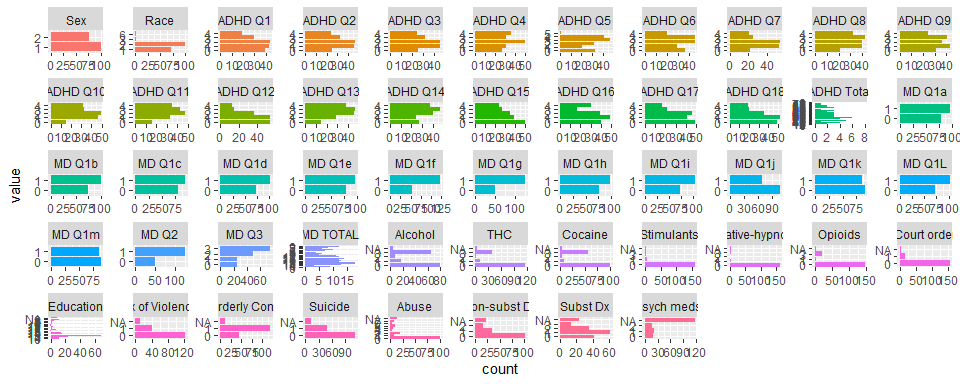
# Data Exploration

| Columns | Variable | Description |
| --- | --- | --- |
| C | Sex | Male-1, Female-2 |
| D | Race | White-1, African American-2, Hispanic-3, Asian-4, Native American-5, Other or missing data -6 |
| E - W | ADHD self-report scale | Never-0, rarely-1, sometimes-2, often-3, very often-4 |
| X - AM | Mood disorder questions | No-0, yes-1; question 3: no problem-0, minor-1, moderate-2, serious-3 |
| AN - AS | Individual substances misuse | no use-0, use-1, abuse-2, dependence-3 |
| AT | Court Order | No-0, Yes-1 |
| AU | Education | 1-12 grade, 13+ college |
| AV | History of Violence | No-0, Yes-1 |
| AW | Disorderly Conduct | No-0, Yes-1 |
| AX | Suicide attempt | No-0, Yes-1 |
| AY | Abuse Hx | No-0, Physical (P)-1, Sexual (S)-2, Emotional (E)-3, P&S-4, P&E-5, S&E-6, P&S&E-7 |
| AZ | Non-substance-related Dx | 0 - none; 1 - one; 2 - More than one |
| BA | Substance-related Dx | 0 - none; 1 - one Substance-related; 2 - two; 3 - three or more |
| BB | Psychiatric Meds | 0 - none; 1 - one psychotropic med; 2 - more than one psychotropic med |

## Data Characteristics

## [1] 175 53

## # A tibble: 175 x 15  
## `MD Q1a` `MD Q1b` `MD Q1c` `MD Q1d` `MD Q1e` `MD Q1f` `MD Q1g` `MD Q1h`  
## <fct> <fct> <fct> <fct> <fct> <fct> <fct> <fct>   
## 1 1 1 1 1 0 1 1 1   
## 2 1 1 1 1 1 1 1 1   
## 3 0 0 0 0 1 1 1 0   
## 4 1 1 0 0 1 1 1 1   
## 5 0 1 0 1 0 1 1 0   
## 6 0 1 0 1 1 1 1 1   
## 7 1 1 0 0 1 1 0 0   
## 8 0 0 0 0 0 1 1 0   
## 9 1 1 0 1 1 1 1 0   
## 10 1 1 0 0 1 0 1 0   
## # ... with 165 more rows, and 7 more variables: MD Q1i <fct>, MD Q1j <fct>,  
## # MD Q1k <fct>, MD Q1L <fct>, MD Q1m <fct>, MD Q2 <fct>, MD Q3 <fct>

