Data622 - Group2 - Homework4

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

10/21/2021

Contents

Overview

In this project, we analyze a real-life mental health dataset to provide context around suicide prediction given a variety of unidentifiable demographic data. Our goals are to understand the variables relationships, identify those variables that influence our target, and develop models that can predict a patient's risk of suicide.

Approach

We will first perform exploratory data analysis (EDA) on the dataset to inform our analysis and build better models. Methods include Clustering, Principal Component Analysis, Gradient Boosting, and Support Vector Machines. This EDA step is crucial to understanding variables' relationships and identifying which variables influence our target.

Once we understand the data, we prepare it for modeling. This includes partitioning the data with a 75-25 train-test split, performing necessary imputations, relevant centering and scaling, and more as outlined in our data exploration and preparation sections. When building our models we focus on using methods that produce real-world accuracy. For this reason, we attempt to select the best generalizable model with accuracy as our primary indicator during model evaluation.

Data Exploration

The dataset with its column IDs, variable names, and variables descriptions are provided below for reference.

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

Columns	Variable	Description
BB	Psychiatric Meds	0 - none; 1 - one psychotropic med; 2 - more than one psychotropic med

Notice how the data is grouped. There are labels for ADHD, Mood disorders, and Individual Substance misuse across a range of columns. These groups are reviewed throughout the exploration process and new features are generated to attempt to improve model performance.

Data Characteristics

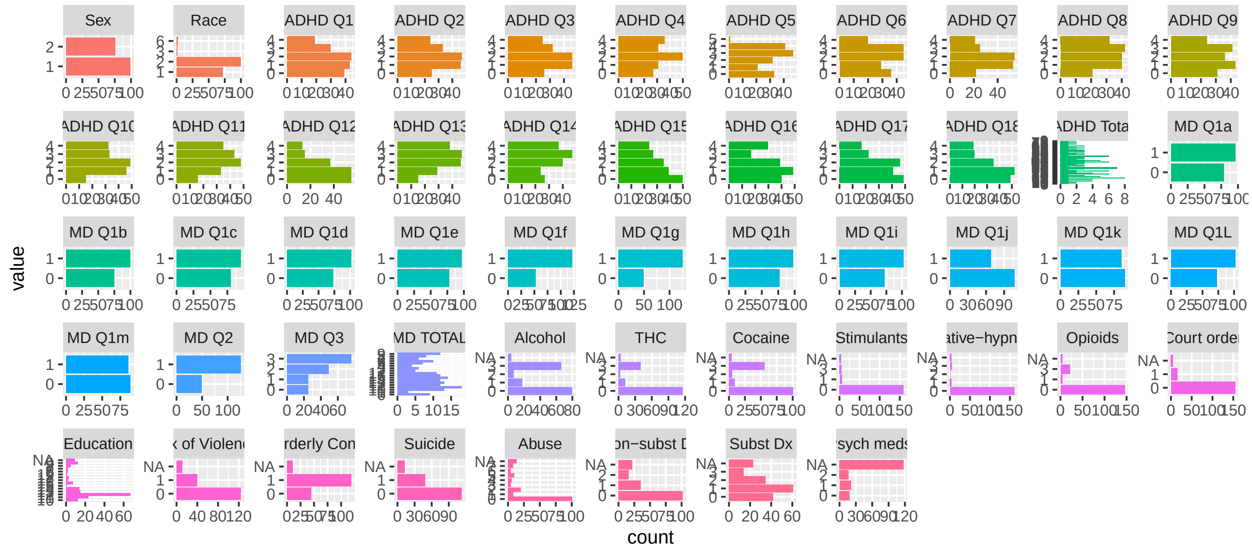
The data contains 175 observations of 53 variables. We import the data from a remote repository and find that 51 of the variables should be of the factor data type given clear levels in their distributions. As is, these variables are interpreted as character strings. This will need to be converted for realistic results. The remaining variables can be numeric for our purposes.

We review one grouped variable set, known as mood disorders (MD), to show what we're working with. These contain a series of associated questions (Q1-Q3) with Q1 containing parts 'a' through 'm.'

```
## # 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>
```

Each part of Q1 'a' through 'm' corresponds with a specific question related to mood disorders for a single patient. In our feature engineering, it may be useful to tally these responses for a more holistic perspective of the patient's overall mood. We repeat this for the other groups to get an sense of the patient well-being which should provide insight into their risk of suicide.

We examine the distribution of responses. This will inform us if there are any questions that are overly represented, missing values, or in need of adjustments or transformations prior to modeling. Assuming that we factorize responses that contain clear levels, we visualize the results as bars for each level of each variable.



The distributions of these values are spread somewhat evenly and should not require transformation at first glance. It is also clear that there are some missing values with various drug use reponses such as THC, Cocaine, Stimulants, Opioids, and more. 'Pschy Meds' med is missing the most values, with nearly all of its values listed as 'NA.' Since there are few missing values relative to the number of observations in the dataset, any basic imputation method should limit the effect on our predictions.

We can visually tell that the data leans in certain directions which may influence how generalizable our results can be. For example, responses only include two racial groups, 1 and 2 with 2 being the most prominent. The rest have so few values their results may be insignificant. The data also leans towards sex 1, of binary-gender categories. There are more unbalanced occurrences such as this in questions for mood, education, abuse, disorderly conduct, substance use, psych med, and our target variable suicide.

Adjustments may be necessary for some sub groups, for example in the responses of those with drug use, but the data are otherwise ready for use. We continue with a summary of the data.

Data summary

We identify any clusters of patients' responses that appear to be imbalanced by counting the frequencies of responses by level. This will provide us percentages of the previous bar chart where we could only compare bar sizes. It also shows if there is any duplication in responses.

```
## 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        |
```

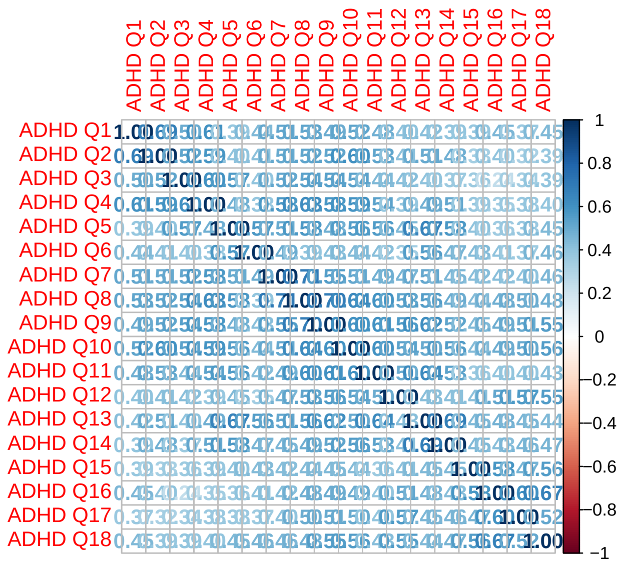
```
## |    | [factor]           | 2. 1                    |  15 ( 8.8%)         | (97.1%)  | (2.9%)  |
## +----+--------------------+-------------------------+---------------------+----------+---------+
## | 46 | Education          | 1. 6                    |  2 ( 1.2%)          | 166      | 9       |
## |    | [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%)          |          |         |
## +----+--------------------+-------------------------+---------------------+----------+---------+
```

Through this table we now know that some groups are more representative than others in the responses. Racial makeup is certainly limited in this study. Age might also be a limiting factor when ensuring predictive accuracy. However, duplication is not an issue. There appear to be no duplicates and the frequencies are close enough to normal that we can build realistic models with the data as it is. This confirms our characteristics.

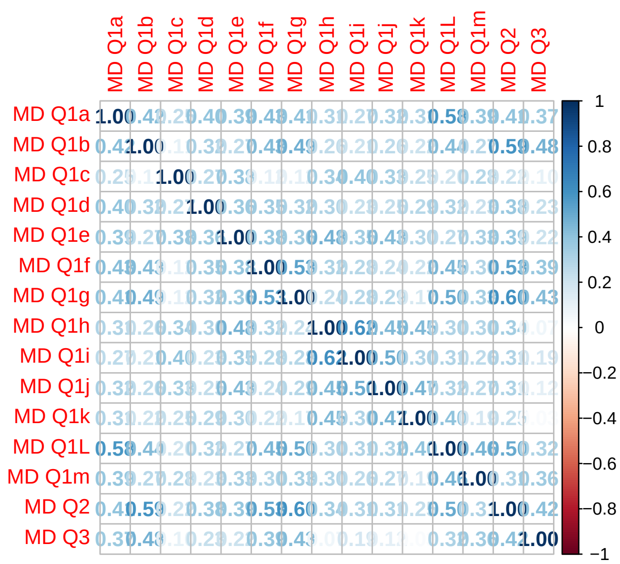
Correlation

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.

ADHD Correlation Plot:



Mood Correlation Plot:



Within both groups, there are few, if any, particularly strong correlations that we could attribute to a patient's risk of suicide. Our most strongly correlated variables are found in the ADHD group which has a maximum correlation value of about 0.71 between Q7 and Q8. Everything else is lower but importantly, not zero.

These correlations imply that the presence of answers to ADHD questions has a slightly higher correlation on the patient's risk of attempting suicide. Mood disorders have a lower set of correlations values with its maximum around 0.60. Although, plenty of these values are close enough to 0, which indicates no correlation, that we can say they have little no connection with a patients risk of attempting suicide.

Data Preparation

Given our EDA, we start preparing the data for modeling by handling missing values, analyzing which factors and factor levels would be most beneficial to the models, then preprocess and transform the data to fit model parameters when appropriate. We also split the data into training and test data sets for evaluating model performance. To complete preparation, we perform a principal component analysis (PCA) to compliment our factor analysis.

Factor Analysis

In Factor Analysis the dimensions of the data are reduced by picking a number of variables that explain most of the variation in the data. Those variables that are not used are called latent variables because their values are inferred from other variables. This will help us further identify underlying factors that explain the correlation among the set of variables. Factor analysis is a great tool for treating multivariate questionnaire studies.

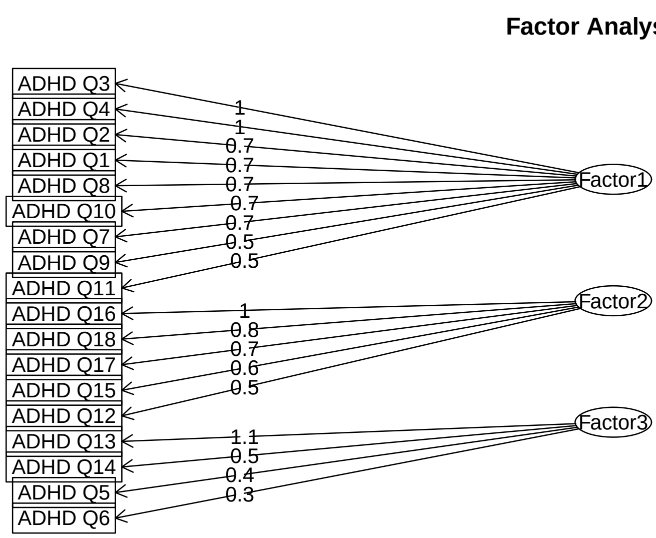
For ADHD questions, a test of the hypothesis determined that 3 factors are sufficient to explain the variation of the data. Our chi square statistic is 197.3 on 102 degrees of freedom. The p-value is 0.0000000476. Which, for our purposes, is higly significant. We used regression factor scores here as they predict the location of each individual on the factor. Results of this process are shown below.

```
##
## Call:
## factanal(x = sapply(adhd_data[, c(4:21)], as.numeric), factors = 3,     scores = "regression", rotati
##
## 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
```

To simplify the results from our regressive factanal call, this visual helps to describe which factors were used to build an our latent variables. The value listed on the arrow pointing to the ADHD question is its correlation with the original patient's response.

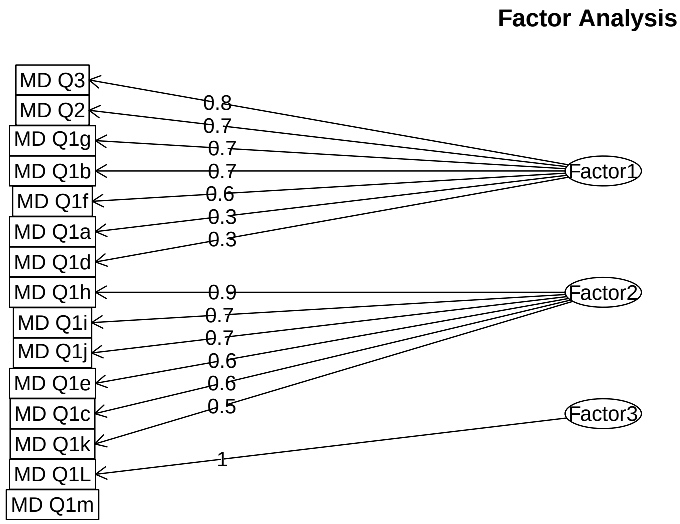
Notice that none of the correlations are below 0.5 in factors 1 and 2. These are our most indicative questions for determining a patient's risk of attempting suicide. If we were looking to shorten the ADHD questionaire, these ADHD questions would be the best selection to retain predictive value in models.



From the source data, we can see that for MD questions only the first MD question has multiple sub questions or parts. Compared to the second and third question, which have no additional parts. We will not need to perform fator analysis on questions with no additional parts because there are no additional factors to analyze.

