Quantifying Happiness: What truly makes us happy?

ST201 - Group Project Report

Word Count: 2316

Candidate IDs: 24369 // 29861 // 34971

02/05/2024

Permission to use as an example: We give permission for our assignment to be used as an example in ST201 (in an electronic form).

Introduction:

"You will never be happy if you continue to search for what happiness consists of. You will never live if you are looking for the meaning of life." — Albert Camus

Understanding the happiness of individuals and the factors correlated with happiness levels can be beneficial in creating healthier and more productive societies. Happy workers are productive, satisfied workers, and their positive effect is associated with good organisational citizenship and good relations (Pavot & Diener, 2004). We understand happiness as a measure of positive emotions marked by contentment, joy, and overall satisfaction with life. Through this research project we hope to further understand:

- What factors in people's lives affect their happiness?
- To what extent do these different factors have an impact on an individual's happiness?

To answer these questions, we use a subset of data from Round 6 of the European Social Survey 2012 containing responses to 20 questions posed to 2,286 UK respondents. Of the 20 associated variables (one of which is the response variable *happy*), we consider 7 of these variables (*rlgblg*, *brncntr*, *domicil*, *marstgb*, *gndr, inprdsc and sclmeet*) categorical variables, while the rest are considered quantitative.

We aim to use linear regression models to model an appropriate subset of variables to explain their effect on happiness. Using relevant selection methods, we are able to find optimal linear regressions, allowing us to make suitable inferences against our research questions.

Data and EDA:

Missing Data

Although we found that there were no missing data points, many of the questions presented to the subjects allowed for responses such as "refusal to answer" and

“don’t know.” We decided to treat these data points as missing values, as we felt attributing meaningfulness to these responses would be pushing the limits of reasonable inference. We also considered how addressing this during our data cleaning affected the viability of their use in our regression model.

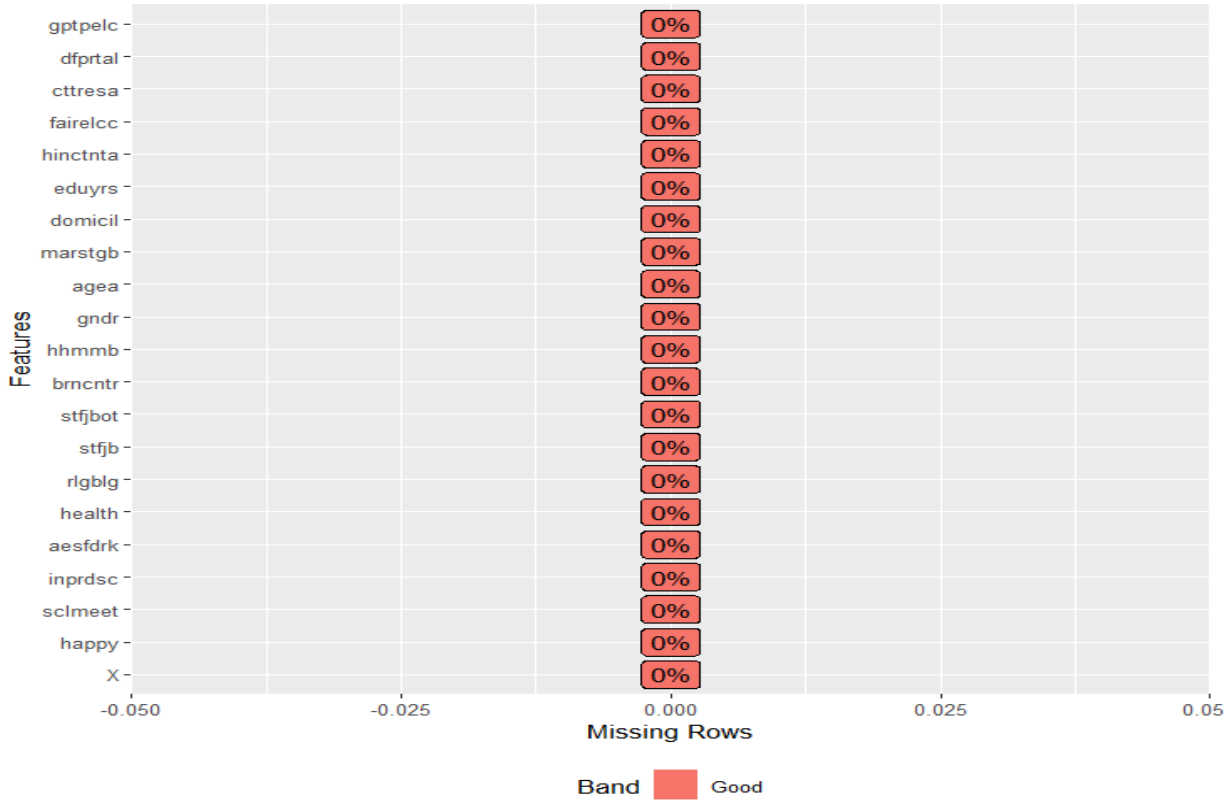


Figure 1: Missing Values

After reviewing the explanatory variables, we noticed these values were concentrated in a few variables, namely those concerning marital status and job satisfaction (Appendix A). We cleaned the dataset by removing any values that indicated a response that did not “directly answer the question,” creating missing values (*NA*) for some explanatory variables involved in our linear regression model. Additionally, when observing correlations within the data, we found that we could not reconcile these *NA* values within R. As a result, we were forced to remove certain subjects from the data when producing our correlation matrix (Appendix B), decreasing the amount of data on which we observed correlations.