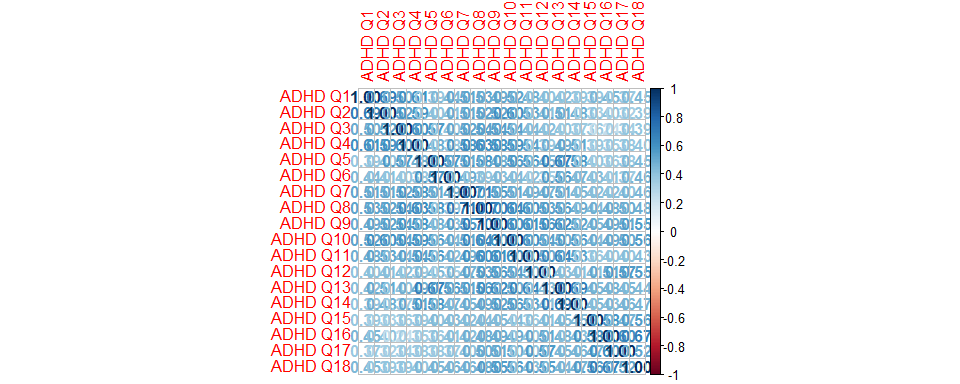


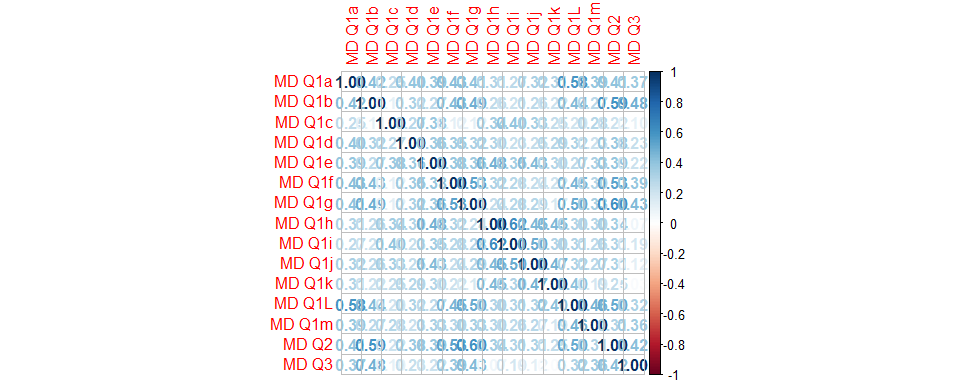
## Data summary

## Data Frame Summary   
## adhd\_data   
## Dimensions: 175 x 53   
## Duplicates: 0   
##   
## +----+--------------------+-------------------------+--------------------+----------+---------+  
## | No | Variable | Stats / Values | Freqs (% of Valid) | Valid | Missing |  
## +====+====================+=========================+====================+==========+=========+  
## | 1 | Age | Mean (sd) : 39.5 (11.2) | 42 distinct values | 175 | 0 |  
## | | [numeric] | min < med < max: | | (100.0%) | (0.0%) |  
## | | | 18 < 42 < 69 | | | |  
## | | | IQR (CV) : 18.5 (0.3) | | | |  
## +----+--------------------+-------------------------+--------------------+----------+---------+  
## | 2 | Sex | 1. 1 | 99 (56.6%) | 175 | 0 |  
## | | [factor] | 2. 2 | 76 (43.4%) | (100.0%) | (0.0%) |  
## +----+--------------------+-------------------------+--------------------+----------+---------+  
## | 3 | Race | 1. 1 | 72 (41.1%) | 175 | 0 |  
## | | [factor] | 2. 2 | 100 (57.1%) | (100.0%) | (0.0%) |  
## | | | 3. 3 | 1 ( 0.6%) | | |  
## | | | 4. 6 | 2 ( 1.1%) | | |  
## +----+--------------------+-------------------------+--------------------+----------+---------+  
## | 4 | ADHD Q1 | 1. 0 | 39 (22.3%) | 175 | 0 |  
## | | [factor] | 2. 1 | 43 (24.6%) | (100.0%) | (0.0%) |  
## | | | 3. 2 | 44 (25.1%) | | |  
## | | | 4. 3 | 30 (17.1%) | | |  
## | | | 5. 4 | 19 (10.9%) | | |  
## +----+--------------------+-------------------------+--------------------+----------+---------+  
## | 5 | ADHD Q2 | 1. 0 | 25 (14.3%) | 175 | 0 |  
## | | [factor] | 2. 1 | 46 (26.3%) | (100.0%) | (0.0%) |  
## | | | 3. 2 | 47 (26.9%) | | |  
## | | | 4. 3 | 33 (18.9%) | | |  
## | | | 5. 4 | 24 (13.7%) | | |  
## +----+--------------------+-------------------------+--------------------+----------+---------+  
## | 6 | ADHD Q3 | 1. 0 | 26 (14.9%) | 175 | 0 |  
## | | [factor] | 2. 1 | 46 (26.3%) | (100.0%) | (0.0%) |  
## | | | 3. 2 | 46 (26.3%) | | |  
## | | | 4. 3 | 32 (18.3%) | | |  
## | | | 5. 4 | 25 (14.3%) | | |  
## +----+--------------------+-------------------------+--------------------+----------+---------+  
## | 7 | ADHD Q4 | 1. 0 | 27 (15.4%) | 175 | 0 |  
## | | [factor] | 2. 1 | 31 (17.7%) | (100.0%) | (0.0%) |  
## | | | 3. 2 | 50 (28.6%) | | |  
## | | | 4. 3 | 31 (17.7%) | | |  
## | | | 5. 4 | 36 (20.6%) | | |  
## +----+--------------------+-------------------------+--------------------+----------+---------+  
## | 8 | ADHD Q5 | 1. 0 | 33 (18.9%) | 175 | 0 |  
## | | [factor] | 2. 1 | 21 (12.0%) | (100.0%) | (0.0%) |  
## | | | 3. 2 | 32 (18.3%) | | |  
## | | | 4. 3 | 47 (26.9%) | | |  
## | | | 5. 4 | 41 (23.4%) | | |  
## | | | 6. 5 | 1 ( 0.6%) | | |  
## +----+--------------------+-------------------------+--------------------+----------+---------+  
## | 9 | ADHD Q6 | 1. 0 | 36 (20.6%) | 175 | 0 |  
## | | [factor] | 2. 1 | 29 (16.6%) | (100.0%) | (0.0%) |  
## | | | 3. 2 | 45 (25.7%) | | |  
## | | | 4. 3 | 45 (25.7%) | | |  
## | | | 5. 4 | 20 (11.4%) | | |  
## +----+--------------------+-------------------------+--------------------+----------+---------+  
## | 10 | ADHD Q7 | 1. 0 | 22 (12.6%) | 175 | 0 |  
## | | [factor] | 2. 1 | 53 (30.3%) | (100.0%) | (0.0%) |  
## | | | 3. 2 | 54 (30.9%) | | |  
## | | | 4. 3 | 25 (14.3%) | | |  
## | | | 5. 4 | 21 (12.0%) | | |  
## +----+--------------------+-------------------------+--------------------+----------+---------+  
## | 11 | ADHD Q8 | 1. 0 | 21 (12.0%) | 175 | 0 |  
## | | [factor] | 2. 1 | 40 (22.9%) | (100.0%) | (0.0%) |  
## | | | 3. 2 | 40 (22.9%) | | |  
## | | | 4. 3 | 42 (24.0%) | | |  
## | | | 5. 4 | 32 (18.3%) | | |  
## +----+--------------------+-------------------------+--------------------+----------+---------+  
## | 12 | ADHD Q9 | 1. 0 | 31 (17.7%) | 175 | 0 |  
## | | [factor] | 2. 1 | 43 (24.6%) | (100.0%) | (0.0%) |  
## | | | 3. 2 | 36 (20.6%) | | |  
## | | | 4. 3 | 41 (23.4%) | | |  
## | | | 5. 4 | 24 (13.7%) | | |  
## +----+--------------------+-------------------------+--------------------+----------+---------+  
## | 13 | ADHD Q10 | 1. 0 | 15 ( 8.6%) | 175 | 0 |  
## | | [factor] | 2. 1 | 46 (26.3%) | (100.0%) | (0.0%) |  
## | | | 3. 2 | 49 (28.0%) | | |  
## | | | 4. 3 | 33 (18.9%) | | |  
## | | | 5. 4 | 32 (18.3%) | | |  
## +----+--------------------+-------------------------+--------------------+----------+---------+  
## | 14 | ADHD Q11 | 1. 0 | 16 ( 9.1%) | 175 | 0 |  
## | | [factor] | 2. 1 | 33 (18.9%) | (100.0%) | (0.0%) |  
## | | | 3. 2 | 48 (27.4%) | | |  
## | | | 4. 3 | 43 (24.6%) | | |  
## | | | 5. 4 | 35 (20.0%) | | |  
## +----+--------------------+-------------------------+--------------------+----------+---------+  
## | 15 | ADHD Q12 | 1. 0 | 55 (31.4%) | 175 | 0 |  
## | | [factor] | 2. 1 | 55 (31.4%) | (100.0%) | (0.0%) |  
## | | | 3. 2 | 37 (21.1%) | | |  
## | | | 4. 3 | 15 ( 8.6%) | | |  
## | | | 5. 4 | 13 ( 7.4%) | | |  
## +----+--------------------+-------------------------+--------------------+----------+---------+  
## | 16 | ADHD Q13 | 1. 0 | 15 ( 8.6%) | 175 | 0 |  
## | | [factor] | 2. 