We apply a similar factor analysis as of ADHD questions. Our hypothesis test showed that 3 factors were again sufficient. The chi square statistic is 88.82 on 63 degrees of freedom. The p-value is 0.0178. Significant to an alpha level below 0.05. Results of this test are shown below.

```
##
## Call:
## factanal(x = sapply(adhd_data[, c(23:37)], as.numeric), factors = 3,     scores = "regression", rota
##
## 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.

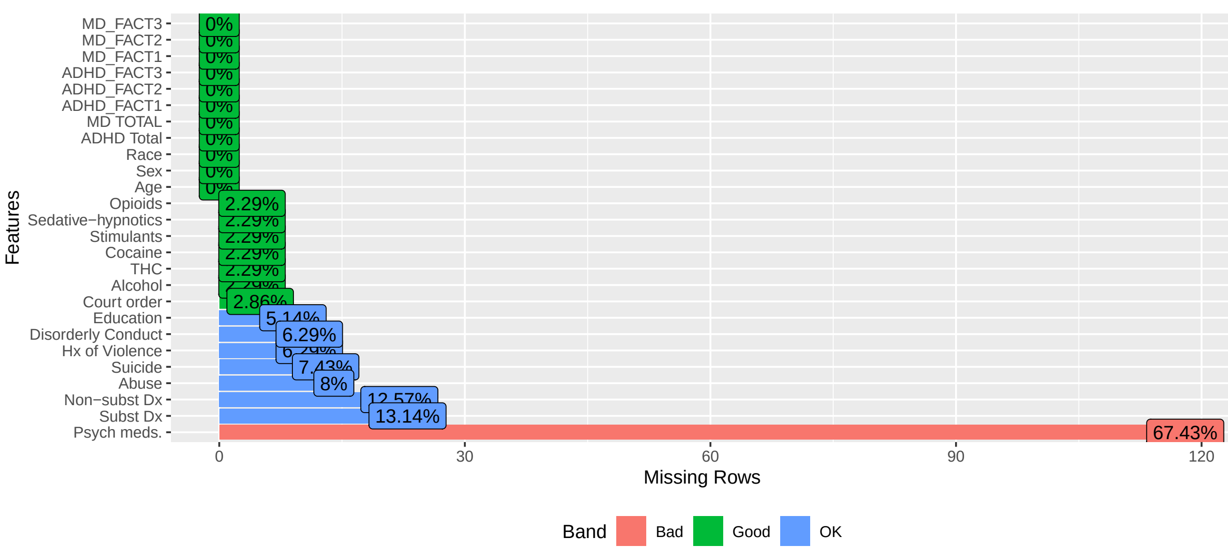
Here is glimpse of new set of data.

```
##   Age Sex Race ADHD Total MD TOTAL Alcohol THC Cocaine Stimulants
## 1  24   1    1         40       15       1   1       1          0
## 2  48   2    1         55       14       0   0       0          0
## 3  51   2    1         31        5       0   0       0          0
## 4  43   1    1         45       13       1   1       1          1
## 5  34   1    1         48        7       1   1       0          0
## 6  39   2    1         55       14       1   0       0          0
##   Sedative-hypnotics Opioids Court order Education Hx of Violence
## 1                  0       0           1        11              0
## 2                  0       0           0        14              0
## 3                  0       0           0        12              0
## 4                  0       0           0        12              0
## 5                  0       0           1         9              1
## 6                  0       0           0        11              0
##   Disorderly Conduct Suicide Abuse Non-subst Dx Subst Dx Psych meds. ADHD_FACT1
## 1                  1       1     0            2        0           2  1.6922046
## 2                  0       1     4            1        0           1  2.0799334
## 3                  0       0     6            2        0           1 -0.5301540
## 4                  0       1     7            2        0           2  0.9321586
## 5                  1       1     0            2        0           0  2.5823393
## 6                  1       1     2            0        0           0 -0.8422991
##   ADHD_FACT2 ADHD_FACT3  MD_FACT1    MD_FACT2   MD_FACT3
## 1  1.6740898 -3.3243648 1.3855809  1.59853502 -2.8509956
## 2  1.5195976 -2.6626233 0.7973360 -0.08361024  0.3704740
## 3  0.3261461 -0.1297720 0.4673052 -0.80624391 -1.0898824
## 4 -0.5385242  0.2275811 0.8725442 -0.45310917  0.5084134
## 5 -1.6535142 -0.2129593 2.1105464 -1.37884110 -1.8594994
## 6  1.3342893  1.0827141 0.2816620  0.44486489  0.4402313
```

The dimensions of this new data set are 175 observations of 25 variables. We have successfully reduced the number of variables by 0.53% while still explaining most of the variation in the data. Future questionnaires could be reduced to those variables listed in our new data set if they were in need of reducing the number of ADHD questions.

Handling missing values

When summarizing the data we discovered missing values for several of the variables. The variable psych med stood out the most since it was missing more than half of its data. Here, we visualize the percentage of missing values for all variables to compare to one another. Green bars indicate an insignificant level of missing values, blue could use imputation or another method to handle, and red is practically useless unless special precautions are taken during analysis.



We can see from this chart that `Psych meds`, in red, contributes to 67.43% of the missing data which is the 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) for columns that contain missing values.

```
##  Alcohol THC     Cocaine Stimulants Sedative_hypnotics Opioids Court_order
##  0:80    0:118   0:102   0:163      0:162              0:147   0:158
##  1:21    1: 13   1: 10   1:  7      1:  5              1:  6   1: 17
##  2: 8    2:  4   2:  7   3:  5      2:  1              3: 22
##  3:66    3: 40   3: 56              3:  7
##
##
##
##    Education  Hx_of_Violence Disorderly_Conduct Suicide     Abuse
##  12      :68  0:132          0: 47              0:124   0      :108
##  11      :25  1: 43          1:128              1: 51   2      : 21
##  13      :16                                            5      : 12
##  14      :14                                            7      : 11
##  9       :13                                            1      :  8
##  10      :13                                            4      :  6
##  (Other):26                                             (Other):  9
##  Non_subst_Dx Subst_Dx
```

```
##  0:114       0:51
##  1: 39       1:66
##  2: 22       2:38
##              3:20
##
##
##
```

```
##   Alcohol THC Cocaine Stimulants Sedative_hypnotics Opioids Court_order
## 1       1   1       1          0                  0       0           1
## 2       0   0       0          0                  0       0           0
## 3       0   0       0          0                  0       0           0
## 4       1   1       1          1                  0       0           0
## 5       1   1       0          0                  0       0           1
## 6       1   0       0          0                  0       0           0
##   Education Hx_of_Violence Disorderly_Conduct Suicide Abuse Non_subst_Dx
## 1        11              0                  1       1     0            2
## 2        14              0                  0       1     4            1
## 3        12              0                  0       0     6            2
## 4        12              0                  0       1     7            2
## 5         9              1                  1       1     0            2
## 6        11              0                  1       1     2            0
##   Subst_Dx ADHD_Total ADHD_FACT1 ADHD_FACT2 ADHD_FACT3 MD_Total  MD_FACT1
## 1        0         40  1.6922046  1.6740898 -3.3243648       15 1.3855809
## 2        0         55  2.0799334  1.5195976 -2.6626233       14 0.7973360
## 3        0         31 -0.5301540  0.3261461 -0.1297720        5 0.4673052
## 4        0         45  0.9321586 -0.5385242  0.2275811       13 0.8725442
## 5        0         48  2.5823393 -1.6535142 -0.2129593        7 2.1105464
## 6        0         55 -0.8422991  1.3342893  1.0827141       14 0.2816620
##      MD_FACT2   MD_FACT3 Race Sex Age
## 1  1.59853502 -2.8509956    1   1  24
## 2 -0.08361024  0.3704740    1   2  48
## 3 -0.80624391 -1.0898824    1   2  51
## 4 -0.45310917  0.5084134    1   1  43
## 5 -1.37884110 -1.8594994    1   1  34
## 6  0.44486489  0.4402313    1   2  39
```

From here, we have imputed or removed variables that contained missing values and identified those factors that could explain most of the data with fewer variables. We continue with some preprocessing steps.

Preprocess using transformation

There are minor changes necessary to fit and feed the models in our analysis. In our first transformation, we use dummyVars to create dummy variables for categorical features. Next we center and scale the data so that models can evaluate the data on the same dimensional planes. Otherwise, we lose any chance predicting with accuracy in real-world settings.

We can review the transformations with the following table of values. These show all the variable's on the same scale with a center at zero. With this dataset it would be possible to supply data to the support vector machine and perform a proper clustering algorithm.

```
##    Alcohol.0  Alcohol.1 Alcohol.2  Alcohol.3     THC.0      THC.1      THC.2
## 1 -0.9150373  2.7002645 -0.218244 -0.7759153 -1.434694  3.5199900 -0.1525062
```

```
## 2  1.0866068 -0.3682179 -0.218244 -0.7759153  0.693030 -0.2824683 -0.1525062
## 3  1.0866068 -0.3682179 -0.218244 -0.7759153  0.693030 -0.2824683 -0.1525062
## 4 -0.9150373  2.7002645 -0.218244 -0.7759153 -1.434694  3.5199900 -0.1525062
## 5 -0.9150373  2.7002645 -0.218244 -0.7759153 -1.434694  3.5199900 -0.1525062
## 6 -0.9150373  2.7002645 -0.218244 -0.7759153  0.693030 -0.2824683 -0.1525062
##        THC.3  Cocaine.0  Cocaine.1  Cocaine.2  Cocaine.3 Stimulants.0
## 1 -0.5427736 -1.1786755  4.0503968 -0.2035401 -0.6840315     0.270553
## 2 -0.5427736  0.8435619 -0.2454786 -0.2035401 -0.6840315     0.270553
## 3 -0.5427736  0.8435619 -0.2454786 -0.2035401 -0.6840315     0.270553
## 4 -0.5427736 -1.1786755  4.0503968 -0.2035401 -0.6840315    -3.675012
## 5 -0.5427736  0.8435619 -0.2454786 -0.2035401 -0.6840315     0.270553
## 6 -0.5427736  0.8435619 -0.2454786 -0.2035401 -0.6840315     0.270553
##   Stimulants.1 Stimulants.3 Sedative_hypnotics.0 Sedative_hypnotics.1
## 1   -0.2035401   -0.1710079            0.2824683           -0.1710079
## 2   -0.2035401   -0.1710079            0.2824683           -0.1710079
## 3   -0.2035401   -0.1710079            0.2824683           -0.1710079
## 4    4.8849623   -0.1710079            0.2824683           -0.1710079
## 5   -0.2035401   -0.1710079            0.2824683           -0.1710079
## 6   -0.2035401   -0.1710079            0.2824683           -0.1710079
##   Sedative_hypnotics.2 Sedative_hypnotics.3 Opioids.0  Opioids.1  Opioids.3
## 1          -0.07559289           -0.2035401  0.435187 -0.1878832 -0.3781127
## 2          -0.07559289           -0.2035401  0.435187 -0.1878832 -0.3781127
## 3          -0.07559289           -0.2035401  0.435187 -0.1878832 -0.3781127
## 4          -0.07559289           -0.2035401  0.435187 -0.1878832 -0.3781127
## 5          -0.07559289           -0.2035401  0.435187 -0.1878832 -0.3781127
## 6          -0.07559289           -0.2035401  0.435187 -0.1878832 -0.3781127
##   Court_order.0 Court_order.1 Education.6 Education.7 Education.8 Education.9
## 1    -3.0399027     3.0399027  -0.1316898   -0.107213  -0.1878832  -0.2824683
## 2     0.3270781    -0.3270781  -0.1316898   -0.107213  -0.1878832  -0.2824683
## 3     0.3270781    -0.3270781  -0.1316898   -0.107213  -0.1878832  -0.2824683
## 4     0.3270781    -0.3270781  -0.1316898   -0.107213  -0.1878832  -0.2824683
## 5    -3.0399027     3.0399027  -0.1316898   -0.107213  -0.1878832   3.5199900
## 6     0.3270781    -0.3270781  -0.1316898   -0.107213  -0.1878832  -0.2824683
##   Education.10 Education.11 Education.12 Education.13 Education.14 Education.15
## 1   -0.2824683    2.4424812   -0.7949104    -0.316313   -0.2940402  -0.07559289
## 2   -0.2824683   -0.4070802   -0.7949104    -0.316313    3.3814621  -0.07559289
## 3   -0.2824683   -0.4070802    1.2508149    -0.316313   -0.2940402  -0.07559289
## 4   -0.2824683   -0.4070802    1.2508149    -0.316313   -0.2940402  -0.07559289
## 5   -0.2824683   -0.4070802   -0.7949104    -0.316313   -0.2940402  -0.07559289
## 6   -0.2824683    2.4424812   -0.7949104    -0.316313   -0.2940402  -0.07559289
##   Education.16 Education.17 Education.18 Education.19 Hx_of_Violence.0
## 1    -0.218244    -0.107213   -0.1316898  -0.07559289        0.5691187
## 2    -0.218244    -0.107213   -0.1316898  -0.07559289        0.5691187
## 3    -0.218244    -0.107213   -0.1316898  -0.07559289        0.5691187
## 4    -0.218244    -0.107213   -0.1316898  -0.07559289        0.5691187
## 5    -0.218244    -0.107213   -0.1316898  -0.07559289       -1.7470621
## 6    -0.218244    -0.107213   -0.1316898  -0.07559289        0.5691187
##   Hx_of_Violence.1 Disorderly_Conduct.0 Disorderly_Conduct.1    Abuse.0
## 1       -0.5691187           -0.6042262            0.6042262  0.7853823
## 2       -0.5691187            1.6455522           -1.6455522 -1.2659894
## 3       -0.5691187            1.6455522           -1.6455522 -1.2659894
## 4       -0.5691187            1.6455522           -1.6455522 -1.2659894
## 5        1.7470621           -0.6042262            0.6042262  0.7853823
## 6       -0.5691187           -0.6042262            0.6042262 -1.2659894
```

