UNIVERSITY OF GOTHENBURG

CHALMERS

MSG 500
Individual project report

Investigating alcohol consumption and driving behaviour and their relationship to risk taking and impulsivity

Fionn Delahunty
Applied Data Science Program
Gusdelfi@student.gu.se

11th January 2018

Table of contents

Note: Although this project is written in the academic general tense (we), I am the only contributor to this project.

Introduction

The Dublin science gallery is a free public museum based in the capital of Ireland and hosted by trinity college Dublin. The purpose of the museum is to enhance the publics scientific knowledge through interactive and fun exhibitions. Research teams can submit different exhibitions based on the current theme of the gallery. During 2015 the gallery hosted a risk-taking theme, as part of this exhibition participants where offered the ability to fill out two psychometric risk-taking questionnaires and discover their own risk-taking attitudes. A total of 1326 individuals participated in the questionnaire over the two months it was left on display.

In total participants where asked to complete 93 questions. The first 10 of these questions related to demographic characteristics. The next thirteen questions related to participants driving behaviour while the following eight questions related to their alcohol consumption behaviour. Participants then completed the 30 item Barratt Impulsiveness Scale (BIS) in its standardised format and the revised 30-Item Domain-Specific Risk-Taking (DOSPERT) Scale in its respective standard format.

Since the data was collected no analysis has been completed on the dataset. This report is the first primary analysis of the raw dataset. We will discuss the implications of working on such a raw dataset later in the report. For the purpose of this assignment we investigated four different hypothesises related to this dataset. These hypotheses are roughly based on existing psychological research on DOSPERT and BIS scales, but are not an extension of any specific published study. Rather they are designed for the purpose of self-teaching for the MSG500 module.

The four hypotheses are follow.

1. Predicting the number of cigarettes smoked per week from a measure of risk taking in a social context (Domain-Specific Risk-Taking (DOSPERT) Scale)
2. Predicting the number of car crashes from a measure of self-control (Barratt Impulsiveness Scale (BIS))
3. Building a prediction model to predict if a participant in the dataset has smoked during their lifetime.
4. Building a prediction model to predict the age that a participant started drinking alcohol.

Following the establishment of the four hypotheses, our first step was to convert the raw dataset into a useable set of variables. Primarily this involved scoring the two psychometric questionnaires into a set of final variables. A number of questions required reversed scoring as an intermediate process. The 30 item BIS questionnaire was converted into eight final factors, while the DOSPERT questionnaire was reduced into six final factors. A brief description of these factors is provided in the appendix.

Overall this process was considerably more time consuming then initially expected. The majority of issues arose with the collection method. In a minor issue this included variables having to be reordered and re-indexed for scoring. In a more major sense the data collection had little or no response validation, which meant that a single question could contain numbers (1), strings (one) and missing or random values. This was definitely the most time-consuming issue of this project, and in hindsight was because of a lack of experience on our part. Given the time limitation of this project, choosing a clean and neat dataset would have been a much better idea.

A final and unexpected issue that arose was that the data collection system had failed for one specific question on the BIS scale ("I buy things on impulse?") meaning two of our final variables on the BIS (BIS Sum & BIS Motor) where invalid since no data was collected for an aspect of these

factors. Since our hypotheses didn't involve this factor specifically we where easily able to move on with creating the dataset without these two factors.

Following data cleaning, a final dataset was created with all necessary factors.

Methods

Throughout this project R programming language was used for all statistical analysis. R code was written and executed in a Jupyter notebook environment. Datasets and code are stored on a private GitHub repository for backup. All code (in notebook format) are attached in the appendix.

All four different hypothesis required the use of different statistical methods in their respective analysis. We will investigate each of the four different methods employed individually.