1 | 29 (16.6%) | (100.0%) | (0.0%) |  
## | | | 3. 2 | 46 (26.3%) | | |  
## | | | 4. 3 | 47 (26.9%) | | |  
## | | | 5. 4 | 38 (21.7%) | | |  
## +----+--------------------+-------------------------+--------------------+----------+---------+  
## | 17 | ADHD Q14 | 1. 0 | 27 (15.4%) | 175 | 0 |  
## | | [factor] | 2. 1 | 24 (13.7%) | (100.0%) | (0.0%) |  
## | | | 3. 2 | 40 (22.9%) | | |  
## | | | 4. 3 | 47 (26.9%) | | |  
## | | | 5. 4 | 37 (21.1%) | | |  
## +----+--------------------+-------------------------+--------------------+----------+---------+  
## | 18 | ADHD Q15 | 1. 0 | 50 (28.6%) | 175 | 0 |  
## | | [factor] | 2. 1 | 39 (22.3%) | (100.0%) | (0.0%) |  
## | | | 3. 2 | 35 (20.0%) | | |  
## | | | 4. 3 | 27 (15.4%) | | |  
## | | | 5. 4 | 24 (13.7%) | | |  
## +----+--------------------+-------------------------+--------------------+----------+---------+  
## | 19 | ADHD Q16 | 1. 0 | 40 (22.9%) | 175 | 0 |  
## | | [factor] | 2. 1 | 49 (28.0%) | (100.0%) | (0.0%) |  
## | | | 3. 2 | 39 (22.3%) | | |  
## | | | 4. 3 | 17 ( 9.7%) | | |  
## | | | 5. 4 | 30 (17.1%) | | |  
## +----+--------------------+-------------------------+--------------------+----------+---------+  
## | 20 | ADHD Q17 | 1. 0 | 49 (28.0%) | 175 | 0 |  
## | | [factor] | 2. 1 | 41 (23.4%) | (100.0%) | (0.0%) |  
## | | | 3. 2 | 46 (26.3%) | | |  
## | | | 4. 3 | 22 (12.6%) | | |  
## | | | 5. 4 | 17 ( 9.7%) | | |  
## +----+--------------------+-------------------------+--------------------+----------+---------+  
## | 21 | ADHD Q18 | 1. 0 | 49 (28.0%) | 175 | 0 |  
## | | [factor] | 2. 1 | 52 (29.7%) | (100.0%) | (0.0%) |  
## | | | 3. 2 | 35 (20.0%) | | |  
## | | | 4. 3 | 20 (11.4%) | | |  
## | | | 5. 4 | 19 (10.9%) | | |  
## +----+--------------------+-------------------------+--------------------+----------+---------+  
## | 22 | ADHD Total | 1. 0 | 1 ( 0.6%) | 175 | 0 |  
## | | [factor] | 2. 1 | 2 ( 1.1%) | (100.0%) | (0.0%) |  
## | | | 3. 3 | 1 ( 0.6%) | | |  
## | | | 4. 5 | 1 ( 0.6%) | | |  
## | | | 5. 6 | 3 ( 1.7%) | | |  
## | | | 6. 7 | 2 ( 1.1%) | | |  
## | | | 7. 8 | 1 ( 0.6%) | | |  
## | | | 8. 9 | 2 ( 1.1%) | | |  
## | | | 9. 10 | 2 ( 1.1%) | | |  
## | | | 10. 11 | 1 ( 0.6%) | | |  
## | | | [ 52 others ] | 159 (90.9%) | | |  
## +----+--------------------+-------------------------+--------------------+----------+---------+  
## | 23 | MD Q1a | 1. 0 | 79 (45.1%) | 175 | 0 |  
## | | [factor] | 2. 1 | 96 (54.9%) | (100.0%) | (0.0%) |  
## +----+--------------------+-------------------------+--------------------+----------+---------+  
## | 24 | MD Q1b | 1. 0 | 75 (42.9%) | 175 | 0 |  
## | | [factor] | 2. 1 | 100 (57.1%) | (100.0%) | (0.0%) |  
## +----+--------------------+-------------------------+--------------------+----------+---------+  
## | 25 | MD Q1c | 1. 0 | 80 (45.7%) | 175 | 0 |  
## | | [factor] | 2. 1 | 95 (54.3%) | (100.0%) | (0.0%) |  
## +----+--------------------+-------------------------+--------------------+----------+---------+  
## | 26 | MD Q1d | 1. 0 | 73 (41.7%) | 175 | 0 |  
## | | [factor] | 2. 1 | 102 (58.3%) | (100.0%) | (0.0%) |  
## +----+--------------------+-------------------------+--------------------+----------+---------+  
## | 27 | MD Q1e | 1. 0 | 78 (44.6%) | 175 | 0 |  
## | | [factor] | 2. 1 | 97 (55.4%) | (100.0%) | (0.0%) |  
## +----+--------------------+-------------------------+--------------------+----------+---------+  
## | 28 | MD Q1f | 1. 0 | 53 (30.3%) | 175 | 0 |  
## | | [factor] | 2. 1 | 122 (69.7%) | (100.0%) | (0.0%) |  
## +----+--------------------+-------------------------+--------------------+----------+---------+  
## | 29 | MD Q1g | 1. 0 | 49 (28.0%) | 175 | 0 |  
## | | [factor] | 2. 1 | 126 (72.0%) | (100.0%) | (0.0%) |  
## +----+--------------------+-------------------------+--------------------+----------+---------+  
## | 30 | MD Q1h | 1. 0 | 77 (44.0%) | 175 | 0 |  
## | | [factor] | 2. 1 | 98 (56.0%) | (100.0%) | (0.0%) |  
## +----+--------------------+-------------------------+--------------------+----------+---------+  
## | 31 | MD Q1i | 1. 0 | 72 (41.1%) | 175 | 0 |  
## | | [factor] | 2. 1 | 103 (58.9%) | (100.0%) | (0.0%) |  
## +----+--------------------+-------------------------+--------------------+----------+---------+  
## | 32 | MD Q1j | 1. 0 | 107 (61.1%) | 175 | 0 |  
## | | [factor] | 2. 1 | 68 (38.9%) | (100.0%) | (0.0%) |  
## +----+--------------------+-------------------------+--------------------+----------+---------+  
## | 33 | MD Q1k | 1. 0 | 90 (51.4%) | 175 | 0 |  
## | | [factor] | 2. 1 | 85 (48.6%) | (100.0%) | (0.0%) |  
## +----+--------------------+-------------------------+--------------------+----------+---------+  
## | 34 | MD Q1L | 1. 0 | 73 (41.7%) | 175 | 0 |  
## | | [factor] | 2. 1 | 102 (58.3%) | (100.0%) | (0.0%) |  
## +----+--------------------+-------------------------+--------------------+----------+---------+  
## | 35 | MD Q1m | 1. 0 | 89 (50.9%) | 175 | 0 |  
## | | [factor] | 2. 1 | 86 (49.1%) | (100.0%) | (0.0%) |  
## +----+--------------------+-------------------------+--------------------+----------+---------+  
## | 36 | MD Q2 | 1. 0 | 49 (28.0%) | 175 | 0 |  
## | | [factor] | 2. 