```
##     Abuse.1    Abuse.2    Abuse.3    Abuse.4   Abuse.5    Abuse.6    Abuse.7
## 1 -0.218244 -0.3682179 -0.1710079 -0.1878832 -0.270553 -0.1525062 -0.2582439
## 2 -0.218244 -0.3682179 -0.1710079  5.2920425 -0.270553 -0.1525062 -0.2582439
## 3 -0.218244 -0.3682179 -0.1710079 -0.1878832 -0.270553  6.5196407 -0.2582439
## 4 -0.218244 -0.3682179 -0.1710079 -0.1878832 -0.270553 -0.1525062  3.8501813
## 5 -0.218244 -0.3682179 -0.1710079 -0.1878832 -0.270553 -0.1525062 -0.2582439
## 6 -0.218244  2.7002645 -0.1710079 -0.1878832 -0.270553 -0.1525062 -0.2582439
##   Non_subst_Dx.0 Non_subst_Dx.1 Non_subst_Dx.2 Subst_Dx.0 Subst_Dx.1 Subst_Dx.2
## 1     -1.3631483      -0.533972      2.6296017   1.554824 -0.7759153 -0.5251545
## 2     -1.3631483       1.862056     -0.3781127   1.554824 -0.7759153 -0.5251545
## 3     -1.3631483      -0.533972      2.6296017   1.554824 -0.7759153 -0.5251545
## 4     -1.3631483      -0.533972      2.6296017   1.554824 -0.7759153 -0.5251545
## 5     -1.3631483      -0.533972      2.6296017   1.554824 -0.7759153 -0.5251545
## 6      0.7294039      -0.533972     -0.3781127   1.554824 -0.7759153 -0.5251545
##   Subst_Dx.3 ADHD_Total.0 ADHD_Total.1 ADHD_Total.3 ADHD_Total.5 ADHD_Total.6
## 1 -0.3581828  -0.07559289    -0.107213  -0.07559289  -0.07559289   -0.1316898
## 2 -0.3581828  -0.07559289    -0.107213  -0.07559289  -0.07559289   -0.1316898
## 3 -0.3581828  -0.07559289    -0.107213  -0.07559289  -0.07559289   -0.1316898
## 4 -0.3581828  -0.07559289    -0.107213  -0.07559289  -0.07559289   -0.1316898
## 5 -0.3581828  -0.07559289    -0.107213  -0.07559289  -0.07559289   -0.1316898
## 6 -0.3581828  -0.07559289    -0.107213  -0.07559289  -0.07559289   -0.1316898
##   ADHD_Total.7 ADHD_Total.8 ADHD_Total.9 ADHD_Total.10 ADHD_Total.11
## 1    -0.107213  -0.07559289    -0.107213     -0.107213   -0.07559289
## 2    -0.107213  -0.07559289    -0.107213     -0.107213   -0.07559289
## 3    -0.107213  -0.07559289    -0.107213     -0.107213   -0.07559289
## 4    -0.107213  -0.07559289    -0.107213     -0.107213   -0.07559289
## 5    -0.107213  -0.07559289    -0.107213     -0.107213   -0.07559289
## 6    -0.107213  -0.07559289    -0.107213     -0.107213   -0.07559289
##   ADHD_Total.12 ADHD_Total.13 ADHD_Total.14 ADHD_Total.16 ADHD_Total.17
## 1    -0.1525062   -0.07559289    -0.1525062   -0.07559289     -0.218244
## 2    -0.1525062   -0.07559289    -0.1525062   -0.07559289     -0.218244
## 3    -0.1525062   -0.07559289    -0.1525062   -0.07559289     -0.218244
## 4    -0.1525062   -0.07559289    -0.1525062   -0.07559289     -0.218244
## 5    -0.1525062   -0.07559289    -0.1525062   -0.07559289     -0.218244
## 6    -0.1525062   -0.07559289    -0.1525062   -0.07559289     -0.218244
##   ADHD_Total.18 ADHD_Total.19 ADHD_Total.20 ADHD_Total.21 ADHD_Total.23
## 1   -0.07559289    -0.1710079    -0.1316898    -0.1316898   -0.07559289
## 2   -0.07559289    -0.1710079    -0.1316898    -0.1316898   -0.07559289
## 3   -0.07559289    -0.1710079    -0.1316898    -0.1316898   -0.07559289
## 4   -0.07559289    -0.1710079    -0.1316898    -0.1316898   -0.07559289
## 5   -0.07559289    -0.1710079    -0.1316898    -0.1316898   -0.07559289
## 6   -0.07559289    -0.1710079    -0.1316898    -0.1316898   -0.07559289
##   ADHD_Total.24 ADHD_Total.25 ADHD_Total.26 ADHD_Total.27 ADHD_Total.28
## 1    -0.1878832    -0.1525062   -0.07559289     -0.107213    -0.1878832
## 2    -0.1878832    -0.1525062   -0.07559289     -0.107213    -0.1878832
## 3    -0.1878832    -0.1525062   -0.07559289     -0.107213    -0.1878832
## 4    -0.1878832    -0.1525062   -0.07559289     -0.107213    -0.1878832
## 5    -0.1878832    -0.1525062   -0.07559289     -0.107213    -0.1878832
## 6    -0.1878832    -0.1525062   -0.07559289     -0.107213    -0.1878832
##   ADHD_Total.29 ADHD_Total.30 ADHD_Total.31 ADHD_Total.32 ADHD_Total.33
## 1     -0.107213    -0.1316898    -0.2035401    -0.2035401    -0.1316898
## 2     -0.107213    -0.1316898    -0.2035401    -0.2035401    -0.1316898
## 3     -0.107213    -0.1316898     4.8849623    -0.2035401    -0.1316898
## 4     -0.107213    -0.1316898    -0.2035401    -0.2035401    -0.1316898
```

```
## 5     -0.107213    -0.1316898    -0.2035401    -0.2035401    -0.1316898
## 6     -0.107213    -0.1316898    -0.2035401    -0.2035401    -0.1316898
##   ADHD_Total.34 ADHD_Total.35 ADHD_Total.36 ADHD_Total.37 ADHD_Total.38
## 1   -0.07559289    -0.1316898    -0.1316898     -0.107213    -0.1316898
## 2   -0.07559289    -0.1316898    -0.1316898     -0.107213    -0.1316898
## 3   -0.07559289    -0.1316898    -0.1316898     -0.107213    -0.1316898
## 4   -0.07559289    -0.1316898    -0.1316898     -0.107213    -0.1316898
## 5   -0.07559289    -0.1316898    -0.1316898     -0.107213    -0.1316898
## 6   -0.07559289    -0.1316898    -0.1316898     -0.107213    -0.1316898
##   ADHD_Total.39 ADHD_Total.40 ADHD_Total.41 ADHD_Total.42 ADHD_Total.43
## 1    -0.1316898     5.2920425    -0.1316898    -0.1710079    -0.1316898
## 2    -0.1316898    -0.1878832    -0.1316898    -0.1710079    -0.1316898
## 3    -0.1316898    -0.1878832    -0.1316898    -0.1710079    -0.1316898
## 4    -0.1316898    -0.1878832    -0.1316898    -0.1710079    -0.1316898
## 5    -0.1316898    -0.1878832    -0.1316898    -0.1710079    -0.1316898
## 6    -0.1316898    -0.1878832    -0.1316898    -0.1710079    -0.1316898
##   ADHD_Total.44 ADHD_Total.45 ADHD_Total.46 ADHD_Total.47 ADHD_Total.48
## 1     -0.107213    -0.1316898    -0.1316898    -0.1316898    -0.1878832
## 2     -0.107213    -0.1316898    -0.1316898    -0.1316898    -0.1878832
## 3     -0.107213    -0.1316898    -0.1316898    -0.1316898    -0.1878832
## 4     -0.107213     7.5502129    -0.1316898    -0.1316898    -0.1878832
## 5     -0.107213    -0.1316898    -0.1316898    -0.1316898     5.2920425
## 6     -0.107213    -0.1316898    -0.1316898    -0.1316898    -0.1878832
##   ADHD_Total.49 ADHD_Total.50 ADHD_Total.51 ADHD_Total.52 ADHD_Total.53
## 1    -0.1878832    -0.1316898     -0.107213    -0.1316898   -0.07559289
## 2    -0.1878832    -0.1316898     -0.107213    -0.1316898   -0.07559289
## 3    -0.1878832    -0.1316898     -0.107213    -0.1316898   -0.07559289
## 4    -0.1878832    -0.1316898     -0.107213    -0.1316898   -0.07559289
## 5    -0.1878832    -0.1316898     -0.107213    -0.1316898   -0.07559289
## 6    -0.1878832    -0.1316898     -0.107213    -0.1316898   -0.07559289
##   ADHD_Total.54 ADHD_Total.55 ADHD_Total.56 ADHD_Total.57 ADHD_Total.58
## 1    -0.1316898    -0.1316898    -0.1316898     -0.107213   -0.07559289
## 2    -0.1316898     7.5502129    -0.1316898     -0.107213   -0.07559289
## 3    -0.1316898    -0.1316898    -0.1316898     -0.107213   -0.07559289
## 4    -0.1316898    -0.1316898    -0.1316898     -0.107213   -0.07559289
## 5    -0.1316898    -0.1316898    -0.1316898     -0.107213   -0.07559289
## 6    -0.1316898     7.5502129    -0.1316898     -0.107213   -0.07559289
##   ADHD_Total.62 ADHD_Total.63 ADHD_Total.65 ADHD_Total.67 ADHD_Total.69
## 1     -0.107213   -0.07559289    -0.1316898   -0.07559289   -0.07559289
## 2     -0.107213   -0.07559289    -0.1316898   -0.07559289   -0.07559289
## 3     -0.107213   -0.07559289    -0.1316898   -0.07559289   -0.07559289
## 4     -0.107213   -0.07559289    -0.1316898   -0.07559289   -0.07559289
## 5     -0.107213   -0.07559289    -0.1316898   -0.07559289   -0.07559289
## 6     -0.107213   -0.07559289    -0.1316898   -0.07559289   -0.07559289
##   ADHD_Total.71 ADHD_Total.72 ADHD_FACT1 ADHD_FACT2  ADHD_FACT3 MD_Total.0
## 1   -0.07559289     -0.107213  1.0783202  1.2371956 -2.13013067 -0.2321789
## 2   -0.07559289     -0.107213  1.3253918  1.1230219 -1.70611102 -0.2321789
## 3   -0.07559289     -0.107213 -0.3378290  0.2410304 -0.08315315 -0.2321789
## 4   -0.07559289     -0.107213  0.5939976 -0.3979833  0.14582561 -0.2321789
## 5   -0.07559289     -0.107213  1.6455389 -1.2219897 -0.13645646 -0.2321789
## 6   -0.07559289     -0.107213 -0.5367366  0.9860743  0.69376338 -0.2321789
##   MD_Total.1 MD_Total.2 MD_Total.3 MD_Total.4 MD_Total.5 MD_Total.6 MD_Total.7
## 1 -0.1316898 -0.1710079 -0.1878832 -0.1525062 -0.2035401 -0.2454786 -0.1878832
## 2 -0.1316898 -0.1710079 -0.1878832 -0.1525062 -0.2035401 -0.2454786 -0.1878832
```

```
## 3 -0.1316898 -0.1710079 -0.1878832 -0.1525062  4.8849623 -0.2454786 -0.1878832
## 4 -0.1316898 -0.1710079 -0.1878832 -0.1525062 -0.2035401 -0.2454786 -0.1878832
## 5 -0.1316898 -0.1710079 -0.1878832 -0.1525062 -0.2035401 -0.2454786  5.2920425
## 6 -0.1316898 -0.1710079 -0.1878832 -0.1525062 -0.2035401 -0.2454786 -0.1878832
##   MD_Total.8 MD_Total.9 MD_Total.10 MD_Total.11 MD_Total.12 MD_Total.13
## 1  -0.218244  -0.270553  -0.2824683  -0.3376308   -0.270553  -0.2824683
## 2  -0.218244  -0.270553  -0.2824683  -0.3376308   -0.270553  -0.2824683
## 3  -0.218244  -0.270553  -0.2824683  -0.3376308   -0.270553  -0.2824683
## 4  -0.218244  -0.270553  -0.2824683  -0.3376308   -0.270553   3.5199900
## 5  -0.218244  -0.270553  -0.2824683  -0.3376308   -0.270553  -0.2824683
## 6  -0.218244  -0.270553  -0.2824683  -0.3376308   -0.270553  -0.2824683
##   MD_Total.14 MD_Total.15 MD_Total.16 MD_Total.17  MD_FACT1    MD_FACT2
## 1   -0.270553   3.3814621   -0.270553  -0.2582439 1.2270551  1.42214452
## 2    3.675012  -0.2940402   -0.270553  -0.2582439 0.7061119 -0.07438426
## 3   -0.270553  -0.2940402   -0.270553  -0.2582439 0.4138403 -0.71727885
## 4   -0.270553  -0.2940402   -0.270553  -0.2582439 0.7727155 -0.40311079
## 5   -0.270553  -0.2940402   -0.270553  -0.2582439 1.8690765 -1.22669275
## 6    3.675012  -0.2940402   -0.270553  -0.2582439 0.2494367  0.39577623
##     MD_FACT3   Race.1    Race.2      Race.3    Race.6      Sex.1      Sex.2
## 1 -2.2790099 1.192636 -1.151397 -0.07559289 -0.107213  0.8736647 -0.8736647
## 2  0.2961471 1.192636 -1.151397 -0.07559289 -0.107213 -1.1380633  1.1380633
## 3 -0.8712229 1.192636 -1.151397 -0.07559289 -0.107213 -1.1380633  1.1380633
## 4  0.4064121 1.192636 -1.151397 -0.07559289 -0.107213  0.8736647 -0.8736647
## 5 -1.4864343 1.192636 -1.151397 -0.07559289 -0.107213  0.8736647 -0.8736647
## 6  0.3519091 1.192636 -1.151397 -0.07559289 -0.107213 -1.1380633  1.1380633
##           Age Suicide
## 1 -1.38522071       1
## 2  0.76320137       1
## 3  1.03175413       0
## 4  0.31561343       1
## 5 -0.49004485       1
## 6 -0.04245691       1
```

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 the `caret` package. We will reserve 75% for training and rest 25% for validation purpose.

Principal Component Analysis

Principal Compenent Analysis (PCA) is another way in which we can reduce the dimensionality of a data set while increasing the interpretability of the data and minimizing information loss. We are going to perform PCA for ADHD and MD response questions below while using scree plots to determine the numberof PCA's to keep. The Scree plot will display the eigenvalues in a downward curve, and order them from largest to smallest.

Groups:

```
* All ADHD Questions
* All MD Questions
* All Other
```

Table 2: ADHD Correlations

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	
ADHD Q1	0.6985307	-0.2193486	0.3373110	-0.3701134	0.1620365	-0.0553265	-0.1155856	0.0634832	-0.
ADHD Q2	0.7094961	-0.3217329	0.2195271	-0.2415871	0.3149977	-0.1249863	-0.0479395	0.0317923	-0.
ADHD Q3	0.6729396	-0.3603530	0.1289838	0.0346323	-0.4064821	-0.2440550	0.2269925	0.0498093	-0.
ADHD Q4	0.7369016	-0.3570819	0.1752141	-0.0289810	0.0048257	0.1540025	0.2389656	0.0989201	0.
ADHD Q5	0.7270038	-0.1316262	-0.4393256	0.0626412	-0.2317282	-0.1653866	0.0161188	-0.0428040	0.
ADHD Q6	0.6478223	0.0330650	-0.3661242	-0.4081271	-0.2010428	-0.1415404	-0.2965562	0.1158451	-0.
ADHD Q7	0.7397991	-0.1570714	0.0745633	-0.0177672	-0.2719875	0.3665232	-0.2695678	0.0441089	0.
ADHD Q8	0.8036604	-0.1261053	0.1172968	0.2398635	-0.1092407	0.2986436	-0.0941213	-0.0499510	0.
ADHD Q9	0.7964059	-0.0211553	0.0758966	0.2369039	0.0344579	0.1119800	-0.0114321	-0.1023371	-0.
ADHD Q10	0.7928193	-0.0575642	0.1026147	0.0988034	0.0887157	-0.2189485	0.1349158	-0.0761935	0.
ADHD Q11	0.7417771	-0.1597164	-0.1426989	0.2224023	0.2630911	-0.0408302	-0.1102323	-0.2676399	-0.
ADHD Q12	0.6870249	0.2504992	0.1759628	0.3410893	-0.0587711	-0.2751124	-0.2002291	0.0034243	-0.
ADHD Q13	0.7625738	-0.0172434	-0.4488216	0.0218549	0.1970229	0.1218555	-0.0469157	0.0246575	-0.
ADHD Q14	0.7182484	0.0234819	-0.3752626	0.0082586	0.2726111	0.0471124	0.2687958	0.1484807	0.
ADHD Q15	0.6366225	0.3890284	-0.0003455	-0.2716084	-0.1802961	0.1957894	0.3451969	-0.1635894	-0.
ADHD Q16	0.6730194	0.5370634	0.1543455	-0.1542437	0.1306438	0.0783525	-0.0724010	-0.0541541	0.
ADHD Q17	0.6545364	0.4299515	0.1096686	0.2022074	0.0183908	-0.0287934	0.0308355	0.5043290	-0.
ADHD Q18	0.7097528	0.4135904	0.1118259	-0.1089439	-0.0884321	-0.1447675	0.0202057	-0.2629143	0.

We begin with the ADHD Questions and produce appropriately titled visualizations to explain each variable's contribution to predicting the risk of attempting suicide within this analysis.

ADHD

First we will use the prcomp function to perform a principal component analysis on the adhd response questions. We will also center and scale once again to ensure normality and proper dimensionality.

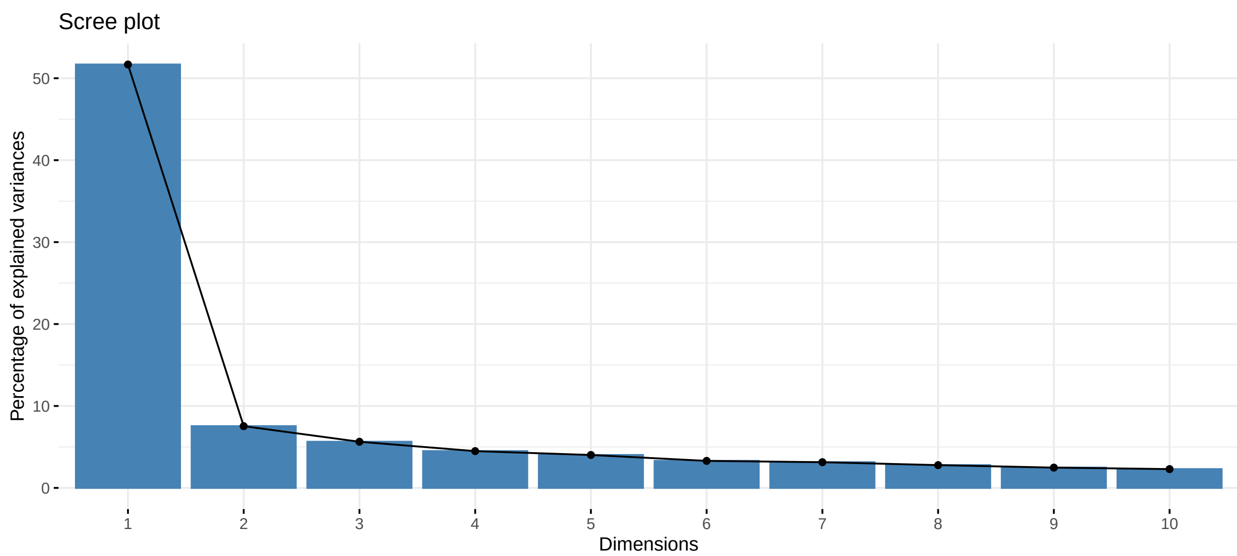
We will use the `factoextra` library to display the results of our PCA. this library specializes in extracting and visualizing the out put of exploratory multivariate analysis. Through this and a correlation table we can see the relationship between each ADHD response score and the Principle Components.

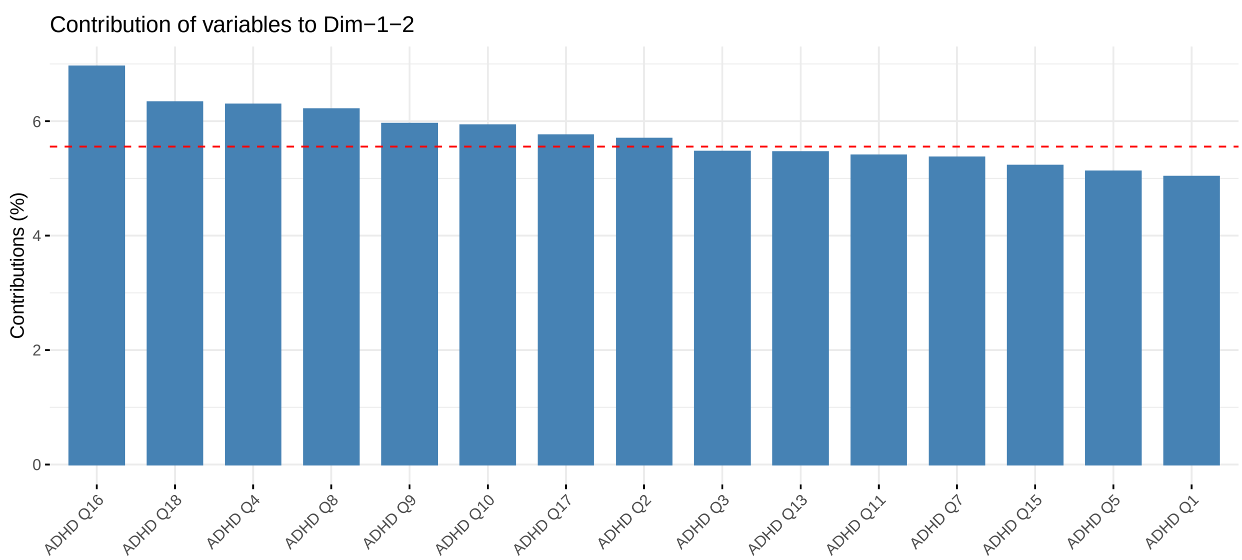
The list of PC's (sorted by descending impact on the variance of score) shows us the components that are the most impactful in grouping the respondents. By viewing the associated plots and correlations we can see the ADHd response questions `4,8,9,10,16,17,18` are the most impactful on plot of PC1 and PC2 which indicates they should be used in initial modeling of this dataset.

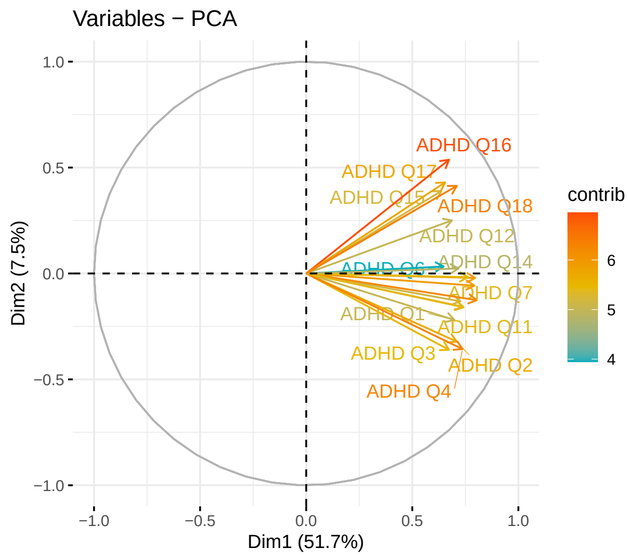
We can also see the factors that are most impactful for other principal components below. Ensure to read the plot titles for a high-level view of what each plot describes.

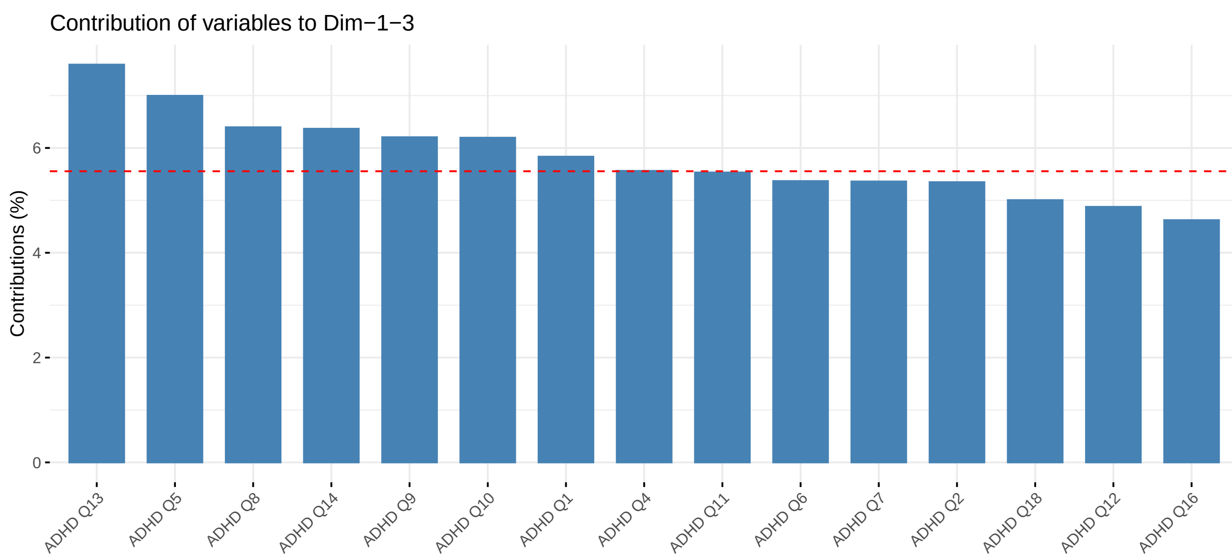
```
## Importance of components:
##                           PC1     PC2     PC3    PC4     PC5    PC6     PC7
## Standard deviation     3.0498 1.16471 1.00693 0.8990 0.85050 0.7707 0.75154
## Proportion of Variance 0.5168 0.07536 0.05633 0.0449 0.04019 0.0330 0.03138
## Cumulative Proportion  0.5168 0.59211 0.64844 0.6933 0.73353 0.7665 0.79791
##                            PC8     PC9    PC10    PC11    PC12    PC13    PC14
## Standard deviation     0.70763 0.66788 0.64291 0.63647 0.60782 0.59495 0.53747
## Proportion of Variance 0.02782 0.02478 0.02296 0.02251 0.02052 0.01966 0.01605
## Cumulative Proportion  0.82573 0.85051 0.87347 0.89598 0.91650 0.93617 0.95222
```

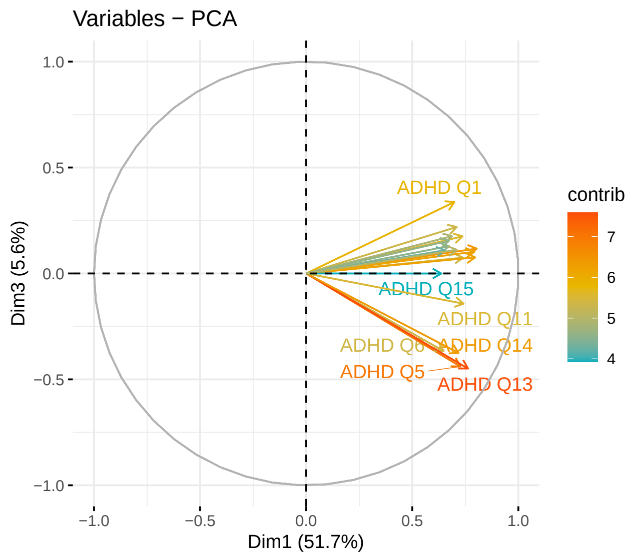
```
##                           PC15    PC16   PC17    PC18
## Standard deviation     0.52569 0.46267 0.4529 0.40566
## Proportion of Variance 0.01535 0.01189 0.0114 0.00914
## Cumulative Proportion  0.96757 0.97946 0.9909 1.00000
```






Variables − PCA
Dim1 (51.7%)
Dim2 (7.5%)
contrib
ADHD Q16
ADHD Q17
ADHD Q15
ADHD Q18
ADHD Q12
ADHD Q6
ADHD Q14
ADHD Q7
ADHD Q1
ADHD Q11
ADHD Q3
ADHD Q2
ADHD Q4


Contribution of variables to Dim−1−3
Contributions (%)
ADHD Q13
ADHD Q5
ADHD Q8
ADHD Q14
ADHD Q9
ADHD Q10
ADHD Q1
ADHD Q4
ADHD Q11
ADHD Q6
ADHD Q7
ADHD Q2
ADHD Q18
ADHD Q12
ADHD Q16


Variables − PCA
ADHD Q1
ADHD Q15
ADHD Q11
ADHD Q6
ADHD Q14
ADHD Q5
ADHD Q13
contrib
7
6
5
4
1.0
0.5
0.0
−0.5
−1.0
Dim3 (5.6%)
Dim1 (51.7%)

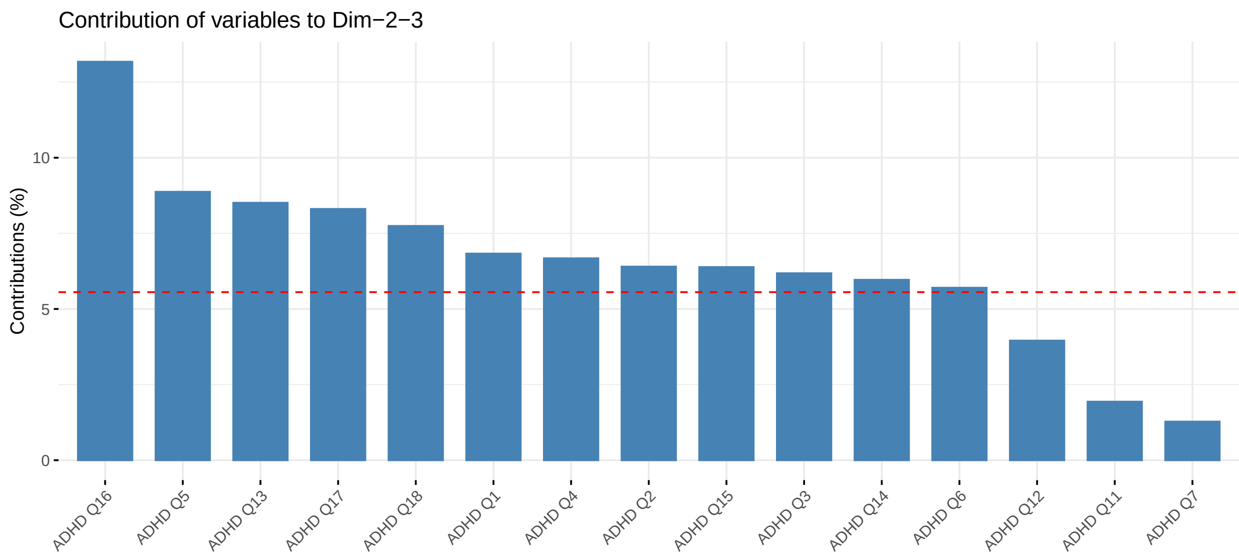
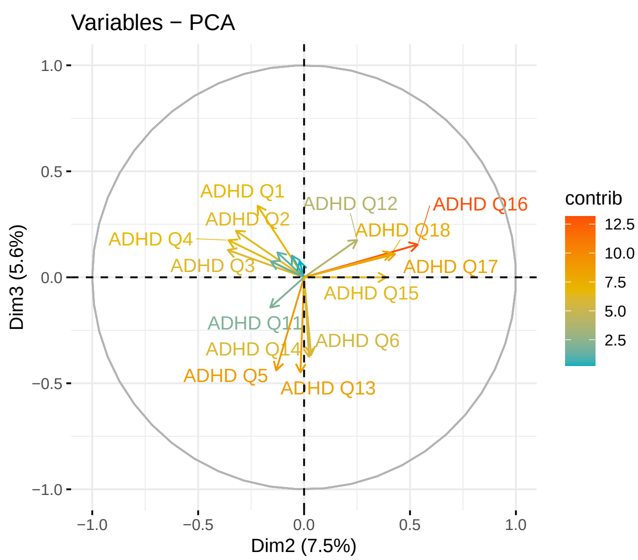

Contribution of variables to Dim−2−3
Contributions (%)
10
5
0
ADHD Q16
ADHD Q5
ADHD Q13
ADHD Q17
ADHD Q18
ADHD Q1
ADHD Q4
ADHD Q2
ADHD Q15
ADHD Q3
ADHD Q14
ADHD Q6
ADHD Q12
ADHD Q11
ADHD Q7

Table 3: md Correlations

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	
MD Q1a	-0.6888686	-0.1322144	0.0229725	-0.3776545	0.1155431	-0.0404785	0.0388949	-0.1162723	0.44