1. *Predicting the number of cigarettes smoked per week from a measure of risk taking in a social context (Domain-Specific Risk-Taking (DOSPERT) Scale)*

   This hypothesis involved one dependent variable (cigarettes smoked per week) and one independent variable (domain-specific risk-taking). Domain specific risk taking is a two-factor measure, composed of social risking taking and recreational risk taking. For the purpose of investigation, we employed the social risking taking measure as the IV. We found this variable to be normally distributed with no extreme outliers.

   Cigarettes smoked per week was a continuous variable (0 to 425), the data was skewed to the left-hand side due to the presence of a few extreme outlying values. The majority of the data itself was clustered into groups rather then having an even spread throughout the range. We felt the best way to represent this data was by dividing it into seven levels (spilt into groups of 25). The highest level (150 to 175 cigarettes smoked per week) also included the values 345,390 and 425 cigarettes smoked per week. These where extreme outlying values, and the for the purpose of our work we considered the best solution was to include them in the highest level. These seven levels only included values above 0, the majority of the sample (1064) did not smoke, we talk about this later in method and result section.

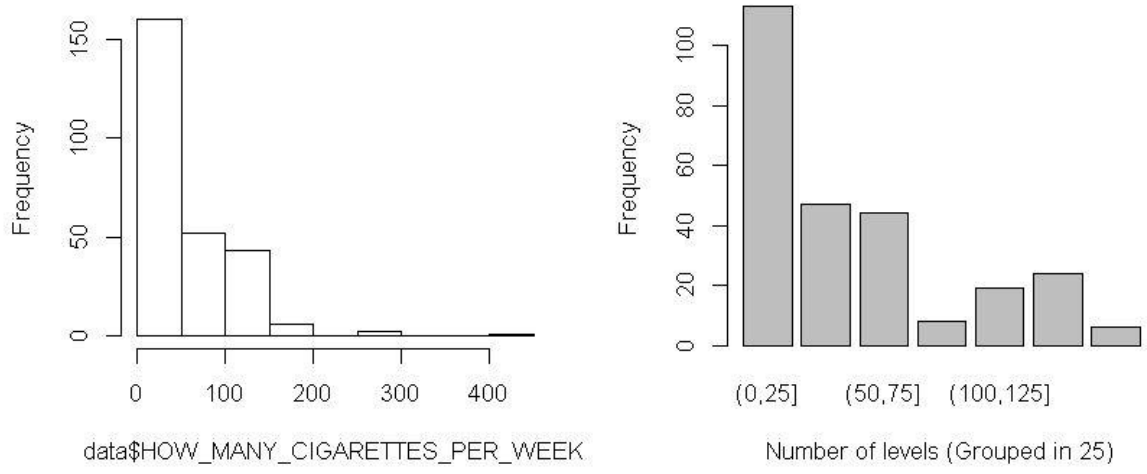

Figure 1 – Raw data on the left-hand side, data divided into 7 levels on the left hand side.

We began investigating the data by first employing an ANOVA to compare the means of cigarettes smoked per week across domain specific risk taking. We also performed a Kruskal-Wallis rank sum test. We have uneven sample size in each of the 7 levels, generally ANOVAs are quite robust to violations of this assumption, but we still preformed the non-parametric test regardless.

To further test the relationship between risk taking and the seven levels we created a binary variable and performed a linear regression model which held the first level as a baseline. We preform a second linear regression with recreational risk taking as well, both social and recreational risk taking should not be highly correlated factors.

2. *Predicting the number of car crashes from a measure of self-control (Barratt Impulsiveness Scale (BIS))*

This hypothesis involved one dependent variable (number of car crashes in the last two years) and one independent variable (self-control (BIS Scale)). Self-control variable has a normal distribution with no extreme outliers. The number of car crashes in the last two years variable however is not normally distributed, rather we see it contains a majority (91%) of 0 values, with some 1,2,3 and 4 values. Following some more investigation we grouped the variable in a dichotomise fashion with 0 as no crashes and 1 as having crashed at least once. Our sample sizes where still extremely uneven 1187 in the 0 group and 141 in the 1 group. We performed both an ANOVA and Kruskal-Wallis rank sum test on the dataset to investigate mean differences between the two levels across self-control. We also preformed a logistical regression with the same variables to investigate the relationship further.