1 | 126 (72.0%) | (100.0%) | (0.0%) |  
## +----+--------------------+-------------------------+--------------------+----------+---------+  
## | 37 | MD Q3 | 1. 0 | 25 (14.3%) | 175 | 0 |  
## | | [factor] | 2. 1 | 25 (14.3%) | (100.0%) | (0.0%) |  
## | | | 3. 2 | 49 (28.0%) | | |  
## | | | 4. 3 | 76 (43.4%) | | |  
## +----+--------------------+-------------------------+--------------------+----------+---------+  
## | 38 | MD TOTAL | 1. 0 | 9 ( 5.1%) | 175 | 0 |  
## | | [factor] | 2. 1 | 3 ( 1.7%) | (100.0%) | (0.0%) |  
## | | | 3. 2 | 5 ( 2.9%) | | |  
## | | | 4. 3 | 6 ( 3.4%) | | |  
## | | | 5. 4 | 4 ( 2.3%) | | |  
## | | | 6. 5 | 7 ( 4.0%) | | |  
## | | | 7. 6 | 10 ( 5.7%) | | |  
## | | | 8. 7 | 6 ( 3.4%) | | |  
## | | | 9. 8 | 8 ( 4.6%) | | |  
## | | | 10. 9 | 12 ( 6.9%) | | |  
## | | | [ 8 others ] | 105 (60.0%) | | |  
## +----+--------------------+-------------------------+--------------------+----------+---------+  
## | 39 | Alcohol | 1. 0 | 80 (46.8%) | 171 | 4 |  
## | | [factor] | 2. 1 | 18 (10.5%) | (97.7%) | (2.3%) |  
## | | | 3. 2 | 7 ( 4.1%) | | |  
## | | | 4. 3 | 66 (38.6%) | | |  
## +----+--------------------+-------------------------+--------------------+----------+---------+  
## | 40 | THC | 1. 0 | 116 (67.8%) | 171 | 4 |  
## | | [factor] | 2. 1 | 12 ( 7.0%) | (97.7%) | (2.3%) |  
## | | | 3. 2 | 3 ( 1.8%) | | |  
## | | | 4. 3 | 40 (23.4%) | | |  
## +----+--------------------+-------------------------+--------------------+----------+---------+  
## | 41 | Cocaine | 1. 0 | 101 (59.1%) | 171 | 4 |  
## | | [factor] | 2. 1 | 9 ( 5.3%) | (97.7%) | (2.3%) |  
## | | | 3. 2 | 5 ( 2.9%) | | |  
## | | | 4. 3 | 56 (32.7%) | | |  
## +----+--------------------+-------------------------+--------------------+----------+---------+  
## | 42 | Stimulants | 1. 0 | 160 (93.6%) | 171 | 4 |  
## | | [factor] | 2. 1 | 6 ( 3.5%) | (97.7%) | (2.3%) |  
## | | | 3. 3 | 5 ( 2.9%) | | |  
## +----+--------------------+-------------------------+--------------------+----------+---------+  
## | 43 | Sedative-hypnotics | 1. 0 | 161 (94.2%) | 171 | 4 |  
## | | [factor] | 2. 1 | 4 ( 2.3%) | (97.7%) | (2.3%) |  
## | | | 3. 2 | 1 ( 0.6%) | | |  
## | | | 4. 3 | 5 ( 2.9%) | | |  
## +----+--------------------+-------------------------+--------------------+----------+---------+  
## | 44 | Opioids | 1. 0 | 146 (85.4%) | 171 | 4 |  
## | | [factor] | 2. 1 | 4 ( 2.3%) | (97.7%) | (2.3%) |  
## | | | 3. 3 | 21 (12.3%) | | |  
## +----+--------------------+-------------------------+--------------------+----------+---------+  
## | 45 | Court order | 1. 0 | 155 (91.2%) | 170 | 5 |  
## | | [factor] | 2. 1 | 15 ( 8.8%) | (97.1%) | (2.9%) |  
## +----+--------------------+-------------------------+--------------------+----------+---------+  
## | 46 | Education | 1. 6 | 2 ( 1.2%) | 166 | 9 |  
## | | [factor] | 2. 7 | 2 ( 1.2%) | (94.9%) | (5.1%) |  
## | | | 3. 8 | 5 ( 3.0%) | | |  
## | | | 4. 9 | 12 ( 7.2%) | | |  
## | | | 5. 10 | 12 ( 7.2%) | | |  
## | | | 6. 11 | 23 (13.9%) | | |  
## | | | 7. 12 | 67 (40.4%) | | |  
## | | | 8. 13 | 15 ( 9.0%) | | |  
## | | | 9. 14 | 14 ( 8.4%) | | |  
## | | | 10. 15 | 1 ( 0.6%) | | |  
## | | | [ 4 others ] | 13 ( 7.8%) | | |  
## +----+--------------------+-------------------------+--------------------+----------+---------+  
## | 47 | Hx of Violence | 1. 0 | 124 (75.6%) | 164 | 11 |  
## | | [factor] | 2. 1 | 40 (24.4%) | (93.7%) | (6.3%) |  
## +----+--------------------+-------------------------+--------------------+----------+---------+  
## | 48 | Disorderly Conduct | 1. 0 | 45 (27.4%) | 164 | 11 |  
## | | [factor] | 2. 1 | 119 (72.6%) | (93.7%) | (6.3%) |  
## +----+--------------------+-------------------------+--------------------+----------+---------+  
## | 49 | Suicide | 1. 0 | 113 (69.8%) | 162 | 13 |  
## | | [factor] | 2. 1 | 49 (30.2%) | (92.6%) | (7.4%) |  
## +----+--------------------+-------------------------+--------------------+----------+---------+  
## | 50 | Abuse | 1. 0 | 101 (62.7%) | 161 | 14 |  
## | | [factor] | 2. 1 | 8 ( 5.0%) | (92.0%) | (8.0%) |  
## | | | 3. 2 | 20 (12.4%) | | |  
## | | | 4. 3 | 4 ( 2.5%) | | |  
## | | | 5. 4 | 6 ( 3.7%) | | |  
## | | | 6. 5 | 10 ( 6.2%) | | |  
## | | | 7. 6 | 4 ( 2.5%) | | |  
## | | | 8. 7 | 8 ( 5.0%) | | |  
## +----+--------------------+-------------------------+--------------------+----------+---------+  
## | 51 | Non-subst Dx | 1. 0 | 102 (66.7%) | 153 | 22 |  
## | | [factor] | 2. 1 | 35 (22.9%) | (87.4%) | (12.6%) |  
## | | | 3. 2 | 16 (10.5%) | | |  
## +----+--------------------+-------------------------+--------------------+----------+---------+  
## | 52 | Subst Dx | 1. 0 | 42 (27.6%) | 152 | 23 |  
## | | [factor] | 2. 1 | 61 (40.1%) | (86.9%) | (13.1%) |  
## | | | 3. 2 | 35 (23.0%) | | |  
## | | | 4. 3 | 14 ( 9.2%) | | |  
## +----+--------------------+-------------------------+--------------------+----------+---------+  
## | 53 | Psych meds. | 1. 0 | 19 (33.3%) | 57 | 118 |  
## | | [factor] | 2. 1 | 21 (36.8%) | (32.6%) | (67.4%) |  
## | | | 3. 2 | 17 (29.8%) | | |  
## +----+--------------------+-------------------------+--------------------+----------+---------+

