Mind Matters: A Data-driven Examination of College Students' Mental Health

Statistical Inference and Learning project

Giovanni Costa - 880892

AY 2023/24

Contents

Introduction

This project aims to delve into the pressing issue of mental health among college students, a topic of significant importance in today's fast-paced, high-stress academic environment. The analysis seeks to provide a comprehensive understanding of the mental health landscape, by utilizing one dataset coming from a survey, with a particular focus on psychiatric symptoms and other associated health problems.

The dataset, that can be retrieved here, contains 27014 objects with 16 features. The first six information are categorical and related to the general individual condition, instead, the others are numeric variables associated with the mental health of the students, where the higher score indicates more serious symptoms. Additional details regarding the data scales, the data origin, and the retrieving methodology are not provided by the dataset's website. A specific description of these features is provided below:

1. **gender**: 0=Female, 1=Male
2. **whether_only_child**: 0=No, 1=Yes
3. **birth_place**: 0=Countryside, 1=Town, 2=SmallCity, 3=MediumToLargeCities

4. **family_economic_status**: 0=ExtremelyPoor, 1=Poor, 2=Average, 3=Good, 4=Rich
5. **major**: 0=Liberal, 1=Science, 2=Art
6. **grade**: 0=Postgraduate, 1=UndergraduateGradeFive, 2=Junior, 3=Sophomore, 4=Freshman, 5=Senior
7. **psychiatric_symptoms**: range from 4 to 16
8. **suicide**: range from 4 to 16
9. **dependence**: range from 4 to 16
10. **impulsivity**: range from 4 to 16
11. **compulsion**: range from 4 to 16
12. **sleeping_disturbance**: range from 4 to 16
13. **internet_addiction**: range from 5 to 20
14. **hostile_aggression**: range from 4 to 16
15. **self_injury_behaviors:** range from 4 to 16
16. **eating_problems**: range from 4 to 16

The purposes of this analysis are inspecting the possible hidden patterns present in the data and predicting the risk of observing a suicide, modeled for convenience as a binary variable.

Prerequirements

```
# Configure the environment
# options(digits = 4, scipen = 999)
rm(list = ls())
# Set seed for reproducibility
seed <- 123
num_k_fold <- 5
set.seed(seed)

# Load all the packages and install them if they are not present
requirements <- c(
  "summarytools", "MASS", "effects", "pROC", "mgcv",
  "glmnetUtils", "e1071", "class", "car", "brglm2",
  "gridExtra", "grid", "stringr", "ggplot2", "reshape2", "scales", "comprehenr",
  "dplyr", "ggcorrplot"
)
for (library_name in requirements) {
  if (!require(library_name, character.only = TRUE)) {
    install.packages(library_name, repos = "https://cloud.r-project.org")
    library(library_name, character.only = TRUE)
  }
}
# Import user defined functions
source("src/utils.R")
source("src/plotting.R")
# library(styler)
# style_file("analysis_report.Rmd")
```

First overview

In this section, some operations are performed to understand the dataset structure and identify any potential issues that need to be addressed. Also, variables are transformed into the appropriate format.

```
# Read the dataset
df <- read.csv("data/mental_health_data.csv")
# Remove the column representing the row index
df <- df[, -1]

# Rename existing columns
```

```r
colnames(df)[colnames(df) == "whether.only.child"] <- "whether_only_child"
colnames(df)[colnames(df) == "birth.place"] <- "birth_place"
colnames(df)[colnames(df) == "family.economic.status"] <- "family_economic_status"
colnames(df)[colnames(df) == "psychiatric.symptoms"] <- "psychiatric_symptoms"
colnames(df)[colnames(df) == "sleeping.disturbance"] <- "sleeping_disturbance"
colnames(df)[colnames(df) == "internet.addiction"] <- "internet_addiction"
colnames(df)[colnames(df) == "hostile.aggression"] <- "hostile_aggression"
colnames(df)[colnames(df) == "self.injury.behaviors"] <- "self_injury_behaviors"
colnames(df)[colnames(df) == "eating.problems"] <- "eating_problems"

# Transforming variables into factors
df$gender <- factor(df$gender,
  levels = c(0, 1),
  labels = c("Female", "Male")
)
df$whether_only_child <- factor(df$whether_only_child,
  levels = c(0, 1),
  labels = c("No", "Yes")
)
df$birth_place <- factor(
  df$birth_place,
  levels = 0:3,
  labels = c("Countryside", "Town", "SmallCity", "MediumToLargeCities")
)
df$family_economic_status <- factor(
  df$family_economic_status,
  levels = 0:4,
  labels = c("ExtremelyPoor", "Poor", "Average", "Good", "Rich")
)
df$major <- factor(df$major,
  levels = 0:2,
  labels = c("Liberal", "Science", "Art")
)
df$grade <- factor(
  df$grade,
  levels = 0:5,
  labels = c(
    "Postgraduate",
    "UndergraduateGradeFive",
    "Junior",
    "Sophomore",
    "Freshman",
    "Senior"
  )
)
general_cols <- c(
  "gender", "whether_only_child", "birth_place",
  "family_economic_status", "major", "grade"
)
symptoms_cols <- c(
  "psychiatric_symptoms", "dependence", "impulsivity", "compulsion", "sleeping_disturbance",
  "internet_addiction", "hostile_aggression", "self_injury_behaviors", "eating_problems"
)
```

```r
print_summary_custom(df, method = "pander")
```

```
##
## ------------------------------------------------------------------------------------------
## No   Variable                Stats / Values             Freqs (% of Valid)   Graph
## ---- ----------------------- -------------------------- -------------------- ------------------
```

```
## 1    gender                  1. Female                 18222 (67.5%)       IIIIIIIIIIII
##      [factor]                2. Male                    8792 (32.5%)       IIIIII
##
## 2    whether_only_child      1. No                     19663 (72.8%)       IIIIIIIIIIIII
##      [factor]                2. Yes                     7351 (27.2%)       IIIII
##
## 3    birth_place             1. Countryside            10613 (39.3%)       IIIIIII
##      [factor]                2. Town                    7222 (26.7%)       IIIII
##                              3. SmallCity               6434 (23.8%)       IIII
##                              4. MediumToLargeCities     2745 (10.2%)       II
##
## 4    family_economic_status  1. ExtremelyPoor            569 ( 2.1%)
##      [factor]                2. Poor                    4459 (16.5%)       III
##                              3. Average                18058 (66.8%)       IIIIIIIIIIII
##                              4. Good                    3813 (14.1%)       II
##                              5. Rich                     115 ( 0.4%)
##
## 5    major                   1. Liberal                15804 (58.5%)       IIIIIIIIII
##      [factor]                2. Science                 7269 (26.9%)       IIIII
##                              3. Art                     3941 (14.6%)       II
##
## 6    grade                   1. Postgraduate           5010 (18.5%)        III
##      [factor]                2. UndergraduateGradeFive   68 ( 0.3%)
##                              3. Junior                 5667 (21.0%)        IIII
##                              4. Sophomore              5625 (20.8%)        IIII
##                              5. Freshman               5676 (21.0%)        IIII
##                              6. Senior                 4968 (18.4%)        III
##
## 7    psychiatric_symptoms    Mean (sd) : 5.1 (1.6)     4 : 15773 (58.4%)   IIIIIIIIII
##      [integer]               min < med < max:          5 :  3647 (13.5%)   II
##                              4 < 4 < 16                6 :  2678 ( 9.9%)   I
##                              IQR (CV) : 2 (0.3)        7 :  1250 ( 4.6%)
##                                                        8 :  2771 (10.3%)   II
##                                                        9 :   467 ( 1.7%)
##                                                        10 :   205 ( 0.8%)
##                                                        11 :    98 ( 0.4%)
##                                                        12 :    72 ( 0.3%)
##                                                        13 :    19 ( 0.1%)
##                                                        14 :    14 ( 0.1%)
##                                                        15 :     8 ( 0.0%)
##                                                        16 :    12 ( 0.0%)
##
## 8    suicide                 Mean (sd) : 4.8 (1.6)     4 : 20126 (74.5%)   IIIIIIIIIIIII
##      [integer]               min < med < max:          5 :  1642 ( 6.1%)   I
##                              4 < 4 < 16                6 :  1191 ( 4.4%)
##                              IQR (CV) : 1 (0.3)        7 :   815 ( 3.0%)
##                                                        8 :  2394 ( 8.9%)   I
##                                                        9 :   346 ( 1.3%)
##                                                        10 :   219 ( 0.8%)
##                                                        11 :   123 ( 0.5%)
##                                                        12 :    88 ( 0.3%)
##                                                        13 :    33 ( 0.1%)
##                                                        14 :    15 ( 0.1%)
##                                                        15 :     6 ( 0.0%)
##                                                        16 :    16 ( 0.1%)
##
## 9    dependence              Mean (sd) : 6.9 (2.3)     4 : 5777 (21.4%)    IIII
##      [integer]               min < med < max:          5 : 3260 (12.1%)    II
##                              4 < 7 < 16                6 : 3144 (11.6%)    II
##                              IQR (CV) : 3 (0.3)        7 : 3268 (12.1%)    II
##                                                        8 : 5774 (21.4%)    IIII
```

```
##                                                    9 : 2571 ( 9.5%)     I
##                                                    10 : 1521 ( 5.6%)    I
##                                                    11 :  819 ( 3.0%)
##                                                    12 :  584 ( 2.2%)
##                                                    13 :  176 ( 0.7%)
##                                                    14 :   64 ( 0.2%)
##                                                    15 :   33 ( 0.1%)
##                                                    16 :   23 ( 0.1%)
##
## 10   impulsivity             Mean (sd) : 7 (2.2)   4 : 4892 (18.1%)     III
##      [integer]               min < med < max:      5 : 3127 (11.6%)     II
##                              4 < 7 < 16            6 : 3155 (11.7%)     II
##                              IQR (CV) : 3 (0.3)    7 : 3563 (13.2%)     II
##                                                    8 : 5971 (22.1%)     IIII
##                                                    9 : 2767 (10.2%)     II
##                                                    10 : 1733 ( 6.4%)    I
##                                                    11 : 1016 ( 3.8%)
##                                                    12 :  500 ( 1.9%)
##                                                    13 :  169 ( 0.6%)
##                                                    14 :   63 ( 0.2%)
##                                                    15 :   37 ( 0.1%)
##                                                    16 :   21 ( 0.1%)
##
## 11   compulsion              Mean (sd) : 7.1 (2.4) 4 : 5293 (19.6%)     III
##      [integer]               min < med < max:      5 : 3321 (12.3%)     II
##                              4 < 7 < 16            6 : 3308 (12.2%)     II
##                              IQR (CV) : 3 (0.3)    7 : 3189 (11.8%)     II
##                                                    8 : 5185 (19.2%)     III
##                                                    9 : 2617 ( 9.7%)     I
##                                                    10 : 1711 ( 6.3%)    I
##                                                    11 : 1152 ( 4.3%)
##                                                    12 :  723 ( 2.7%)
##                                                    13 :  276 ( 1.0%)
##                                                    14 :  132 ( 0.5%)
##                                                    15 :   67 ( 0.2%)
##                                                    16 :   40 ( 0.1%)
##
## 12   sleeping_disturbance    Mean (sd) : 7.1 (2.4) 4 : 5207 (19.3%)     III
##      [integer]               min < med < max:      5 : 3019 (11.2%)     II
##                              4 < 7 < 16            6 : 3451 (12.8%)     II
##                              IQR (CV) : 4 (0.3)    7 : 3271 (12.1%)     II
##                                                    8 : 4785 (17.7%)     III
##                                                    9 : 2751 (10.2%)     II
##                                                    10 : 1956 ( 7.2%)    I
##                                                    11 : 1266 ( 4.7%)
##                                                    12 :  701 ( 2.6%)
##                                                    13 :  322 ( 1.2%)
##                                                    14 :  159 ( 0.6%)
##                                                    15 :   68 ( 0.3%)
##                                                    16 :   58 ( 0.2%)
##
## 13   internet_addiction      Mean (sd) : 9.9 (3.3) 5 : 3824 (14.2%)     II
##      [integer]               min < med < max:      6 : 1608 ( 6.0%)     I
##                              5 < 10 < 20           7 : 1807 ( 6.7%)     I
##                              IQR (CV) : 5 (0.3)    8 : 2026 ( 7.5%)     I
##                                                    9 : 2302 ( 8.5%)     I
##                                                    10 : 4289 (15.9%)    III
##                                                    11 : 2636 ( 9.8%)    I
##                                                    12 : 2296 ( 8.5%)    I
##                                                    13 : 1979 ( 7.3%)    I
##                                                    14 : 1750 ( 6.5%)    I
```

```
##                                                        15 : 1324 ( 4.9%)
##                                                        16 :  479 ( 1.8%)
##                                                        17 :  299 ( 1.1%)
##                                                        18 :  170 ( 0.6%)
##                                                        19 :  120 ( 0.4%)
##                                                        20 :  105 ( 0.4%)
##
## 14   hostile_aggression       Mean (sd) : 5.6 (1.8)      4 : 10899 (40.3%)    IIIIIIII
##      [integer]                min < med < max:           5 :  4717 (17.5%)    III
##                               4 < 5 < 16                 6 :  3425 (12.7%)    II
##                               IQR (CV) : 3 (0.3)         7 :  2732 (10.1%)    II
##                                                          8 :  3802 (14.1%)    II
##                                                          9 :   799 ( 3.0%)
##                                                          10 :   338 ( 1.3%)
##                                                          11 :   167 ( 0.6%)
##                                                          12 :    79 ( 0.3%)
##                                                          13 :    29 ( 0.1%)
##                                                          14 :     9 ( 0.0%)
##                                                          15 :     4 ( 0.0%)
##                                                          16 :    14 ( 0.1%)
##
## 15   self_injury_behaviors   Mean (sd) : 4.8 (1.4)      4 : 17540 (64.9%)    IIIIIIIIIIII
##      [integer]                min < med < max:           5 :  4072 (15.1%)    III
##                               4 < 4 < 16                 6 :  1536 ( 5.7%)    I
##                               IQR (CV) : 1 (0.3)         7 :  1249 ( 4.6%)
##                                                          8 :  2201 ( 8.1%)    I
##                                                          9 :   212 ( 0.8%)
##                                                          10 :    98 ( 0.4%)
##                                                          11 :    55 ( 0.2%)
##                                                          12 :    26 ( 0.1%)
##                                                          13 :    13 ( 0.0%)
##                                                          14 :     5 ( 0.0%)
##                                                          16 :     7 ( 0.0%)
##
## 16   eating_problems         Mean (sd) : 5.5 (1.5)      4 : 8968 (33.2%)     IIIIIII
##      [integer]                min < med < max:           5 : 7202 (26.7%)     IIIII
##                               4 < 5 < 16                 6 : 4720 (17.5%)     III
##                               IQR (CV) : 2 (0.3)         7 : 2492 ( 9.2%)     I
##                                                          8 : 2595 ( 9.6%)     I
##                                                          9 :  697 ( 2.6%)
##                                                          10 :  185 ( 0.7%)
##                                                          11 :   77 ( 0.3%)
##                                                          12 :   50 ( 0.2%)
##                                                          13 :   16 ( 0.1%)
##                                                          14 :    5 ( 0.0%)
##                                                          15 :    1 ( 0.0%)
##                                                          16 :    6 ( 0.0%)
## ---------------------------------------------------------------------------------------------
```

```r
# "Dataset dimensions:"
dim(df)
```

```
## [1] 27014    16
```

```r
# "N. of missing values:"
sum(is.na(df))
```

```
## [1] 0
```

```r
# "Example of some objects:"
head(df, 3)
```

```
##   gender whether_only_child birth_place family_economic_status   major  grade
## 1 Female                 No   SmallCity                   Good Liberal Senior
## 2 Female                 No Countryside                   Good Liberal Senior
## 3 Female                 No        Town                Average Liberal Senior
##   psychiatric_symptoms suicide dependence impulsivity compulsion
## 1                    4       4          9           6         10
## 2                    4       4          5           5          5
## 3                    4       4          5           6          5
##   sleeping_disturbance internet_addiction hostile_aggression
## 1                    7                 10                  6
## 2                    4                  8                  6
## 3                    6                 11                  4
##   self_injury_behaviors eating_problems
## 1                     6              7
## 2                     4              6
## 3                     5              4
```

From the above table, it can be seen that all the columns have the expected format now. Also can be observed that there are no missing values.

Talking about the summary of the variables, different aspects can be observed:

- The majority of the students are female (67.5%)
- A lot of people have at least one brother or sister (72.8%)
- Fewer students in the dataset are from medium to large cities (10.2%)
- The number of students with average family economic status is the highest (66.8%), instead the number of students with rich family it quite low (0.4%)
- The most common major is liberal (58.5%)
- Only 0.3% of the people are undergraduates with grade five, instead the other grades are quite equally distributed
- The mean value of the variable Internet addiction is near 10 (scale [5, 20]) while the mean for sleeping disturbance, impulsivity, compulsion, and dependence are around 7 (scale [4, 16]). These symptom scores are a bit higher with respect to the other variables values
- No peoples with self-injury behaviors has symptoms of value 15 and they have the lowest average score (4.82), suggesting it might be the least prevalent issue
- In general students with very high symptoms are fortunately the minority among the observed data sample, so the numeric distributions are right skewed.

As anticipated, in this analysis, it's decided to transform the variable *suicide* into a binary feature, considering the current problem as a classification task. In this sense, the values of *suicide* lower or equal to 9 take the value *FALSE*, instead the other takes the value *TRUE*. The choice of this threshold is thought reasonable because the purpose of this analysis is to consider the higher severity of suicide but at the same time don't underestimate a real risk even if the value is not too big.

```r
print(summary(df$suicide))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   4.000   4.000   4.000   4.791   5.000  16.000
```

```r
threshold_suicide <- 9
numeric_suicide <- df$suicide
factor_suicide <- as.factor(df$suicide > threshold_suicide)
df$suicide <- factor_suicide

# Number of suicide
print(table(df$suicide))
```

```
##
## FALSE  TRUE
## 26514   500
```

```r
# Percentage of suicide
print(table(df$suicide) / nrow(df) * 100)
```

```
##
##     FALSE      TRUE
## 98.149108  1.850892
```

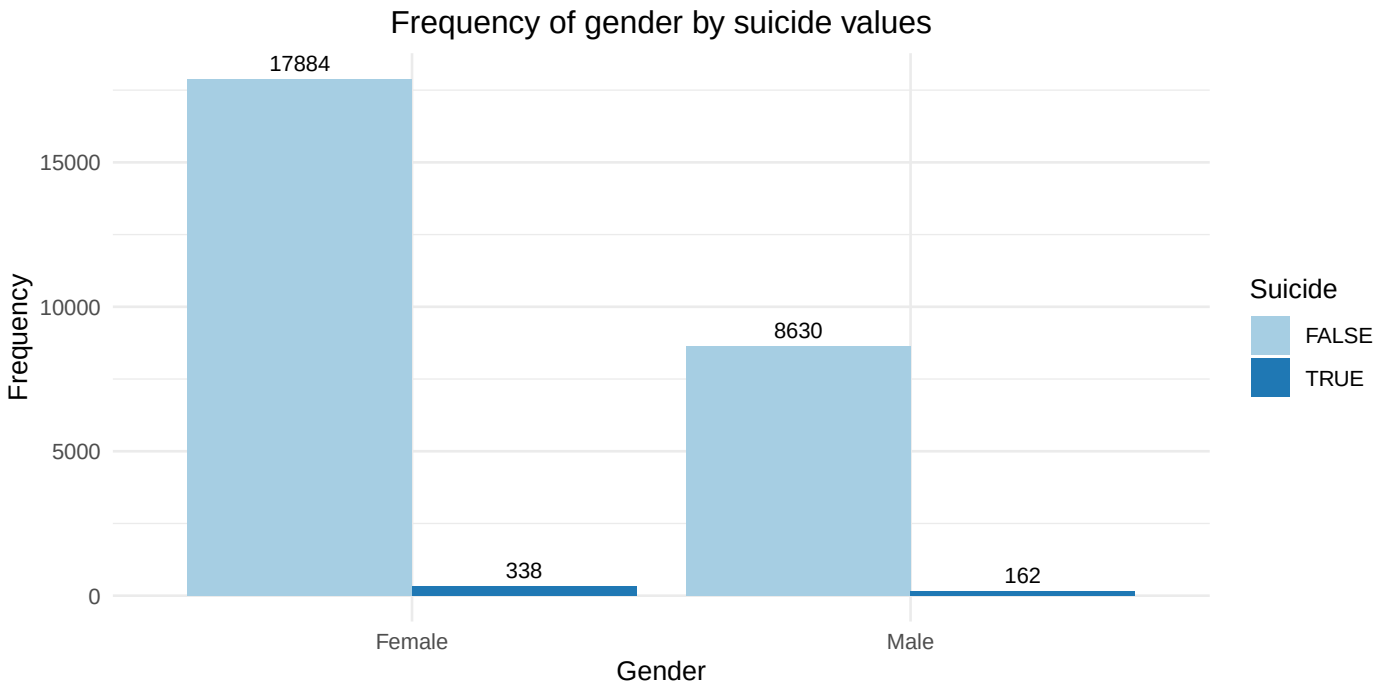
Imposing this threshold for the response variable, the classes appear quite unbalanced: indeed the proportion of negative suicide in the dataset is % and the proportion of positive suicide is %. Furthermore, in the model development section, some considerations on that will be stated.

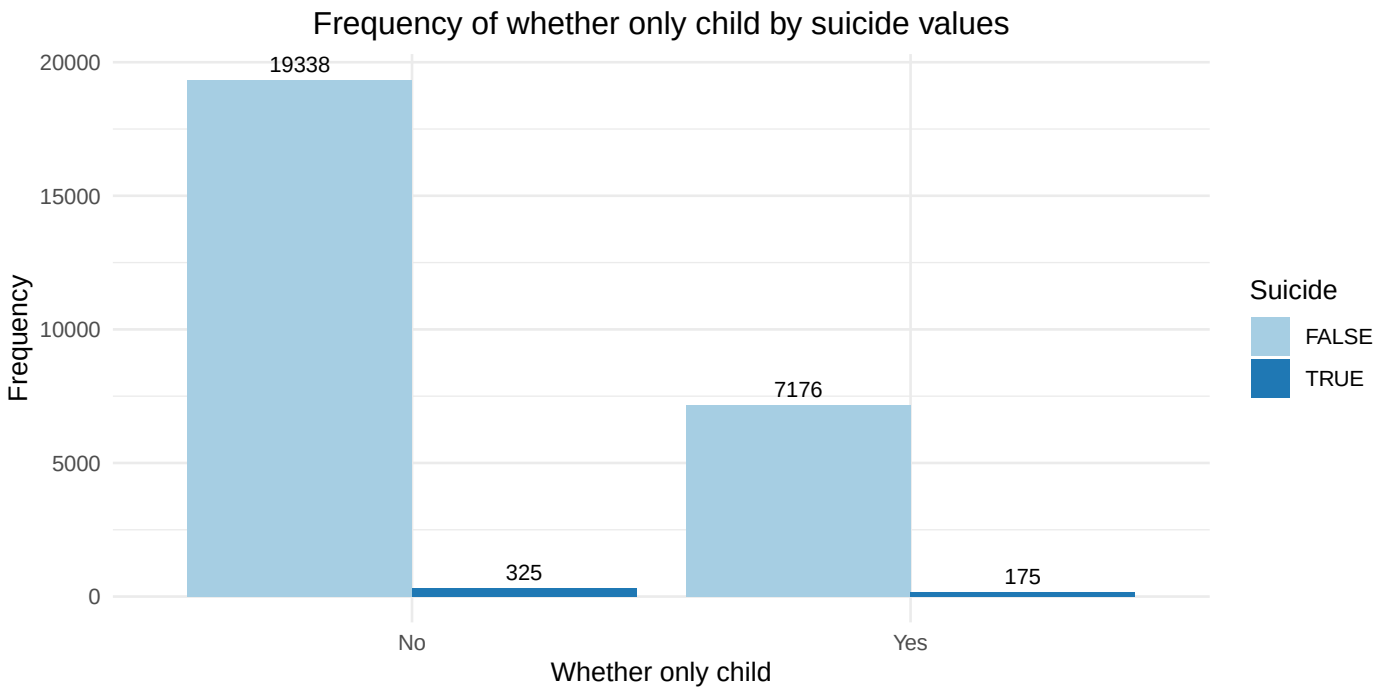
Exploratory Data Analysis

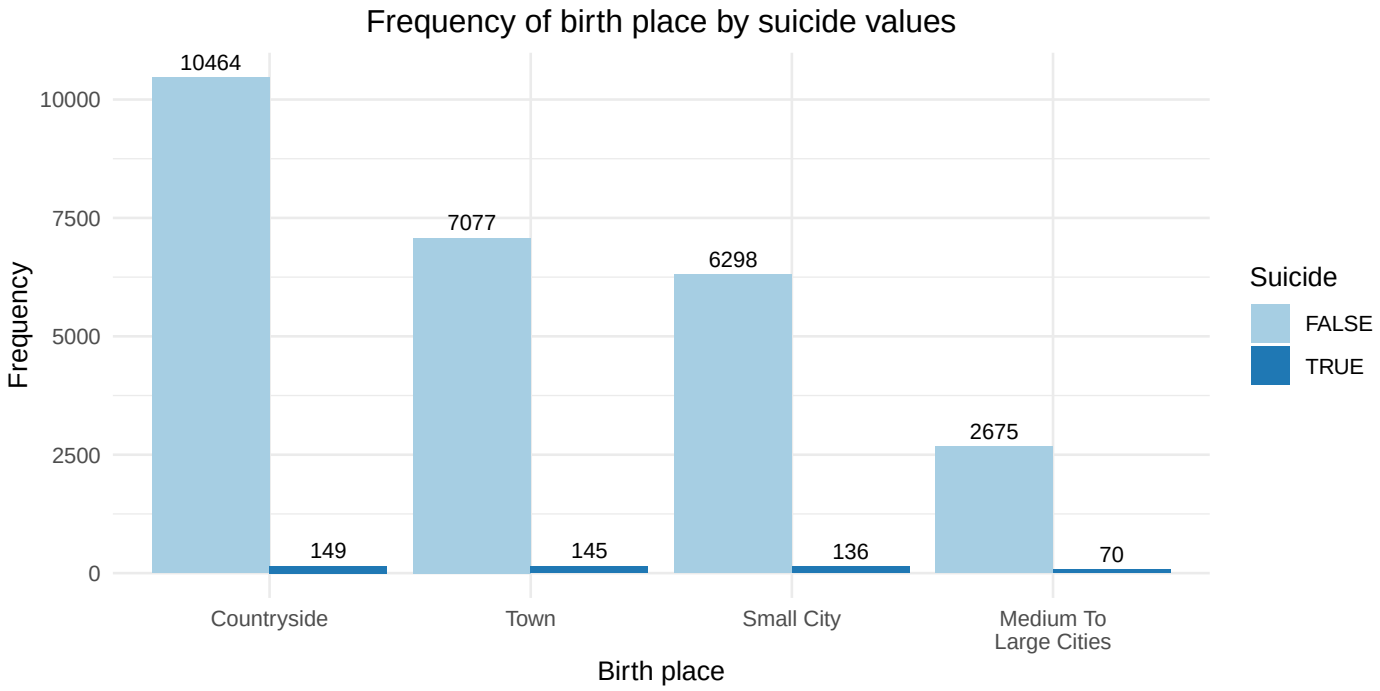
After a first overview of the data, it's needed to inspect some possible relations and hidden patterns between the variables. Some hypothesized interaction terms are also checked.