MD Q1b	-0.6421749	-0.3857186	0.1381774	0.1459230	-0.0448104	0.3624483	-0.0835332	-0.0859054	-0.16
MD Q1c	-0.4439895	0.4544153	-0.4763749	-0.0977568	0.2182164	0.2983609	-0.0077300	0.4015333	0.04
MD Q1d	-0.5655577	0.0049581	0.0644164	-0.0442116	0.6697469	0.0871900	0.2433963	-0.1694837	-0.11
MD Q1e	-0.6391312	0.2391336	-0.1887304	0.1745264	0.2662259	-0.3369547	-0.4017349	-0.0905138	0.01
MD Q1f	-0.6668512	-0.2825469	0.1009022	0.2032375	0.0540147	-0.3484816	0.1227297	0.0355070	0.10
MD Q1g	-0.6818994	-0.3687898	0.0536512	0.1911240	-0.0795600	-0.1339641	-0.0784985	0.2639884	0.02
MD Q1h	-0.6193596	0.4973240	0.0584859	0.2126773	-0.1071923	-0.1494907	0.2548126	-0.1564241	-0.16
MD Q1i	-0.5873223	0.4451927	-0.1405849	0.2802170	-0.3060325	0.0937060	0.3489444	-0.0655044	0.17
MD Q1j	-0.5954591	0.4356557	0.1328873	0.0517699	-0.1947518	0.1288149	-0.3882469	-0.0806025	0.16
MD Q1k	-0.5118310	0.4082574	0.5106816	-0.2784104	-0.0101545	0.1258235	-0.1131692	-0.0723359	-0.17
MD Q1L	-0.7087339	-0.1724858	0.1944088	-0.4085494	-0.1796511	-0.0457976	0.1524746	0.2247044	0.06
MD Q1m	-0.5682204	-0.0522540	-0.4267685	-0.4284171	-0.2416706	-0.2261652	-0.0162980	-0.0879603	-0.37
MD Q2	-0.7328573	-0.2802758	0.0969890	0.2440009	0.0151007	0.0994268	-0.0339440	0.2755074	-0.17
MD Q3	-0.5118573	-0.5064956	-0.3385010	0.0527350	-0.1123956	0.2676439	-0.0517879	-0.3591049	0.06

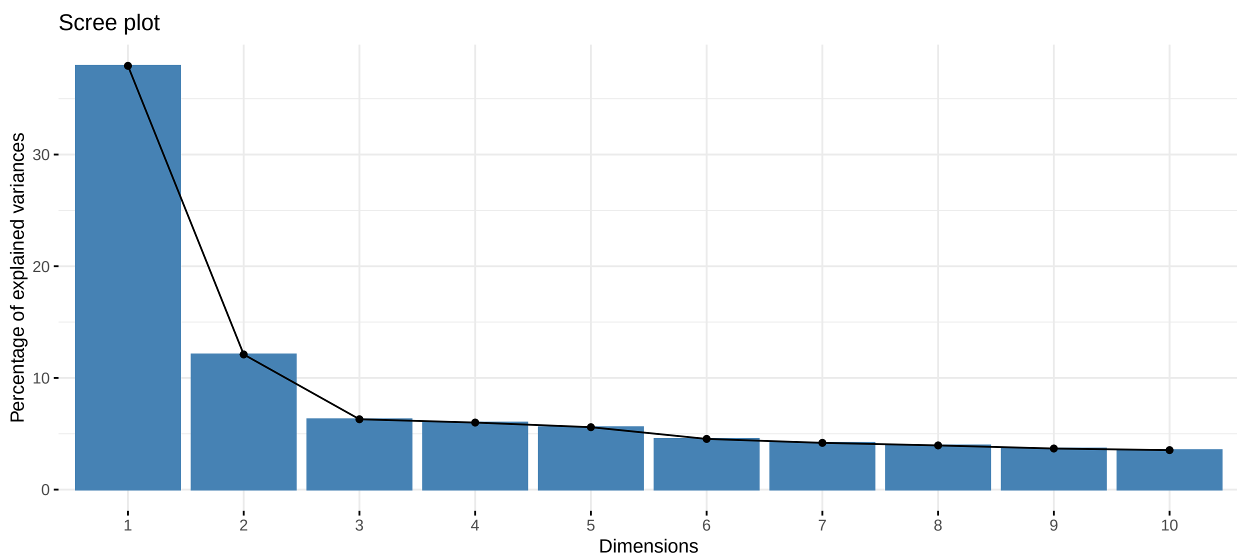


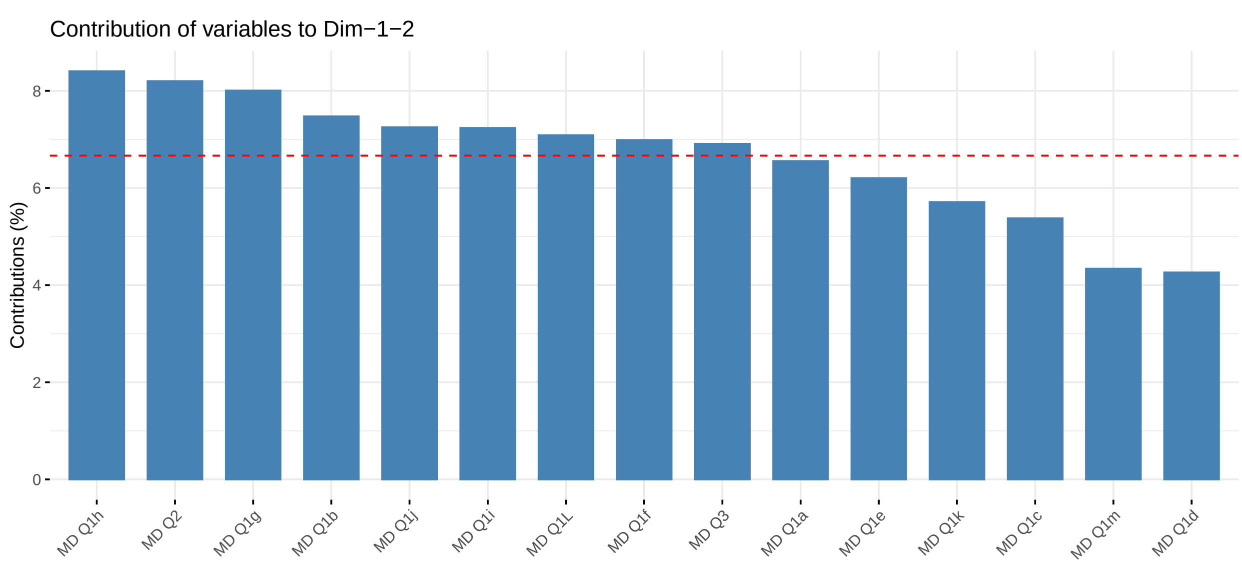
We will repeat the process above on MD response questions to get a better understanding of which of these questions are the most impactful. We can see that for PC1 and PC2 MD Questions 'Q1h', 'Q1j', 'Q1g', and 'Q2' have the greatest impact.

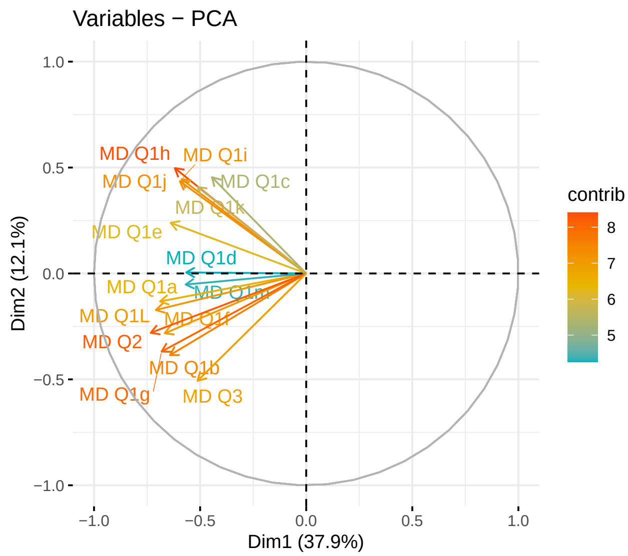
MD

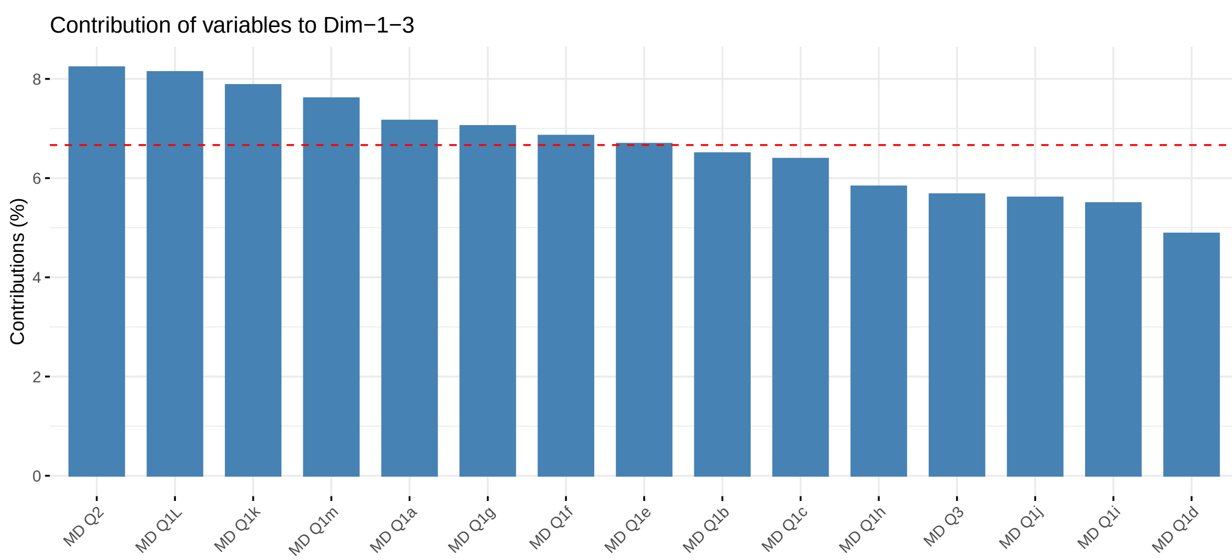
```
## Importance of components:
##                           PC1    PC2    PC3     PC4     PC5     PC6     PC7
## Standard deviation     2.3857 1.3474 0.9721 0.94891 0.91563 0.82515 0.79246
## Proportion of Variance 0.3794 0.1210 0.0630 0.06003 0.05589 0.04539 0.04187
## Cumulative Proportion  0.3794 0.5004 0.5635 0.62348 0.67937 0.72476 0.76663
##                            PC8     PC9    PC10    PC11    PC12   PC13    PC14
## Standard deviation     0.77044 0.74286 0.72814 0.70302 0.65162 0.5835 0.55685
```

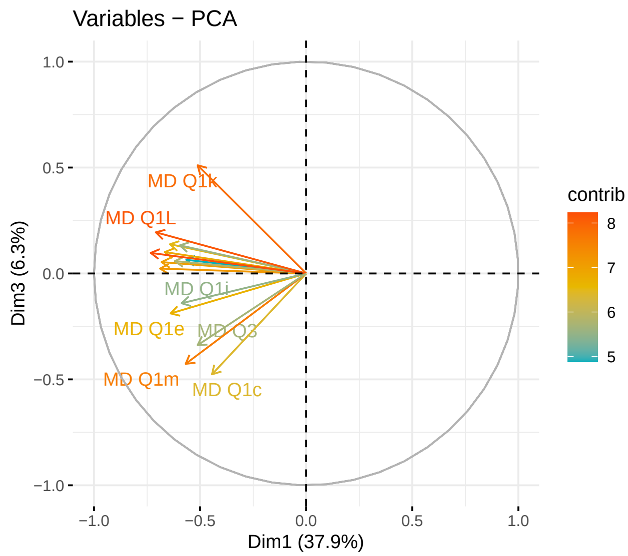
```
## Proportion of Variance 0.03957 0.03679 0.03535 0.03295 0.02831 0.0227 0.02067
## Cumulative Proportion  0.80620 0.84299 0.87834 0.91128 0.93959 0.9623 0.98296
##                           PC15
## Standard deviation     0.50553
## Proportion of Variance 0.01704
## Cumulative Proportion  1.00000
```



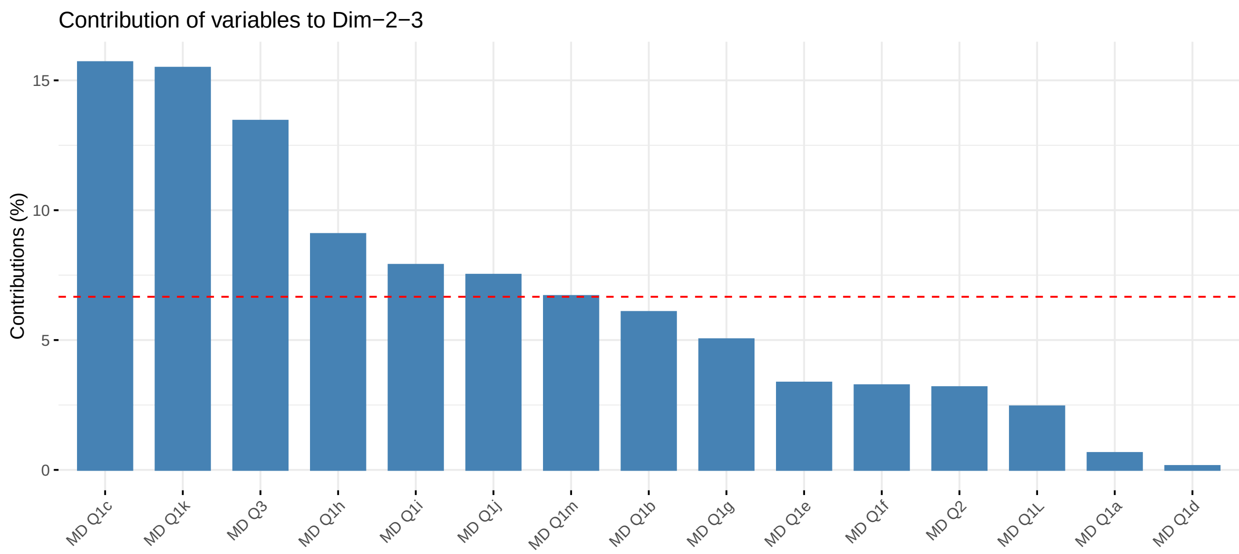



Variables − PCA
Dim1 (37.9%)
Dim2 (12.1%)
contrib
MD Q1h
MD Q1i
MD Q1j
MD Q1c
MD Q1k
MD Q1e
MD Q1d
MD Q1a
MD Q1m
MD Q1L
MD Q1f
MD Q2
MD Q1b
MD Q1g
MD Q3

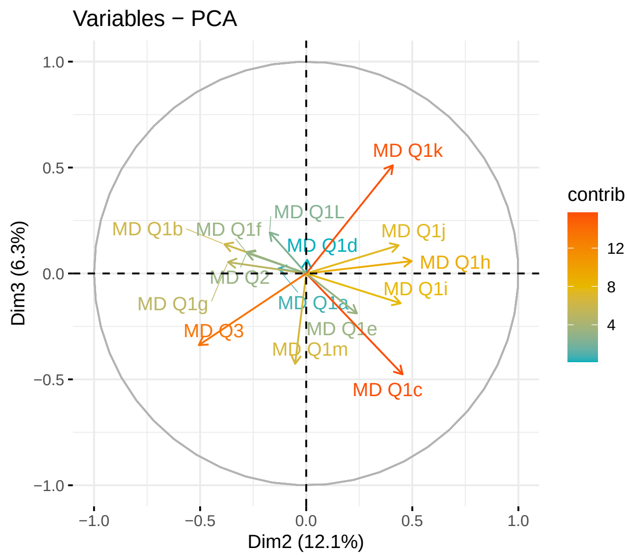

Contribution of variables to Dim−1−3
Contributions (%)
MD Q2
MD Q1L
MD Q1k
MD Q1m
MD Q1a
MD Q1g
MD Q1f
MD Q1e
MD Q1b
MD Q1c
MD Q1h
MD Q3
MD Q1j
MD Q1i
MD Q1d



Variables − PCA
contrib
8
7
6
5
MD Q1k
MD Q1L
MD Q1i
MD Q1e
MD Q3
MD Q1m
MD Q1c
Dim3 (6.3%)
Dim1 (37.9%)
1.0
0.5
0.0
−0.5
−1.0



Contribution of variables to Dim−2−3
Contributions (%)
15
10
5
0
MD Q1c
MD Q1k
MD Q3
MD Q1h
MD Q1i
MD Q1j
MD Q1m
MD Q1b
MD Q1g
MD Q1e
MD Q1f
MD Q2
MD Q1L
MD Q1a
MD Q1d



Gradient Boosting: Suicide

We assume we are modeling whether a patient attempted suicide codified in column AX. This known as a binary target variable. We use Gradient Boosting to predict whether a patient attempts suicides. Our specific approach uses an `XGBoost` of the `XGBoost` package library.

Initially, we remove the rows null values in the target column and drop the Non-subset Dx column because it had a lot of nulls as well. XGBoost needs data to be in a matrix so we convert the dataframes to numeric matrices.

We choose `XGBoost` because it is one of the more efficient gradient boosting methods that have some natural strengths at improving accuracy when used with questionnaire data such as ours. When constructing decision trees, variables that improve accuracy are selected for and 'boosted' to create trees that maximize accuracy. It does this without concern for how the leaves of the decision tree, exemplified by our sub-questions, get split into the branches of the decisions. Ideally, it will help us boost the accuracy of our model when trying to predict if a patient may attempt suicide. We continue this method with the first step, a cross validation (CV) split.

CV Split

We split the data into three folds for cross validation to improve the ability of the model to generalize and help with overfitting. We create a function to help with parameter tunning and make use of the `bayesOpt` package.

Our `bayesOpt` package contained a set of functions that allowed us to optimize the model further, in a sense to improve the model's ability to reduce error and thus, improve predictive accuracy. We also create a scoring function that we use to perform the boosting on our target and evaluate its performance over rounds of testing. The results are shown for reference.