## Coorelation

Next we will see the correlation among ADHD questions and MD questions. As we can deduce from below 2 correlation plots, ADHD questions are highly correlated and MD questions comparatively shows moderate correlation.





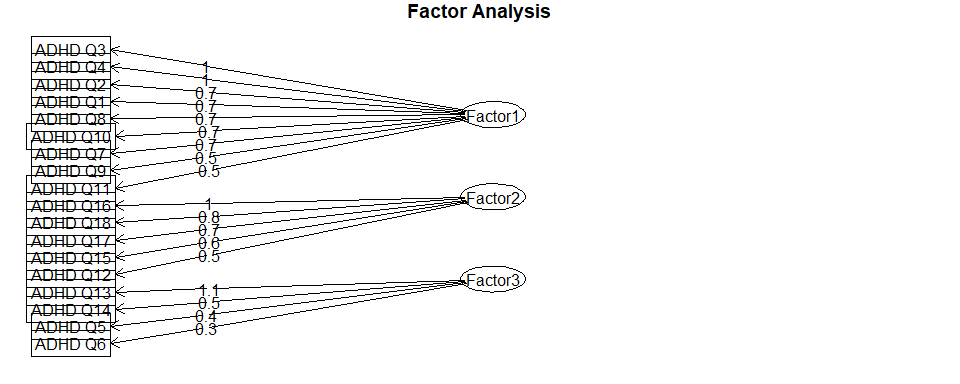
# Data Preparation

## Factor Analysis

Like PCA, Factor Analysis too, reduces larger number of variables into smaller number of variables, called latent variables.It is used to identify underlying factors that explain the correlation among set of variables. Factor analysis is a great tool for treating multivariate questionnaire studies.

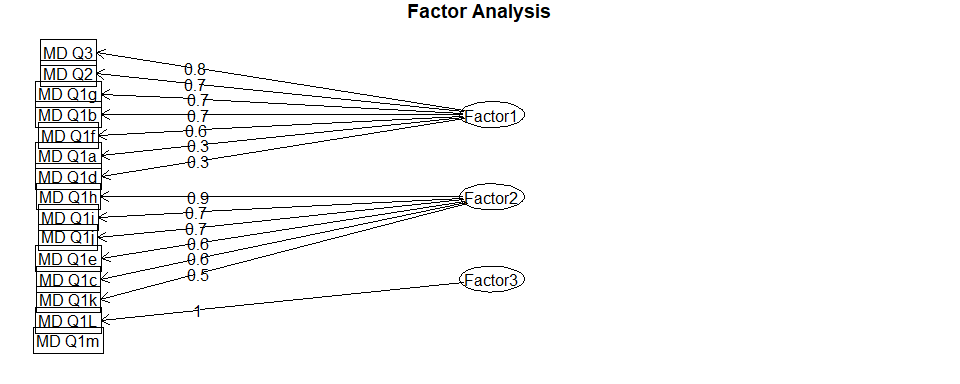
For ADHD questions, test of the hypothesis that 3 factors are sufficient. The chi square statistic is 197.3 on 102 degrees of freedom. The p-value is 0.0000000476. We have used regression factor scores here as they predict the location of each individual on the factor.

##   
## Call:  
## factanal(x = sapply(adhd\_data[, c(4:21)], as.numeric), factors = 3, scores = "regression", rotation = "promax")  
##   
## Uniquenesses:  
## ADHD Q1 ADHD Q2 ADHD Q3 ADHD Q4 ADHD Q5 ADHD Q6 ADHD Q7 ADHD Q8   
## 0.493 0.470 0.447 0.360 0.454 0.605 0.457 0.344   
## ADHD Q9 ADHD Q10 ADHD Q11 ADHD Q12 ADHD Q13 ADHD Q14 ADHD Q15 ADHD Q16   
## 0.378 0.372 0.444 0.516 0.008 0.460 0.538 0.266   
## ADHD Q17 ADHD Q18   
## 0.496 0.360   
##   
## Loadings:  
## Factor1 Factor2 Factor3  
## ADHD Q1 0.738 0.102 -0.142   
## ADHD Q2 0.743   
## ADHD Q3 0.972 -0.186 -0.144   
## ADHD Q4 0.967 -0.164   
## ADHD Q5 0.379 0.447   
## ADHD Q6 0.173 0.185 0.332   
## ADHD Q7 0.675   
## ADHD Q8 0.731 0.110   
## ADHD Q9 0.500 0.194 0.159   
## ADHD Q10 0.687 0.237 -0.113   
## ADHD Q11 0.480 0.327   
## ADHD Q12 0.302 0.511   
## ADHD Q13 -0.163 1.142   
## ADHD Q14 0.158 0.122 0.512   
## ADHD Q15 0.638   
## ADHD Q16 -0.241 1.014   
## ADHD Q17 0.682   
## ADHD Q18 0.823 -0.116   
##   
## Factor1 Factor2 Factor3  
## SS loadings 5.298 3.079 2.095  
## Proportion Var 0.294 0.171 0.116  
## Cumulative Var 0.294 0.465 0.582  
##   
## Factor Correlations:  
## Factor1 Factor2 Factor3  
## Factor1 1.000 0.765 -0.685  
## Factor2 0.765 1.000 -0.748  
## Factor3 -0.685 -0.748 1.000  
##   
## Test of the hypothesis that 3 factors are sufficient.  
## The chi square statistic is 197.3 on 102 degrees of freedom.  
## The p-value is 4.76e-08



For MD questions we could see that 1st MD question has multiple sub questions as compared to 2nd and 3rd question. Now for these set of MD questions too, we will apply similar factor analysis as of ADHD questions. Test of the hypothesis that 3 factors are sufficient. The chi square statistic is 88.82 on 63 degrees of freedom. The p-value is 0.0178.

##   
## Call:  
## factanal(x = sapply(adhd\_data[, c(23:37)], as.numeric), factors = 3, scores = "regression", rotation = "promax")  
##   
## Uniquenesses:  
## MD Q1a MD Q1b MD Q1c MD Q1d MD Q1e MD Q1f MD Q1g MD Q1h MD Q1i MD Q1j MD Q1k   
## 0.562 0.506 0.736 0.735 0.564 0.536 0.446 0.388 0.507 0.567 0.638   
## MD Q1L MD Q1m MD Q2 MD Q3   
## 0.005 0.719 0.394 0.601   
##   
## Loadings:  
## Factor1 Factor2 Factor3  
## MD Q1a 0.345 0.117 0.308   
## MD Q1b 0.732   
## MD Q1c 0.565   
## MD Q1d 0.342 0.257   
## MD Q1e 0.283 0.568 -0.194   
## MD Q1f 0.632   
## MD Q1g 0.735   
## MD Q1h 0.856   
## MD Q1i 0.738   
## MD Q1j 0.662   
## MD Q1k -0.172 0.515 0.265   
## MD Q1L 0.133 -0.124 0.981   
## MD Q1m 0.228 0.158 0.240   
## MD Q2 0.738   
## MD Q3 0.751 -0.184   
##   
## Factor1 Factor2 Factor3  
## SS loadings 3.009 2.790 1.241  
## Proportion Var 0.201 0.186 0.083  
## Cumulative Var 0.201 0.387 0.469  
##   
## Factor Correlations:  
## Factor1 Factor2 Factor3  
## Factor1 1.000 0.550 -0.587  
## Factor2 0.550 1.000 -0.563  
## Factor3 -0.587 -0.563 1.000  
##   
## Test of the hypothesis that 3 factors are sufficient.  
## The chi square statistic is 88.82 on 63 degrees of freedom.  
## The p-value is 0.0178