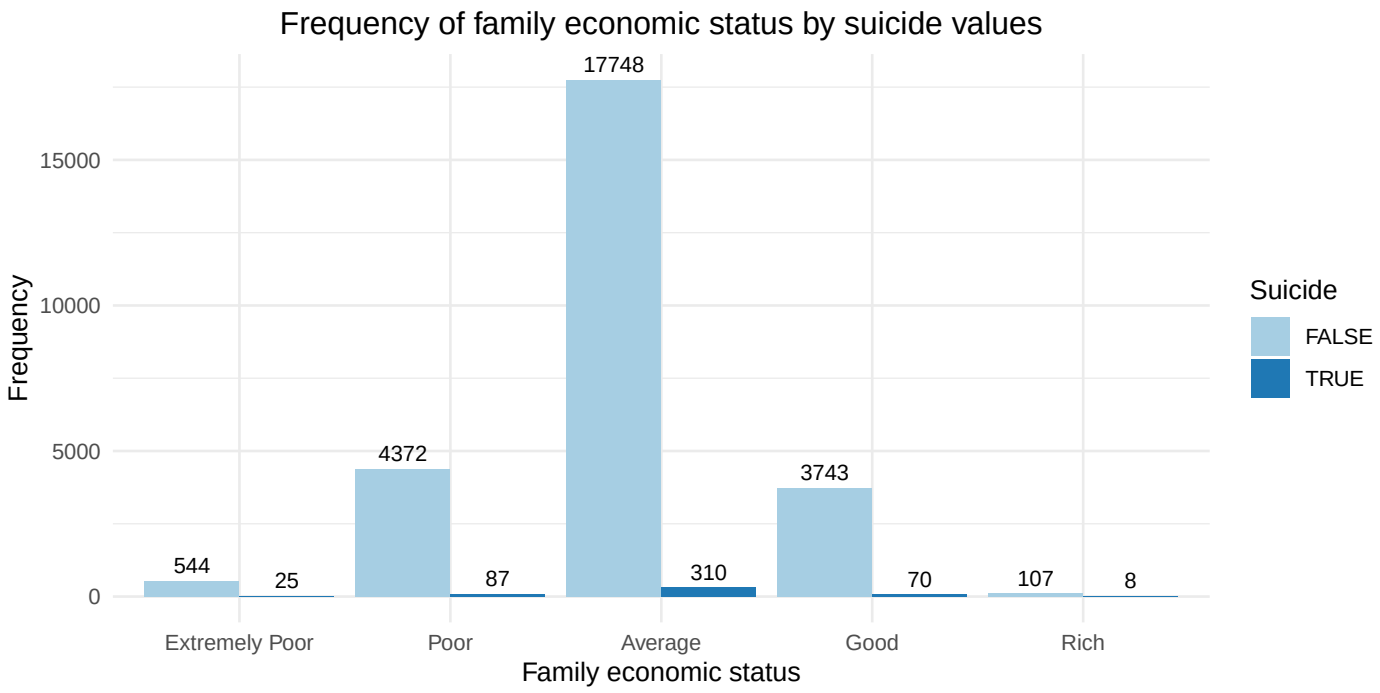
Frequencies concerning response variable

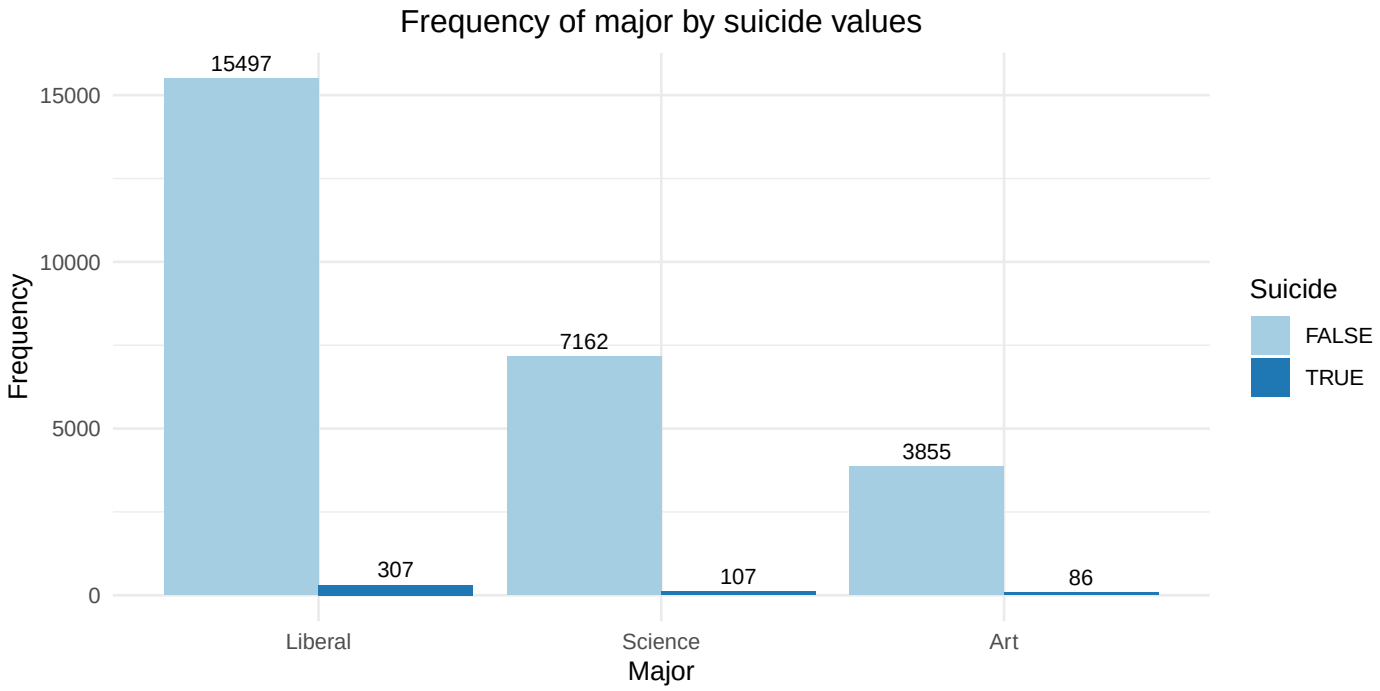
General information columns

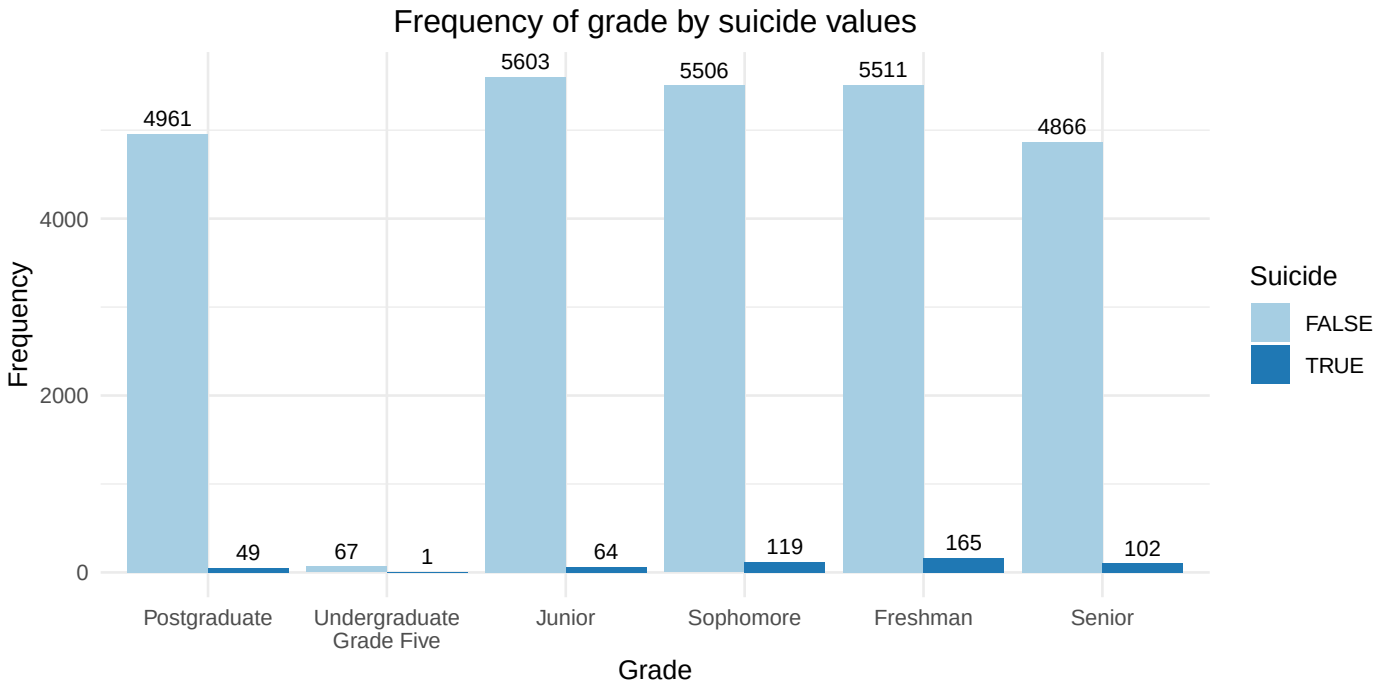



Frequency of whether only child by suicide values
20000
15000
10000
5000
0
Frequency
19338
325
7176
175
No
Yes
Whether only child
Suicide
FALSE
TRUE

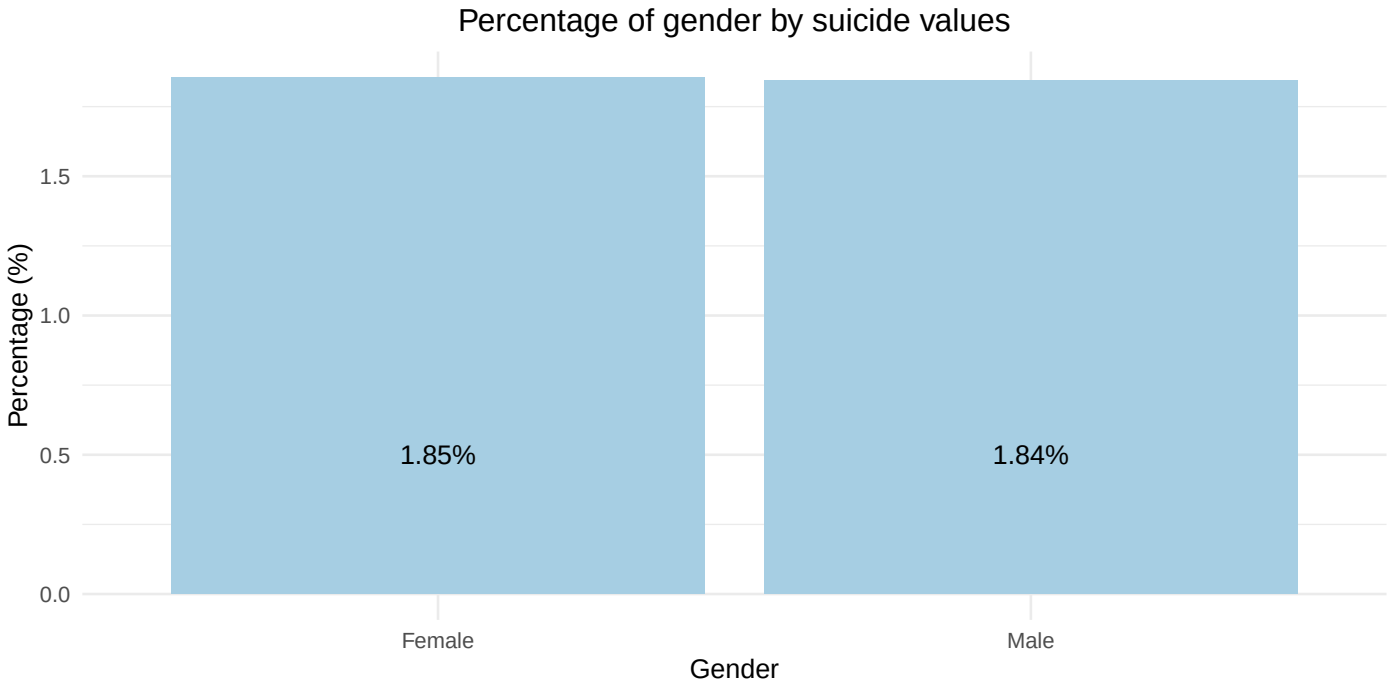

Frequency of birth place by suicide values
10000
7500
5000
2500
0
Frequency
10464
149
7077
145
6298
136
2675
70
Countryside
Town
Small City
Medium To
Large Cities
Birth place
Suicide
FALSE
TRUE


Frequency of family economic status by suicide values
Frequency
15000
10000
5000
0
544
25
4372
87
17748
310
3743
70
107
8
Extremely Poor
Poor
Average
Good
Rich
Family economic status
Suicide
FALSE
TRUE

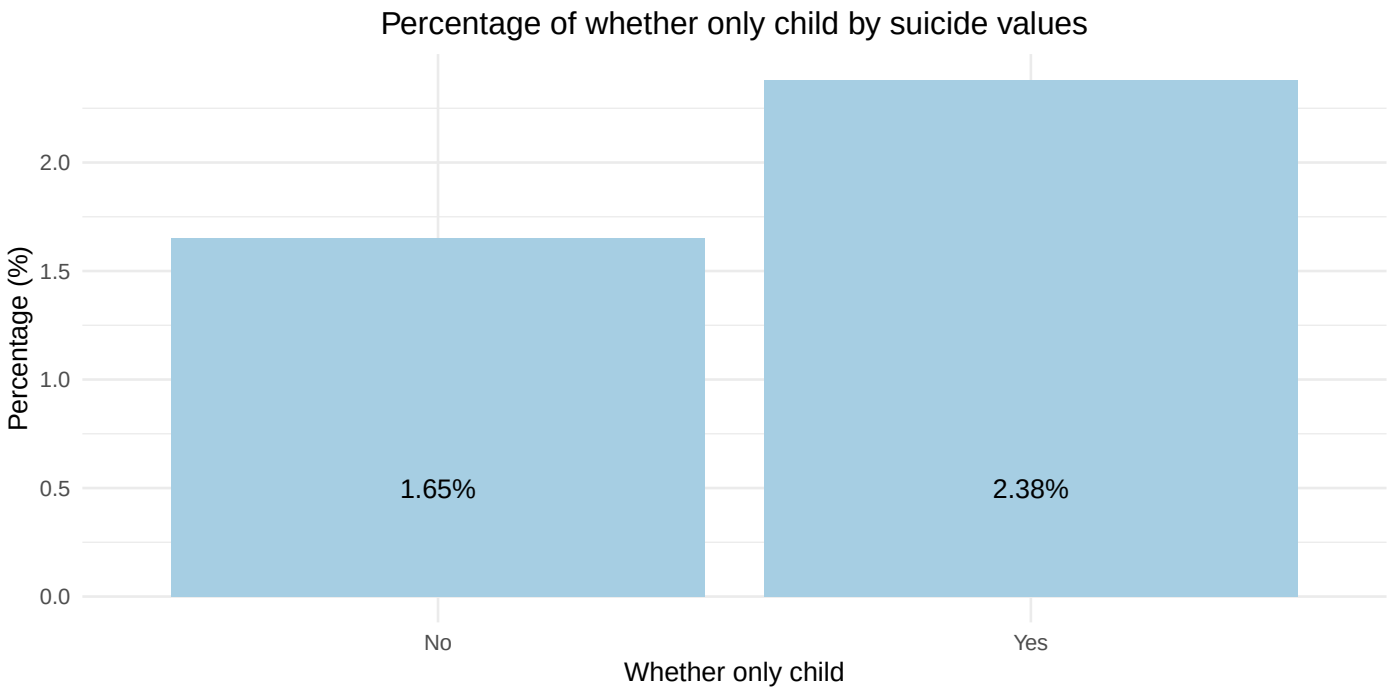

Frequency of major by suicide values
Frequency
15000
10000
5000
0
15497
307
7162
107
3855
86
Liberal
Science
Art
Major
Suicide
FALSE
TRUE



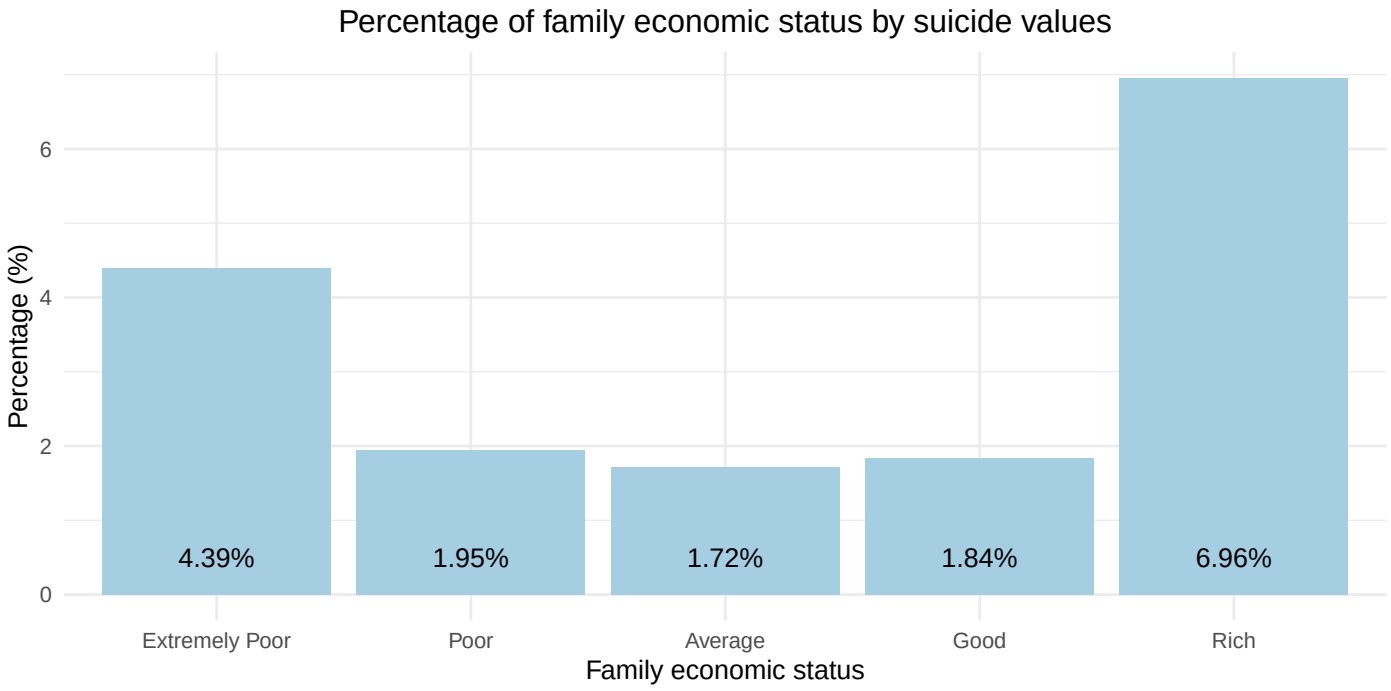
Looking at the bar plots referring to the general information about the students, it seems that there are no strange relationships between the categories and the values of the frequencies for the variable *suicide*: in particular the higher number of suicides for a categorical level is usually related with the higher number of person in that level. Just for the grade[1] *Freshmen* the number of suicides is higher than the other levels, and in particular, there are fewer suicides for older grading (e.g. postgraduate and undergraduate). Also, the *grade* variable is possibly related to some age information that in the dataframe is not present and it could suggest that younger people are less prone to suicide ideal.

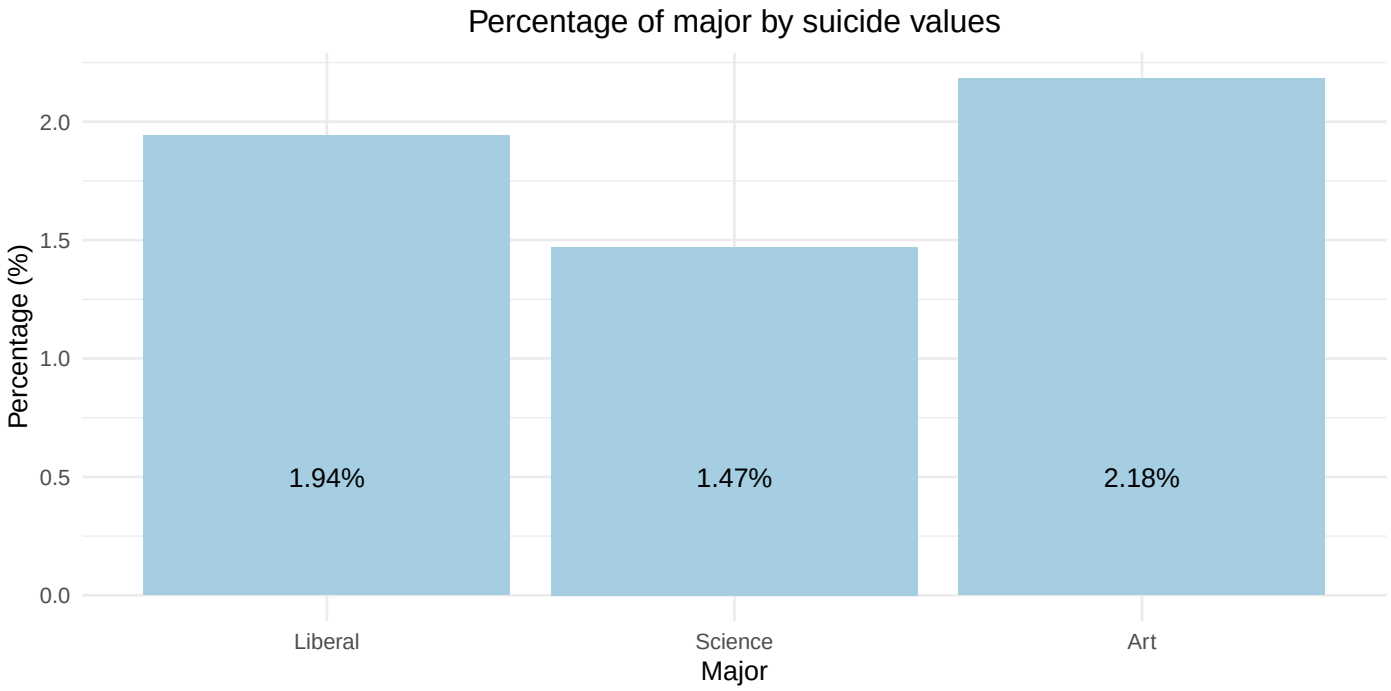


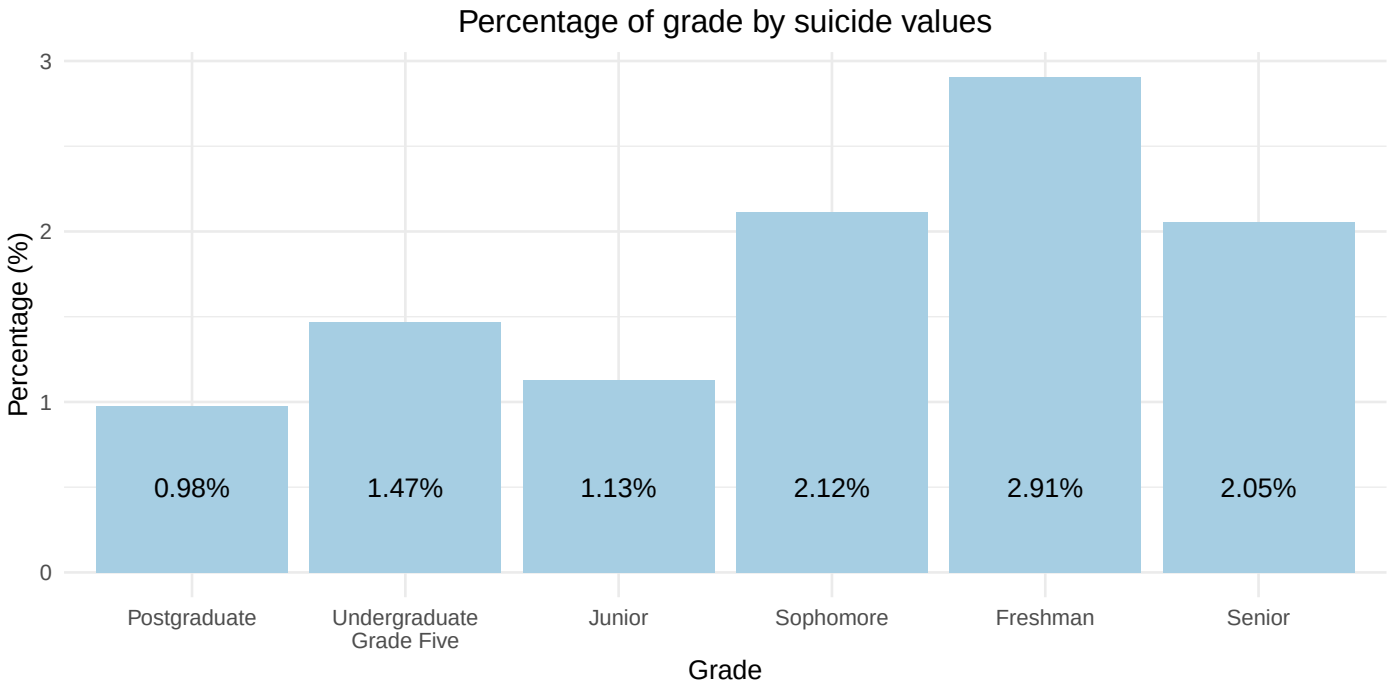
[1] https://uniexperts.com/en/news/freshman-sophomore-junior-senior-an-explanation/