3. *Building a prediction model to predict if a participant in the dataset has smoked during their lifetime.*

This hypothesis involved one dependent variable (Have you smoked?) and 11 independent variables. These IVs included age as a demographic variable, all four BIS factors and all six

DOSPERT factors. The model employed logistical regression as a means of identifying which factors contributed to the prediction of dichotomise smoking variable. In order to test the effeteness of the model the data was divided into test and training datasets.

4. *Building a prediction model to predict the age that a participant started drinking alcohol.*

   This hypothesis involved one dependent variable (The age a participant first consumed alcohol) and 11 independent variables. These IVs included age as a demographic variable, all four BIS factors and all six DOSPERT factors. This hypothesis was not actually tested. On investigation of the age started drinking variable it became clear that the variable was not normally distributed. The majority of participated had first consumed alcohol at the legal age in Ireland (18 years old). Some participants had first consumed it earlier, and some later in life. In order to make the variable investigable we divided it into four levels, those who had never consumed alcohol (*n= 202*), those who had first consumed alcohol under the legal drinking age (*n=510*), those who had consumed alcohol at the legal drinking age *(n=473*) and those who had first consumed alcohol later then the legal drinking age (*n=178*). As a result of dependent variable became a multinomial variable, the best method of investigation would have been multinomial logistic regression. However, given this was not in the scope of our course, along with general time constrictions we did not proceed any further with this investigation.

Results

Predicting the number of cigarettes smoked per week from a measure of risk taking in a social context (Domain-Specific Risk-Taking (DOSPERT) Scale)

1. **Model building**
   Our first step in building this model was to employ both a ANOVA and Kruskal-Wallis rank sum test. Neither of these tests returned statically significant results. By investigating the boxplot in figure 2 we can see an overall difference in the means level of social risk taking, but no clear upwards or downwards tread across the levels. There is no visible suggestion that those who smoke a high number of cigarettes per week have a higher risk-taking score.

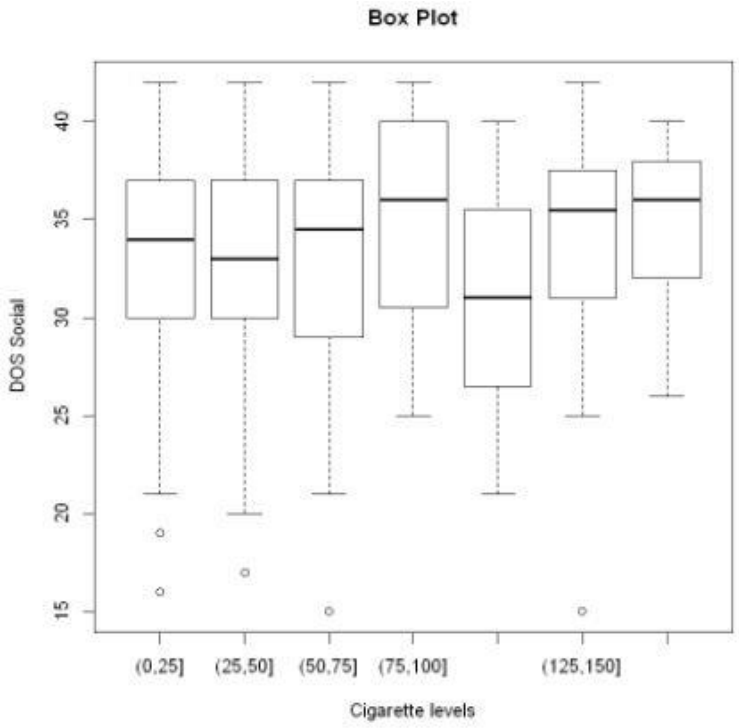

Figure 2 Boxplot of all seven levels of the number of cigarettes smoked per week

2. **Model selection**
   In order to further investigate which cigarettes levels contributed to social risk taking we developed a linear regression model after having converted the cigarette levels into a set of dummy variables. In order for regression to occur, we took the first level as a baseline measure. We began model selection with a standard multiple regression, which suggested none of the five levels where significant predictors. The overall model itself was also nonsignificant. We employed a backward model sectional approach which also resulted in an overall insignificant result. No combinations of the variables it employed produced anything significant.