Transformations

We found several continuous variables to not follow normal distributions. To account for this and improve the validity of our regression model, we tried applying log transformations to remove the skewed nature of the data. However, we realised that our data was not sufficiently continuous for this to be effective, and so we decided against applying non-linear transformations to our data.

Data Types

Importantly, we noted that some variables in the dataset used for the analysis were ordinal rather than truly continuous. This suggests that the values were meaningfully ranked or ordered, but the spacing of these ranks were not always clear or consistent. We treat these variables as continuous data to allow us to run smoother linear regressions. This can be done where we find that the interval points are equally spaced out and we have a suitable number of interval points. However, in some cases we were unable to make this assumption, leading us to choose to treat the data as categorical. For instance, the intimate and personal discussions variable (*inprdsc*) has an unequal jump between 2 to 3 and 3 to 4, with the former representing one extra discussion and the latter indicating an additional one to three discussions. We therefore treat this variable as categorical in our regressions.

The table below shows the data type we chose to treat each variable as.

.

Variables	Data Type	Variables	Data Type	Variables	Data Type
happy	Continuous	*stfjbot*	Continuous	*eduyrs*	Continuous
sclmeet	Categorical	*brncntr*	Categorical	*hinctnta*	Continuous
inprdsc	Categorical	*hhmmb*	Continuous	*fairelcc*	Continuous
aesfdrk	Continuous	*gndr*	Categorical	*cttresa*	Continuous
health	Continuous	*agea*	Continuous	*dfprtal*	Continuous
rlgblg	Categorical	*marstgb*	Categorical	*gptpelc*	Continuous
stfjb	Continuous	*domicil*	Categorical		

Figure 2: Variable Types

Regression Analysis and Results:

Having cleaned our data and applied the necessary preprocessing along with the transformations on it, we first specified a regression model that regresses happiness on all the variables. Following this specification, in order to identify the most significant features to improve the model performance and ultimately to enhance the interpretability, we have implemented linear model selection methods.

We have initialised our analysis with the Best Subset Selection with the regression *hap_reg_sub*. Implementing the Forward Selection and Backward Selection algorithms, we have identified that the best three-variable models suggested from all are the same.(Appendix I and J) This model specification included *health, stfjb* and *hinctnta*.

intercept	***health***	***stfjb***	***hinctnta***
5.9860676	-0.5672498	0.2277866	0.1116497

Figure 3: The Best Three-Variable Models

To observe where the difference comes at, we checked the estimations provided with 4 variables from *hap_reg_sub*, *hap_reg_bwd* and *hap_reg_fwd* that corresponds to the models from best subset selection, backward stepwise selection and forward stepwise selection respectively.

Best Subset Selection	***Intercept***	***health***	***stfjb***	***hinctnta***	***marstgb6***
	5.76929694	-0.54409554	0.23408056	0.09922745	0.34835954
Forward Selection	***Intercept***	***health***	***stfjb***	***hinctnta***	***marstgb6***
	5.76929694	-0.54409554	0.23408056	0.09922745	0.34835954

	Intercept	*health*	*stfjb*	*hinctnta*	*domicil5*
Backward Selection	5.8997909	-0.5566338	0.2372456	0.1065459	0.8761092

Figure 4: The Best Four-Variable Models

The best four-variable models identified, best subset and forward selection gave the identical results.

Lastly, best five-variable models specifications all yielded distinct outcomes.

	Intercept	*health*	*stfjb*	*hinctnta*	*marstgb3*	*marstgb4*
Best Subset Selection	6.16937118	-0.55186652	0.23109121	0.09642039	-0.63026272	-0.44419519
	Intercept	***health***	***stfjb***	***hinctnta***	***marstgb6***	***stfjbot***
Forward Selection	5.43470655	-0.50528033	0.20115324	0.10295904	0.36806471	0.07127983
	Intercept	***health***	***stfjb***	***hinctnta***	***domicil5***	***fairelecc***
Backward Selection	5.30127305	-0.55290199	0.23287548	0.09267750	0.96663131	0.08505272

Figure 5: The Best Five-Variable Models

We obtained the highest adjusted R-squared measure from the backward selection's proposed variables. Furthermore, to enhance the model specification, we have added the quadratic terms and interaction terms to improve our model fit. We have included *agea* and *eduyrs* as quadratic terms to capture their potential diminishing marginal contributions to happiness through time. Interaction terms were identified based on a priori expectations and the correlation analysis between the parameters as well. (Appendix: F, G and H)

We then obtained our ultimate model named *hap_fin_reg* with 32% of explanatory power and F-statistics was significant, indicating the overall model significance.