Percentage of whether only child by suicide values
Percentage (%)
2.0
1.5
1.0
0.5
0.0
1.65%
2.38%
No
Yes
Whether only child


Percentage of birth place by suicide values
Percentage (%)
2
1
0
1.4%
2.01%
2.11%
2.55%
Countryside
Town
Small City
Medium To Large Cities
Birth place


Percentage of family economic status by suicide values
Percentage (%)
6
4
2
0
4.39%
1.95%
1.72%
1.84%
6.96%
Extremely Poor
Poor
Average
Good
Rich
Family economic status


Percentage of major by suicide values
Percentage (%)
2.0
1.5
1.0
0.5
0.0
1.94%
1.47%
2.18%
Liberal
Science
Art
Major

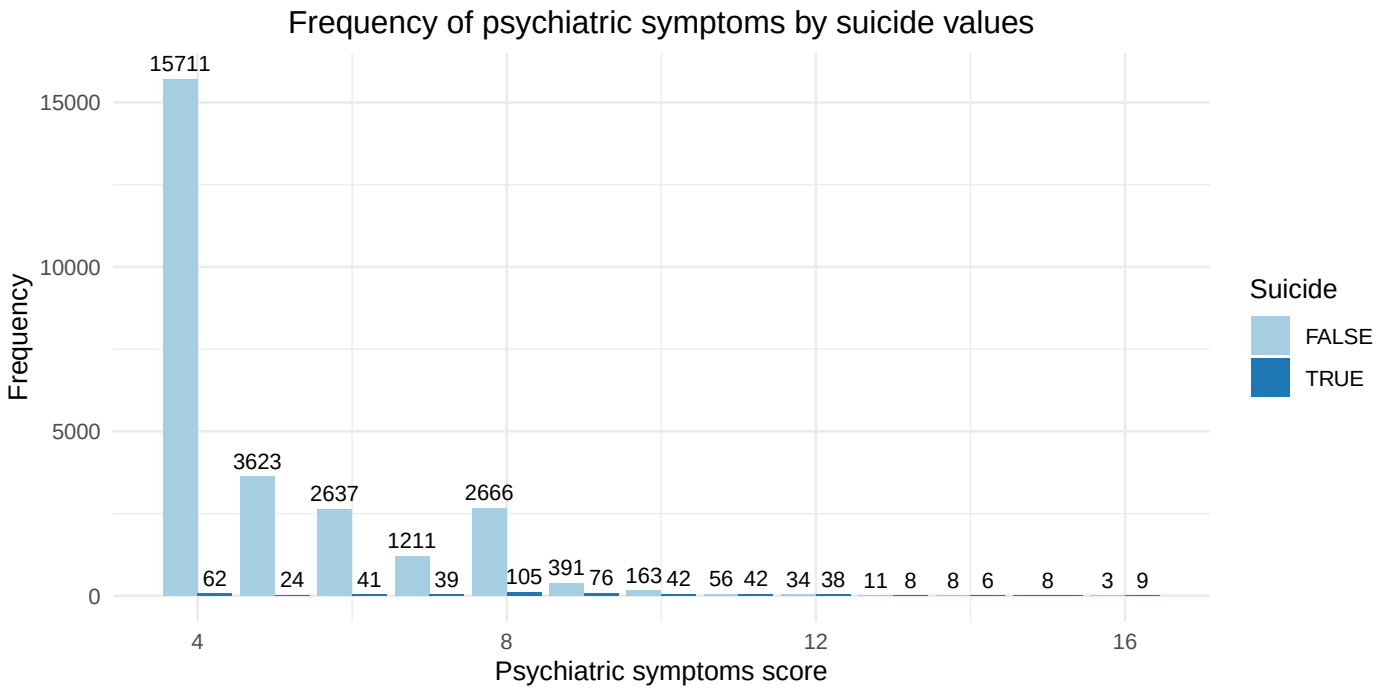


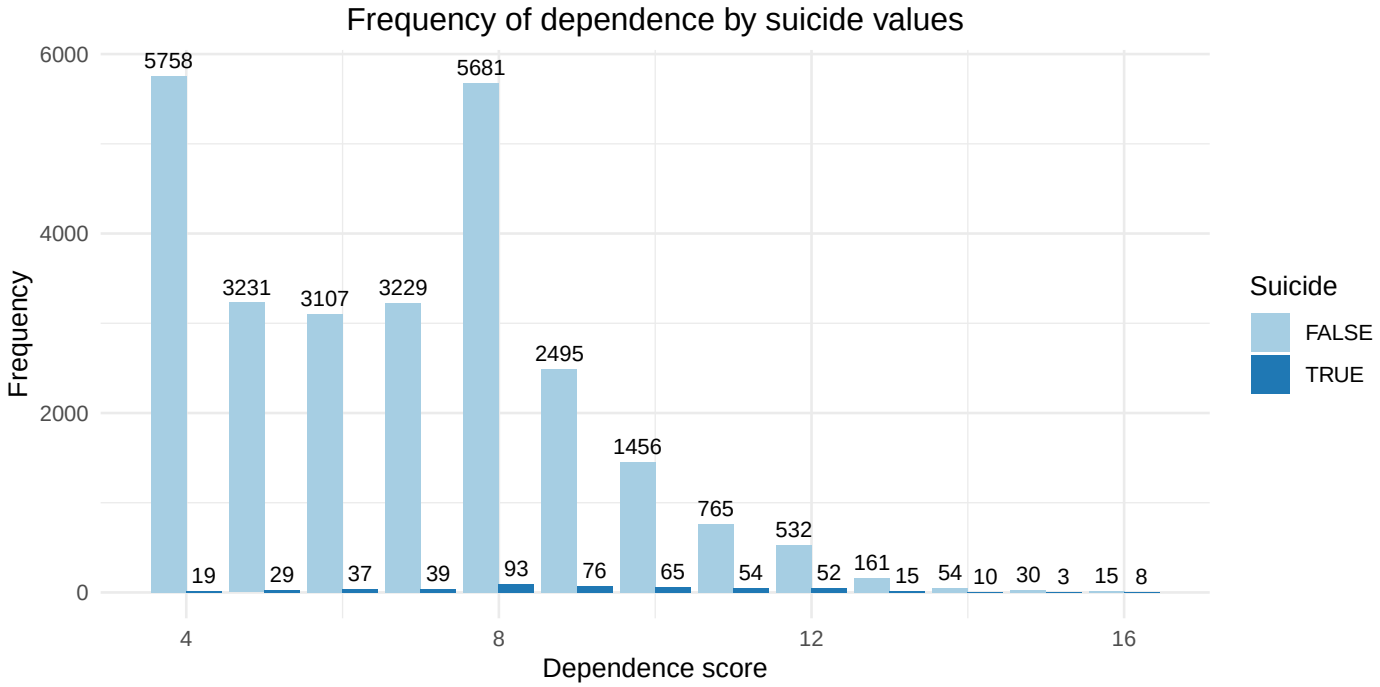
The plots above regarding the proportions of students for each category divided by the response values instead highlight interesting patterns: indeed it's more clear to understand the relationships considering ratios and not the frequencies, also because the y scale is different.

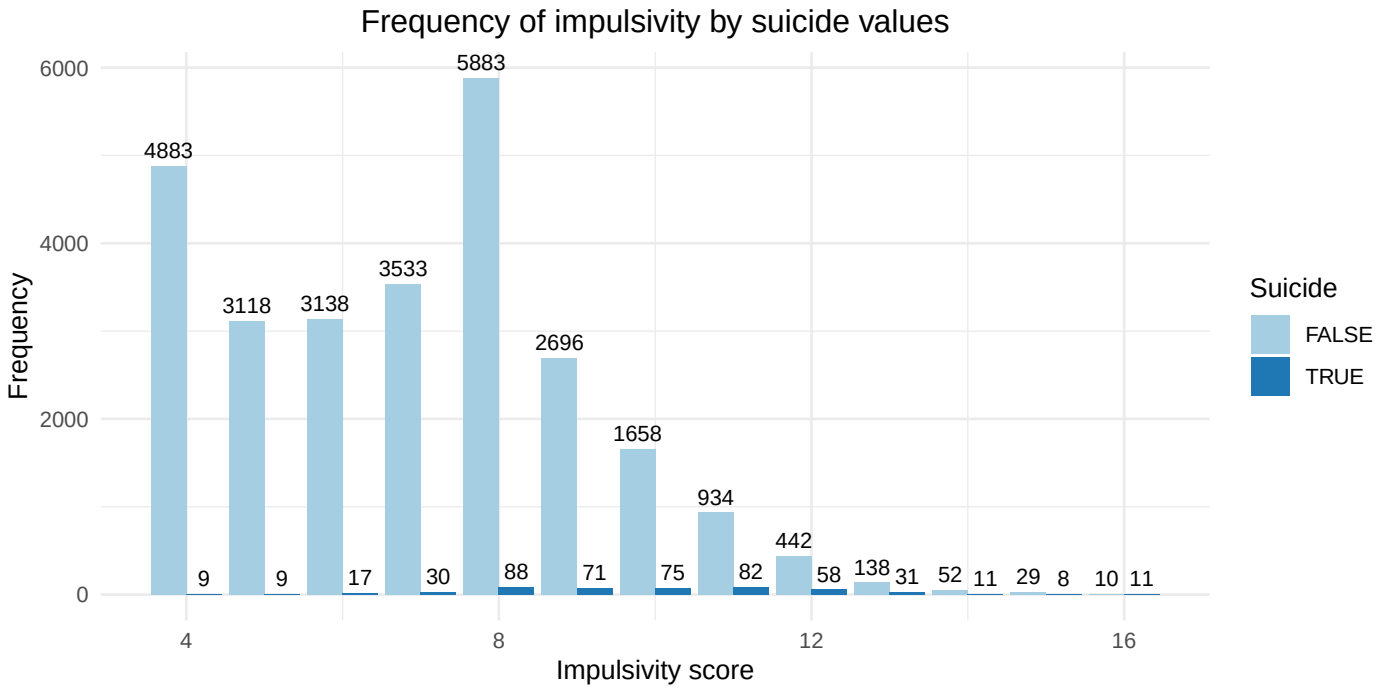
There are a few visible things in these visualizations:

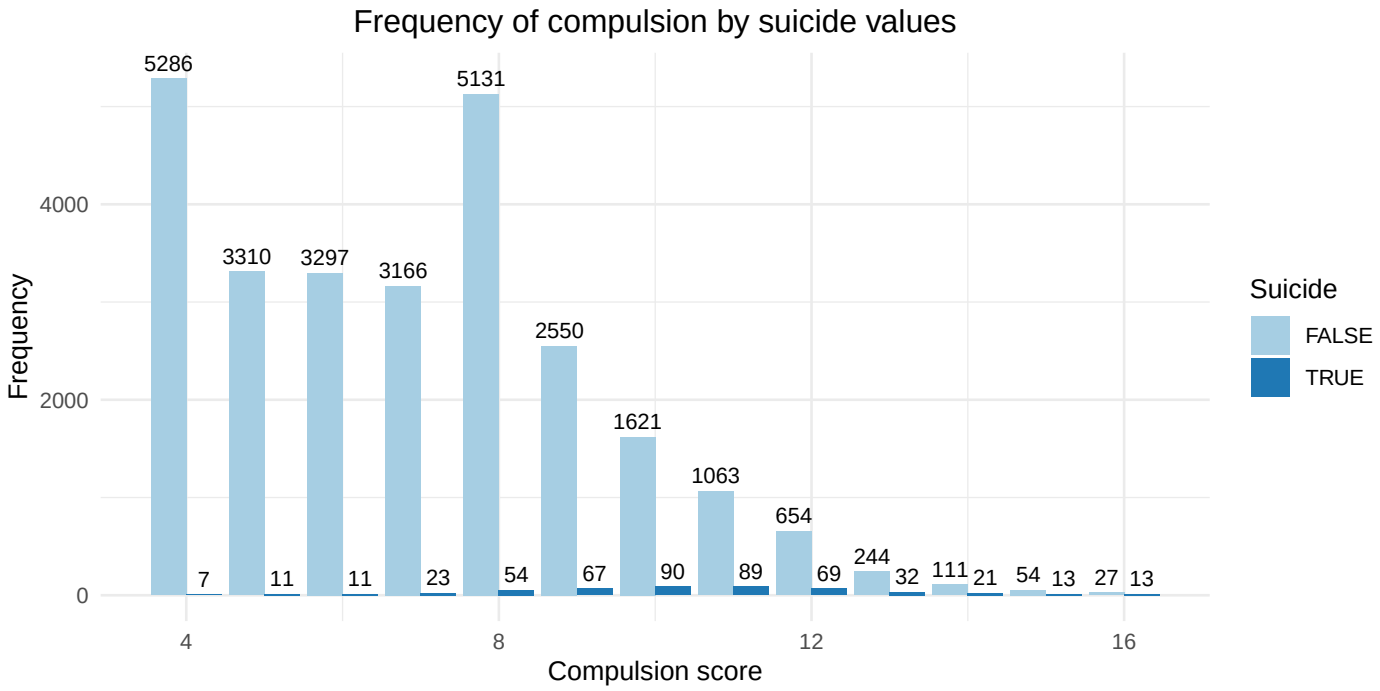
- The relative difference in suicide frequency considering the gender seems not so significant based on these data, instead, the gap is more noticeable compared with *whether only child* variable
- It seems that the percentage of suicide in medium to large cities is higher.
- Being in a rich family leads to a higher proportion of suicide based on this dataset, even more, that being in an extremely poor family's economic condition
- Attempting the art major seems to have a higher risk compared with liberal and science courses
- As observed before, the percentage of positive values in the response variable is higher for the freshman category

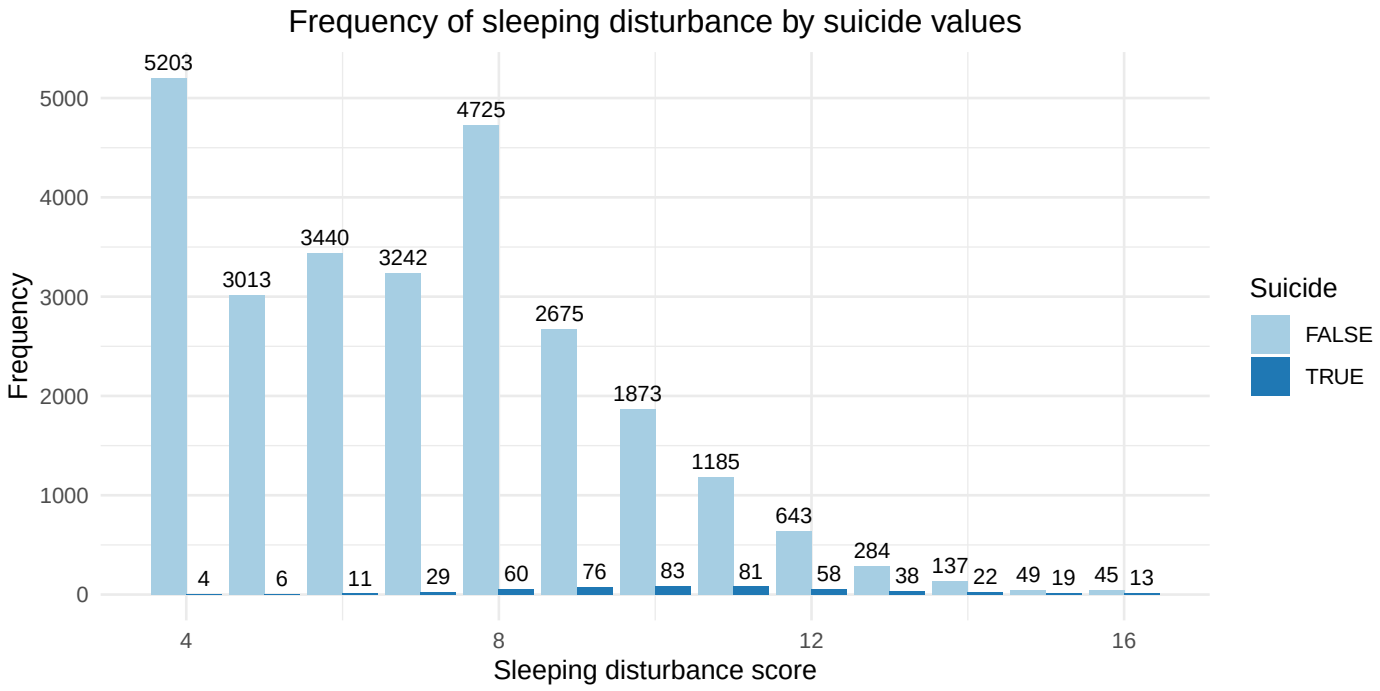
Symptoms information columns

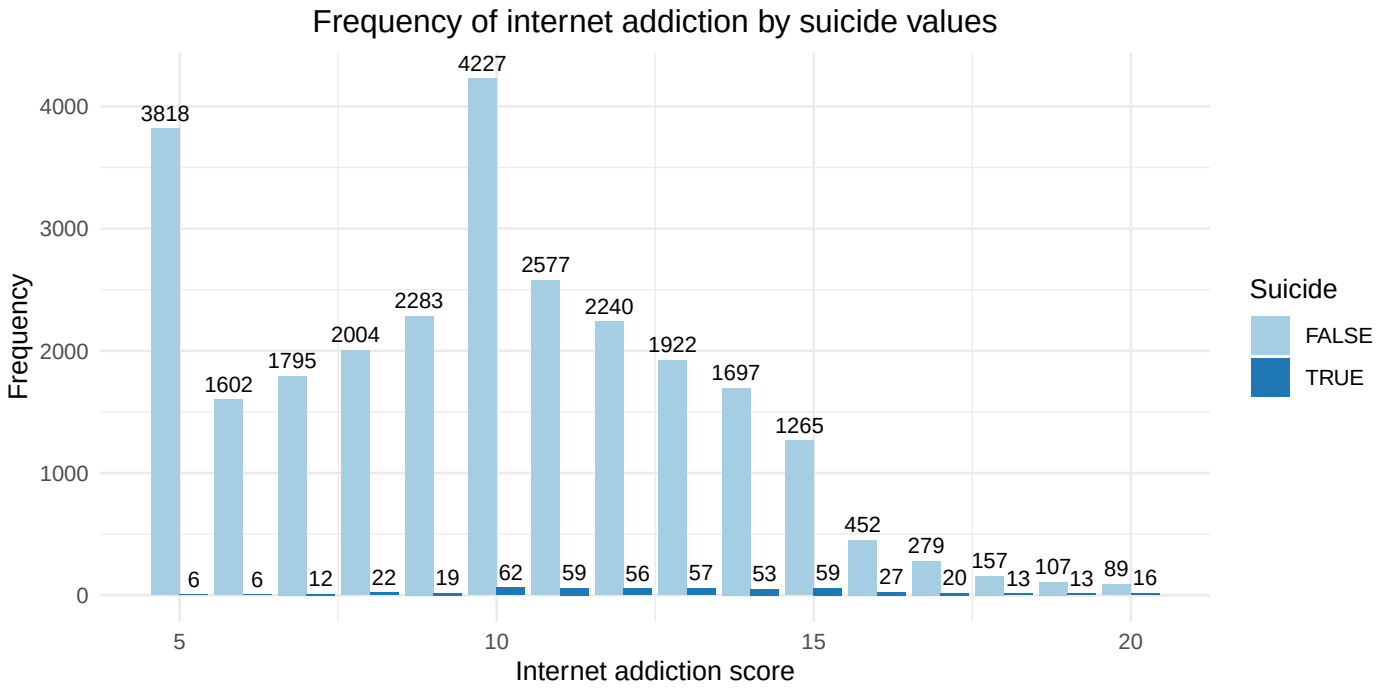


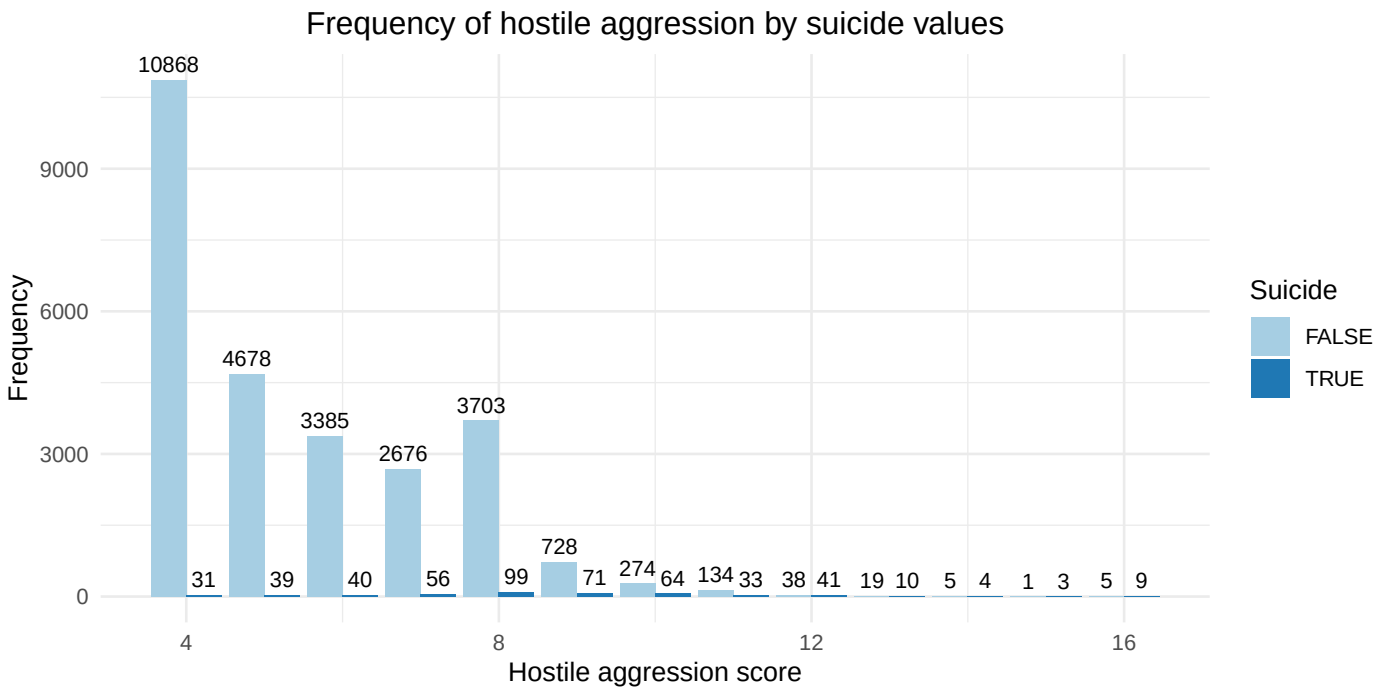


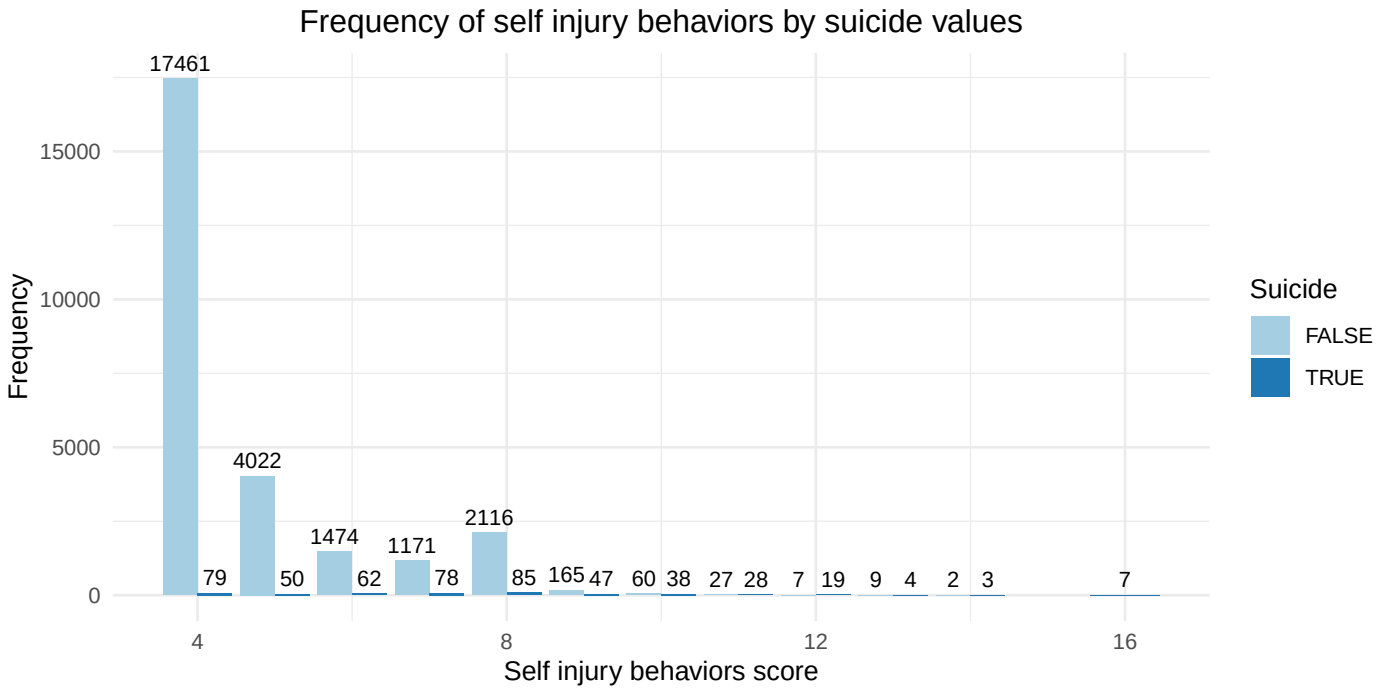

Frequency of impulsivity by suicide values
Frequency
6000
4000
2000
0
4883
9
3118
9
3138
17
3533
30
5883
88
2696
71
1658
75
934
82
442
58
138
31
52
11
29
8
10
11
4
8
12
16
Impulsivity score
Suicide
FALSE
TRUE