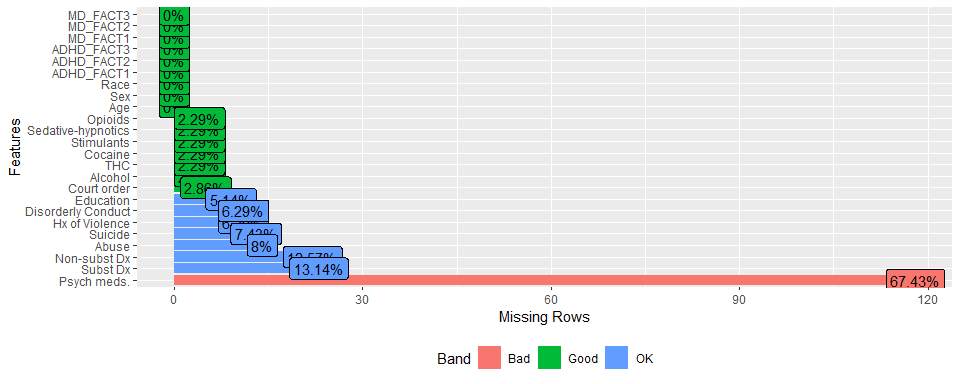


In the next step we will remove all ADHD Question columns, ADHD Total, MD questions columns and MD TOTAL columns. Then we will add the new factors found above for ADHD and MD questions.

Here is glimpse of new set of data.

## Age Sex Race Alcohol THC Cocaine Stimulants Sedative-hypnotics Opioids  
## 1 24 1 1 1 1 1 0 0 0  
## 2 48 2 1 0 0 0 0 0 0  
## 3 51 2 1 0 0 0 0 0 0  
## 4 43 1 1 1 1 1 1 0 0  
## 5 34 1 1 1 1 0 0 0 0  
## 6 39 2 1 1 0 0 0 0 0  
## Court order Education Hx of Violence Disorderly Conduct Suicide Abuse  
## 1 1 11 0 1 1 0  
## 2 0 14 0 0 1 4  
## 3 0 12 0 0 0 6  
## 4 0 12 0 0 1 7  
## 5 1 9 1 1 1 0  
## 6 0 11 0 1 1 2  
## Non-subst Dx Subst Dx Psych meds. ADHD\_FACT1 ADHD\_FACT2 ADHD\_FACT3 MD\_FACT1  
## 1 2 0 2 1.6922046 1.6740898 -3.3243648 1.3855809  
## 2 1 0 1 2.0799334 1.5195976 -2.6626233 0.7973360  
## 3 2 0 1 -0.5301540 0.3261461 -0.1297720 0.4673052  
## 4 2 0 2 0.9321586 -0.5385242 0.2275811 0.8725442  
## 5 2 0 0 2.5823393 -1.6535142 -0.2129593 2.1105464  
## 6 0 0 0 -0.8422991 1.3342893 1.0827141 0.2816620  
## MD\_FACT2 MD\_FACT3  
## 1 1.59853502 -2.8509956  
## 2 -0.08361024 0.3704740  
## 3 -0.80624391 -1.0898824  
## 4 -0.45310917 0.5084134  
## 5 -1.37884110 -1.8594994  
## 6 0.44486489 0.4402313

## Handling missing values



We can see from this chart that Psych meds. contributes to 67.43% of missing data which is maximum among all missing data in other columns. We will remove this column before imputation. We then impute values using MICE (Multivariate Imputation by Chained Equations).

Here is the summary after imputation.

## Age Sex Race Alcohol THC Cocaine Stimulants  
## Min. :18.00 1:99 1: 72 0:80 0:117 0:102 0:162   
## 1st Qu.:29.50 2:76 2:100 1:18 1: 13 1: 10 1: 8   
## Median :42.00 3: 1 2: 8 2: 4 2: 5 3: 5   
## Mean :39.47 6: 2 3:69 3: 41 3: 58   
## 3rd Qu.:48.00   
## Max. :69.00   
##   
## Sedative\_hypnotics Opioids Court\_order Education Hx\_of\_Violence  
## 0:163 0:149 0:159 12 :69 0:131   
## 1: 6 1: 4 1: 16 11 :23 1: 44   
## 2: 1 3: 22 13 :17   
## 3: 5 10 :15   
## 14 :15   
## 9 :13   
## (Other):23   
## Disorderly\_Conduct Suicide Abuse Non\_subst\_Dx Subst\_Dx  
## 0: 49 0:120 0 :107 0:112 0:47   
## 1:126 1: 55 2 : 20 1: 44 1:70   
## 1 : 10 2: 19 2:38   
## 5 : 10 3:20   
## 7 : 9   
## 4 : 8   
## (Other): 11   
## ADHD\_FACT1 ADHD\_FACT2 ADHD\_FACT3 MD\_FACT1   
## Min. :-4.3630 Min. :-4.13016 Min. :-4.17808 Min. :-2.5372   
## 1st Qu.:-1.0233 1st Qu.:-0.81557 1st Qu.:-0.87811 1st Qu.:-0.8859   
## Median :-0.1190 Median :-0.08667 Median :-0.04472 Median : 0.2680   
## Mean : 0.0000 Mean : 0.00000 Mean : 0.00000 Mean : 0.0000   
## 3rd Qu.: 0.9766 3rd Qu.: 0.81132 3rd Qu.: 0.92458 3rd Qu.: 0.7981   
## Max. : 5.4096 Max. : 4.23133 Max. : 4.28336 Max. : 2.4166   
##   
## MD\_FACT2 MD\_FACT3   
## Min. :-2.4903 Min. :-2.9969   
## 1st Qu.:-0.8573 1st Qu.:-0.7034   
## Median : 0.1479 Median : 0.2110   
## Mean : 0.0000 Mean : 0.0000   
## 3rd Qu.: 0.9501 3rd Qu.: 0.7913   
## Max. : 2.7187 Max. : 2.7210   
##

## Preprocess using transformation

## Training and Test Partition

In this step for data preparation we will partition the training dataset in training and validation sets using createDataPartition method from caret package. We will reserve 75% for training and rest 25% for validation purpose.

# Clustering Models

## K-means

## Hierarchical

# Principal Component Analysis

## Individual Substance Misuse

# Gradient Boosting: Suicide

The gradient boosting algorithm we will use for this example is XGBoost. XGBoost is an extremely popular machine learning model because of its speed and accuracy. XGBoost is a tree-based model that leverages gradient boosting to improve performance.