After finalising our model for interpretation, we followed our analysis with the diagnostic tests.
The Gauss-Markov assumptions are vital for our model's validity.

- Linearity in parameters,
- Homoscedasticity,
- Exogeneity (corr(regressors, u) = 0),
- No perfect multicollinearity,
- No Serial Correlation of Error Terms (Autocorrelation),
- The regression model is correctly specified, without a specification bias.

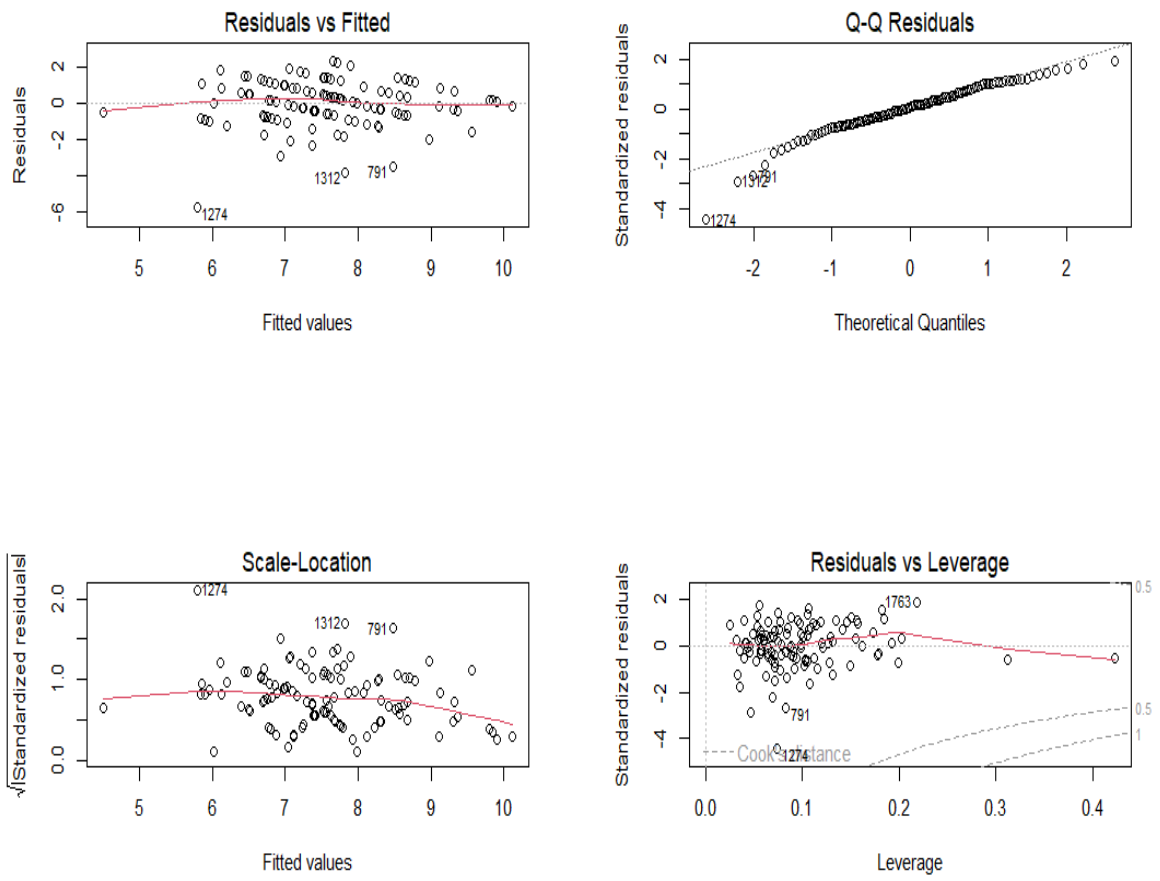


Figure 6: Diagnostics of *hap_fin_reg*

The regression model assumes that there is a linear relationship between the regressors and the regressand. This is crucial as if the true relationship is far from the linear, the inferential conclusions would be biassed. We can check the true relationship from the Residuals vs. Fitted plot. The residual plot resulting from hap_fin_reg suggests that there is a linear pattern in the residuals. It indicates that our additions of quadratic terms to the model specification has improved the fit.

Another underlying assumption of the linear regression model is the normality. We observe the distributional property of our model through the QQ-plot of residuals. The theoretical quantiles resemble the standard normal distribution, our residuals closely approximate the normal distribution with a very slight fat tails.