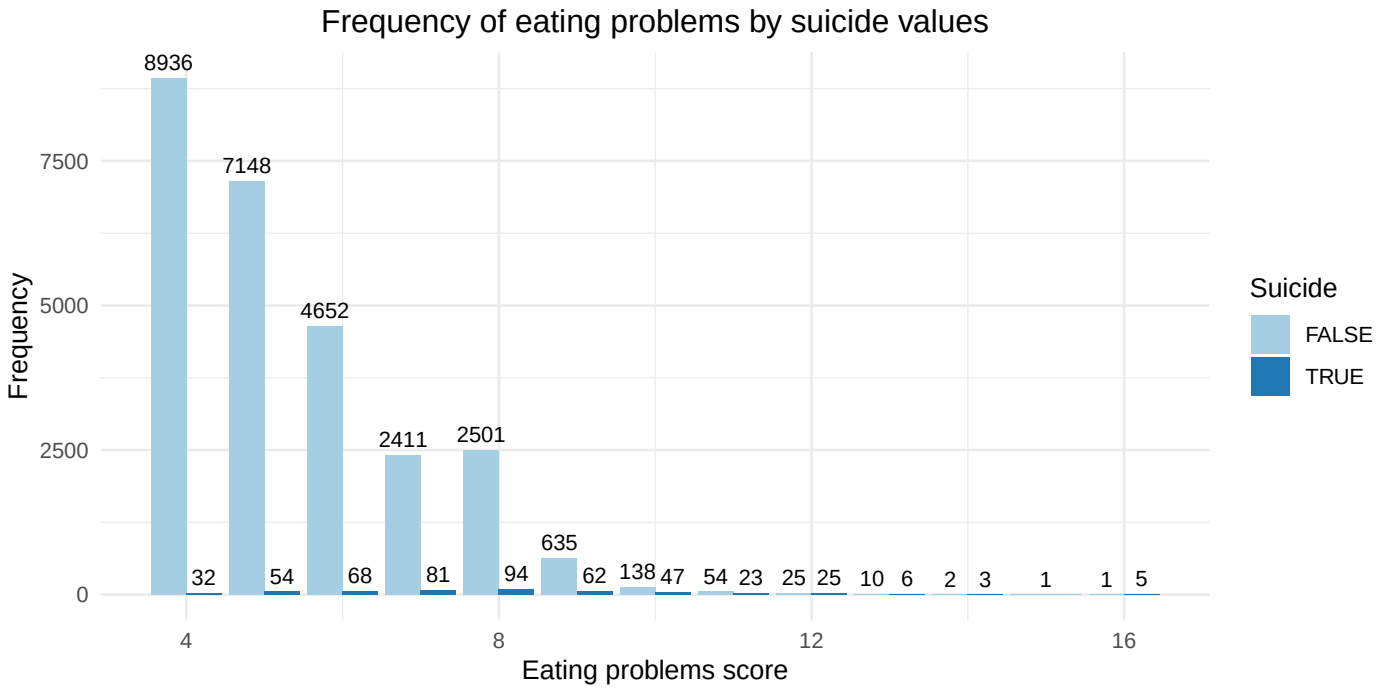

Frequency of compulsion by suicide values
Frequency
4000
2000
0
5286
7
3310
11
3297
11
3166
23
5131
54
2550
67
1621
90
1063
89
654
69
244
32
111
21
54
13
27
13
4
8
12
16
Compulsion score
Suicide
FALSE
TRUE


Frequency of sleeping disturbance by suicide values
Frequency
Sleeping disturbance score
Suicide
FALSE
TRUE
5203
4
3013
6
3440
11
3242
29
4725
60
2675
76
1873
83
1185
81
643
58
284
38
137
22
49
19
45
13

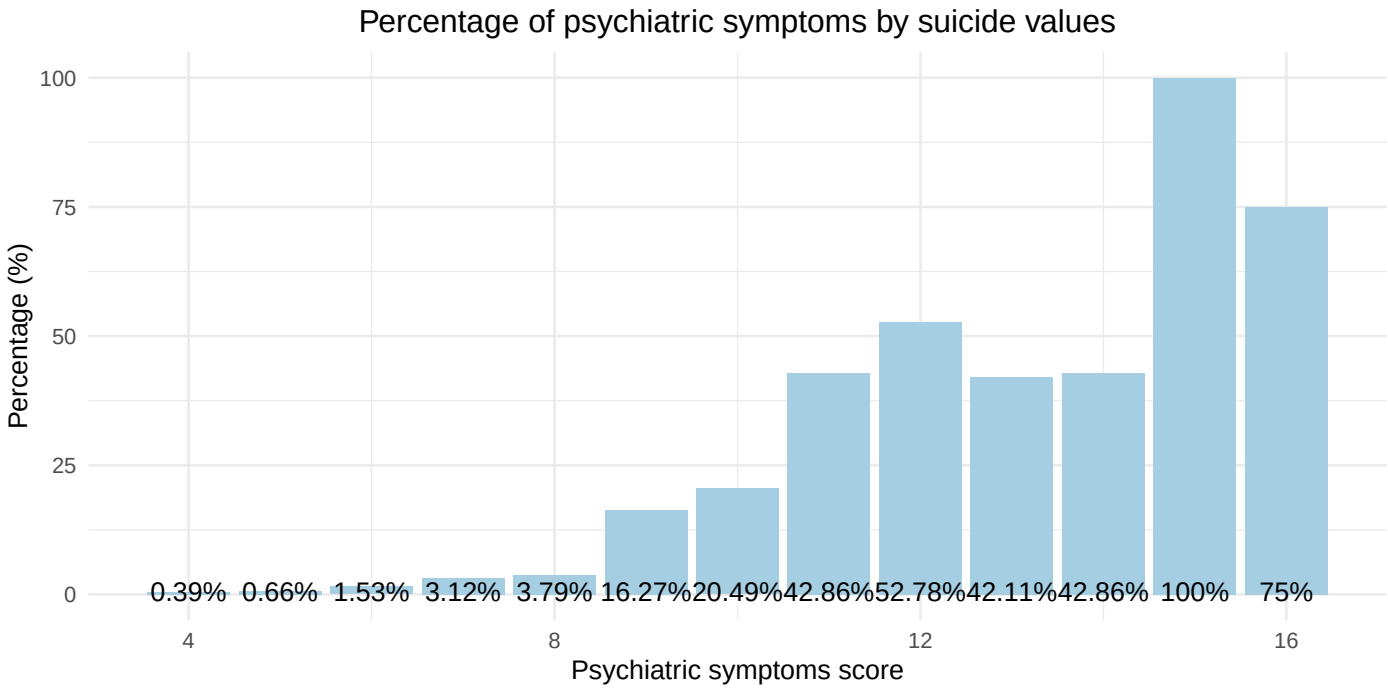

Frequency of internet addiction by suicide values
Frequency
Internet addiction score
Suicide
FALSE
TRUE
3818
6
1602
6
1795
12
2004
22
2283
19
4227
62
2577
59
2240
56
1922
57
1697
53
1265
59
452
27
279
20
157
13
107
13
89
16

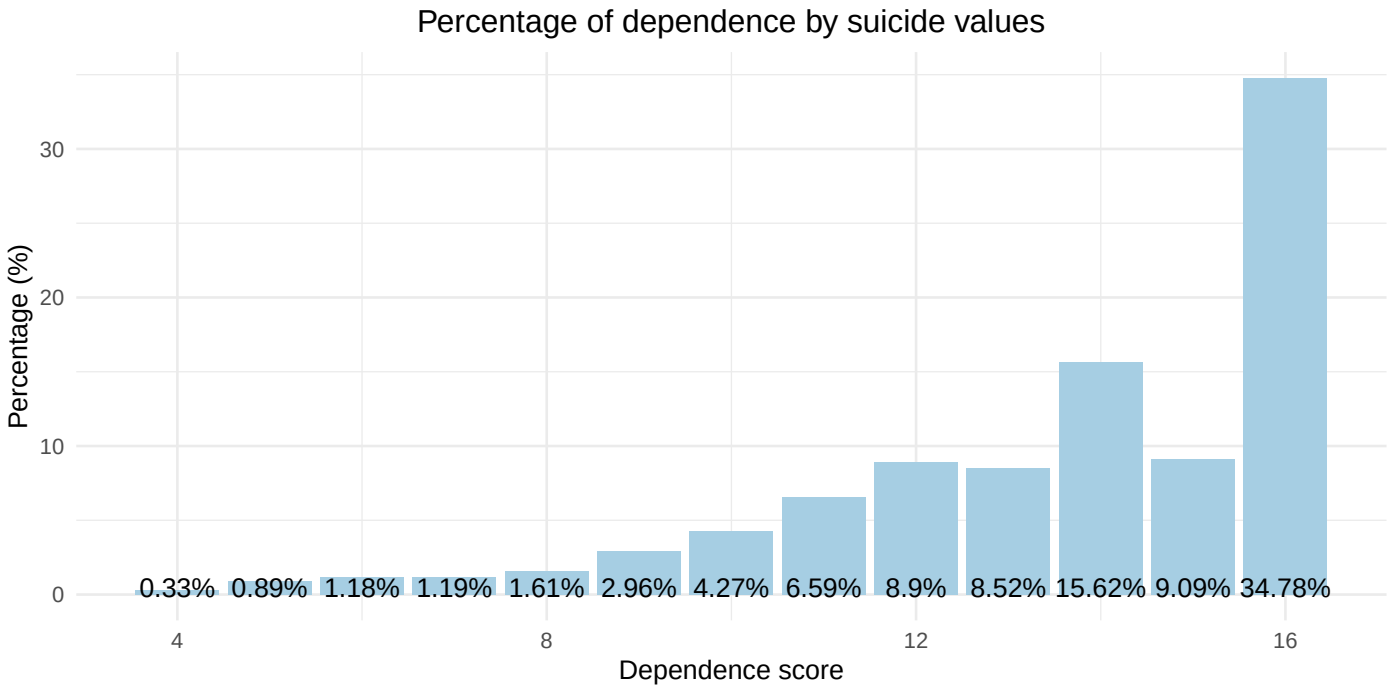

Frequency of hostile aggression by suicide values
Frequency
Hostile aggression score
Suicide
FALSE
TRUE
10868
31
4678
39
3385
40
2676
56
3703
99
728
71
274
64
134
33
38
41
19
10
5
4
1
3
5
9
0
3000
6000
9000
4
8
12
16

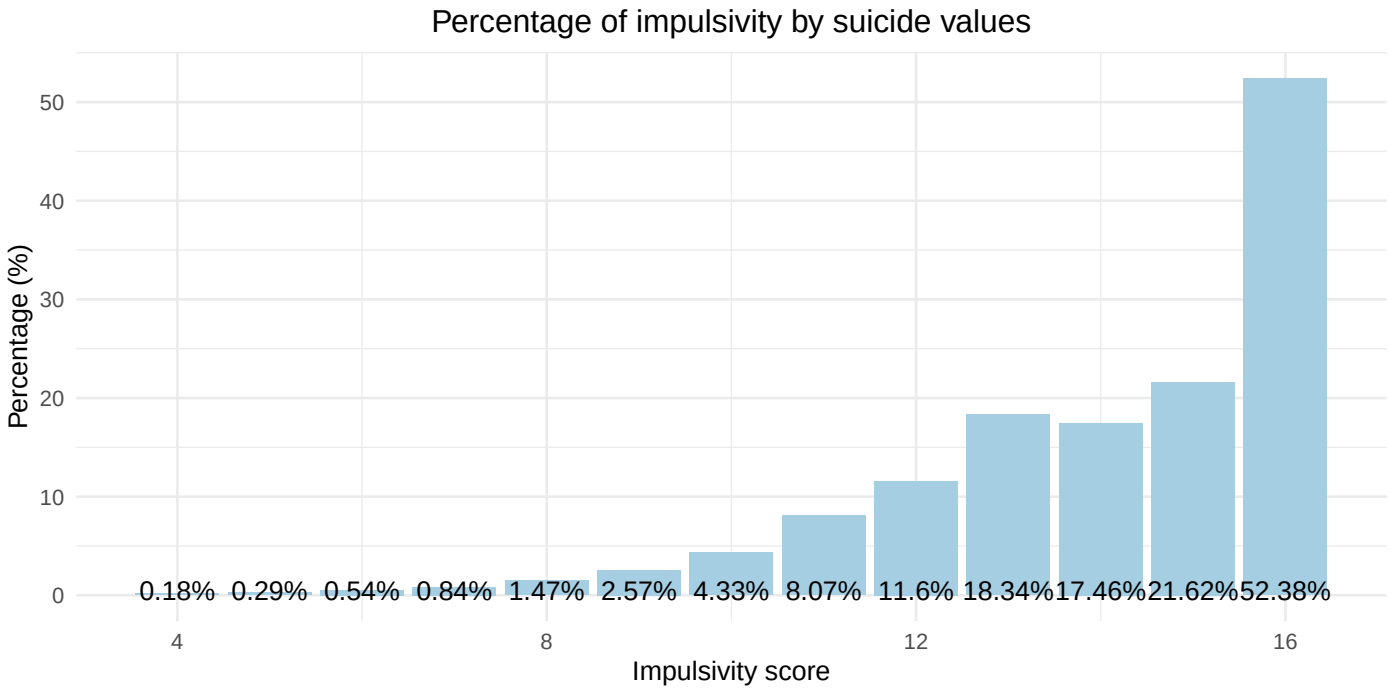

Frequency of self injury behaviors by suicide values
Frequency
Self injury behaviors score
Suicide
FALSE
TRUE
17461
79
4022
50
1474
62
1171
78
2116
85
165
47
60
38
27
28
7
19
9
4
2
3
7
0
5000
10000
15000
4
8
12
16

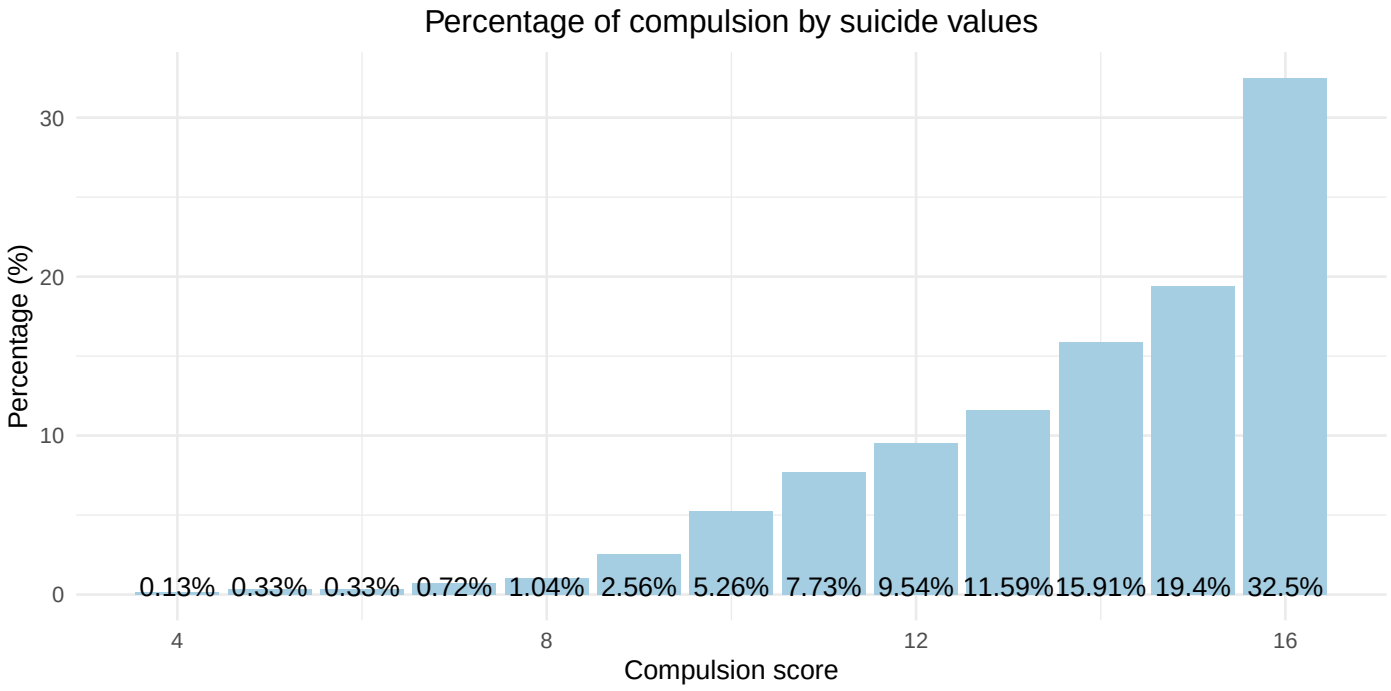


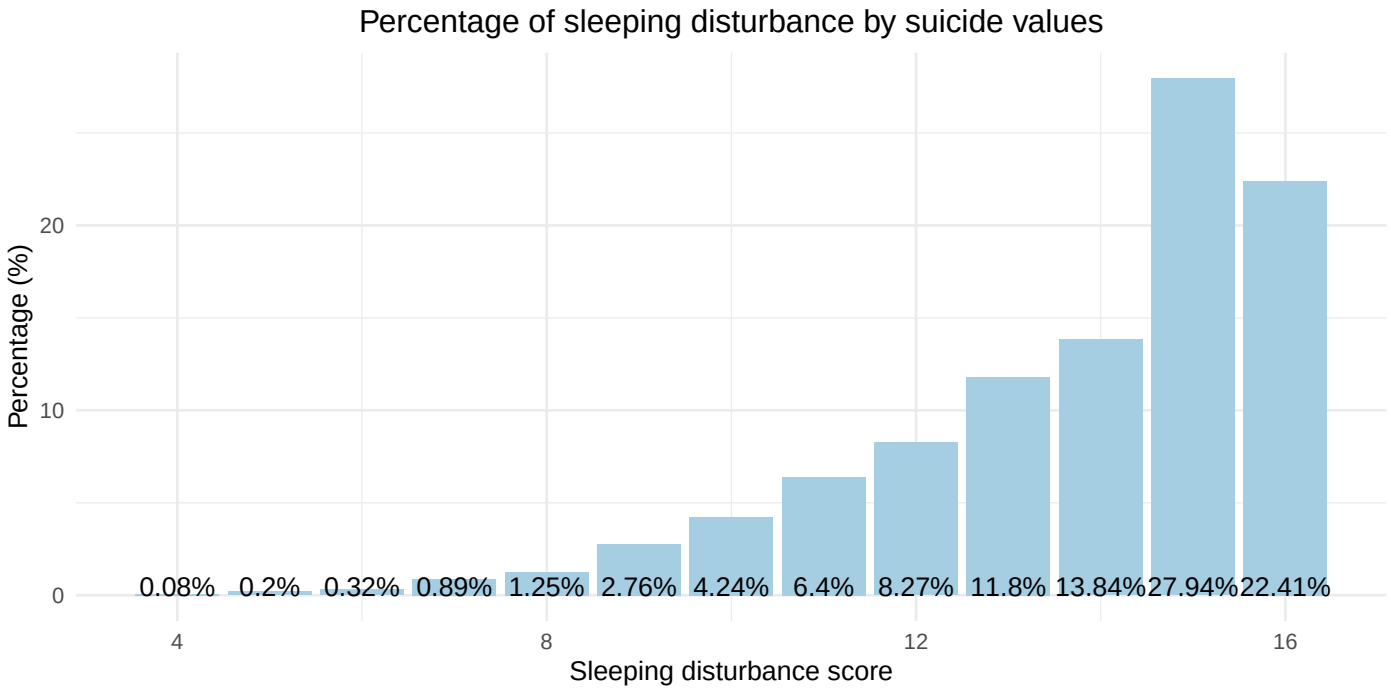
The histograms of the symptoms features show a slightly different behavior compared with the previous categorical predictors' plots. As observed in the beginning, the number of high-severity symptoms in general is low but the frequency of suicide gets bigger when these symptoms increase.

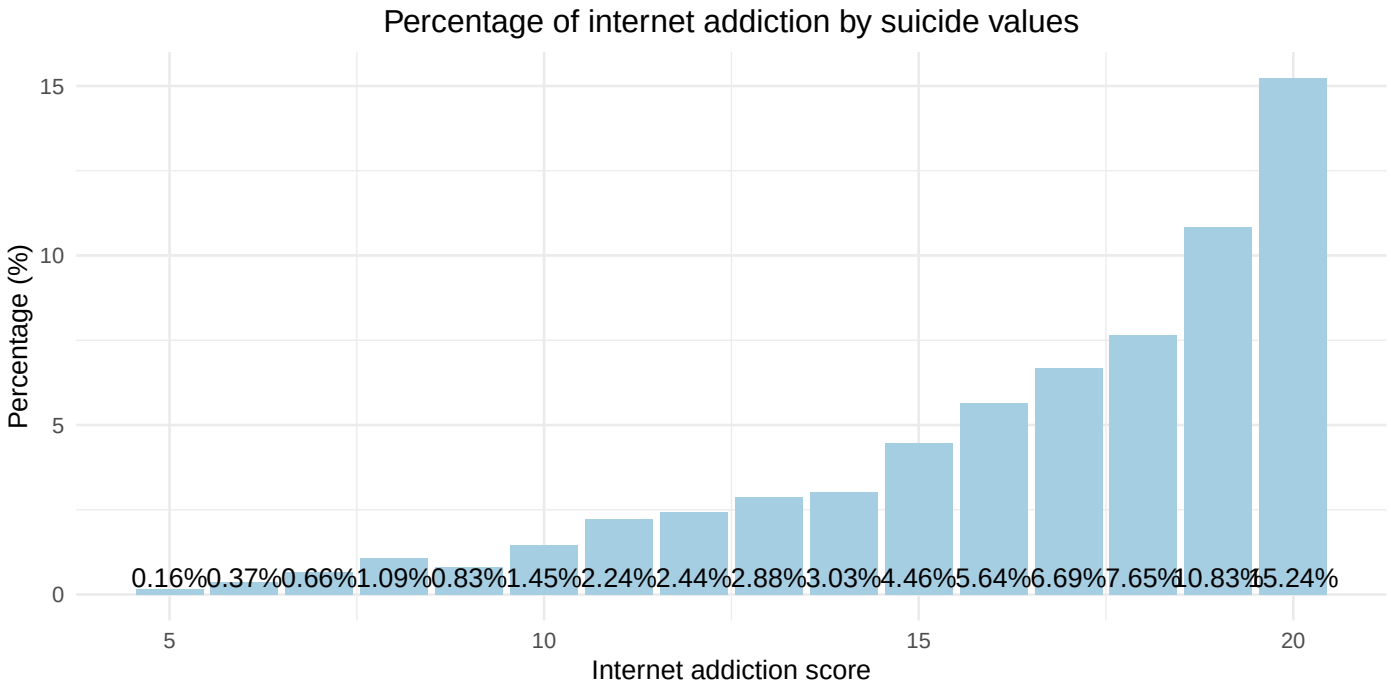


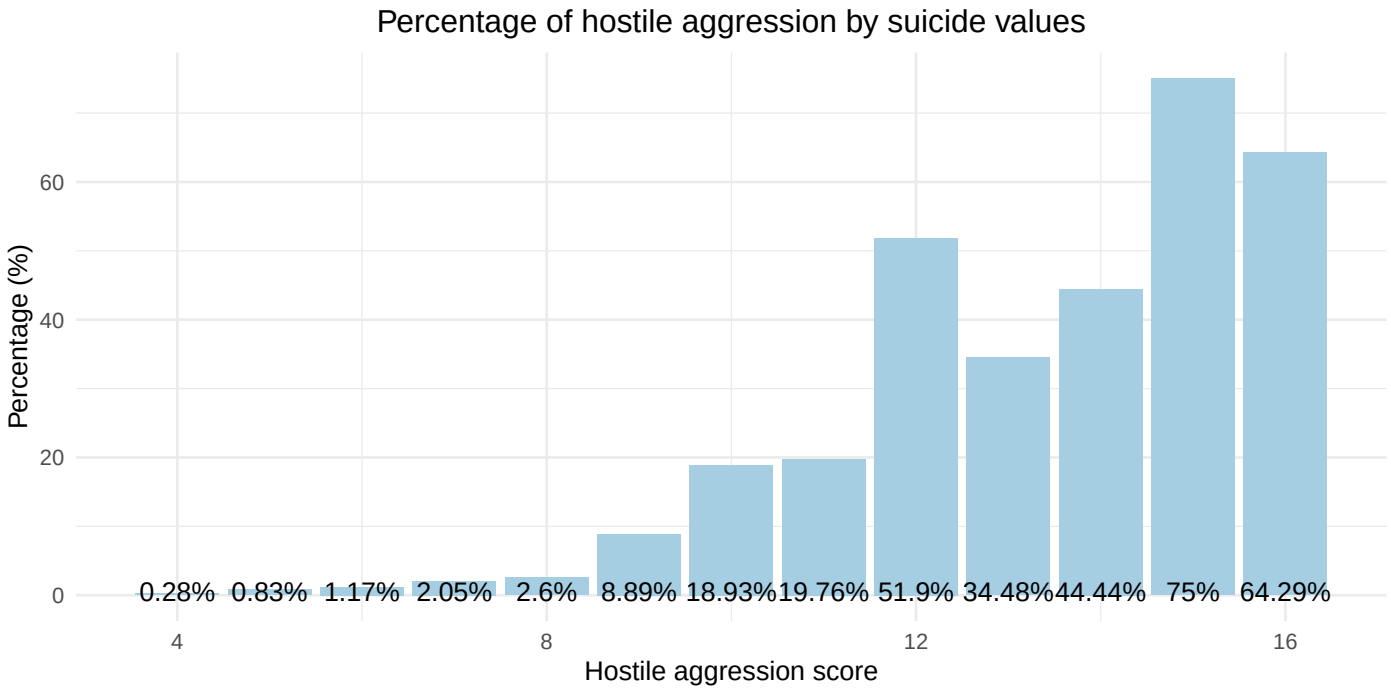

Percentage of dependence by suicide values
Percentage (%)
30
20
10
0
0.33%
0.89%
1.18%
1.19%
1.61%
2.96%
4.27%
6.59%
8.9%
8.52%
15.62%
9.09%
34.78%
4
8
12
16
Dependence score