To prepare the data for model training we drop the rows that have null values contained in the Target column, in this case we are looking to predict Suicide. We will also remove the Non-subst Dx column because it contains a lot of null values. The XGBoost package in R expects the data to be converted to a numeric matrix prior to modeling.

| ```{r} gb\_\_train <-subset(training[complete.cases(training$Suicide), ], select= -`Non-subst Dx`) gb\_\_test <-subset(testing[complete.cases(testing$Suicide), ], select= -`Non-subst Dx`) y\_label\_tr <- as.matrix(gb\_\_train$Suicide) y\_label\_test <- as.matrix(gb\_\_test$Suicide) gb\_\_train <- sapply(subset(gb\_\_train, select = -Suicide), as.numeric) gb\_test <- sapply(subset(gb\_\_test, select = -Suicide), as.numeric) ``` |
| --- |

One of the advantages of using XGBoost is the large number of parameters you can tune. Hyperparameter tuning allows you to make sure the model performs well on both trained data and unseen data. In this case we have no prior domain expertise. We built a function to iterate over three parameters. We make use of the ParBayesianOptimization package to help us with tuning. This package uses bayes theory to help the model select what features to test while the model is fitting different variations. We also split the data into the three folds for crossfold validation to prevent overfitting.

| scoringFunction <- function(max\_depth, min\_child\_weight, subsample) {  dtrain <- xgb.DMatrix(gb\_\_train, label=y\_label\_tr)  Pars <- list(booster = "gbtree", eta = 0.01, max\_depth = max\_depth, min\_child\_weight = min\_child\_weight, subsample = subsample, objective = "binary:logistic", eval\_metric = "auc"  )  xgbcv <- xgb.cv(params = Pars, data = dtrain, nround = 100, folds = Folds, prediction = TRUE, showsd = TRUE, early\_stopping\_rounds = 5, maximize = TRUE, verbose = 0)  return(  list(   Score = max(xgbcv$evaluation\_log$test\_auc\_mean)  , nrounds = xgbcv$best\_iteration  ) )} |
| --- |

For the best model we used a max\_depth of 10, min child weight of 1, and subsample of .5. This model had an AUC of .76 on the training data and .8 level of accuracy on unseen data.

| Confusion Matrix and Statistics  y\_label\_test xgbpred 0 1  0 27 7  1 1 z   Accuracy : 0.8   95% CI : (0.6435, 0.9095)  No Information Rate : 0.7   P-Value [Acc > NIR] : 0.1110   Kappa : 0.4444   Mcnemar's Test P-Value : 0.0771   Sensitivity : 0.9643   Specificity : 0.4167   Pos Pred Value : 0.7941   Neg Pred Value : 0.8333   Prevalence : 0.7000   Detection Rate : 0.6750   Detection Prevalence : 0.8500   Balanced Accuracy : 0.6905     'Positive' Class : 0 |
| --- |

# Support Vector

# Build Models

## Linear Discriminant Analysis (LDA)

## Clustering Method

We use K-nearest neighbor (KNN) to identify clusters of patients that share similar patterns that could help us predict our target variable.

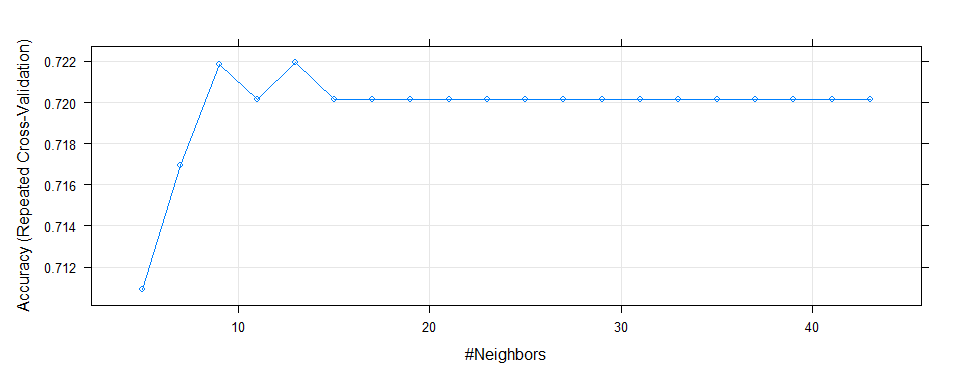
## [1] <NA>  
## Levels: 0 1 2

## [1] 0

## [1] <NA>  
## Levels: 0 1 2

## [1] 0

## k-Nearest Neighbors   
##   
## 132 samples  
## 50 predictor  
## 2 classes: '0', '1'   
##   
## Pre-processing: centered (157), scaled (157)   
## Resampling: Cross-Validated (10 fold, repeated 5 times)   
## Summary of sample sizes: 119, 119, 119, 118, 119, 118, ...   
## Resampling results across tuning parameters:  
##   
## k Accuracy Kappa   
## 5 0.7109158 0.0329247345  
## 7 0.7169597 0.0002244742  
## 9 0.7218315 0.0085714286  
## 11 0.7201648 0.0000000000  
## 13 0.7219414 0.0123065729  
## 15 0.7201648 0.0000000000  
## 17 0.7201648 0.0000000000  
## 19 0.7201648 0.0000000000  
## 21 0.7201648 0.0000000000  
## 23 0.7201648 0.0000000000  
## 25 0.7201648 0.0000000000  
## 27 0.7201648 0.0000000000  
## 29 0.7201648 0.0000000000  
## 31 0.7201648 0.0000000000  
## 33 0.7201648 0.0000000000  
## 35 0.7201648 0.0000000000  
## 37 0.7201648 0.0000000000  
## 39 0.7201648 0.0000000000  
## 41 0.7201648 0.0000000000  
## 43 0.7201648 0.0000000000  
##   
## Accuracy was used to select the optimal model using the largest value.  
## The final value used for the model was k = 13.



## [1] 0.7209302

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 0 1  
## 0 31 12  
## 1 0 0  
##   
## Accuracy : 0.7209   
## 95% CI : (0.5633, 0.8467)  
## No Information Rate : 0.7209   
## P-Value [Acc > NIR] : 0.576988   
##   
## Kappa : 0   
##   
## Mcnemar's Test P-Value : 0.001496   
##   
## Sensitivity : 1.0000   
## Specificity : 0.0000   
## Pos Pred Value : 0.7209   
## Neg Pred Value : NaN   
## Prevalence : 0.7209   
## Detection Rate : 0.7209   
## Detection Prevalence : 1.0000   
## Balanced Accuracy : 0.5000   
##   
## 'Positive' Class : 0   
##

Our KNN model accuracy comes out to 72.1%

## Decision Trees

## Random Forests

# Model Performance

# Conclusion

# References

<https://towardsdatascience.com/what-is-the-difference-between-pca-and-factor-analysis-5362ef6fa6f9>

<https://scholarworks.umass.edu/cgi/viewcontent.cgi?article=1226&context=pare>

# Code Appendix

knitr::opts\_chunk$set(echo=FALSE, error=FALSE, warning=FALSE, message=FALSE, fig.align="center", fig.width = 10)  
*# Libraries*  
  
library(summarytools)  
library(tidyverse)  
library(DataExplorer)  
library(reshape2)  
library(mice)  
library(caret)  
library(MASS)  
library(e1071)  
library(tree)  
library(corrplot)  
library(kableExtra)  
library(htmltools)  
library(readxl)  
library(psych)  
  
set.seed(622)  
  