Next, we identify the homoscedasticity assumption that is vital as the standard errors, confidence intervals and inferential procedures heavily depend on it. The red line which is the smooth fit of residuals is approximately horizontal and we can observe that the fitted values are scattered around the line. This indicates that homoscedasticity assumption is preserved.

Lastly, we wanted to observe whether there are observations that heavily influence our model fit. Following the Cook's Distance thresholds, none of our observations lie in the boundaries. Thus, our final model specification does not have any influential points that lie outside of the normal range of observations. We think this is thanks to the 0-10 scale providing a healthy range for parameter estimation.

Discussion and Limitations:

Initially, we checked the normality of the disturbance term. This is crucial since the t (that we use for parameter significance) and F (that we use for overall model significance) tests are used under the normality assumption. Otherwise, the inferential methods will not be valid.

We should also consider the economic significance, whether the sign of the estimated parameters are suitable with a priori expectations or not, as well as the statistical significance.

R-squared (The Coefficient of Determination) is a measure that explains the predictive power of regressors for the regressand. While potentially meaningful, it has certain drawbacks including the fact that it is a monotonically increasing function of explanatory variables. Although our R-squared values seem to be lower, we assert that a low R-squared value does not necessarily denote a bad model. Instead, we associate this with the unpredictability of individual behaviour, answers and actions. We must rather focus on the regression coefficients, statistical significance and signs of the variables. Ultimately, the theoretical compliance along with the correct model specification is what matters the most for us.

As mentioned in our discussion of our exploratory data analysis, we were forced to confront a large amount of data that was not significant to our research. While there was no missing data in the traditional sense, several survey responses did not directly answer the question (ex. “Do not know,” “Prefer not to answer,” etc.). By cleaning our data to account for such values, we had a significant number of missing values for select explanatory variables. A more complete dataset without missing values would have added to the validity of our findings.

Interpretation and Conclusions:

The purpose of the statistical research we conducted was to better understand the factors in people's lives which contribute to their happiness and to understand the scale of impact that each of these variables have on their happiness levels. Through our statistical models we are able to identify the variables that are linked to an individual's happiness. If policies are introduced to maximise the variables which increase happiness and minimise the ones which decrease happiness, we will be able to foster a happier and more productive society.

The model demonstrates a notable correlation between an individual's self-reported health status and their level of happiness. The coefficient for health in the model is 0.655 (reversed the scale to make 5 best health and 0 least and removed negative), which means that a healthier status is linked to happiness. Each incremental rise of one unit in health on a scale from 0 to 5 is associated with a corresponding increase of 0.655 units in happiness. This highlights the significant influence that health has on one's entire state of well-being. It implies that implementing policies focused on enhancing public health, such as expanding the availability of medical services, preventive care, and wellness initiatives, could greatly improve the overall satisfaction of individuals. Furthermore, this discovery provides evidence for promoting endeavours that advocate for a well-balanced way of living, encompassing regular exercise and proper dietary choices, as these factors can directly enhance overall levels of happiness. Investing in health yields not only personal well-being but also social advantages in terms of enhanced production and lower healthcare expenses.

The model found that a person's place of residence can have a great effect on their happiness levels. In particular we found that those who reside in a farm or a home in the countryside tend to be much happier than other domiciles. A person living in a farm is reported to be an average of 1.355 points happier than that of someone living in a big city considering happiness on a scale of 0-10. This is interesting to note however in practice it is difficult to introduce policy to take advantage of this. Though an individual is seeking to improve their happiness can consider relocating to farmlands or to the countryside away from the city. It would be worth further understanding the effect of population densities on happiness levels as the effect of many people relocating may potentially cause an inverse effect on general happiness due to changing densities.