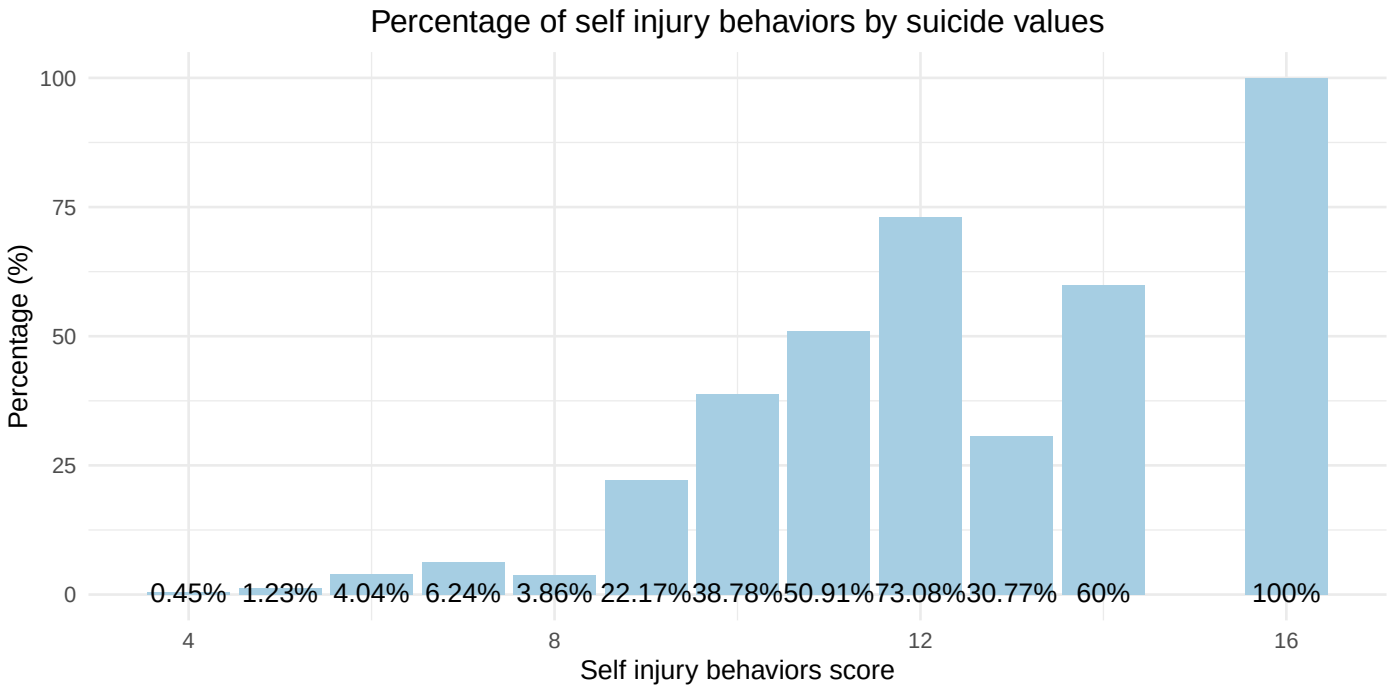

Percentage of impulsivity by suicide values
Percentage (%)
50
40
30
20
10
0
0.18%
0.29%
0.54%
0.84%
1.47%
2.57%
4.33%
8.07%
11.6%
18.34%
17.46%
21.62%
52.38%
4
8
12
16
Impulsivity score

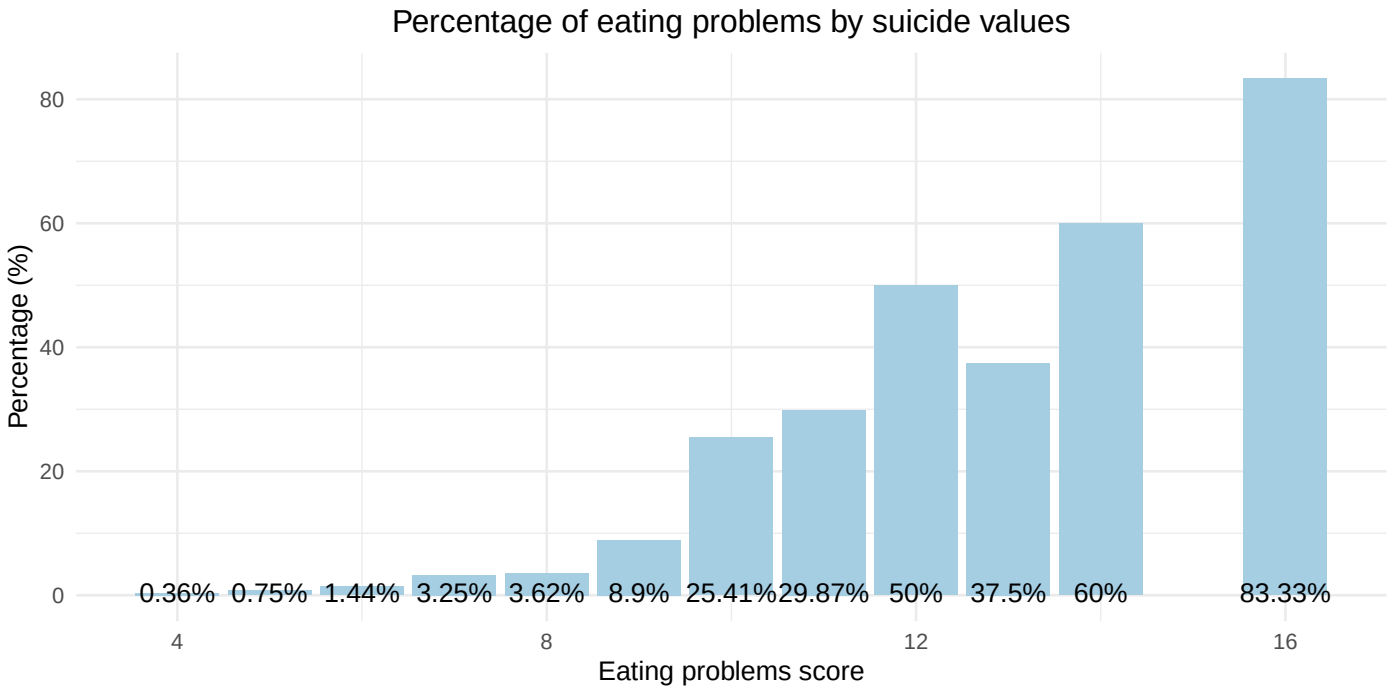

Percentage of compulsion by suicide values
Percentage (%)
30
20
10
0
0.13%
0.33%
0.33%
0.72%
1.04%
2.56%
5.26%
7.73%
9.54%
11.59%
15.91%
19.4%
32.5%
4
8
12
16
Compulsion score


Percentage of sleeping disturbance by suicide values
Percentage (%)
20
10
0
0.08%
0.2%
0.32%
0.89%
1.25%
2.76%
4.24%
6.4%
8.27%
11.8%
13.84%
27.94%
22.41%
4
8
12
16
Sleeping disturbance score


Percentage of internet addiction by suicide values
Percentage (%)
15
10
5
0
0.16%
0.37%
0.66%
1.09%
0.83%
1.45%
2.24%
2.44%
2.88%
3.03%
4.46%
5.64%
6.69%
7.65%
10.83%
15.24%
5
10
15
20
Internet addiction score

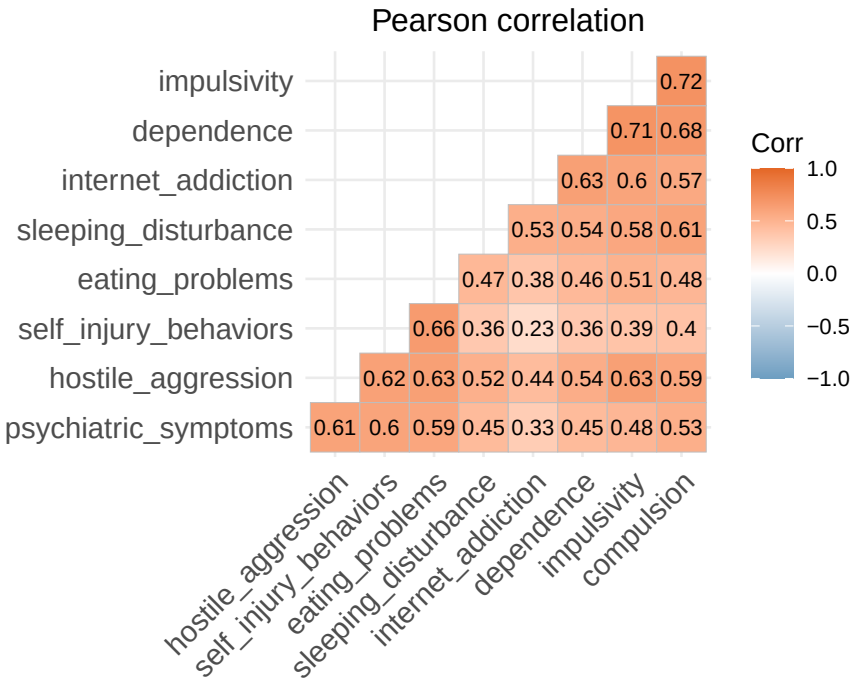

Percentage of hostile aggression by suicide values
Percentage (%)
60
40
20
0
0.28%
0.83%
1.17%
2.05%
2.6%
8.89%
18.93%
19.76%
51.9%
34.48%
44.44%
75%
64.29%
4
8
12
16
Hostile aggression score





Checking the proportion histograms divided by the response variable what was stated previously becomes more evident: now it can be seen a noticeable increasing trend in the suicide ratios when the symptoms severity increases and in particular the percentage is higher for the variables psychiatric symptoms, hostile aggression, self-injury behaviors and eating problems, even if the frequency of the higher symptoms for them is quite low (except for impulsivity and compulsion variables)

Numeric variables



The heatmap of the correlation matrix shows that all the variables are quite highly positively correlated with each other. By the plot can be seen that the most correlated variables are *impulsivity* and *compulsion* (0.72), *dependence* and *impulsivity* (0.71), *dependence* and *compulsion* (0.68). Also *self-injury behaviors* and *eating problems* seem to be correlated in a significant way (0.66).

Inspecting the correlation matrix, it seems that some variables are quite high positive correlated. however this information is not fully considered because numeric variables have all positive variables and they are almost all in the same range.

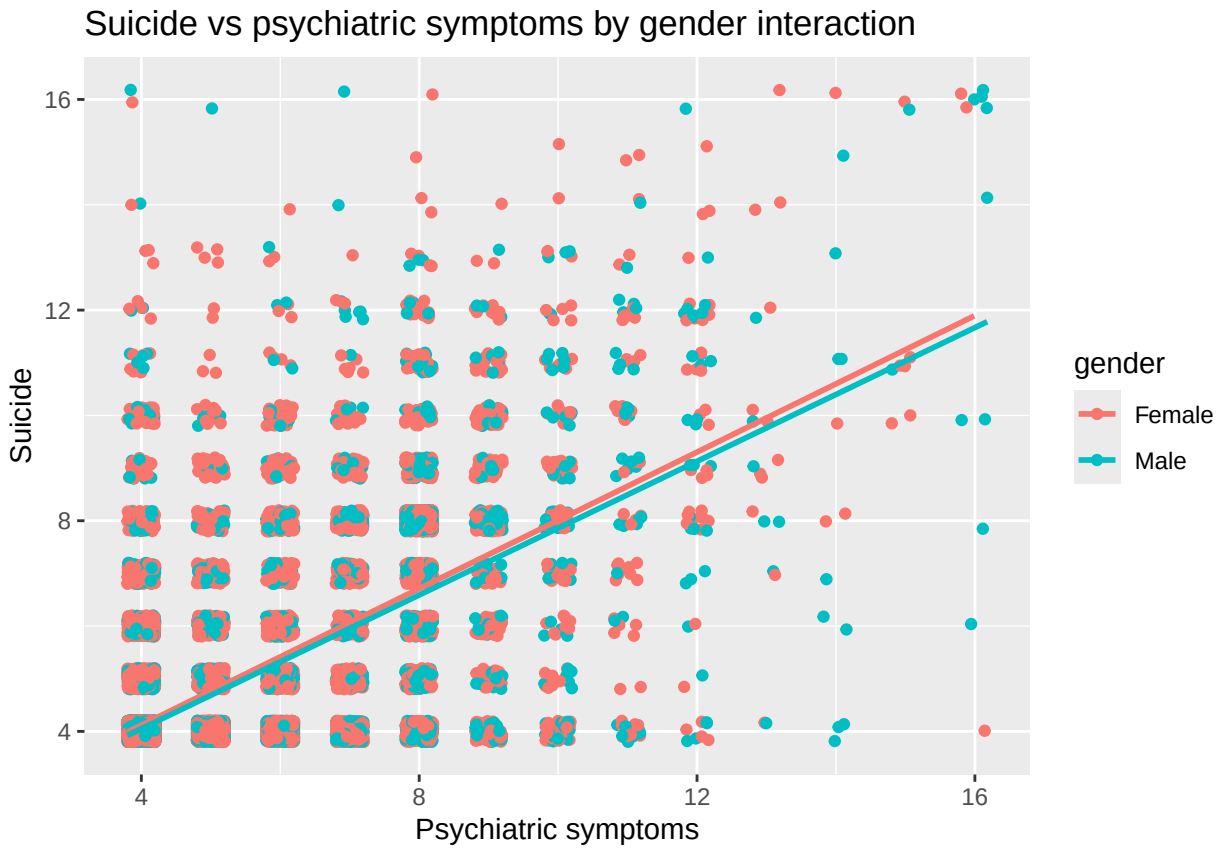
Interaction terms

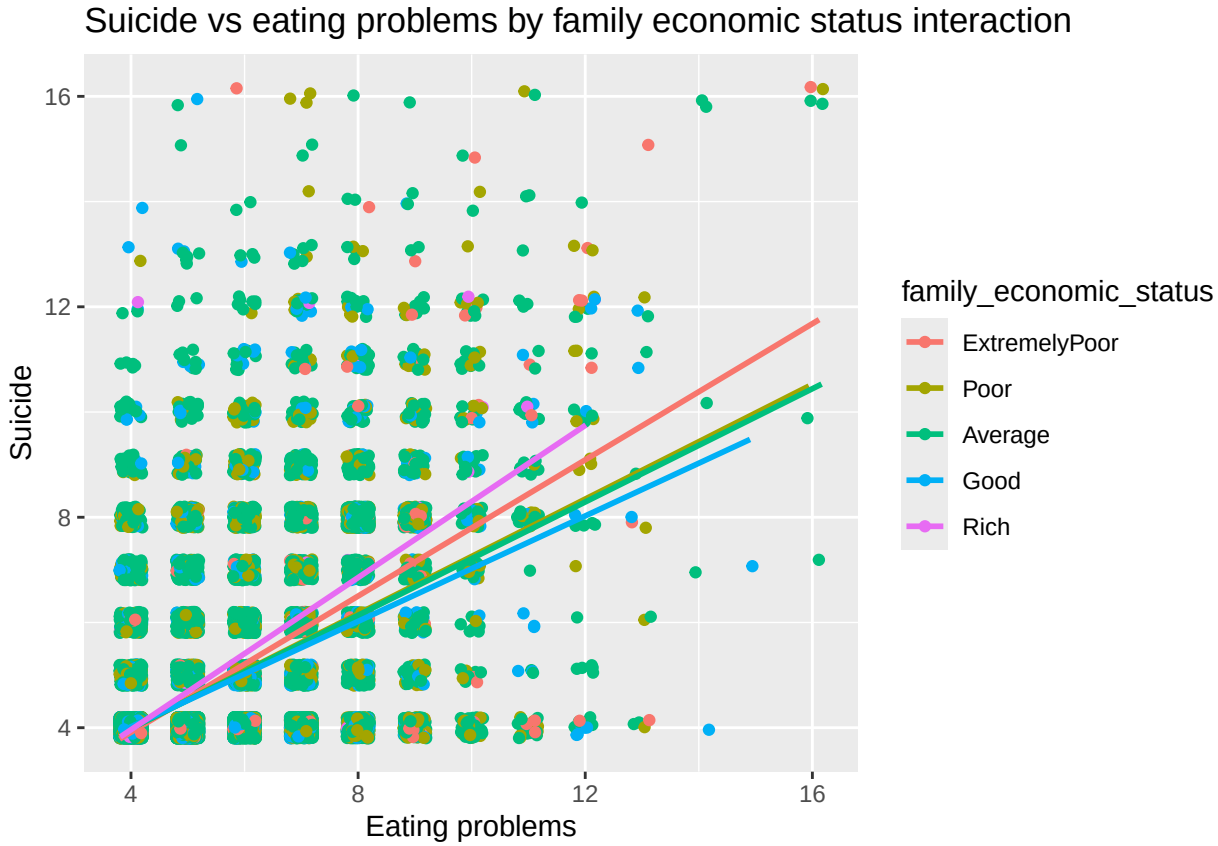
Just for a better understanding of the possible interaction between the variables, some scatter plots are drawn to check if some evident patterns can be useful for model development. For this purpose, the response variable is considered as a numeric variable and a line is fitted to highlight the linear relation between variables divided by categories.

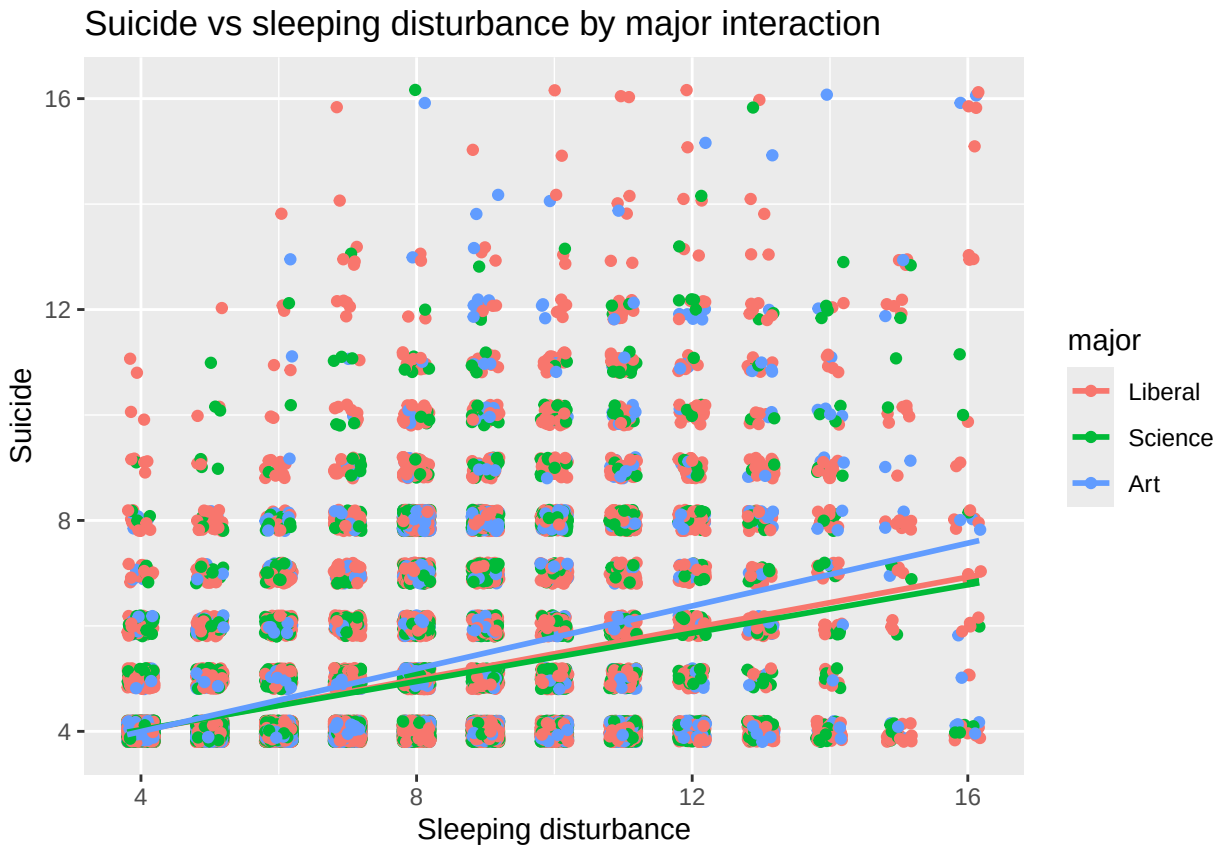
In particular, the following interactions are considered:

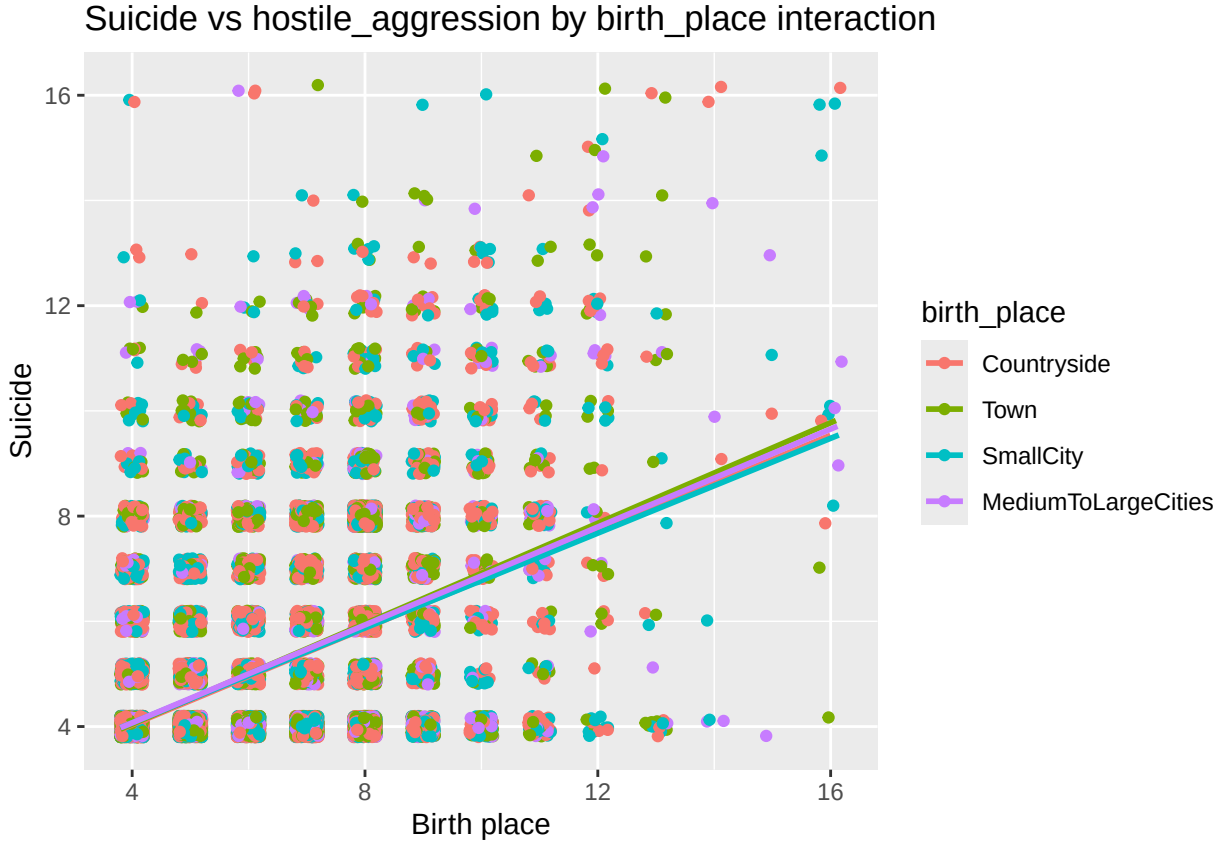
- **Psychiatric symptoms and gender**: It's important to inspect these variables because psychiatric conditions often manifest differently across genders, potentially leading to gender-specific patterns in symptoms and treatment outcomes.
- **Eating problems and family economic status**: Economic factors can affect access to resources and support for managing eating problems, influencing the severity and consequences of these issues in different socioeconomic groups.
- **Sleeping disturbance and major**: Different majors may have varying stress levels and workload demands, which can interact with sleep disturbances to affect academic performance and mental health uniquely in each field of study.
- **Hostile aggression and birth place**: Cultural norms around aggression and conflict resolution can vary significantly by location, impacting how hostile aggression influences social relationships and legal issues in different birthplaces.

```
df$suicide <- numeric_suicide
jitter_val <- 1
```


Suicide vs psychiatric symptoms by gender interaction
Suicide
16
12
8
4
4
8
12
16
Psychiatric symptoms
gender
Female
Male


Suicide vs eating problems by family economic status interaction
Suicide
16
12
8
4
4
8
12
16
Eating problems
family_economic_status
ExtremelyPoor
Poor
Average
Good
Rich






```
df$suicide <- factor_suicide
```

For these data, it seems that just the interaction between *eating problems and family economic status* and *sleeping disturbance and major* is relevant.

Models development

In a statistical analysis, splitting data into training and testing sets is crucial for model development. This process allows building the model using the training set and then assessing its quality on the unseen data at the very late stage using the test set. The hyperparameter tuning, where required, and the performance metrics are instead achieved by cross-validation on the training partition. This is needed to check if the model generalizes well to new data and provides an unbiased estimate of its predictive accuracy.

The reference level for the categorical predictors in a full model is a female student with brothers/sisters, from the countryside with an extremely poor family, liberal major, and postgraduate.

```
idx <- sample(nrow(df), 0.8 * nrow(df))
df_train <- df[idx, ]
df_test <- df[-idx, ]

print(paste0("Train size: ", nrow(df_train)))
```

```
## [1] "Train size: 21611"
```

```
print(paste0("Test size: ", nrow(df_test)))
```

```
## [1] "Test size: 5403"
```

```
# Reference levels
for (i in general_cols) {
  print(paste0(i, ": ", levels(df[, i])[1], sep = ""))
}
```

```
## [1] "gender: Female"
## [1] "whether_only_child: No"
## [1] "birth_place: Countryside"
## [1] "family_economic_status: ExtremelyPoor"
## [1] "major: Liberal"
## [1] "grade: Postgraduate"
```

```
# Class frequency in the training set
print(table(df_train$suicide))
```

```
##
## FALSE  TRUE
## 21203   408
```

```
# Class proportion in the training set
print(table(df_train$suicide) / nrow(df_train) * 100)
```

```
##
##     FALSE      TRUE
## 98.112073  1.887927
```

```
# Class frequency in the test set
print(table(df_test$suicide))
```

```
##
## FALSE  TRUE
##  5311    92
```

```r
# Class proportion in the test set
print(table(df_test$suicide) / nrow(df_test) * 100)
```

```
##
##     FALSE      TRUE
## 98.297242  1.702758
```

Logistic regression

Logistic regression is used for modeling directly the probability that the response variable belongs to a particular class. The coefficients of the generalized linear models in this case represent the log-odds change of the outcome (suicide) for a one-unit increase in the predictor variable, holding all other variables constant. A positive coefficient indicates an increase in the log-odds of suicide, while a negative coefficient means a decrease.