   As mentioned previously domain specific risk taking is composed of two different factors, social and recreational risk taking. We decided to add recreational risk taking to the model

to see if it would improve prediction. The two factors are not strongly correlated (R^2= *0.347*) since they measure different constructs of the same domain. We began by plotting the two scales against each other and overlaying that with the smoking grouping we had earlier established. Via a visual inspection we see no clear pattern of smoking groups across the spread of risk taking. We employed an additive model in addition to our backward selection with the addition of recreational risk taking as a factor.

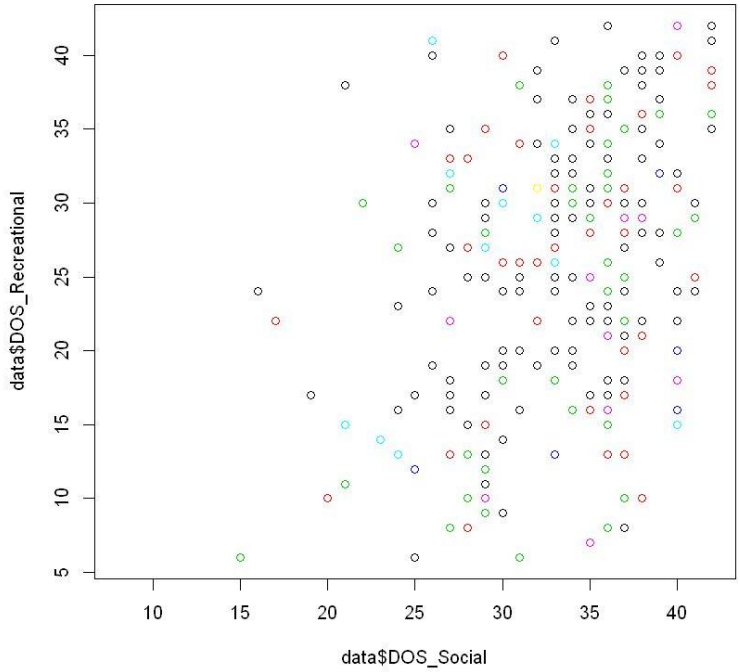

Figure 3 Social risk taking verse recreational risk taking. Colours = smoking levels

3. Model comparison

In summary our additive model performed a better fit then our backward selection. Our backward selection found no significant combination of the variables. However, our additive model found recreational risk taking strongly contributed to the model, along with the 3rd and 4th grouping having a mild positive significant effect. The predictive power of this model is quite strong (Adj R^2 of 81%) however, when we investigate the residuals plots we see this is most likely an inflated value. The fixed residuals plot does not show constant variance, or a linear distribution. Rather it is unevenly spread across the fit. We also see tails in the QQ suggesting as the fixed residuals did an overall poor linear fit. Finally, there is no constant error variance.

Predicting the number of car crashes from a measure of self-control (Barratt Impulsiveness Scale (BIS))

1. **Model building**

Following splitting the data into no car crashes (0) and at least one car crash (1) we began model building by performing an ANOVA and Kruskal-Wallis rank sum test. Both of these returned insignificant values meaning we don't reject the null hypothesis. According to a boxplot we see only a very small difference in self-control measure between the two car crash groups. We move onto our main prediction test for this hypothesis which is logistical regression.