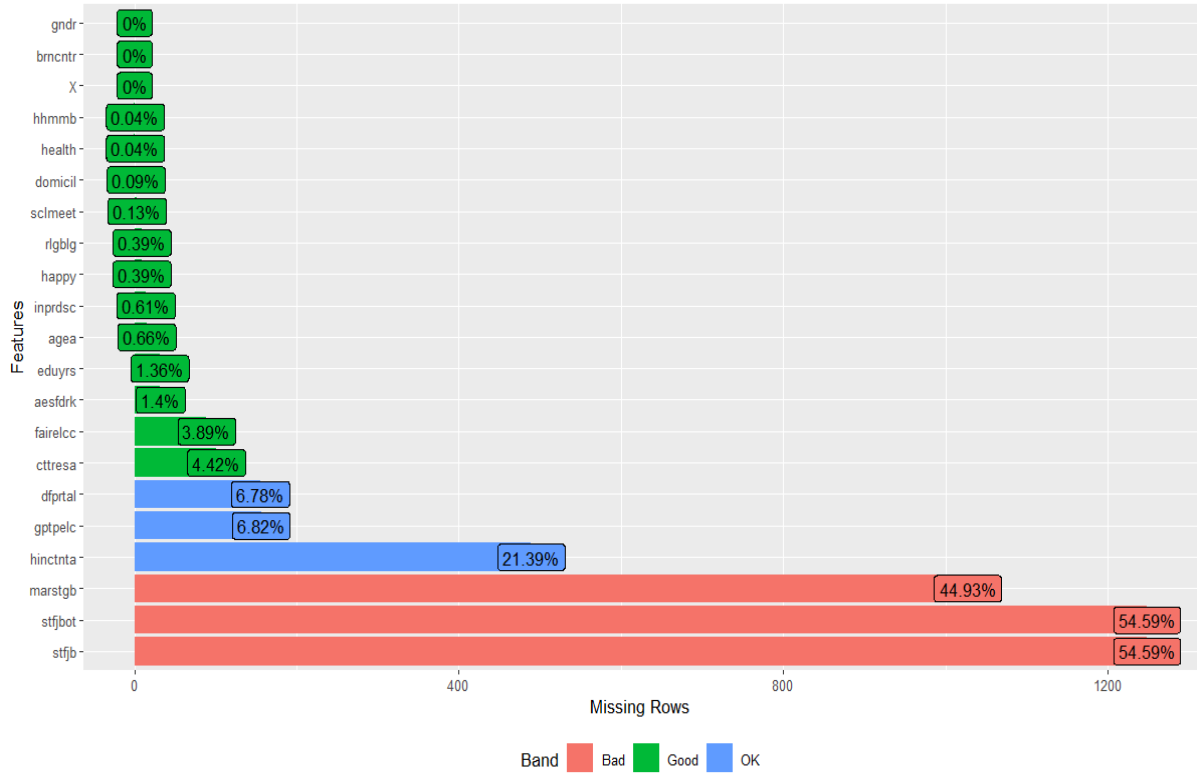
There was also an observed relationship between an individual's belief that national elections are free and fair and their happiness levels. For an extra unit of belief in the integrity of the elections we estimate an individual is 0.229 units happier (both measured on a scale of 0-10). This suggests some effort should be made in building the public's trust in the election system which in return can have benefits on the productivity of the nation. Those in positions of power should also be more cautious in avoiding scandals and other negative publicity which can affect the public's perception.

After interpreting the final model, there are still some statistical relationships which can further be explored to gain an understanding of the general themes which can have an effect on individuals happiness. We propose these themes as an extension to be explored.

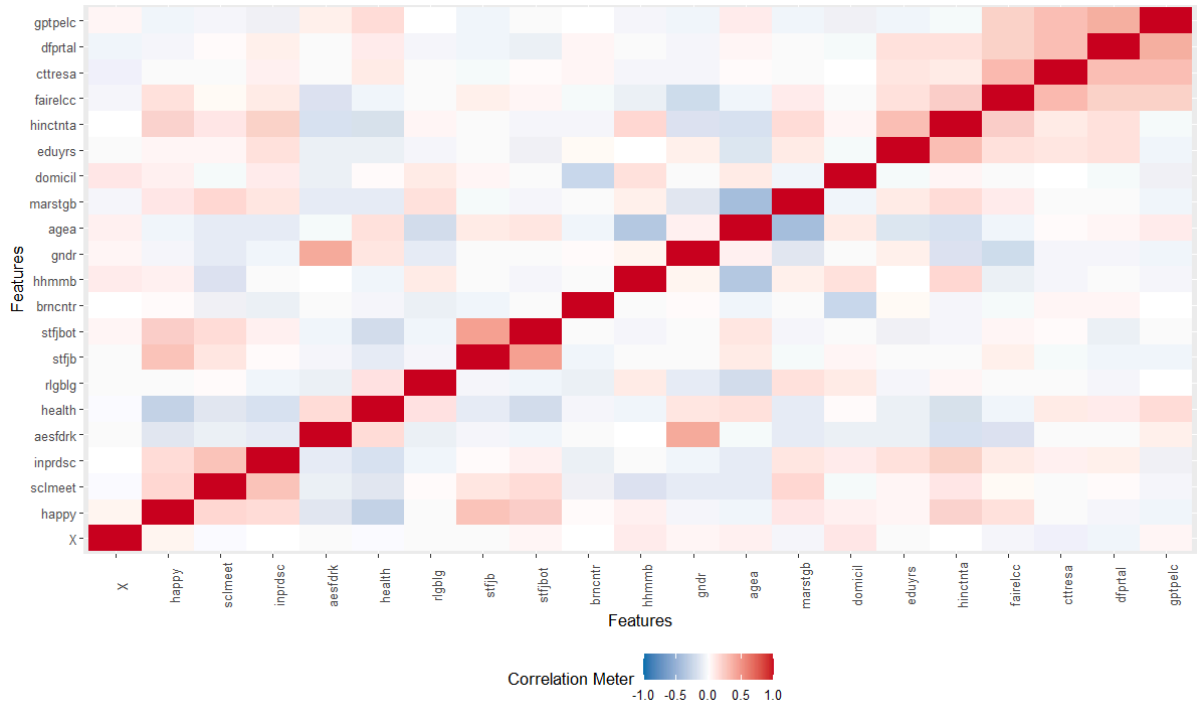
How does an individual's political beliefs affect their happiness?
How does an individual's mental health affect their happiness?
How does an individual’s sense of belonging to a community affect their happiness?
How does an individual's household factors affect their happiness?