Full model

```r
mod_full_std <- glm(suicide ~ ., data = df_train, family = "binomial")
summary(mod_full_std)
```

```
##
## Call:
## glm(formula = suicide ~ ., family = "binomial", data = df_train)
##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   -11.88714    0.49495 -24.017  < 2e-16 ***
## genderMale                     -0.25844    0.13309  -1.942 0.052164 .
## whether_only_childYes           0.14780    0.13181   1.121 0.262145
## birth_placeTown                 0.32325    0.15355   2.105 0.035281 *
## birth_placeSmallCity            0.17414    0.16796   1.037 0.299819
## birth_placeMediumToLargeCities  0.42616    0.20867   2.042 0.041128 *
## family_economic_statusPoor     -0.92264    0.31047  -2.972 0.002961 **
## family_economic_statusAverage  -0.84997    0.29010  -2.930 0.003390 **
## family_economic_statusGood     -0.81134    0.32844  -2.470 0.013501 *
## family_economic_statusRich      0.56664    0.65252   0.868 0.385184
## majorScience                   -0.21291    0.14863  -1.432 0.152014
## majorArt                       -0.22460    0.16935  -1.326 0.184762
## gradeUndergraduateGradeFive    -0.49144    1.33470  -0.368 0.712725
## gradeJunior                    -0.13454    0.23724  -0.567 0.570649
## gradeSophomore                  0.28366    0.21338   1.329 0.183737
## gradeFreshman                   0.95035    0.20603   4.613 3.97e-06 ***
## gradeSenior                     0.35933    0.22045   1.630 0.103109
## psychiatric_symptoms            0.31778    0.03260   9.748  < 2e-16 ***
## dependence                     -0.13880    0.03671  -3.781 0.000156 ***
## impulsivity                     0.08895    0.04095   2.172 0.029856 *
## compulsion                      0.10920    0.03547   3.079 0.002080 **
## sleeping_disturbance            0.22632    0.03015   7.506 6.11e-14 ***
## internet_addiction              0.05235    0.02444   2.142 0.032182 *
## hostile_aggression              0.10354    0.03841   2.696 0.007026 **
## self_injury_behaviors           0.31238    0.03712   8.415  < 2e-16 ***
## eating_problems                 0.08033    0.04221   1.903 0.057012 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4047.5  on 21610  degrees of freedom
```

```
## Residual deviance: 2576.1  on 21585  degrees of freedom
## AIC: 2628.1
##
## Number of Fisher Scoring iterations: 8
```

Examining the model that includes all the variables, several predictors appear to be strongly significant. Notably, the intercept, which corresponds to the reference level and numeric predictors with coefficients of 0, has a log-odds value of -11.8871419. When converted, this log-odds value represents a probability of 6.8783277×10^{-6} for the response variable $\mathcal{P}(suicide) = \frac{e^{\beta_0}}{1+e^{\beta_0}}$.

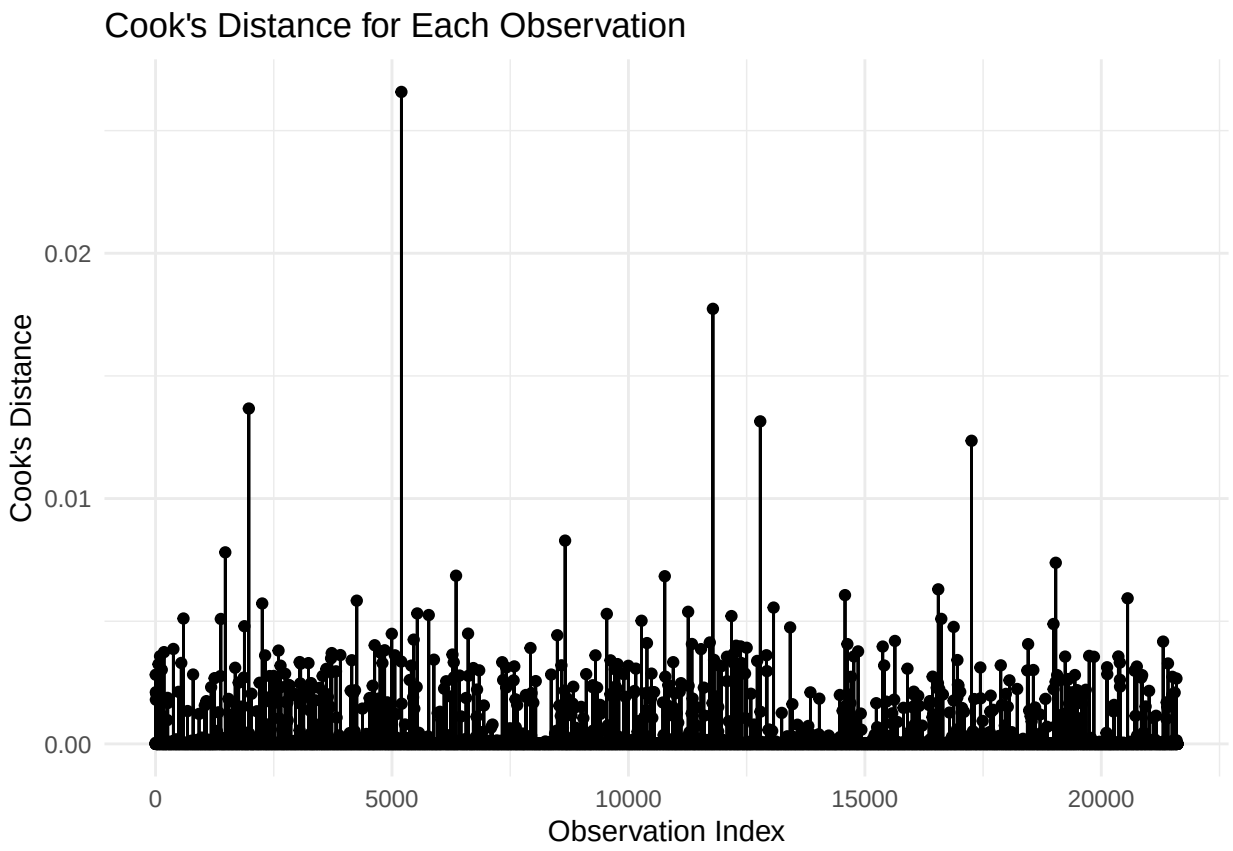
In particular, the following model interpretation is considered:

- Being male decreases the log-odds of suicide compared to being female. This suggests that males are less likely to report suicidal tendencies than females in this dataset
- Students from towns, small cities, and medium to large cities have higher log-odds of suicide compared to students from rural areas. This indicates that urbanization might be associated with higher suicidal tendencies.
- Being in a family with poor, average, or good economic status leads to a large decrease in the log-odds compared with the base level, instead being in a rich family seems to be not significant in this model
- Freshmen grade has the highest value for the coefficient and it's strongly significant.
- All the symptom predictors are statistically significant or strongly significant except for *eating_problems* in this model. Higher levels of psychiatric symptoms, impulsivity, compulsion, sleeping disturbances, internet addiction, hostile aggression, and self-injury behaviors are associated with increased log-odds of suicide. Dependence, however, shows a negative relationship, suggesting that higher dependence decreases the log-odds of suicide for these data.
- The levels in whether only child and in major have quite high p-values for the significance test and they seem not influential

```r
vif(mod_full_std)
```

```
##                            GVIF Df GVIF^(1/(2*Df))
## gender                 1.166777  1        1.080174
## whether_only_child     1.163804  1        1.078798
## birth_place            1.445289  3        1.063308
## family_economic_status 1.370377  4        1.040172
## major                  1.179966  2        1.042239
## grade                  1.192608  5        1.017770
## psychiatric_symptoms   1.541675  1        1.241642
## dependence             1.769891  1        1.330372
## impulsivity            2.096339  1        1.447874
## compulsion             2.007265  1        1.416780
## sleeping_disturbance   1.479705  1        1.216431
## internet_addiction     1.536539  1        1.239572
## hostile_aggression     1.802896  1        1.342719
## self_injury_behaviors  1.764043  1        1.328173
## eating_problems        1.795025  1        1.339785
```

```r
cook_dist <- cooks.distance(mod_full_std)
# Plot using ggplot2
ggplot(
  data.frame(Index = 1:length(cook_dist), Cook = cook_dist),
  aes(x = Index, y = Cook)
) +
  geom_point() +
  geom_line() +
  labs(x = "Observation Index", y = "Cook's Distance") +
  ggtitle("Cook's Distance for Each Observation") +
  theme_minimal()
```



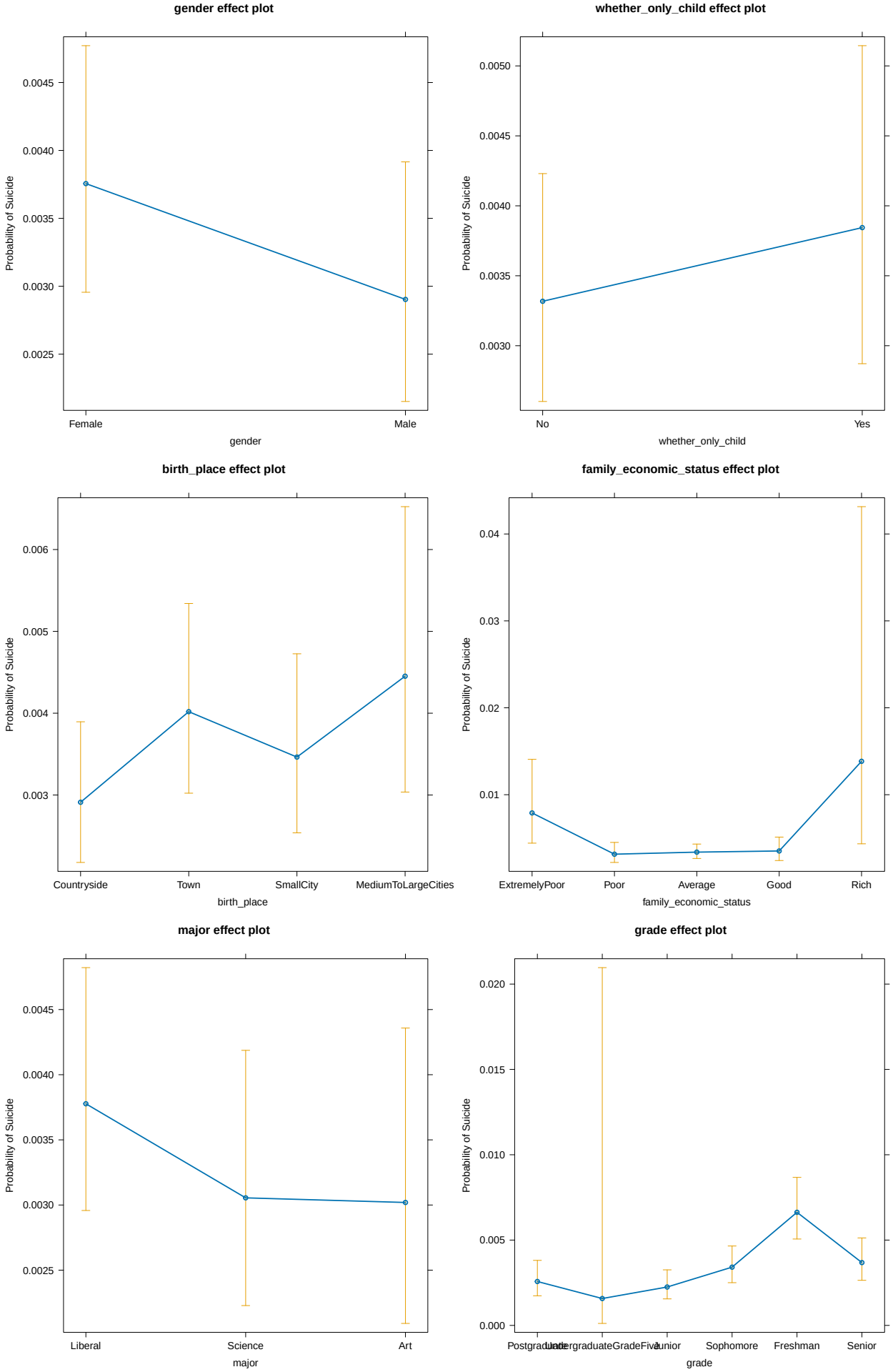

```
cook_dist <- NULL
```

The model doesn't show significant signs of multicollinearity or influential points

The above interpretation can be also visualized by checking the effect plots, paying attention to the fact that now the y-axis is represented in the response scale (not the log-odds one)

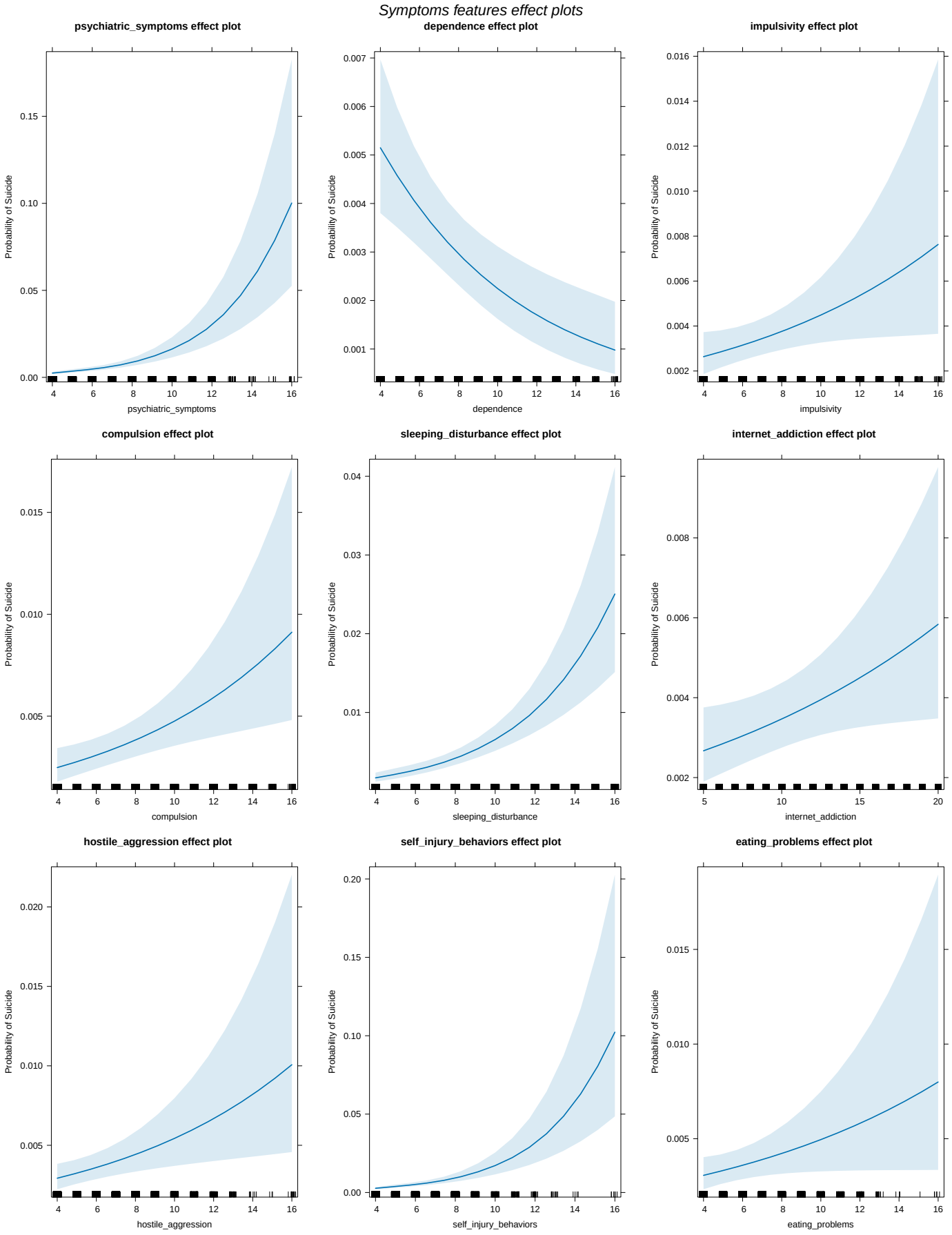
```
plts <- list()
for (i in general_cols) {
  p <- plot(effect(i, mod_full_std), ylab = "Probability of Suicide", rescale.axis = FALSE)
  plts[[i]] <- p
}
grid.arrange(grobs = plts, top = textGrob("General features effect plots", gp = gpar(fontsize = 20, font = 3)))
```

General features effect plots



```r
plts <- NULL
```

```r
plts <- list()
for (i in symptoms_cols) {
  p <- plot(effect(i, mod_full_std), ylab = "Probability of Suicide", rescale.axis = FALSE)
  plts[[i]] <- p
}
grid.arrange(grobs = plts, top = textGrob("Symptoms features effect plots", gp = gpar(fontsize = 20, font = 3)))
```




```
plts <- NULL
```

Models fitted using the standard Maximum likelihood method where the proportion of classes in the response variable is

unbalanced could not be optimal. Specifically, to overcome this problem, bias-corrected maximum likelihood[2] technique is utilized for estimating the coefficients of the logistic regression models robustly.

```
mod_zero_br <- glm(suicide ~ 1, data = df_train, family = "binomial", method = "brglmFit")
mod_full_br <- glm(suicide ~ ., data = df_train, family = "binomial", method = "brglmFit")
```

```
tab <- cbind(coef(mod_full_std), coef(mod_full_br), abs(coef(mod_full_std) - coef(mod_full_br)))
colnames(tab) <- c("Vanilla Model", "Full Model", "Abs diff.")
tab
```

```
##                                Vanilla Model   Full Model    Abs diff.
## (Intercept)                     -11.88714188 -11.78422672 0.1029151592
## genderMale                       -0.25843546  -0.25101351 0.0074219431
## whether_only_childYes             0.14779891   0.14746706 0.0003318491
## birth_placeTown                   0.32324920   0.32132514 0.0019240577
## birth_placeSmallCity              0.17413954   0.17335934 0.0007801917
## birth_placeMediumToLargeCities    0.42615507   0.42431977 0.0018352947
## family_economic_statusPoor       -0.92263892  -0.91883300 0.0038059149
## family_economic_statusAverage    -0.84997470  -0.85129266 0.0013179549
## family_economic_statusGood       -0.81133535  -0.80586527 0.0054700787
## family_economic_statusRich        0.56664010   0.61852762 0.0518875175
## majorScience                     -0.21290660  -0.20868792 0.0042186796
## majorArt                         -0.22459513  -0.21601445 0.0085806803
## gradeUndergraduateGradeFive      -0.49143602  -0.10885594 0.3825800793
## gradeJunior                      -0.13453561  -0.13417801 0.0003576007
## gradeSophomore                    0.28365928   0.27793153 0.0057277431
## gradeFreshman                     0.95034874   0.93766240 0.0126863389
## gradeSenior                       0.35933304   0.35333144 0.0060015998
## psychiatric_symptoms              0.31778235   0.31427988 0.0035024694
## dependence                       -0.13880022  -0.13791081 0.0008894104
## impulsivity                       0.08894817   0.08930182 0.0003536538
## compulsion                        0.10920039   0.10874986 0.0004505231
## sleeping_disturbance              0.22632042   0.22584999 0.0004704265
## internet_addiction                0.05235154   0.05180719 0.0005443460
## hostile_aggression                0.10354394   0.10197755 0.0015663910
## self_injury_behaviors             0.31238034   0.30900233 0.0033780037
## eating_problems                   0.08033333   0.07937898 0.0009543492
```

```
tab <- NULL
```

Models present slight differences in the coefficient estimations in this case, but due to the better theoretical properties of the bias reduction methods for logistic regression, this one is utilized for the analysis.

Manual selection

Given the available information, an attempt to develop a model manually is proposed. Specifically, the first model contains some significant variables present in the full model, the previously checked interaction terms that seem influential and their main effects, impulsivity feature that in the corr matrix seems the most correlated variable. The second model instead contains some possible variables that appear more discriminative

```
mod_manual1_interaction <- glm(
  suicide ~ impulsivity + self_injury_behaviors + family_economic_status +
    eating_problems + major + sleeping_disturbance +
    family_economic_status:eating_problems + major:sleeping_disturbance,
  data = df_train,
  family = "binomial",
  method = "brglmFit"
)
summary(mod_manual1_interaction)
```

[2]Kosmidis, I., Kenne Pagui, E.C. & Sartori, N. Mean and median bias reduction in generalized linear models. Stat Comput 30, 43–59 (2020)

```
##
## Call:
## glm(formula = suicide ~ impulsivity + self_injury_behaviors +
##     family_economic_status + eating_problems + major + sleeping_disturbance +
##     family_economic_status:eating_problems + major:sleeping_disturbance,
##     family = "binomial", data = df_train, method = "brglmFit")
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1923  -0.1599  -0.0859  -0.0486   3.7784
##
## Coefficients:
##                                           Estimate Std. Error z value
## (Intercept)                              -14.39739    1.48046  -9.725
## impulsivity                                0.24407    0.03029   8.059
## self_injury_behaviors                      0.42657    0.03303  12.915
## family_economic_statusPoor                 2.12351    1.55237   1.368
## family_economic_statusAverage              2.95201    1.46538   2.014
## family_economic_statusGood                 3.32255    1.53432   2.165
## family_economic_statusRich                 6.61449    2.18961   3.021
## eating_problems                            0.52720    0.16314   3.232
## majorScience                              -1.14733    0.60790  -1.887
## majorArt                                  -0.47606    0.69608  -0.684
## sleeping_disturbance                       0.27544    0.03177   8.671
## family_economic_statusPoor:eating_problems -0.34253    0.17967  -1.906
## family_economic_statusAverage:eating_problems -0.42678  0.16793  -2.541
## family_economic_statusGood:eating_problems -0.44533    0.17872  -2.492
## family_economic_statusRich:eating_problems -0.69967    0.27056  -2.586
## majorScience:sleeping_disturbance          0.08271    0.05899   1.402
## majorArt:sleeping_disturbance              0.02680    0.06724   0.399
##                                           Pr(>|z|)
## (Intercept)                               < 2e-16 ***
## impulsivity                               7.71e-16 ***
## self_injury_behaviors                     < 2e-16 ***
## family_economic_statusPoor                0.17134
## family_economic_statusAverage             0.04396 *
## family_economic_statusGood                0.03035 *
## family_economic_statusRich                0.00252 **
## eating_problems                           0.00123 **
## majorScience                              0.05911 .
## majorArt                                  0.49402
## sleeping_disturbance                      < 2e-16 ***
## family_economic_statusPoor:eating_problems 0.05659 .
## family_economic_statusAverage:eating_problems 0.01104 *
## family_economic_statusGood:eating_problems 0.01271 *
## family_economic_statusRich:eating_problems 0.00971 **
## majorScience:sleeping_disturbance         0.16086
## majorArt:sleeping_disturbance             0.69016
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4047.5  on 21610  degrees of freedom
## Residual deviance: 2804.6  on 21594  degrees of freedom
## AIC:  2838.6
##
## Type of estimator: AS_mixed (mixed bias-reducing adjusted score equations)
## Number of Fisher Scoring iterations: 4
```

```r
mod_manual2 <- glm(
  suicide ~ gender + grade + birth_place + impulsivity + self_injury_behaviors +
    hostile_aggression + psychiatric_symptoms,
  data = df_train,
  family = "binomial",
  method = "brglmFit"
)
summary(mod_manual2)
```

```
##
## Call:
## glm(formula = suicide ~ gender + grade + birth_place + impulsivity +
##     self_injury_behaviors + hostile_aggression + psychiatric_symptoms,
##     family = "binomial", data = df_train, method = "brglmFit")
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9921  -0.1420  -0.0773  -0.0512   3.6835
##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   -11.54813    0.35166 -32.839  < 2e-16 ***
## genderMale                     -0.34388    0.12258  -2.805  0.00503 **
## gradeUndergraduateGradeFive     0.28595    1.19355   0.240  0.81066
## gradeJunior                    -0.16172    0.22980  -0.704  0.48160
## gradeSophomore                  0.33190    0.20608   1.611  0.10729
## gradeFreshman                   0.82554    0.19941   4.140 3.48e-05 ***
## gradeSenior                     0.32628    0.21396   1.525  0.12726
## birth_placeTown                 0.35721    0.14303   2.497  0.01251 *
## birth_placeSmallCity            0.26185    0.14966   1.750  0.08019 .
## birth_placeMediumToLargeCities  0.58237    0.18144   3.210  0.00133 **
## impulsivity                     0.19707    0.03336   5.908 3.47e-09 ***
## self_injury_behaviors           0.32107    0.03246   9.890  < 2e-16 ***
## hostile_aggression              0.16213    0.03815   4.250 2.14e-05 ***
## psychiatric_symptoms            0.38375    0.03112  12.331  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4047.5  on 21610  degrees of freedom
## Residual deviance: 2707.9  on 21597  degrees of freedom
## AIC:  2735.9
##
## Type of estimator: AS_mixed (mixed bias-reducing adjusted score equations)
## Number of Fisher Scoring iterations: 7
```

```r
print(BIC(mod_full_br, mod_manual1_interaction, mod_manual2))
```

```
##                           df      BIC
## mod_full_br               26 2836.049
## mod_manual1_interaction   17 2974.259
## mod_manual2               14 2847.662
```

Including the interaction terms and the variables considered in the scatter plots seems to lead to no benefit to the model because they appear not statistically significant based on these data. Also the BIC score, which penalizes more complex models, is higher compared with the GLM with all the predictors. The second manual has all significant predictors, except for some levels in the variable *grade*. The information criteria is higher also in this case to the full model.