set.seed(622)  
*# read data*  
adhd\_data <- read\_excel("ADHD\_data.xlsx", sheet = "Data") %>% na\_if("") %>% dplyr::select(-1)  
*#columns <- list(dimnames(adhd\_data)[2])*  
*#df <- adhd\_data[,2:53]*  
adhd\_data[,2:53] <- lapply(adhd\_data[,2:53], factor)  
dim(adhd\_data)  
adhd\_data[,c(23:37)]  
*# select categorical columns*  
cat\_cols <- dimnames(adhd\_data[,2:53])[[2]]  
adhd\_fact <- adhd\_data[cat\_cols]  
*# long format*  
adhd\_factm <- melt(adhd\_fact, measure.vars = cat\_cols, variable.name = 'metric', value.name = 'value')  
*# plot categorical columns*  
ggplot(adhd\_factm, aes(x = value)) +   
 geom\_bar(aes(fill = metric)) +   
 facet\_wrap( ~ metric, nrow = 5L, scales = 'free') + coord\_flip() +   
 theme(legend.position = "none")  
dfSummary(adhd\_data, style = 'grid', graph.col = FALSE)  
adhds <- sapply(adhd\_data[,c(4:21)], as.numeric) %>% cor()  
corrplot::corrplot(adhds, method="number")  
mds <- sapply(adhd\_data[,c(23:37)], as.numeric) %>% cor()  
corrplot::corrplot(mds, method="number")  
adhd\_ques\_fa <- factanal(sapply(adhd\_data[,c(4:21)], as.numeric),   
 factors = 3,   
 rotation = "promax",   
 scores = "regression")  
adhd\_ques\_fa  
fa.diagram(adhd\_ques\_fa$loadings)  
md\_ques\_fa <- factanal(sapply(adhd\_data[,c(23:37)], as.numeric),   
 factors = 3,   
 rotation = "promax",   
 scores = "regression")  
md\_ques\_fa  
fa.diagram(md\_ques\_fa$loadings)  
*# ADHD question scores dataframe*  
adhd\_ques\_fa <- as.data.frame(adhd\_ques\_fa$scores)   
names(adhd\_ques\_fa) <- c('ADHD\_FACT1','ADHD\_FACT2','ADHD\_FACT3')  
  
*# MD questions scores dataframe*  
md\_ques\_fa <- as.data.frame(md\_ques\_fa$scores)  
names(md\_ques\_fa) <- c('MD\_FACT1','MD\_FACT2','MD\_FACT3')  
  
*# remove ADHD and MD columns*  
adhd\_newdata <- adhd\_data %>% dplyr::select(-c(starts\_with('ADHD'), starts\_with('MD')))  
  
*# Add new factor columns created*  
adhd\_newdata <- cbind(adhd\_newdata, adhd\_ques\_fa, md\_ques\_fa)  
head(adhd\_newdata)  
*# plot missing values*  
plot\_missing(adhd\_newdata)  
*# rename columns to apply mice*  
adhd\_newdata <- adhd\_newdata %>%   
 rename('Sedative\_hypnotics'='Sedative-hypnotics',   
 'Court\_order' = 'Court order',   
 'Hx\_of\_Violence'='Hx of Violence',   
 'Disorderly\_Conduct'='Disorderly Conduct',   
 'Non\_subst\_Dx'='Non-subst Dx',  
 'Subst\_Dx'='Subst Dx',   
 'Psych\_meds'='Psych meds.') %>%   
 dplyr::select(-Psych\_meds)  
*# impute predictors using mice*  
adhd\_mice <- complete(mice(data=adhd\_newdata, print=FALSE))  
summary(adhd\_mice)  
*# Filter out*   
*#adhd\_data <- adhd\_data %>% filter(!is.na(Alcohol) &*  
*# !is.na(THC) &*  
*# !is.na(Cocaine) &*  
*# !is.na(Stimulants) &*  
*# !is.na(`Sedative-hypnotics`) &*  
*# !is.na(Opioids) &*  
*# !is.na(`Court order`) &*  
*# !is.na(Education) &*  
*# !is.na(`Hx of Violence`) &*  
*# !is.na(`Disorderly Conduct`) &*  
*# !is.na(Suicide) &*  
*# !is.na(Abuse) &*  
*# !is.na(`Non-subst Dx`) &*  
*# !is.na(`Subst Dx`) &*  
*# !is.na(`Psych meds.`))*  
*# impute numeric predictors using mice*  
*#adhd\_data <- complete(mice(data=adhd\_data[,:53], method="pmm", print=FALSE))*  
set.seed(622)  
adhd\_transform <- adhd\_mice %>%   
 dplyr::select(c("Age","ADHD\_FACT1","ADHD\_FACT2","ADHD\_FACT3",,"MD\_FACT1","MD\_FACT2","MD\_FACT3")) %>%  
 preProcess(method = c("center","scale")) %>%   
 predict(adhd\_mice)  
set.seed(622)  
partition <- createDataPartition(adhd\_data$Suicide, p=0.75, list = FALSE)  
training <- adhd\_data[partition,]  
testing <- adhd\_data[-partition,]  
*# training/validation partition for independent variables*  
*#X.train <- ld.clean[partition, ] %>% dplyr::select(-Loan\_Status)*  
*#X.test <- ld.clean[-partition, ] %>% dplyr::select(-Loan\_Status)*  
*# training/validation partition for dependent variable Loan\_Status*  
*#y.train <- ld.clean$Loan\_Status[partition]*  
*#y.test <- ld.clean$Loan\_Status[-partition]*  
set.seed(622)  
mode <- **function**(x){  
 levels <- unique(x)  
 indicies <- tabulate(match(x, levels))  
 levels[which.max(indicies)]  
}  
*# Clean up training data*  
training\_factors <- training %>%   
 dplyr::select(-Age, -`ADHD Total`, `MD TOTAL`)   
training\_factors <- data.frame(lapply(training\_factors, as.factor))  
train\_knn <- training\_factors %>%   
 mutate(across(everything(), ~replace\_na(., mode(.))))  
mode(train\_knn$Psych.meds.)   
train\_knn$Psych.meds.[which(is.na(train\_knn$Psych.meds.))] <- 0  
sum(is.na(train\_knn$Psych.meds.))  
  
*# Clean up testing data*  
testing\_factors <- testing %>%   
 dplyr::select(-Age, -`ADHD Total`, `MD TOTAL`)   
testing\_factors <- data.frame(lapply(testing\_factors, as.factor))  
test\_knn <- testing\_factors %>%   
 mutate(across(everything(), ~replace\_na(., mode(.))))  
mode(test\_knn$Psych.meds.)   
test\_knn$Psych.meds.[which(is.na(test\_knn$Psych.meds.))] <- 0  
sum(is.na(test\_knn$Psych.meds.))  
  
*# Train KNN model*  
train.knn <- (train\_knn[, names(train\_knn) != "Suicide"])  
prep <- preProcess(x = train.knn, method = c("center", "scale"))  
cl <- trainControl(method="repeatedcv", repeats = 5)   
knn\_model <- train(Suicide ~ ., data = train\_knn,   
 method = "knn",   
 trControl = cl,   
 preProcess = c("center","scale"),   
 tuneLength = 20)  
knn\_model   
*# Evaluate Model*  
plot(knn\_model)  
knn\_predict <- predict(knn\_model, newdata = test\_knn)  
mean(knn\_predict == test\_knn$Suicide) *# accuracy*  
conf.mat.knn <- confusionMatrix(knn\_predict, test\_knn$Suicide)  
accuracy <- round(conf.mat.knn$overall[[1]], 3)\*100  
conf.mat.knn