Using our findings we build a strong understanding of what can affect happiness in society and we have scope to develop our understanding further and build happier and more productive societies.

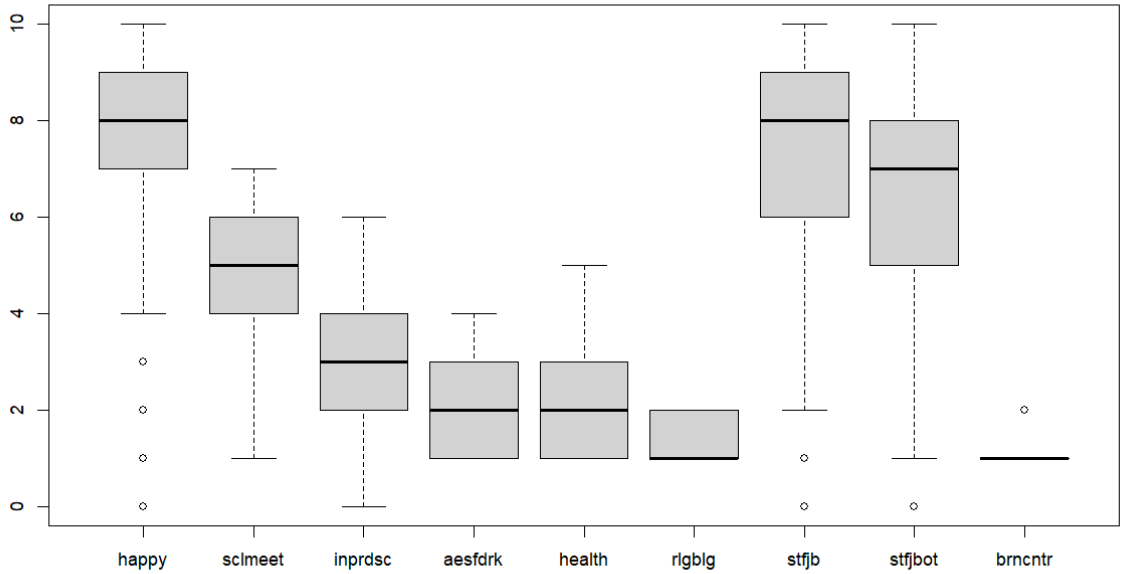
Appendix:



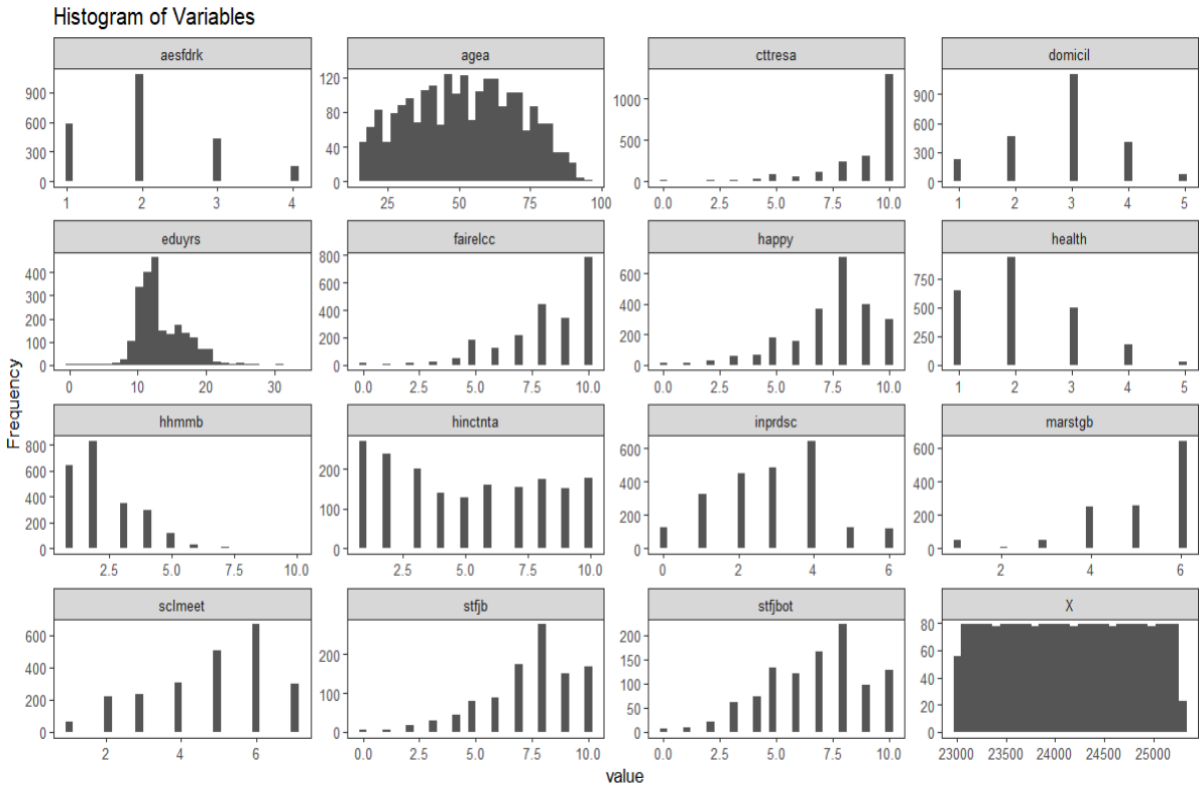
Appendix A: Missing Data after NA imputation