Stepwise regression

Bidirectional Stepwise Regression, is a variable selection method that combines both forward and backward stepwise regression approaches. It iteratively adds and removes predictors based on a predefined information criterion until no further improvements are observed. This method provides a balance between computational efficiency and model selection performance, making it suitable for situations where automated variable selection is desired while controlling for overfitting.

```
# AIC stepwise regression
mod_step_aic <- stepAIC(mod_zero_br, direction = "both", scope = list(upper = formula(mod_full_br)), trace = 0)
summary(mod_step_aic)
```

```
##
## Call:
## glm(formula = suicide ~ psychiatric_symptoms + sleeping_disturbance +
##     self_injury_behaviors + grade + hostile_aggression + compulsion +
##     family_economic_status + gender + dependence + internet_addiction +
##     impulsivity + birth_place + eating_problems, family = "binomial",
##     data = df_train, method = "brglmFit")
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.1975  -0.1393  -0.0707  -0.0418   3.7832
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    -11.854298   0.486934 -24.345  < 2e-16 ***
## psychiatric_symptoms             0.314522   0.032283   9.743  < 2e-16 ***
## sleeping_disturbance             0.223959   0.029687   7.544 4.56e-14 ***
## self_injury_behaviors            0.306755   0.036771   8.342  < 2e-16 ***
## gradeUndergraduateGradeFive      0.007164   1.193119   0.006 0.995209
## gradeJunior                     -0.140008   0.233909  -0.599 0.549467
## gradeSophomore                   0.273813   0.210699   1.300 0.193758
## gradeFreshman                    0.929230   0.203329   4.570 4.88e-06 ***
## gradeSenior                      0.354041   0.217864   1.625 0.104151
## hostile_aggression               0.103006   0.038105   2.703 0.006867 **
## compulsion                       0.113252   0.034981   3.238 0.001206 **
## family_economic_statusPoor      -0.905250   0.307873  -2.940 0.003279 **
## family_economic_statusAverage   -0.826819   0.287816  -2.873 0.004069 **
## family_economic_statusGood      -0.753687   0.324900  -2.320 0.020354 *
## family_economic_statusRich       0.707106   0.635322   1.113 0.265714
## genderMale                      -0.299166   0.124684  -2.399 0.016422 *
## dependence                      -0.139456   0.036370  -3.834 0.000126 ***
## internet_addiction               0.054705   0.024039   2.276 0.022867 *
## impulsivity                      0.087175   0.040495   2.153 0.031340 *
## birth_placeTown                  0.352781   0.149518   2.359 0.018302 *
## birth_placeSmallCity             0.210775   0.160337   1.315 0.188654
## birth_placeMediumToLargeCities   0.481534   0.199897   2.409 0.016000 *
## eating_problems                  0.075901   0.041863   1.813 0.069820 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4047.5  on 21610  degrees of freedom
## Residual deviance: 2580.8  on 21588  degrees of freedom
## AIC:  2626.8
##
## Type of estimator: AS_mixed (mixed bias-reducing adjusted score equations)
## Number of Fisher Scoring iterations: 5
```

The logistic regression model identifies several significant predictors of suicide among students. Psychiatric symptoms, sleeping disturbance, self-injury behaviors, compulsion, grade (sophomores, freshmen and senior), hostile aggression, birth

places, gender, family economic status, dependence and impulsivity show statistically significant associations with the likelihood of suicide. Notably, the predictors associated with a lower likelihood of suicide with respect to the reference level are gender, the other family economic conditions (except *Rich*) and dependence. The model seems to have a good explainable power as indicated the the slight reduction of the AIC score with respect to the full model, however, the predictors are still a bit (12 features selected vs 15 total features).

It's possible to try stepwise regression using BIC penalty to select a more parsimonious model.

```
# BIC stepwise regression
mod_step_bic <- stepAIC(mod_zero_br, direction = "both", scope = list(upper = formula(mod_full_br)), trace = 0,
summary(mod_step_bic)
```

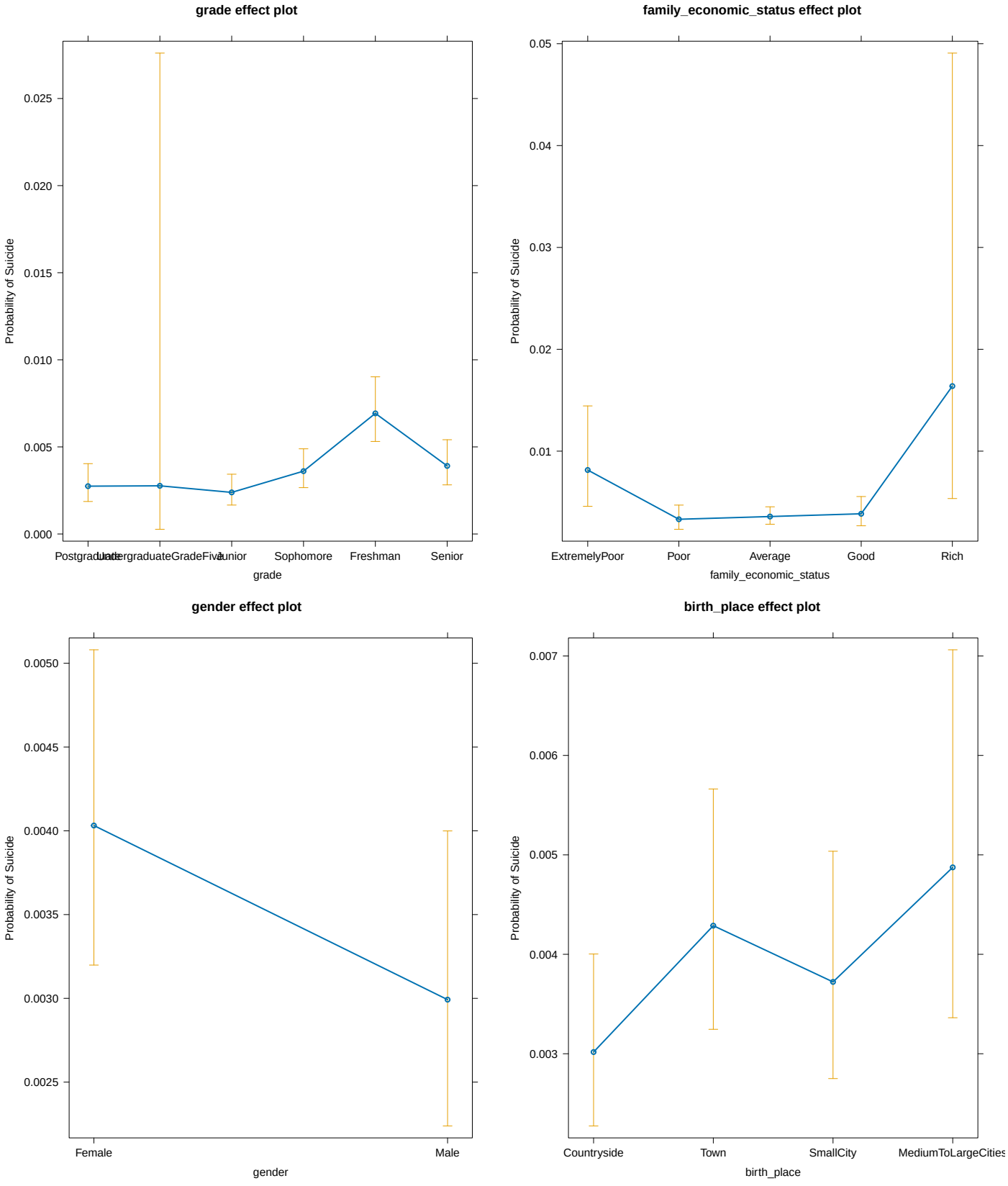
```
##
## Call:
## glm(formula = suicide ~ psychiatric_symptoms + sleeping_disturbance +
##     self_injury_behaviors + compulsion + hostile_aggression,
##     family = "binomial", data = df_train, method = "brglmFit")
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7750  -0.1437  -0.0737  -0.0453   3.7895
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -11.93186    0.32621 -36.577  < 2e-16 ***
## psychiatric_symptoms    0.32888    0.03215  10.229  < 2e-16 ***
## sleeping_disturbance    0.23490    0.02854   8.231  < 2e-16 ***
## self_injury_behaviors   0.29668    0.03205   9.257  < 2e-16 ***
## compulsion              0.14336    0.03102   4.621 3.82e-06 ***
## hostile_aggression      0.11876    0.03540   3.355 0.000794 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4047.5  on 21610  degrees of freedom
## Residual deviance: 2682.4  on 21605  degrees of freedom
## AIC:  2694.4
##
## Type of estimator: AS_mixed (mixed bias-reducing adjusted score equations)
## Number of Fisher Scoring iterations: 3
```

Both models consistently identify psychiatric symptoms, sleeping disturbance, self-injury behaviors, compulsion and hostile aggression as significant predictors and here there's strong significant evidence on all. The signs of the estimated coefficients are the same for both. The choice between the models depends on the preference for model complexity versus simplicity and the specific goals of the analysis.

For completeness, effect plots from the stepwise regression model with AIC are provided:

```
plts <- list()
for (i in extract_predictors_in_vec(mod_step_aic$formula, general_cols)) {
  p <- plot(effect(i, mod_step_aic), ylab = "Probability of Suicide", rescale.axis = FALSE)
  plts[[i]] <- p
}
grid.arrange(grobs = plts, top = textGrob("General features effect plots", gp = gpar(fontsize = 20, font = 3)))
```

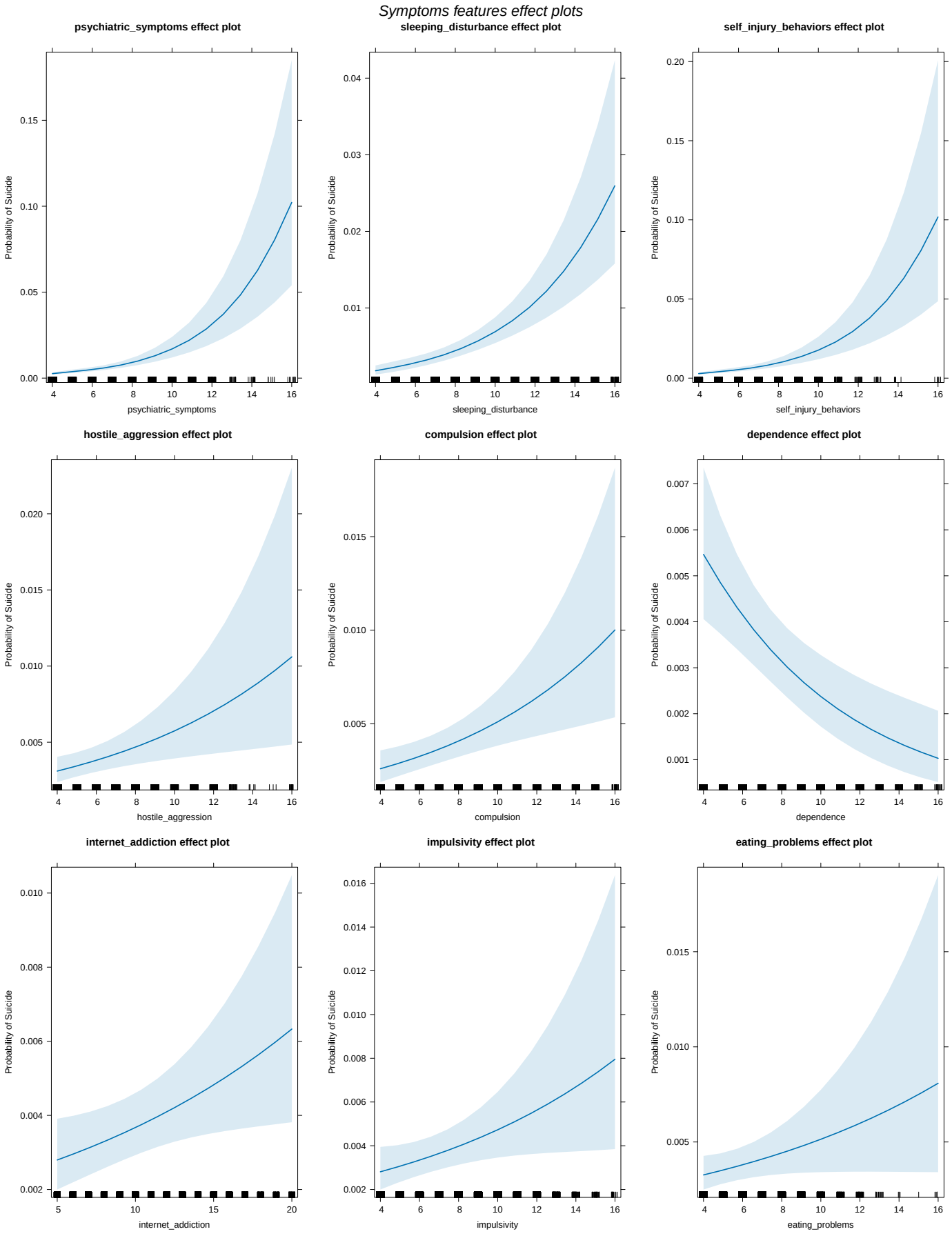
General features effect plots




```
plts <- NULL
```

```
plts <- list()
for (i in extract_predictors_in_vec(mod_step_aic$formula, symptoms_cols)) {
  p <- plot(effect(i, mod_step_aic), ylab = "Probability of Suicide", rescale.axis = FALSE)
  plts[[i]] <- p
}
```

```
grid.arrange(grobs = plts, top = textGrob("Symptoms features effect plots", gp = gpar(fontsize = 20, font = 3)))
```



```
plts <- NULL
```

Lasso shrinkage method

Lasso technique introduces a penalty term to the standard regression objective function, encouraging sparsity by shrinking the coefficients of less important predictors toward zero. This regularization performs a feature selection, making it particularly useful when dealing with high-dimensional datasets with potentially correlated predictors.

Notice that the implementation of Lasso in `glmnetUtils` R library[3] deliberately avoids the usual treatment of factors with a reference level "absorbed" by the intercept. The model type, in this case, it's more interpretable because otherwise, the lasso shrinkage could cancel the differences in the factor levels of the predictors.

```
mod_lasso <- cv.glmnet(suicide ~ ., data = df_train, family = "binomial", method = "brglmFit", alpha = 1, nfolds
beta_1se <- coef(mod_lasso, s = mod_lasso$lambda.1se)
beta_1se_idx <- which(beta_1se[, 1] != 0)
print(cbind(beta_1se_idx, beta_1se[beta_1se_idx])[, 2])
```

```
##           (Intercept)         gradeFreshman  psychiatric_symptoms
##          -10.52493686            0.31518575            0.31815800
##           impulsivity            compulsion  sleeping_disturbance
##            0.01829906            0.07273130            0.17832650
##    hostile_aggression self_injury_behaviors       eating_problems
##            0.09211289            0.26204273            0.01289593
```

```
print(sum(beta_1se[, 1] != 0))
```

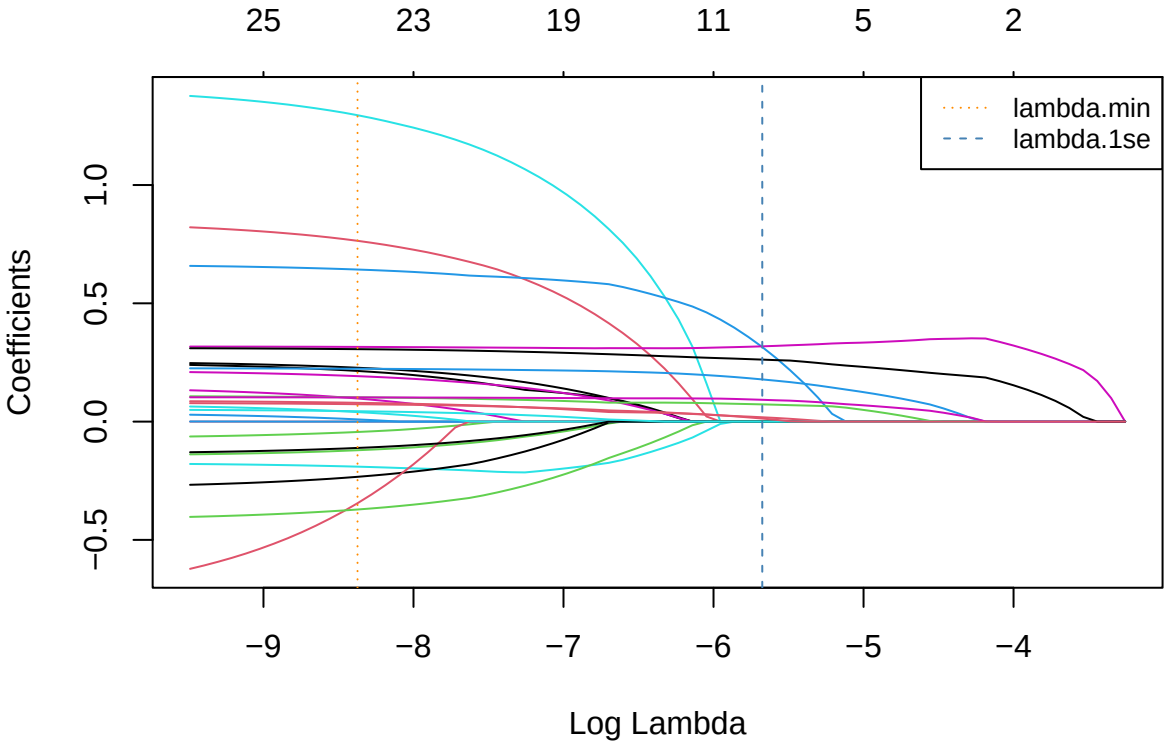
```
## [1] 9
```

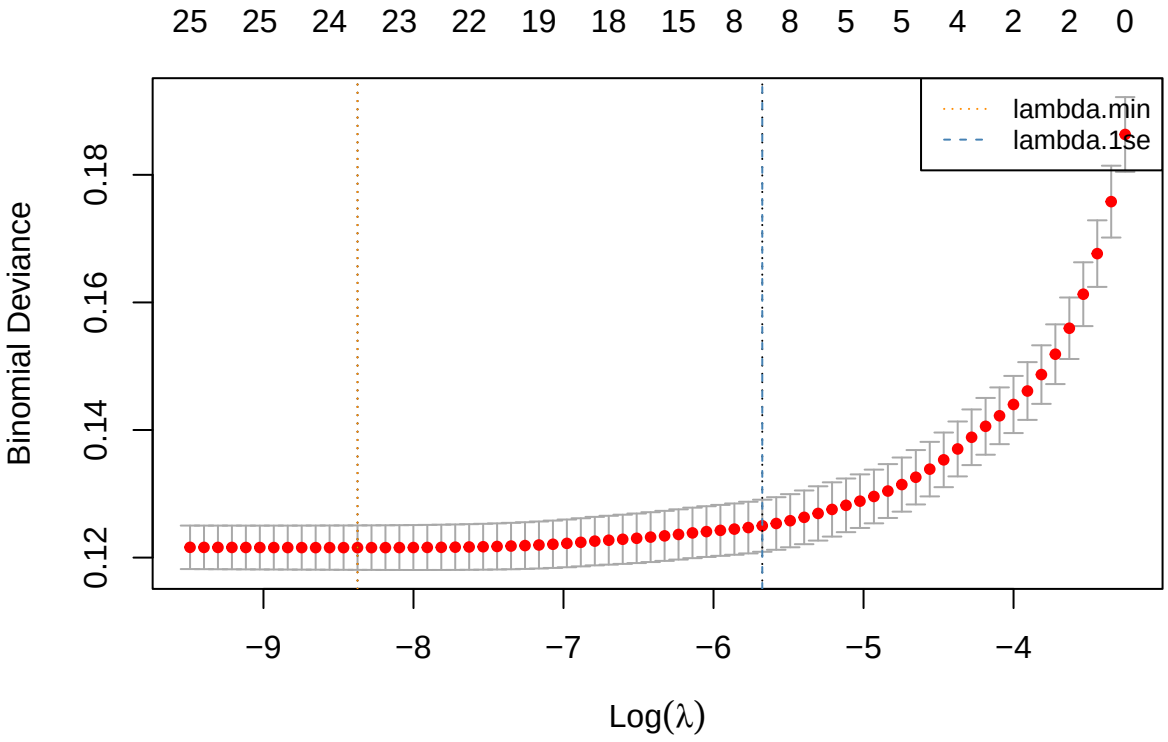
The fitted logistic regression model suggests several factors, all related to symptom variables and they contribute all positively to a higher risk of suicide.

If compared to the coefficients selected from stepwise regression, the magnitude of the common coefficients is usually smaller, indicating that the lasso model tends to shrink the coefficients towards zero as expected. In addition, BIC and Lasso models select the same set of predictors, even if the latter includes also *impulsivity* predictor.

The below plots show the coefficients shrinkage with respect to the log lambda value in the first, and the behavior of the binomial deviance with respect to this hyperparameter in the latter. The dotted line represents the log lambda value that minimizes the cross-validated error, while the dashed line represents the log lambda value that is within one standard error of the minimum. It's decided to take the model with the lambda value plus one standard error because it's more parsimonious and the difference in the binomial deviance is not so high concerning the minimum value.

[3] https://cran.r-project.org/web/packages/glmnetUtils/vignettes/intro.html





Generative models

After modeling directly the response variable by the logistic regression based models, it's possible to consider generative models that estimate the class-specific densities and then use Bayes' theorem to derive the posterior probabilities of the

classes. These types of models are more stable and robust when there's a substantial separation between the classes as probably is in this case.

To enhance interpretability for all the generative models, the preditors selected by stepwise regression with BIC are used in the model fitting.

Linear Discriminant Analysis (LDA)

The LDA model assumes that the predictors are normally distributed and that the covariance matrices are equal across classes.

```
lda_formula <- mod_step_bic$formula
mod_lda <- lda(lda_formula, data = df_train)
mod_lda
```

```
## Call:
## lda(lda_formula, data = df_train)
##
## Prior probabilities of groups:
##      FALSE       TRUE
## 0.98112073 0.01887927
##
## Group means:
##       psychiatric_symptoms sleeping_disturbance self_injury_behaviors
## FALSE             5.017781             7.087959              4.769136
## TRUE              8.416667            10.406863              7.468137
##       compulsion hostile_aggression
## FALSE   6.993774           5.545725
## TRUE   10.306373           8.475490
##
## Coefficients of linear discriminants:
##                              LD1
## psychiatric_symptoms  0.35907864
## sleeping_disturbance  0.06826491
## self_injury_behaviors 0.30571284
## compulsion            0.01788497
## hostile_aggression    0.01766586
```

The printed output of LDA includes the a-priori probabilities of suicide and the group means. The proportion of training observation that belongs to positive suicide variables is quite low. As previously observed different times, when the response variable is true and the predictors are the symptom variables, their mean is higher compared with the case in which the response variable is false.

Quadratic Discriminant Analysis (QDA)

The QDA model generalizes the LDA assuming different covariance matrices for each class, so the flexibility is increased but the number of parameters to estimate is much larger

```
qda_formula <- mod_step_bic$formula
mod_qda <- qda(qda_formula, data = df_train)
mod_qda
```

```
## Call:
## qda(qda_formula, data = df_train)
##
## Prior probabilities of groups:
##      FALSE       TRUE
## 0.98112073 0.01887927
##
```

```
## Group means:
##       psychiatric_symptoms sleeping_disturbance self_injury_behaviors
## FALSE             5.017781             7.087959              4.769136
## TRUE              8.416667            10.406863              7.468137
##       compulsion hostile_aggression
## FALSE   6.993774           5.545725
## TRUE   10.306373           8.475490
```

Naive Bayes

Naive Bayes instead assumes independence between the predictors to estimate the class-specific densities: this assumption usually introduces bias but reduces the variance component of the error. The standard naive Bayes classifier for estimating univariate class densities of quantitative predictors assumes Gaussian distribution (likelihood component in the Bayes' formula)

```
nb_formula <- mod_step_bic$formula
mod_nb <- naiveBayes(nb_formula, data = df_train)
mod_nb
```

```
##
## Naive Bayes Classifier for Discrete Predictors
##
## Call:
## naiveBayes.default(x = X, y = Y, laplace = laplace)
##
## A-priori probabilities:
## Y
##      FALSE       TRUE
## 0.98112073 0.01887927
##
## Conditional probabilities:
##        psychiatric_symptoms
## Y           [,1]     [,2]
##   FALSE 5.017781 1.533826
##   TRUE  8.416667 2.791337
##
##        sleeping_disturbance
## Y           [,1]     [,2]
##   FALSE  7.087959 2.398756
##   TRUE  10.406863 2.433620
##
##        self_injury_behaviors
## Y           [,1]     [,2]
##   FALSE 4.769136 1.335392
##   TRUE  7.468137 2.560130
##
##        compulsion
## Y           [,1]     [,2]
##   FALSE  6.993774 2.351817
##   TRUE  10.306373 2.399172
##
##        hostile_aggression
## Y           [,1]     [,2]
##   FALSE 5.545725 1.708450
##   TRUE  8.475490 2.585042
```

The output of Naive Bayes contains the estimated a-priori probabilities and a table containing the mean and standard deviation of the numeric variables, conditioned by the response.

Other models

Generalized Additive Model (GAM)

GAMs extend traditional linear models by allowing for non-linear relationships between predictors and the response variable through the use of smooth functions. This flexibility enables the model to capture complex patterns in the data without assuming a specific parametric form.

In particular, the function `gam` of the R package `mgcv` by default includes in the nonlinear model the smooth terms using thin plate regression splines[4], where the degree of smoothness is estimated as part of the fitting procedure.