```
##    Epoch Iteration max_depth min_child_weight subsample gpUtility acqOptimum
## 1:     0         1         4        16.900949 0.3980034        NA      FALSE
## 2:     0         2         9        22.465545 0.4598973        NA      FALSE
## 3:     0         3         4         2.543344 0.3380428        NA      FALSE
## 4:     0         4         7         8.946161 0.2773560        NA      FALSE
```

```
## 5:     1         5         2         1.601433 0.4039243 0.5941911       TRUE
## 6:     2         6        10         1.000000 0.5000000 0.5294463       TRUE
## 7:     3         7        10         1.000000 0.2500000 0.3537376       TRUE
##    inBounds Elapsed    Score nrounds errorMessage
## 1:     TRUE    0.05 0.500000       1           NA
## 2:     TRUE    0.03 0.500000       1           NA
## 3:     TRUE    0.10 0.722162      24           NA
## 4:     TRUE    0.03 0.500000       1           NA
## 5:     TRUE    0.10 0.732463      22           NA
## 6:     TRUE    0.11 0.762502      16           NA
## 7:     TRUE    0.03 0.646280       2           NA
```

```
## $max_depth
## [1] 10
##
## $min_child_weight
## [1] 1
##
## $subsample
## [1] 0.5
```

In several cases our model's scoring function appears to have a roughly 50% chance of attempting suicide. However, these cases all had fewer rounds of testing, shorter time elapsed, and used nonoptimum routes in its prediction. This is corrected with other tests in which further testing is completed through more rounds with heavier boosting. This is likely to produce a modest increase in our accuracy during model building.

Build Models

We develop three main models including a K-Clustering Method, Gradient Boosting method, and Support Vector Machines (SVM). Each focuses on a particular strategy. Our clustering attempts to identify groups of variables that are similar and use them to assess a patient's risk of attempting suicide. Gradient Boosting constructs decision trees with an emphasis on returning the most accurate decision path and our Support Vector Machine uses linear regression to classify the factors into groups using hyperplanes and by assessing the classes likelihood of being our target variable.

K-Clustering Method 1

We use K-nearest neighbor (KNN) to identify clusters of patients that share similar patterns that could help us predict our target variable. KNN works by identifying the "k" closest neighbors in the dataset. This works particularly well for classification of factorized data, as our dataset happens to be. Results of this model's construction are shown below.

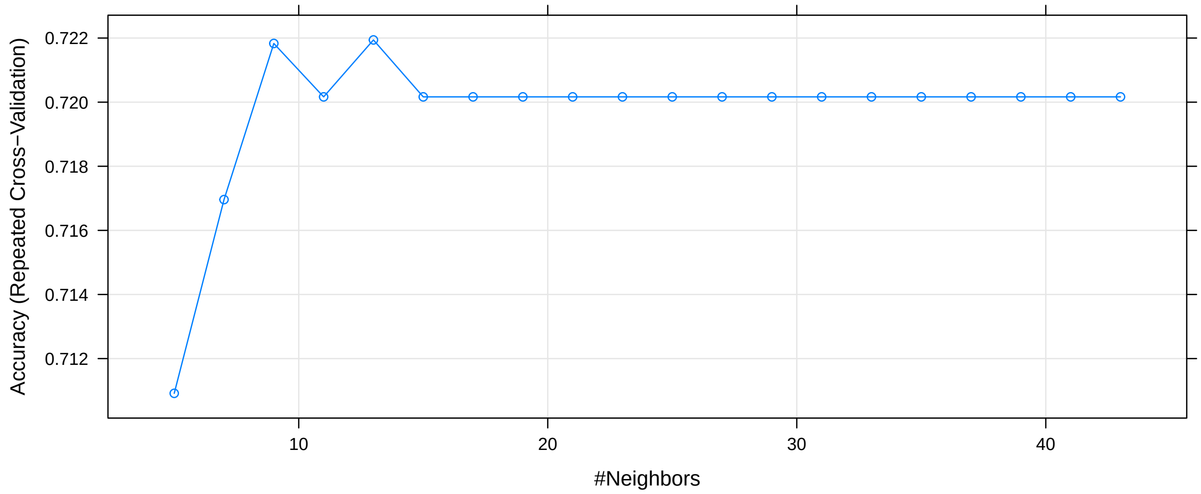
```
## [1] <NA>
## Levels: 0 1 2
```

```
## [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
```

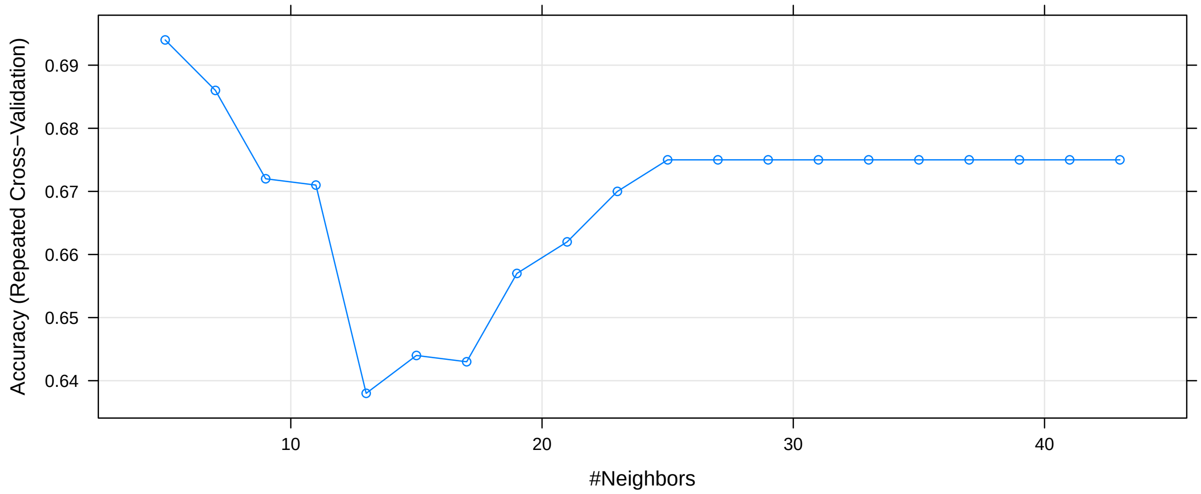
```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 31 12
##          1  0  0
##
##                Accuracy : 0.7209
##                  95% CI : (0.5633, 0.8467)
##     No Information Rate : 0.7209
##     P-Value [Acc > NIR] : 0.576988
##
##                   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%. We note that the ideal number of k was experimentally determined by running through several simulated model and testing the accuracy of each. The most accurate model was our final choice.