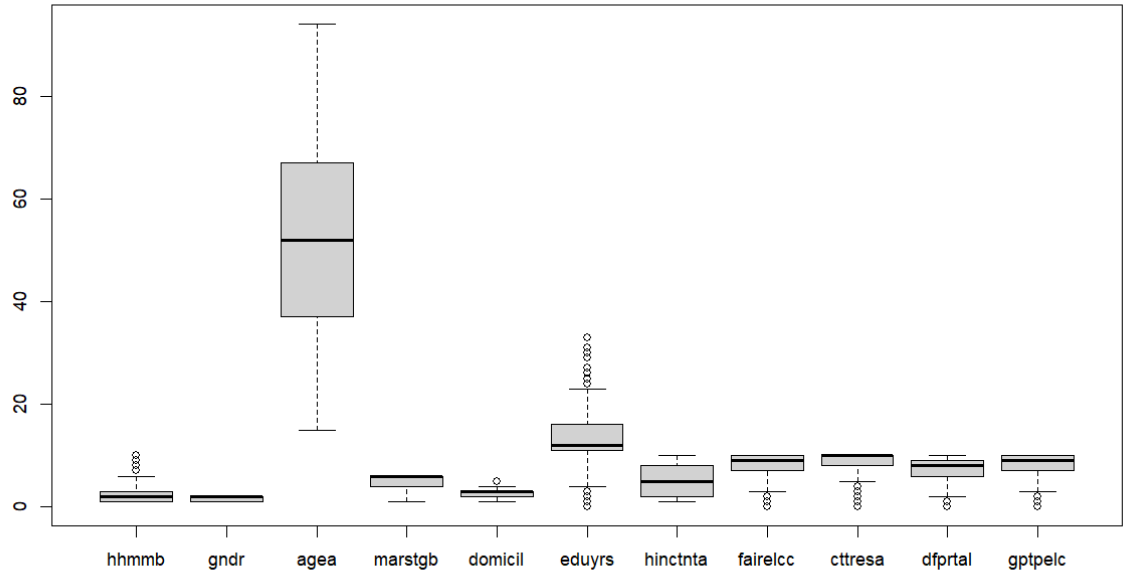
Appendix B: Correlation Matrix



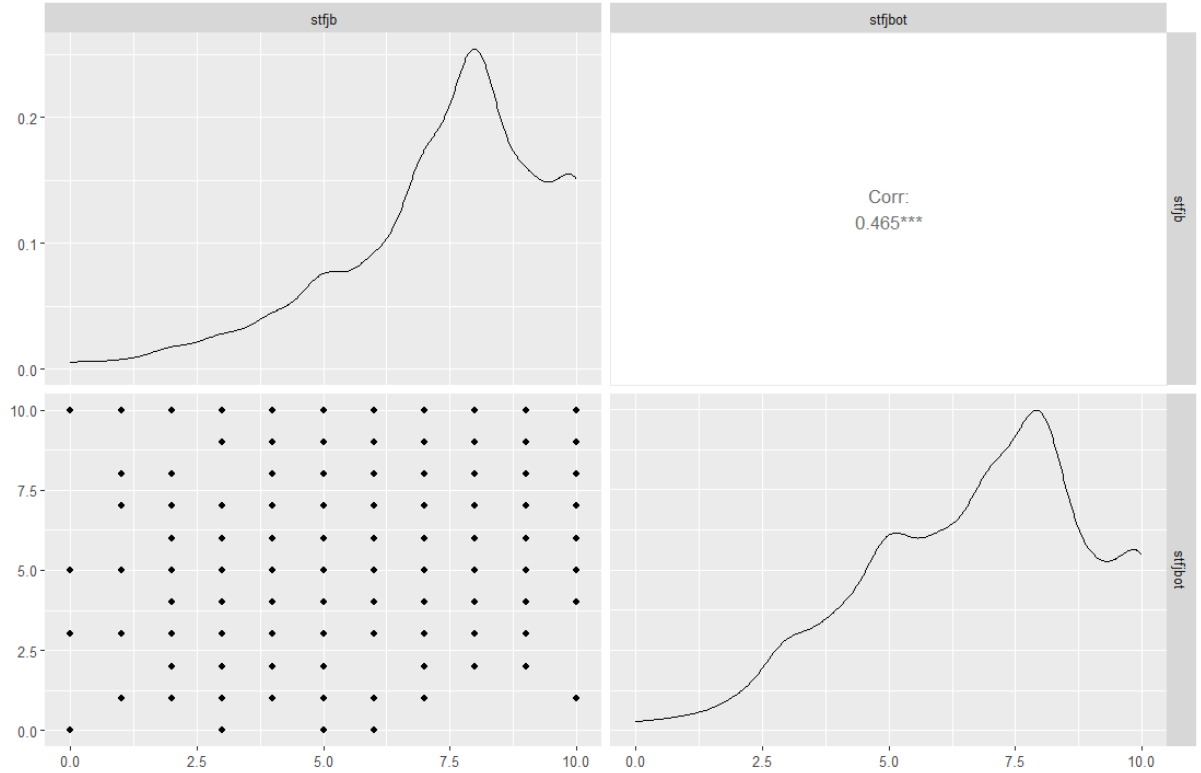