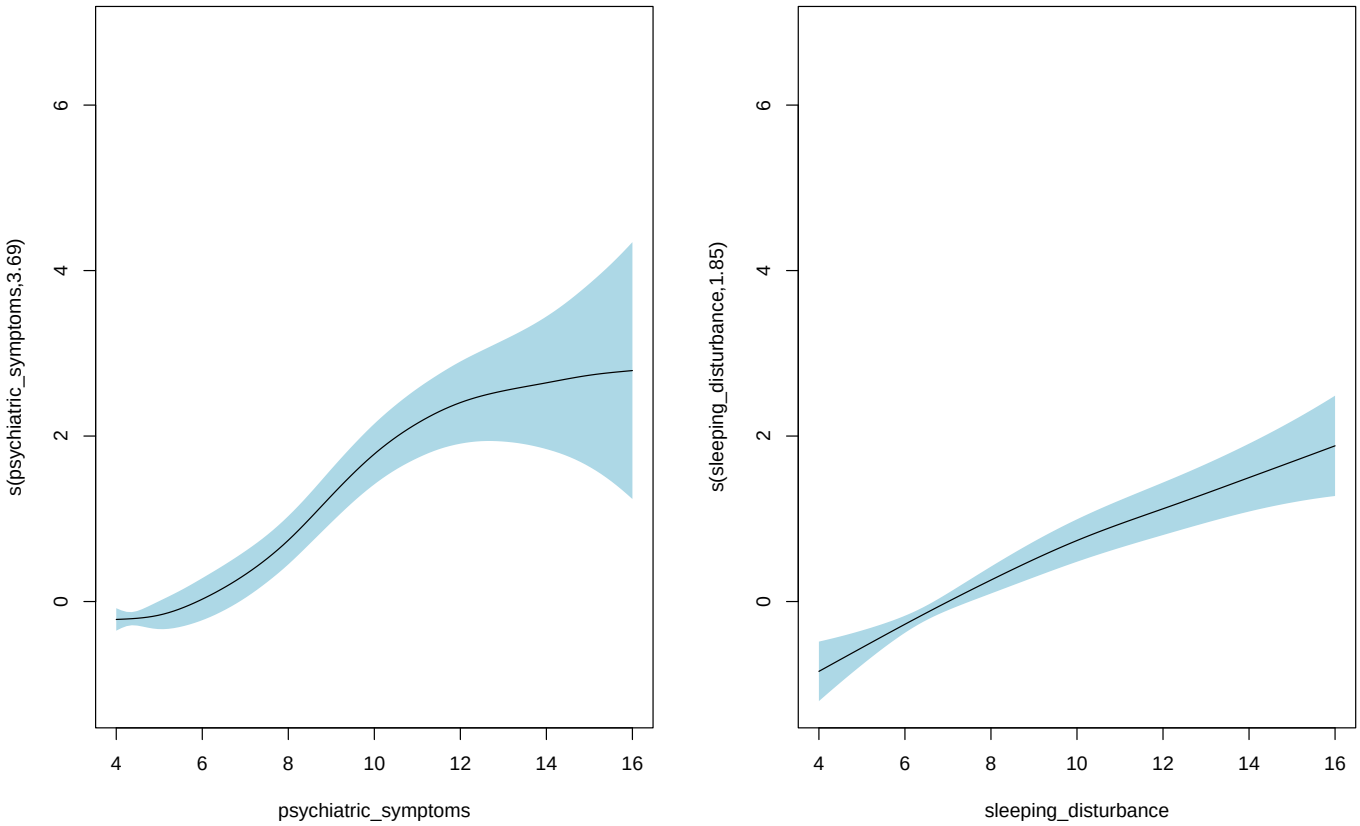
```
# extract only the predictors from the formula and model the numeric one as smooth terms
gam_formula_bic <- formula(paste(
  "suicide~",
  paste(extract_predictors_in_vec(mod_step_bic$formula, general_cols), collapse = " + "),
  " + ",
  paste(
    "s(",
    extract_predictors_in_vec(mod_step_bic$formula, symptoms_cols),
    ")",
    collapse = "+"
  )
))

mod_gam_bic <- gam(gam_formula_bic, family = "binomial", data = df_train)
summary(mod_gam_bic)
```

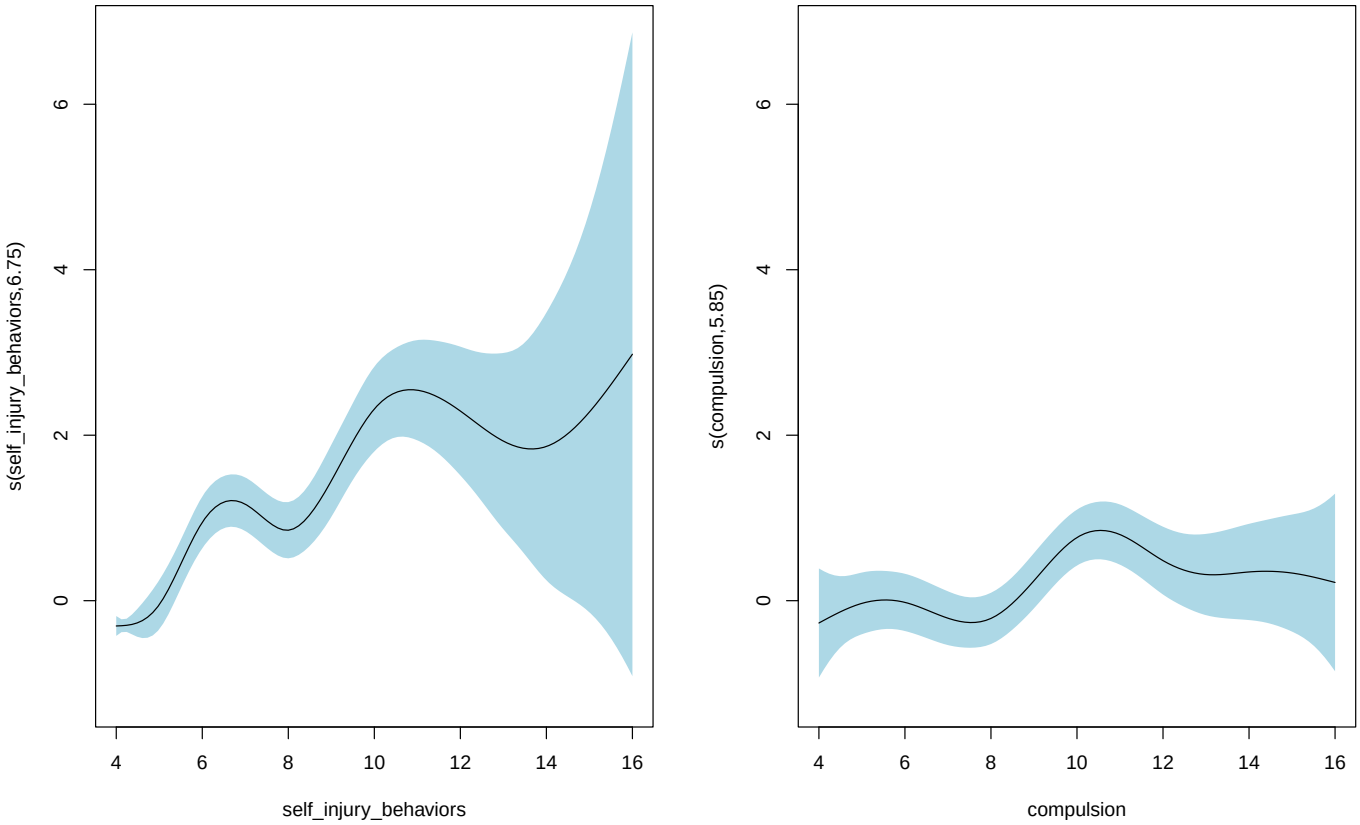
```
##
## Family: binomial
## Link function: logit
##
## Formula:
## suicide ~ +s(psychiatric_symptoms) + s(sleeping_disturbance) +
##     s(self_injury_behaviors) + s(compulsion) + s(hostile_aggression)
##
## Parametric coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -5.3666     0.1265  -42.42   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                            edf Ref.df Chi.sq p-value
## s(psychiatric_symptoms)  3.695  4.586 128.50  <2e-16 ***
## s(sleeping_disturbance)  1.849  2.364  57.69  <2e-16 ***
## s(self_injury_behaviors) 6.749  7.715 130.75  <2e-16 ***
## s(compulsion)            5.853  6.986  41.64  <2e-16 ***
## s(hostile_aggression)    7.243  8.061  39.21  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.234   Deviance explained = 37.3%
## UBRE = -0.88017  Scale est. = 1         n = 21611
```

All five smooth terms are considered statistically significant according to the p-value and the model explains 36.8% of the deviance. In addition, the estimated effective degrees of freedom (edf) propose complex fits for all the predictors, suggesting the presence of nonlinear effects in the data. Notice also that higher edf values suggest more flexibility in capturing non-linear effects.

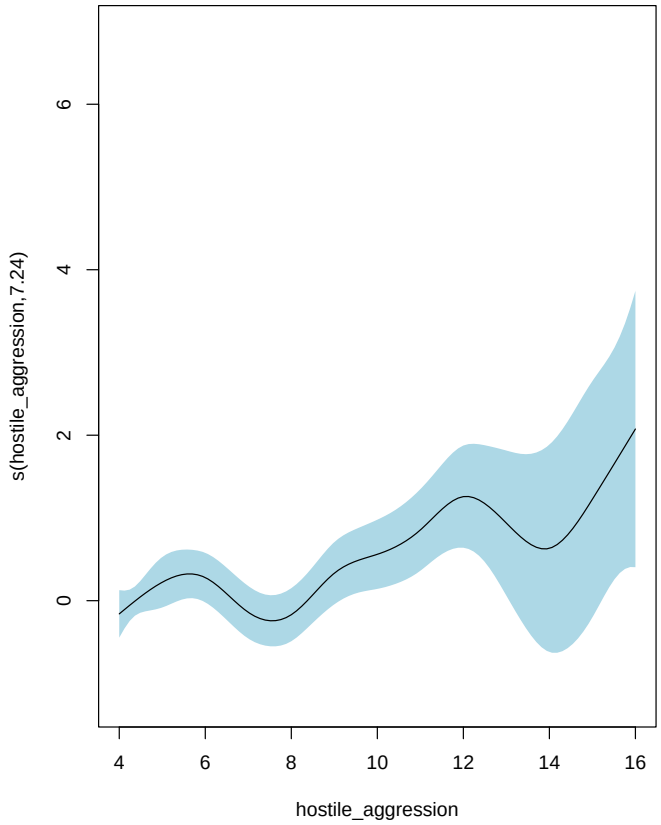
[4]Thin plate regression splines are described in detail in Wood (2003), *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(1), 95-114.

```r
par(mfrow = c(1, 2))
plot(mod_gam_bic, shade = TRUE, shade.col = "lightblue")
```






```
par(mfrow = c(1, 1))
```



It's decided to try including some additional predictors considered significant in other models.

```
mod_gam_manual <- update(mod_gam_bic, . ~ . + grade + birth_place + s(impulsivity))
summary(mod_gam_manual)
```

```
##
## Family: binomial
## Link function: logit
##
## Formula:
## suicide ~ s(psychiatric_symptoms) + s(sleeping_disturbance) +
##     s(self_injury_behaviors) + s(compulsion) + s(hostile_aggression) +
##     grade + birth_place + s(impulsivity)
##
## Parametric coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     -5.8862     0.2210 -26.639  < 2e-16 ***
## gradeUndergraduateGradeFive     -0.7616     1.5953  -0.477 0.633091
## gradeJunior                     -0.2420     0.2435  -0.994 0.320327
## gradeSophomore                   0.2283     0.2168   1.053 0.292392
## gradeFreshman                    0.7927     0.2088   3.796 0.000147 ***
## gradeSenior                      0.2794     0.2274   1.228 0.219338
## birth_placeTown                  0.3691     0.1502   2.457 0.014022 *
## birth_placeSmallCity             0.2175     0.1585   1.373 0.169855
## birth_placeMediumToLargeCities   0.6025     0.1910   3.155 0.001606 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                            edf Ref.df  Chi.sq  p-value
```

```
## s(psychiatric_symptoms)  3.619  4.497 116.696  < 2e-16 ***
## s(sleeping_disturbance)  1.965  2.516  65.675  < 2e-16 ***
## s(self_injury_behaviors) 6.686  7.658 138.592  < 2e-16 ***
## s(compulsion)            5.825  6.958  31.831 3.59e-05 ***
## s(hostile_aggression)    7.485  8.236  35.793 2.37e-05 ***
## s(impulsivity)           2.368  3.053   3.862    0.281
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.239   Deviance explained = 38.8%
## UBRE = -0.88193  Scale est. = 1         n = 21611
```

The *grade* categorical predictor presents some levels not statistically significant also in this model, instead the other terms seem to have an influence. The manual GAM model explains now the 38% of the deviance The nonlinear effect of *impulsivity* appears not statistically significant so it's decided to remove it.

```
mod_gam_manual <- update(mod_gam_manual, . ~ . - s(impulsivity))
summary(mod_gam_manual)
```

```
##
## Family: binomial
## Link function: logit
##
## Formula:
## suicide ~ s(psychiatric_symptoms) + s(sleeping_disturbance) +
##     s(self_injury_behaviors) + s(compulsion) + s(hostile_aggression) +
##     grade + birth_place
##
## Parametric coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    -5.8747     0.2200 -26.699  < 2e-16 ***
## gradeUndergraduateGradeFive    -0.8077     1.6342  -0.494 0.621143
## gradeJunior                    -0.2542     0.2437  -1.043 0.296778
## gradeSophomore                  0.2217     0.2167   1.023 0.306322
## gradeFreshman                   0.8014     0.2082   3.850 0.000118 ***
## gradeSenior                     0.2647     0.2274   1.164 0.244385
## birth_placeTown                 0.3618     0.1500   2.412 0.015867 *
## birth_placeSmallCity            0.2139     0.1579   1.354 0.175665
## birth_placeMediumToLargeCities  0.5909     0.1902   3.108 0.001886 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                            edf Ref.df Chi.sq  p-value
## s(psychiatric_symptoms)  3.706  4.597 119.23  < 2e-16 ***
## s(sleeping_disturbance)  1.600  2.016  66.94  < 2e-16 ***
## s(self_injury_behaviors) 6.720  7.687 139.01  < 2e-16 ***
## s(compulsion)            5.827  6.959  33.92 1.15e-05 ***
## s(hostile_aggression)    7.458  8.219  42.21  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.238   Deviance explained = 38.6%
## UBRE = -0.8819  Scale est. = 1         n = 21611
```

```
AIC(mod_gam_bic, mod_gam_manual)
```

```
##                       df      AIC
## mod_gam_bic    26.38890 2589.590
## mod_gam_manual 34.31007 2552.231
```

As can be observed the new manual-designed model seems to fit better the data due to the lower AIC score.

Model comparison

The performance of the models is evaluated using different metrics but in this analysis, the sensitivity (true positive rate) is considered crucial. Indeed, this score refers to the model's ability to correctly identify individuals at risk of suicide, and missing a positive case (i.e., failing to predict a suicide risk when one exists) can have severe and life-threatening consequences. The optimal threshold for the response variable prediction is determined by the point on the ROC curve that maximizes the Youden's J statistic, $max(sensitivity + specificity)$, which balances the true positive and false positive rates. As anticipated in the previous section, the validation set is used to assess the model metrics to reduce the error variability introduced by the random splitting of the dataset, instead, the test set is utilized at the end to evaluate the final model performance. Note that the evaluations on the test set are performed using the models fitted on the whole training set: indeed validation set is no longer needed and all the possible data are utilized for training purposes.

```
str_row_names <- c(
  "Full std model", "Full br model", "Manual model 2", "Stepwise regression with AIC",
  "Stepwise regression with BIC", "Lasso", "LDA", "QDA", "Naive Bayes", "GAM with BIC", "GAM manual"
)
# Summary of the information criteria when available
info_criteria <- as.data.frame(cbind(
  AIC(mod_full_std, mod_full_br, mod_manual2, mod_step_aic, mod_step_bic, mod_gam_bic, mod_gam_manual)[, 2],
  BIC(mod_full_std, mod_full_br, mod_manual2, mod_step_aic, mod_step_bic, mod_gam_bic, mod_gam_manual)[, 2]
))
colnames(info_criteria) <- c("AIC", "BIC")
rownames(info_criteria) <- str_row_names[c(1, 2, 3, 4, 5, 10, 11)]
```

```
# Sort by AIC
(info_criteria %>% arrange(AIC))[, c("AIC", "BIC")]
```

```
##                                   AIC      BIC
## GAM manual                   2552.231 2826.058
## GAM with BIC                 2589.590 2800.199
## Stepwise regression with AIC 2626.833 2810.395
## Full std model               2628.138 2835.643
## Full br model                2628.545 2836.049
## Stepwise regression with BIC 2694.415 2742.300
## Manual model 2               2735.929 2847.662
```

```
# Sort by BIC
(info_criteria %>% arrange(BIC))[, c("BIC", "AIC")]
```

```
##                                   BIC      AIC
## Stepwise regression with BIC 2742.300 2694.415
## GAM with BIC                 2800.199 2589.590
## Stepwise regression with AIC 2810.395 2626.833
## GAM manual                   2826.058 2552.231
## Full std model               2835.643 2628.138
## Full br model                2836.049 2628.545
## Manual model 2               2847.662 2735.929
```

The information criteria for the models containing all the predictors are very similar. The lower AIC score instead is from the manual-designed GAM, highlighting that this one has a good balance between the goodness of fitting and the complexity of the model. The model with the lower BIC score is instead the linear one selected by stepwise regression with BIC and it's the more parsimonious.

```
metrics_val_full_std <- k_fold_cv(
  model_formula = mod_full_std$formula, model_func = stats::glm, data = df_train,
  n_fold = num_k_fold,
  family = "binomial"
)
metrics_val_full_br <- k_fold_cv(
```

```
  model_formula = mod_full_br$formula, model_func = stats::glm, data = df_train,
  n_fold = num_k_fold,
  family = "binomial", method = "brglmFit"
)
metrics_val_manual2 <- k_fold_cv(
  model_formula = mod_manual2$formula, model_func = stats::glm, data = df_train,
  n_fold = num_k_fold,
  family = "binomial", method = "brglmFit"
)
metrics_val_stepwise_aic <- k_fold_cv(
  model_formula = mod_step_aic$formula, model_func = stats::glm, data = df_train,
  n_fold = num_k_fold,
  family = "binomial", method = "brglmFit"
)
metrics_val_stepwise_bic <- k_fold_cv(
  model_formula = mod_step_bic$formula, model_func = stats::glm, data = df_train,
  n_fold = num_k_fold,
  family = "binomial", method = "brglmFit"
)
metrics_val_lasso <- k_fold_cv(
  model_formula = formula("suicide~."), model_func = glmnetUtils::glmnet, data = df_train,
  n_fold = num_k_fold,
  family = "binomial", method = "brglmFit", lambda = mod_lasso$lambda.1se, alpha = 1
)
metrics_val_lda <- k_fold_cv(
  model_formula = lda_formula, model_func = MASS::lda, data = df_train,
  n_fold = num_k_fold, is_generative = TRUE
)
metrics_val_qda <- k_fold_cv(
  model_formula = qda_formula, model_func = MASS::qda, data = df_train,
  n_fold = num_k_fold, is_generative = TRUE
)
metrics_val_nb <- k_fold_cv(
  model_formula = nb_formula, model_func = e1071::naiveBayes, data = df_train,
  n_fold = num_k_fold, is_nb = TRUE
)
metrics_val_gam_bic <- k_fold_cv(
  model_formula = mod_gam_bic$formula, model_func = mgcv::gam, data = df_train,
  n_fold = num_k_fold, family = "binomial"
)
metrics_val_gam_manual <- k_fold_cv(
  model_formula = mod_gam_manual$formula, model_func = mgcv::gam, data = df_train,
  n_fold = num_k_fold, family = "binomial"
)
```

```
metrics_val_total <- as.data.frame(rbind(
  unlist(metrics_val_full_std),
  unlist(metrics_val_full_br),
  unlist(metrics_val_manual2),
  unlist(metrics_val_stepwise_aic),
  unlist(metrics_val_stepwise_bic),
  unlist(metrics_val_lasso),
  unlist(metrics_val_lda),
  unlist(metrics_val_qda),
  unlist(metrics_val_nb),
  unlist(metrics_val_gam_bic),
  unlist(metrics_val_gam_manual)
))
colnames(metrics_val_total) <- c("threshold", "sensitivity", "auc", "specificity", "accuracy")
rownames(metrics_val_total) <- str_row_names
```

```r
# Sort by best sensitivity (true positive rate)
(metrics_val_total %>% arrange(desc(sensitivity)))[, c("sensitivity", "auc", "specificity", "accuracy", "thresho
```

```
##                              sensitivity       auc specificity  accuracy
## Manual model 2                 0.8851094 0.9067570   0.7743855 0.7764594
## Full br model                  0.8607130 0.9229243   0.8338127 0.8342081
## GAM manual                     0.8522415 0.9311334   0.8645327 0.8642815
## Stepwise regression with BIC   0.8484072 0.9095532   0.8155325 0.8160669
## Lasso                          0.8475158 0.9134681   0.8268128 0.8271717
## Stepwise regression with AIC   0.8454379 0.9217996   0.8476026 0.8475766
## Full std model                 0.8443857 0.9233012   0.8477759 0.8475741
## LDA                            0.8253497 0.8971520   0.8268526 0.8268949
## QDA                            0.8116110 0.9104273   0.8501621 0.8493848
## Naive Bayes                    0.8102868 0.9111183   0.8540254 0.8533154
## GAM with BIC                   0.7941287 0.9206764   0.8949637 0.8931081
##                               threshold
## Manual model 2               0.01394986
## Full br model                0.02419020
## GAM manual                   0.02097350
## Stepwise regression with BIC 0.02200189
## Lasso                        0.02177320
## Stepwise regression with AIC 0.02299416
## Full std model               0.02525382
## LDA                          0.01173920
## QDA                          0.03491888
## Naive Bayes                  0.11318042
## GAM with BIC                 0.03056204
```

Considering the evaluation metrics, some considerations can be stated:

- The model selected with stepwise regression and AIC achieves the highest score
- The models containing all the predictions have a good prediction power based on these tests but have the disadvantage of being less interpretable due to the large number of coefficients. Manual GAM has a sensitivity score very near to the full models
- The generative models perform a little worse compared with the others but the also use less predictors
- In general, all the best probability thresholds selected for the binary predictions are quite low.

```r
metrics_test_full_std <- evaluate_test_set(model = mod_full_std, df_test = df_test, threshold = metrics_val_full
metrics_test_full_br <- evaluate_test_set(mod_full_br, df_test = df_test, threshold = metrics_val_full_br[["mean
metrics_test_manual2 <- evaluate_test_set(mod_manual2, df_test = df_test, threshold = metrics_val_manual2[["mean
metrics_test_stepwise_aic <- evaluate_test_set(mod_step_aic, df_test = df_test, threshold = metrics_val_stepwise
metrics_test_stepwise_bic <- evaluate_test_set(mod_step_bic, df_test = df_test, threshold = metrics_val_stepwise
metrics_test_lasso <- evaluate_test_set(mod_lasso, df_test = df_test, threshold = metrics_val_lasso[["mean_thres
metrics_test_lda <- evaluate_test_set(mod_lda, df_test = df_test, threshold = metrics_val_lda[["mean_threshold"]
metrics_test_qda <- evaluate_test_set(mod_qda, df_test = df_test, threshold = metrics_val_qda[["mean_threshold"]
metrics_test_nb <- evaluate_test_set(mod_nb, df_test = df_test, threshold = metrics_val_nb[["mean_threshold"]],
metrics_test_gam_bic <- evaluate_test_set(mod_gam_bic, df_test = df_test, threshold = metrics_val_gam_bic[["mean
metrics_test_gam_manual <- evaluate_test_set(mod_gam_manual, df_test = df_test, threshold = metrics_val_gam_manu
```

```r
metrics_test_total <- as.data.frame(rbind(
  unlist(metrics_test_full_std),
  unlist(metrics_test_full_br),
  unlist(metrics_test_manual2),
  unlist(metrics_test_stepwise_aic),
  unlist(metrics_test_stepwise_bic),
  unlist(metrics_test_lasso),
  unlist(metrics_test_lda),
  unlist(metrics_test_qda),
  unlist(metrics_test_nb),
```

```
  unlist(metrics_test_gam_bic),
  unlist(metrics_test_gam_manual)
))
colnames(metrics_test_total) <- c("sensitivity", "specificity", "accuracy")
rownames(metrics_test_total) <- str_row_names
```

```
# Sort by best sensitivity (true positive rate)
(metrics_test_total %>% arrange(desc(sensitivity)))[, c("sensitivity", "specificity", "accuracy")]
```

```
##                              sensitivity specificity  accuracy
## GAM with BIC                   0.9081152   0.7065217 0.9046826
## Naive Bayes                    0.8793071   0.6956522 0.8761799
## QDA                            0.8783657   0.7934783 0.8769202
## GAM manual                     0.8672566   0.7500000 0.8652600
## Full std model                 0.8627377   0.7608696 0.8610031
## Full br model                  0.8550179   0.7717391 0.8535999
## Stepwise regression with AIC   0.8508755   0.7608696 0.8493430
## Stepwise regression with BIC   0.8324233   0.7500000 0.8310198
## LDA                            0.8303521   0.7391304 0.8287988
## Lasso                          0.8215025   0.7608696 0.8204701
## Manual model 2                 0.7862926   0.8260870 0.7869702
```

Considering instead the scores obtained on the test set, the situation changes a bit:

- Indeed the GAM model with BIC formula reaches a very high sensitivity score
- The model fitted on all the predictors reach almost a score similar to the previous one but maybe the overfits a bit on the data, considering these results
- Impressively Naive Bayes obtain a high sensitivity score: maybe for this dataset the assumption of the independence between variables are not so incorrect

Conclusions

Summarizing, this analysis highlights various aspects of the data related to the student's mental health, even in the absence of more detailed information about the students and the source of the data. Through exploratory data analysis, it was observed that students experiencing more severe mental symptoms are generally at a higher risk of suicidal ideation. The most significant factors influencing this risk are psychiatric symptoms and self-injurious behaviors.

Additionally, some specific information such as being freshmen, females, and those from extremely poor or wealthy family backgrounds, appear to have a notable impact on the outcome variable.

In terms of modeling, a decision was made to use fewer predictors to enhance interpretability, which was considered crucial for this analysis. This choice may have slightly compromised accuracy, but most models achieved high sensitivity scores. The Generative Additive Model (GAM) emerged as the best compromise between validation and test scores. This model also indicated the presence of nonlinear patterns among the predictors, suggesting a complex interaction mainly between suicide probability and mental symptoms. If greater interpretability and less computational power are preferred, standard logistic regression or Naive Bayes classifier using a subset of predictors can be a viable alternative for this data.