2. **Model selection**

Overall our model did not find that self-control was a significant predictor of car crashes. Here we diagnose perhaps why this might be the case, and can we correct it. We start with investigating our Pearson residuals plots which have a large number of values above the $|ri| > 2$ threshold. The actually fit of the model along the proposed model fit line is quite poor, with tails at either end deviating from the line. We attempted a number of transformations of the x variable (Log/Square root) which had no noticeable effect on either the summary stats or the plot. We suspect the issue here is due to uneven sample size, with the number of 0 values having almost ten times the number of 1 values. We discuss the effect of uneven sample size in logistical regression in the discussion section. Our deviance residuals also have quite a large number of values above the $|di| > 2$ threshold, suggesting that almost all the 1 values have a significant influence on the log-likelihood. Additionally, we have extremely influential outliers as seen as the cook distance plot. The final residual test we would have liked to employee as a link test, however we where unable to get R code to perform such a test to work correctly. Given the above results we would have expected this plot to show a skew rather then a trend.

Building a prediction model to predict if a participant in the dataset has smoked during their lifetime.

1. **Model building**

As mentioned above our prediction variables in this case was a binary outcome if a participant had smoked or not in their lifetime (534 yes and 586 no). We had 11 independent variables and our main method of model selection and building was logistical regression. In order to test the effectiveness of this model our data was also 20% / 80% test and train respectively.

2. **Model selection**

Our initial logistical model had a chi squared cumulative probability of 0, meaning we reject the fit of the binomial model with logit link. Of the eleven variables in the model five had some level of significant prediction value in the model, the strongest of which was health & safety risk taking. The residual plots of this model suggested a slightly non linear fit based on a residual plot, seen by a tail on the right hand side. At this stage we didn't correct this with any transformations. We also see one extreme outlier on the leverage plot, but this wasn't significant on the cook distance plot.

To improve on this general model selection, we ran a backward model selection which reduced the eleven variables down to six. All of these included variables are significant in some regard varying between less than 0.001 and less then 0.05. At this refined stage we ran a correlation matrix to test for multicollinearity, fortunately none of the six variables where highly correlated. Residuals

diagnostics improved for this refined model, we still had an issue with the non-linear fit in the residual plot, but we no longer had any extreme leverage outliers.

Our next step was to test for interactions between the six variables, our R code ran a logistical regression including all interactions up to three-way interactions between the variables. As a result of this we found a strongly positive three-way interaction between cognitive instability, self-control and financial gambling risk taking. We included this in the final model but found no significant improved it also introduced a significant amount of collinearity between the remaining variables, for this reason we dropped it from our final model.

3. Model effectiveness

Variables included in the final model
Age
Cognitive instability
Self-control
Financial gambling risk taking
Financial investment risk taking
Health and safety

Our final model included the above list of variables, in order to test its effectiveness, we applied to predict the scores of the test set. We had a misclassification error of one out of 200 classifications. This is surprisingly good. We discuss this further in the discussion selection.

Building a prediction model to predict the age that a participant started drinking alcohol.

1. Model building

Age started drinking alcohol was not an original factor in the dataset, we manually calculated this by subtracting the reported age of an individual against the reported age they first drunk alcohol. In four cases we had to insert NA values for people who reported they had first drunk alcohol earlier then they had been born (negative value). In the histogram below, we display the distribution of the age started drinking variable, this excludes 217 NA values. Following some work, we found the best way to divide the data was into four groups as outlined in figure 4. Reducing the variable into anything less (binary variable) would have been redundant as we would have no useful information from the prediction. The expectation being people who had started drinking under the legal age and those later, however since the legal age has changed over the last 40 years in Ireland and socio-cultural differences have existed with regard to underage drinking we likely would have had to control for age and perhaps even location or nationality.