K-Clustering Method 2

In this we use a special kind of K-means method where we attempt to identify clear grouping of factors and their respective levels across the dataset. To do this we isolate those factors, center and scale the values, and focus on the mean euclidean distances between values. This method assumes there is an inherent similarity across values in the dataset including similarities with our target variable. The goal is to use these similarities (should there be any) to identify the groups closest to our target. If there are any groups, we would assume that they contain the patients who are at the highest risk of attempting suicide. Based on the data characteristics and summary from our EDA, we should not expect much from this model but it should prove informative.

```
## k-Nearest Neighbors
##
## 45 samples
## 49 predictors
##  2 classes: '0', '1'
##
## Pre-processing: centered (140), scaled (140)
## Resampling: Cross-Validated (10 fold, repeated 5 times)
## Summary of sample sizes: 41, 40, 41, 41, 41, 40, ...
## Resampling results across tuning parameters:
##
##   k   Accuracy  Kappa
##    5  0.694      0.200559441
##    7  0.686      0.126666667
##    9  0.672      0.080303030
##   11  0.671      0.046969697
##   13  0.638     -0.051515152
##   15  0.644     -0.049090909
##   17  0.643     -0.048484848
##   19  0.657     -0.027878788
##   21  0.662     -0.021212121
##   23  0.670     -0.006666667
##   25  0.675      0.000000000
##   27  0.675      0.000000000
##   29  0.675      0.000000000
##   31  0.675      0.000000000
##   33  0.675      0.000000000
##   35  0.675      0.000000000
##   37  0.675      0.000000000
##   39  0.675      0.000000000
##   41  0.675      0.000000000
##   43  0.675      0.000000000
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was k = 5.
```



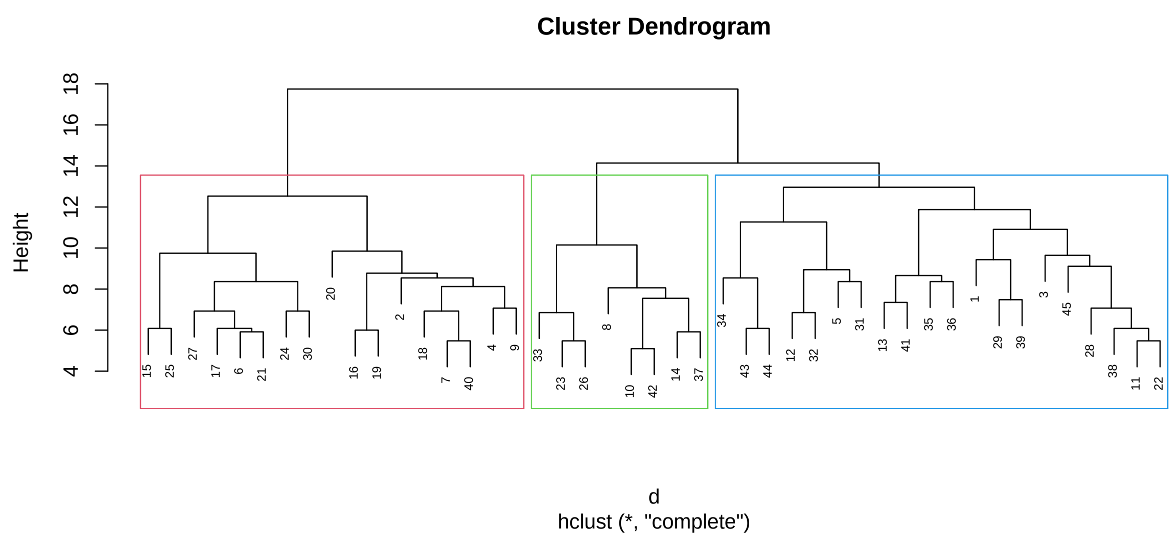
Our k value was experimentally determined to be significantly smaller for this model. As expected, the model's accuracy decreased since we were trying to group variable based similarities that we suspected would not be present in the data. However, this does confirm our expectations that there are fewer patterns between the target and all variables clustered together. When compared with smaller groups and more isolated variables, this tells us there may be a handful of important variables that predict our target better than all of them combined.

Heirarchical Clustering

In this we use a special kind of unsupervised learning method where we attempt to identify clear orders of factor and their respective levels across the dataset. To do this we isolate those factors, center and scale the values, remove our target from the data set, and review the results. Perhaps the easiest way to do this is with a dendrogram, as shown below.

```
##   Sex Race ADHD.Q1 ADHD.Q2 ADHD.Q3 ADHD.Q4 ADHD.Q5 ADHD.Q6 ADHD.Q7 ADHD.Q8
## 1   1    1       1       1       4       2       3       1       1       3
## 2   2    1       3       3       4       4       5       2       2       3
## 3   2    1       2       1       2       1       3       3       3       2
## 4   1    1       3       3       2       2       4       3       2       4
## 5   1    1       4       4       2       4       4       2       3       4
## 6   2    1       2       3       1       4       3       2       3       4
##   ADHD.Q9 ADHD.Q10 ADHD.Q11 ADHD.Q12 ADHD.Q13 ADHD.Q14 ADHD.Q15 ADHD.Q16
## 1       2        4        2        4        1        0        3        1
## 2       2        4        1        4        2        4        4        3
## 3       0        1        2        0        2        2        3        2
## 4       4        2        3        1        3        3        1        2
## 5       4        2        4        1        3        2        1        2
## 6       4        2        4        2        4        4        3        4
##   ADHD.Q17 ADHD.Q18 MD.Q1a MD.Q1b MD.Q1c MD.Q1d MD.Q1e MD.Q1f MD.Q1g MD.Q1h
## 1        3        4      1      1      1      1      0      1      1      1
## 2        1        4      1      1      1      1      1      1      1      1
## 3        1        1      0      0      0      0      1      1      1      0
## 4        1        2      1      1      0      0      1      1      1      1
## 5        1        1      0      1      0      1      0      1      1      0
```

```
## 6        3        3       0       1       0       1       1       1       1       1
##   MD.Q1i MD.Q1j MD.Q1k MD.Q1L MD.Q1m MD.Q2 MD.Q3 Alcohol THC Cocaine Stimulants
## 1      1      1      1      0      1     1     3       1   1       1          0
## 2      1      0      0      1      0     1     3       0   0       0          0
## 3      0      0      0      0      0     0     2       0   0       0          0
## 4      1      0      0      1      1     1     3       1   1       1          1
## 5      0      0      0      0      0     1     2       1   1       0          0
## 6      1      1      1      1      0     1     3       1   0       0          0
##   Sedative.hypnotics Opioids Court.order Education Hx.of.Violence
## 1                  0       0           1        11              0
## 2                  0       0           0        14              0
## 3                  0       0           0        12              0
## 4                  0       0           0        12              0
## 5                  0       0           1         9              1
## 6                  0       0           0        11              0
##   Disorderly.Conduct Abuse Non.subst.Dx Subst.Dx Psych.meds. cluster
## 1                  1     0            2        0           2       1
## 2                  0     4            1        0           1       2
## 3                  0     6            2        0           1       1
## 4                  0     7            2        0           2       2
## 5                  1     0            2        0           0       1
## 6                  1     2            0        0           0       2
```



Given that this dendrogram contains 53 variables, each labeled by their index value, we notice 3 large groups. These generally correspond to ADHD questions, Mood Disorder questions, and all other questions. However, there is a clear seperate branch of the dendrogram boxed in red. This mainly contains ADHD questions. This is interesting since it suggest the patient's responses to these questions have greater impact on the patient's risk of suicide. Alternatively, it could signify that the groups have little in common but at smaller sizes of k there may be significant predictors. We should learn more with the support vector machine.

Support Vector Machine

The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space (N being the number of features) that classifies the data points. Hyperplanes are decision boundaries to classify the data points. Data points that fall on either side of the hyperplane can qualify as different classes. Support vectors are data points that are closer to the hyperplane and effect the position and orientation of the hyperplane. Using these support vectors, we maximize the margin of the classifier.

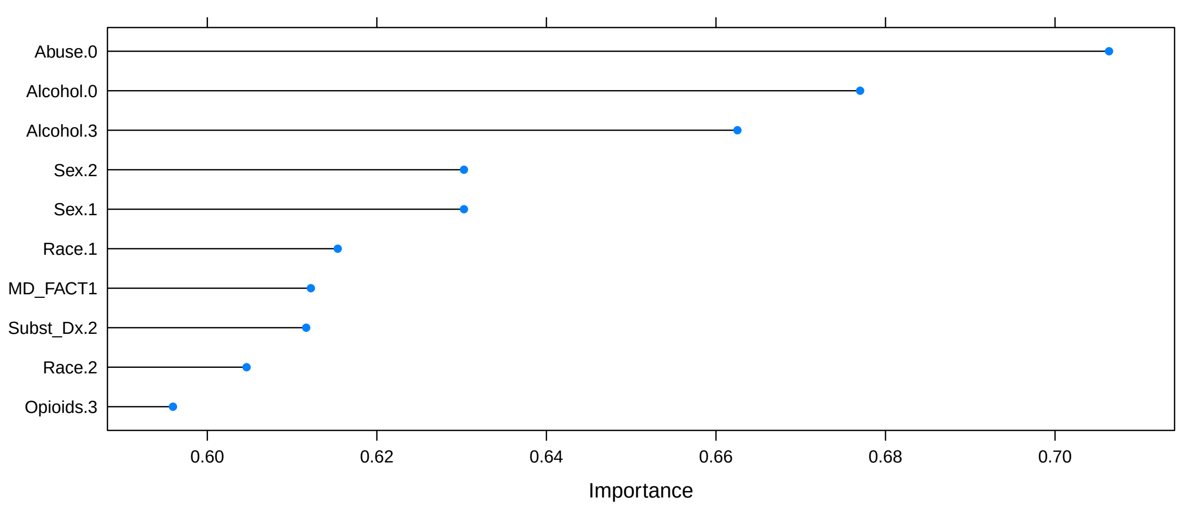
There are number of R packages available to implement SVM. The `train` function can be used for SVM using methods as `svmRadial`, `svmLinear` and `svmPoly` that fit different kernels. The functions we focus on are available in the `caret` package.

```
##
## Call:
## summary.resamples(object = svm_resamps)
##
## Models: Linear, Radial, Poly
## Number of resamples: 10
##
## Accuracy
##             Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## Linear 0.5384615 0.6071429 0.7774725 0.7520147 0.9006410 0.9230769    0
## Radial 0.6153846 0.7115385 0.7857143 0.7787546 0.8461538 0.9230769    0
## Poly   0.7500000 0.8461538 0.8461538 0.8469780 0.8571429 0.9285714    0
##
## Kappa
##              Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## Linear -0.2580645 0.1115854 0.4629695 0.3780025 0.7433665 0.8311688    0
## Radial -0.1428571 0.1030332 0.4251714 0.3781074 0.6243280 0.8059701    0
## Poly    0.2500000 0.5806452 0.5844402 0.5841986 0.6262255 0.8108108    0
```

We can see our Support Vector Machine Linear, Radial, and Poly fit had median accuracy rates of .631, .769 and .769 respectively indicating of radial or poly SVM should be chosen for future modeling.

Most Important Features

To better comprehend how the model came to it conclusion, it is beneficial to know which factors influence the target variable most. A plot of the top variables and their importance is displayed for review and comparison.



Abuse, alcoholism, and Mood Disorder Question 1 were the most important factors in the SVM models.

Gradient Boosted

We use the information from the above function to fit our final model, make predictions, and evaluate results. This gradient boost relys on the data preparation and cross-validation split from our previous sections.

```
## Confusion Matrix and Statistics
##
##        y_label_test
## xgbpred  0  1
##       0 27  8
##       1  1  4
##
##                Accuracy : 0.775
##                  95% CI : (0.6155, 0.8916)
##     No Information Rate : 0.7
##     P-Value [Acc > NIR] : 0.1959
##
##                   Kappa : 0.3571
##
##  Mcnemar's Test P-Value : 0.0455
##
##             Sensitivity : 0.9643
##             Specificity : 0.3333
##          Pos Pred Value : 0.7714
##          Neg Pred Value : 0.8000
##              Prevalence : 0.7000
##          Detection Rate : 0.6750
##    Detection Prevalence : 0.8750
##       Balanced Accuracy : 0.6488
##
##        'Positive' Class : 0
##
```

This produced an accuracy rate of 77.5%

Model Performance

We can see that model Gradient Boosted model has the best accuracy at 77.5% when applied to the test dataset. The Models shorthand names are K-nearest neighbor (KNN), Hierarchical Clustering (HC), linear support vector machine (LIN), radial support vector machine (RAD), polynomial support vector machine (PLY), and the gradient XGBoosted model (XGB). We could improve these models through more through feature selection via PCA or other methods and by focusing on feature engineering by using what was identified by these methods.