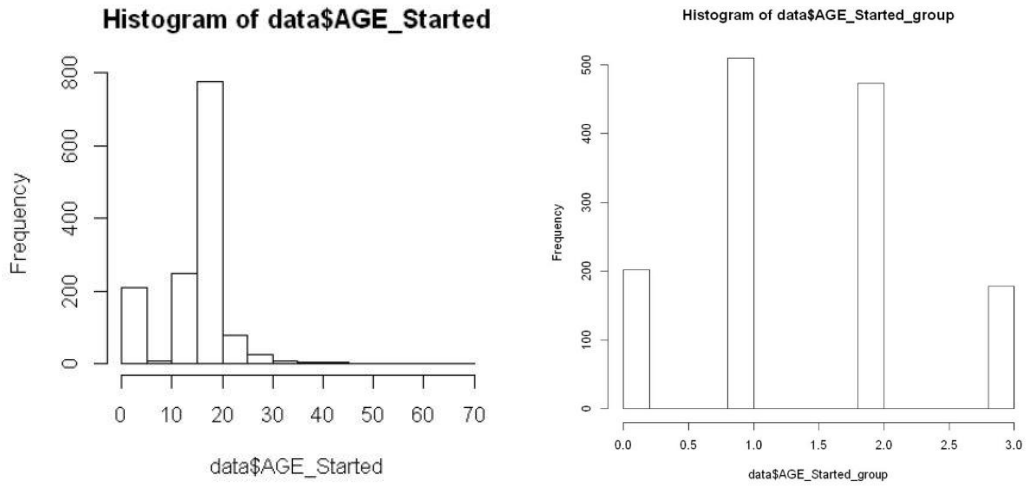

Figure 4 Histogram of data before and after grouping

2. Model selection.

Based on the fact we would have to perform a multinomial logistic regression we choose not to continue this model as it was outside the scope of the course

Discussion

Predicting the number of cigarettes smoked per week from a measure of risk taking in a social context (Domain-Specific Risk-Taking (DOSPERT) Scale)

As established early in the model there was very little difference between social risk-taking levels across the cigarette consumptions groups. From this early onset it was unlikely we where going to find a good prediction model for this dataset. However, surprisingly our model did find an acceptable prediction with the addition of recreational risk taking. However, despite the low collinearity we can't help but consider that this model was simply recreationally risk taking predicting social risk taking with two smoking groups hanging on the side.

Regardless of what the model was actually predicting, the overall model was not a good statistical model due to the poor residual fit noted above. This could possibly have been solved been solved by more advanced transformations outside the scope of this course such as an arcsine transformation. Our take home message from this work was that social risk taking is not a viable predictor of the number of cigarettes smoked by an individual. Self-report bias may have been issue here, given the nature of reporting cigarettes consumption.

Future research considers using a better measure of cigarettes consumption that is not self-reported and explorering more way to group the cigarette data (by box smoked per week perhaps etc). However, we do not see much potential in this area.

Predicting the number of car crashes from a measure of self-control (Barratt Impulsiveness Scale (BIS))

Due to a range of poor diagnose plots we dismissed this model as not being sufficient for prediction of car crashes from self-control. One issue we noted was the uneven sample size between the two groups in the logistical regression, 1187 in the 0 group and 141 in the 1 group. Here we discuss the work of King & Zeng (2001) for logistical regression in rare event data. For the purpose of their publication rare event normally constitute less then 1% of the sample, while ours is almost 10%. Despite this there solution of converting Y to an unobserved continuous variable distributed according to a logistic density with mean μ could still be attempted with out dataset and might yield a better prediction model without such poor diagnose plots. This is a proposal for future research on this data.

The positive take home message from this model is that seemly very messy data can still be grouped and worked on. On our first investigation of the car crash variable we where very concerned on how we'd be able to model the data given the disruption. However grouping it dichotomously seemed to be a very good step in the right direction, with the suggestion from King & Zeng a possible improvement going forward.

Building a prediction model to predict if a participant in the dataset has smoked during their lifetime.

This was arguable our most successfully model, and the only model that showed some positive prediction ability. Firstly, we shall example the variables include in the final model. It is reasonable to assume that age and self-control as both acceptable predicter of smoking behaviour during an individuals lifespan, along with health and safety risk taking. Cognitive instability is an interesting factor to be included and would possible require more reason into the underlying measures within this factor as to example the logic for it's inclusion. It is quite interesting to see the two-financial risk-taking measures included. This was unexpected, and it would be interesting to see if it this is present in future dataset or just a singularity within this population sample. Normally we would not consider an individual financial risk-taking behaviour to so strongly related to a health behaviour such as smoking. Future research might try to graph this data in more detail and estimate if there was specific group in the data that causes these two factors to be included and could any more information be detailed on them.

In summary, according to our misclass cation test our model had quite a strong prediction power. More then we expected. We expect this might due to over fitting the training set, or a miscoding in our R script as we don't understand how we could have such a powerful predicter given the errors in the model (Nonlinear residuals). We would be interested in run this model on a new clean dataset in the future.

This model is close to acceptable, but the non-linear residuals still prevents us from recommending it as a sufficient statistical model. As mentioned for the first model, perhaps complex transformations might solve this issue.

Building a prediction model to predict the age that a participant started drinking alcohol.

As previously mentioned we did not perform any model selection on this hypothesis. However, our learning outcome in this case was on how to group real life data into a statistically manageable format while still keeping the data valid as a source of information. This was a quite challenge as it was our first time experiencing with this issue, we referenced some published work for inspiration.

We feel our final solution (fours levels) is both statistically manageable and keeps the data as a valid source of information.

If we had covered multinomial logistic regression in class we would have been able to explore this model in more detail, and it would be an interesting exploration given the work we put into grouping the data. This is defiantly a suggestion for future research on this dataset. We would attempt to apply the same set of variables we used for the smoking model and hopefully find some useful and interesting predictions.

The main take home message for this model is that grouping data into levels turns seemly incoherent data in manageable and useful data quite quickly, but there is no easy way to know to how to do this. It must be done on a case by case basis, and importantly must be done with some domain knowledge so that the meaning of the data is still kept valid.

Overall conclusion.

This assignment was considerable more challenging then first expected, a lot of time spent in the early stages working on sorting and grouping the data left less time then required for the modelling aspect of the assignment which was unfortunate. We hadn't expected such a poor dataset with such high number of missing values and requirements to recode and group the data. In saying that however the lessons we learnt with this dataset will hopefully stand to us going forward into the field of social science research.

We accept that there is considerable more research to be done on this dataset, looking at more complex interactions of the variables and applying more complex statistical tests and transformations. It is unfortunate we uncovered no positive results, but good science reports both positive and negative results.

In conclusion all three of our tested hypothesis are rejected.

Appendix

Name in R	Full name
ID	ID Code
Gender	Gender
AGE	Age
EVER_SMOKED	Have you ever smoked?
COLLISIONS_LAST_TWO_YEARS	Have you had a car collision in the last two years?
HOW_MANY_YEARS_DRINKING	How many years have you been drinking?
AGE_STARTED	Age started drinking
BIS_Sum	Barratt Impulsiveness Scale Summary
BIS_Attention	Barratt Impulsiveness Scale Attention
BIS_Cog_Instability	Barratt Impulsiveness Scale cognitive instability
BIS_Motor	Barratt Impulsiveness Scale motor
BIS_Preseverance	Barratt Impulsiveness Scale perseverance
BIS_self_control	Barratt Impulsiveness Scale self-control
BIS_Cog_Complexity	Barratt Impulsiveness Scale cognitive complexity
DOS_Ethical	Domain-Specific Risk-Taking Ethical
DOS_Fin_Investment	Domain-Specific Risk-Taking financial investment
DOS_Fin_Gambling	Domain-Specific Risk-Taking financial gambling
DOS_HealthSafety	Domain-Specific Risk-Taking health and safety
DOS_Recreational	Domain-Specific Risk-Taking recreational
DOS_Social	Domain-Specific Risk-Taking Social

Dataset:

Dataset is a private dataset, I can send you the .csv file if you like.

R Code:

Attached as python notebooks. If you have an issue opening these people let me know and I can convert them to R script manually.