Table 4: Model Performance

	Accuracy	Kappa	AccuracyLower	AccuracyUpper	AccuracyNull	AccuracyPValue	McnemarPValue
KNN	0.721	0.000	0.563	0.847	0.721	0.577	0.001
HC	0.600	-0.250	0.262	0.878	0.800	0.967	1.000
LIN	0.581	-0.159	0.421	0.730	0.721	0.984	0.480
RAD	0.628	-0.162	0.467	0.770	0.721	0.934	0.080
PLY	0.628	-0.162	0.467	0.770	0.721	0.934	0.080
XGB	0.775	0.357	0.615	0.892	0.700	0.196	0.046

Note:

Variables from PCA & factor analysis considered in model performance evaluation

Conclusion

Through the use of feature engineering and different models we can see that there are numerous ways to approach a dataset such as this. Both models were better at predicting when a patient would not attempt to commit suicide, and not nearly as good at predicting when a patient would. Going forward it would be best to modify the model to focus on predicting when someone would attempt suicide. It is much more beneficial given the problem at hand to be over cautious and less accurate then to be more accurate but less cautious. Potentially using principle components could improve the model and focusing on feature engineering in regards to "positive" cases where the patient attempted suicide.

References

https://cran.r-project.org/web/packages/ParBayesianOptimization/vignettes/tuningHyperparameters.html

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

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

http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/118-principal-component-analysis-in-r-prcomp-vs-princomp/

https://uc-r.github.io/hc_clustering

https://rdrr.io/r/stats/prcomp.html

Code Appendix

```
knitr::opts_chunk$set(echo=FALSE, error=FALSE, warning=FALSE, message=FALSE, fig.align="center", fig.wi
# Libraries
library(summarytools)
library(tidyverse)
library(DataExplorer)
library(reshape2)
library(mice)
library(caret)
```

```r
library(MASS)
library(e1071)
library(tree)
library(corrplot)
library(kableExtra)
library(htmltools)
library(readxl)
library(psych)
library(xgboost)
library(ParBayesianOptimization)
library(factoextra)
library(kernlab)
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)
adhd_data.dims <- 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 Q'), starts_with('MD Q')))

# Add new factor columns created
adhd_newdata <- cbind(adhd_newdata, adhd_ques_fa, md_ques_fa)
head(adhd_newdata)
adhd_newdata.dims <- dim(adhd_newdata)
dim_reduced <- (53-25)/53
# plot missing values
plot_missing(adhd_newdata)
# rename columns to apply mice
adhd_newdata <- adhd_newdata %>%
  rename('ADHD_Total'='ADHD Total',
         'MD_Total'='MD TOTAL',
         '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)
# select columns with non missing values
temp <- adhd_newdata %>% dplyr::select(c(starts_with('ADHD_'), starts_with('MD_'), 'Race', 'Sex', 'Age'

# impute predictors using mice
adhd_impute <- adhd_newdata %>% dplyr::select(-c(starts_with('ADHD_'), starts_with('MD_'), 'Race', 'Sex
adhd_impute <- complete(mice(data=adhd_impute, print=FALSE))
summary(adhd_impute)
# Merged the imputed dataframe with temp
adhd_newdata <- cbind(adhd_impute, temp)
head(adhd_newdata)
# 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)
```

```r
# create dummy variables for categorical features
adhd_dummy <- dummyVars(Suicide ~ ., data = adhd_newdata)
adhd_dummy <- predict(adhd_dummy, newdata=adhd_newdata)

# center and scaling
adhd_transformed <- adhd_dummy %>%
  preProcess(c("center", "scale")) %>%
  predict(adhd_dummy) %>%
  as.data.frame()

# add Suicide column
adhd_transformed$Suicide <- adhd_newdata$Suicide
head(adhd_transformed)
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]
# create subset of ADHD Questions for PCA
adhd_ques_pca <- sapply(adhd_data[,c(4:21)], as.numeric)

# create subset of MD Questions for PCA
md_ques_pca <- sapply(adhd_data[,c(23:37)], as.numeric)
pca_adhd <- prcomp(adhd_ques_pca, scale. = TRUE, center=TRUE)

cor(adhd_ques_pca, pca_adhd$x[,1:10]) %>%
  kableExtra::kbl(booktabs = T, caption ="ADHD Correlations") %>%
  kable_styling(latex_options = c("striped"), full_width = F)
summary(pca_adhd)
fviz_eig(pca_adhd)
#top 10 contributors to the dimension of PC1 and PC2
fviz_contrib(pca_adhd, choice = "var", axes = c(1,2), top = 15)
fviz_pca_var(pca_adhd,
             col.var ="contrib", # Color by contributions to the PC
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE     # Avoid text overlapping
             ,axes=c(1,2)
             )
#top 10 contributors to the dimension of PC1 and PC3
fviz_contrib(pca_adhd, choice = "var", axes = c(1,3), top = 15)
fviz_pca_var(pca_adhd,
             col.var = "contrib", # Color by contributions to the PC
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE     # Avoid text overlapping
             ,axes=c(1,3)
             )
#top 10 contributors to the dimension of PC2 and PC3
fviz_contrib(pca_adhd, choice = "var", axes = c(2,3), top = 15)
```

```r
fviz_pca_var(pca_adhd,
             col.var = "contrib", # Color by contributions to the PC
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE     # Avoid text overlapping
             ,axes=c(2,3)
             )
pca_md <- prcomp(md_ques_pca, scale. = TRUE, center=TRUE)
cor(md_ques_pca, pca_md$x[,1:10]) %>%
  kableExtra::kbl(booktabs = T, caption ="md Correlations") %>%
  kable_styling(latex_options = c("striped"), full_width = F)
summary(pca_md)
fviz_eig(pca_md)
#top 10 contributors to the dimension of PC1 and PC2
fviz_contrib(pca_md, choice = "var", axes = c(1,2), top = 15)
fviz_pca_var(pca_md,
             col.var ="contrib", # Color by contributions to the PC
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE     # Avoid text overlapping
             ,axes=c(1,2)
             )
#top 10 contributors to the dimension of PC1 and PC3
fviz_contrib(pca_md, choice = "var", axes = c(1,3), top = 15)
fviz_pca_var(pca_md,
             col.var = "contrib", # Color by contributions to the PC
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE     # Avoid text overlapping
             ,axes=c(1,3)
             )
#top 10 contributors to the dimension of PC2 and PC3
fviz_contrib(pca_md, choice = "var", axes = c(2,3), top = 15)
fviz_pca_var(pca_md,
             col.var = "contrib", # Color by contributions to the PC
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE     # Avoid text overlapping
             ,axes=c(2,3)
             )
gb__train <-subset(training[complete.cases(training$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)
Folds <- list(
    Fold1 = as.integer(seq(1,nrow(gb__train),by = 3))
  , Fold2 = as.integer(seq(2,nrow(gb__train),by = 3))
  , Fold3 = as.integer(seq(3,nrow(gb__train),by = 3))
)

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
    )
  )
}

set.seed(50)
bounds <- list(
    max_depth = c(2L, 10L)
  , min_child_weight = c(1, 25)
  , subsample = c(0.25, .5)
)

optObj <- bayesOpt(
    FUN = scoringFunction
  , bounds = bounds
  , initPoints = 4
  , iters.n = 3
)
optObj$scoreSummary
print(getBestPars(optObj))
set.seed(622)
mode <- function(x){
  levels <- unique(x)
  indicies <- tabulate(match(x, levels))
  levels[which.max(indicies)]
}
# Clean up training data
training_factors <- training %>%
  dplyr::select(-Age, -`ADHD Total`, `MD TOTAL`)
training_factors <- data.frame(lapply(training_factors, as.factor))
train_knn <- training_factors %>%
  mutate(across(everything(), ~replace_na(., mode(.))))
mode(train_knn$Psych.meds.)
train_knn$Psych.meds.[which(is.na(train_knn$Psych.meds.))] <- 0
```

```r
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$overall[[1]], 3)*100
conf.mat.knn
set.seed(622)
# Clean up training data
training_factors <- training %>%
  na.omit() %>%
  dplyr::select(-Age, -`ADHD Total`, -`MD TOTAL`)
training_factors <- data.frame(lapply(training_factors, as.factor))
train_heir <- training_factors %>%
  mutate(across(everything(), ~replace_na(., mode(.))))
# mode(train_heir$Psych.meds.)
# train_heir$Psych.meds.[which(is.na(train_heir$Psych.meds.))]
# sum(is.na(train_heir))

# Clean up testing data
testing_factors <- testing %>%
  na.omit() %>%
  dplyr::select(-Age, -`ADHD Total`, -`MD TOTAL`)
testing_factors <- data.frame(lapply(testing_factors, as.factor))
test_heir <- testing_factors %>%
  mutate(across(everything(), ~replace_na(., mode(.))))
# mode(test_heir$Psych.meds.)
# test_heir$Psych.meds.[which(is.na(test_heir$Psych.meds.))]
# sum(is.na(test_heir$Psych.meds.))

# Train Heir model
```

```r
train.heir <- (train_heir[, names(train_heir) != "Suicide"])
prep <- preProcess(x = train.heir, method = c("center", "scale"))
cl <- trainControl(method="repeatedcv", repeats = 5)
heir_model <- train(Suicide ~ ., data = train_heir,
                method = "knn",
                trControl = cl,
                preProcess = c("center","scale"),
                tuneLength = 20)
heir_model
# Evaluate Model
plot(heir_model)
heir_predict <- predict(heir_model, newdata = test_heir)
conf.mat.heir <- confusionMatrix(heir_predict, test_heir$Suicide)
accuracy <- round(conf.mat.heir$overall[[1]], 3)*100
set.seed(622)

# Clean up training data
training_factors <- training %>%
  na.omit() %>%
  dplyr::select(-Age, -`ADHD Total`, -`MD TOTAL`)
training_factors <- data.frame(lapply(training_factors, as.factor))
train_heir <- training_factors %>%
  mutate(across(everything(), ~replace_na(., mode(.))))

# Clean up testing data
testing_factors <- testing %>%
  na.omit() %>%
  dplyr::select(-Age, -`ADHD Total`, -`MD TOTAL`)
testing_factors <- data.frame(lapply(testing_factors, as.factor))
test_heir <- testing_factors %>%
  mutate(across(everything(), ~replace_na(., mode(.))))

# Create Euclidean Dissimilarity Matrix
train.heir <- (train_heir[, names(train_heir) != "Suicide"])
d <- dist(train.heir, method = "euclidean")
# Hierarchical clustering using Complete Linkage
hc1 <- hclust(d, method = "complete" )
# Plot the obtained dendrogram
# Cut tree into groups
sub_grp <- cutree(hc1, k = 3)
train.heir %>%
  mutate(cluster = sub_grp) %>%
  head()
plot(hc1, cex = 0.6)
rect.hclust(hc1, k = 3, border = 2:5)
# partitioning for train and test
partition <- createDataPartition(adhd_transformed$Suicide, p=0.75, list = FALSE)
training <- adhd_transformed[partition,]
testing <- adhd_transformed[-partition,]
set.seed(622)

# fit with svmLinear
svm_lin_fit <- train(Suicide ~ .,
```

```
                data = training,
                method = "svmLinear",
                preProcess = c("center","scale"),
                tuneLength = 5,
                trControl = trainControl(method = "cv"))

pred_lin_suicide <- predict(svm_lin_fit, testing)
cm_lin <- confusionMatrix(testing$Suicide, pred_lin_suicide)

# fit with svmRadial
svm_rad_fit <- train(Suicide ~ .,
                data = training,
                method = "svmRadial",
                preProcess = c("center","scale"),
                tuneLength = 5,
                trControl = trainControl(method = "cv"))

pred_rad_suicide <- predict(svm_rad_fit, testing)
cm_rad <- confusionMatrix(testing$Suicide, pred_rad_suicide)

# fit with svmPoly
svm_poly_fit <- train(Suicide ~ .,
                data = training,
                method = "svmPoly",
                preProcess = c("center","scale"),
                tuneLength = 5,
                trControl = trainControl(method = "cv"))

pred_poly_suicide <- predict(svm_poly_fit, testing)
cm_poly <- confusionMatrix(testing$Suicide, pred_poly_suicide)

# Evaluate models
conf.mat.svnpoly <- confusionMatrix(pred_poly_suicide, testing$Suicide)
conf.mat.svnrad <- confusionMatrix(pred_rad_suicide, testing$Suicide)
conf.mat.svnlin <- confusionMatrix(pred_lin_suicide, testing$Suicide)
#Compare 3 models:
svm_resamps <- resamples(list(Linear = svm_lin_fit, Radial = svm_rad_fit, Poly = svm_poly_fit))
summary(svm_resamps)
# Important features
imp_vars <- varImp(svm_rad_fit, scale=FALSE)
plot(imp_vars, top=10)
dtrain <- xgb.DMatrix(gb__train, label=y_label_tr)
dtest <- xgb.DMatrix(gb_test, label=y_label_test)
xgb <- xgb.train(
      params = list(
                  booster = "gbtree"
                , eta = 0.01
                , max_depth = 10
                , min_child_weight = 1
                , subsample = .5
                , objective = "binary:logistic"
                , eval_metric = "auc"
              )
```

```
    , data = dtrain
    , nround = 100
    , maximize = TRUE
            , verbose = 0)

xgbpred <- predict(xgb,dtest)
xgbpred <- ifelse (xgbpred > 0.5,1,0)
y_label_test <- as.numeric(y_label_test)
conf.mat.xgboost <- confusionMatrix(table(xgbpred, y_label_test))
conf.mat.xgboost
(round(data.frame(conf.mat.knn$overall,
                  conf.mat.heir$overall,
                  conf.mat.svnlin$overall,
                  conf.mat.svnrad$overall,
                  conf.mat.svnpoly$overall,
                  conf.mat.xgboost$overall), 3)) %>%
  rename(KNN = conf.mat.knn.overall,
         HC = conf.mat.heir.overall,
         LIN = conf.mat.svnlin.overall,
         RAD = conf.mat.svnrad.overall,
         PLY = conf.mat.svnpoly.overall,
         XGB = conf.mat.xgboost.overall) %>%
  t() %>%
  kbl(booktabs =  T, caption = "Model Performance") %>%
  kable_styling(latex_options = c("striped", "HOLD_position"), full_width = F) %>%
  footnote(c(
    "Variables from PCA & factor analysis considered in model performance evaluation"
  ))
```