Appendix C: Boxplot of Variables I



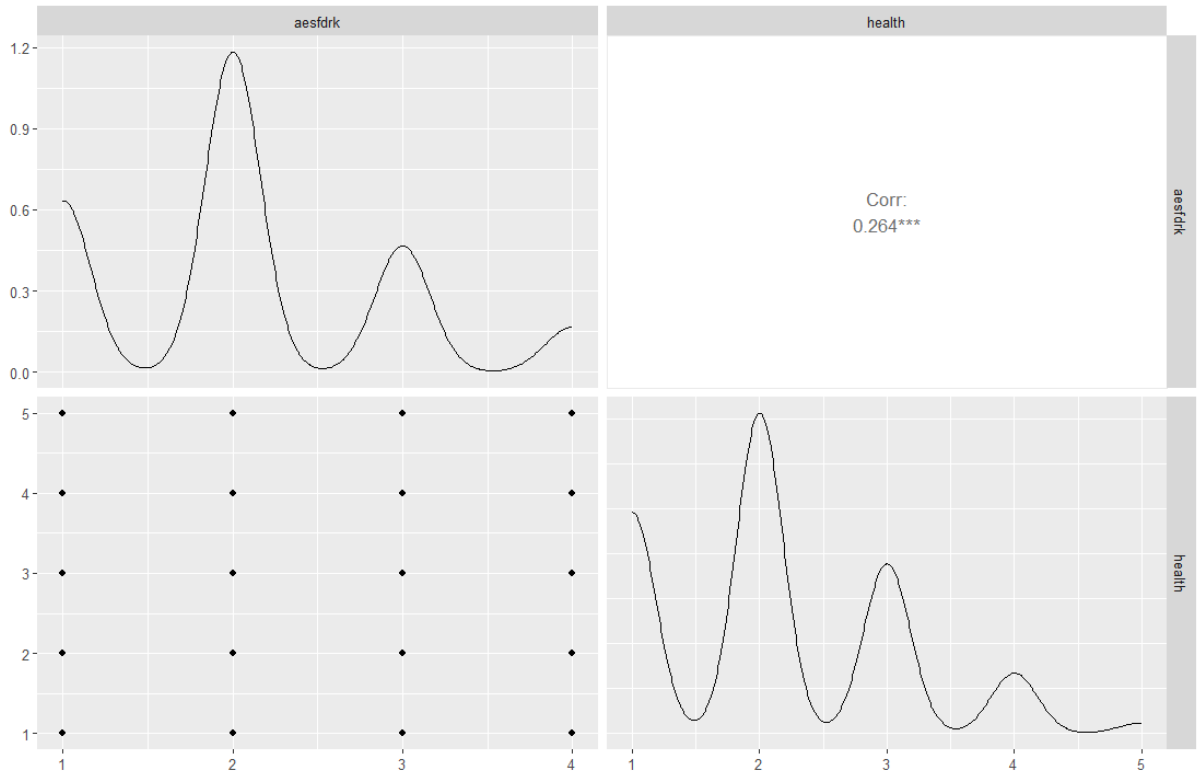
Appendix D: Histograms of Explanatory Variables



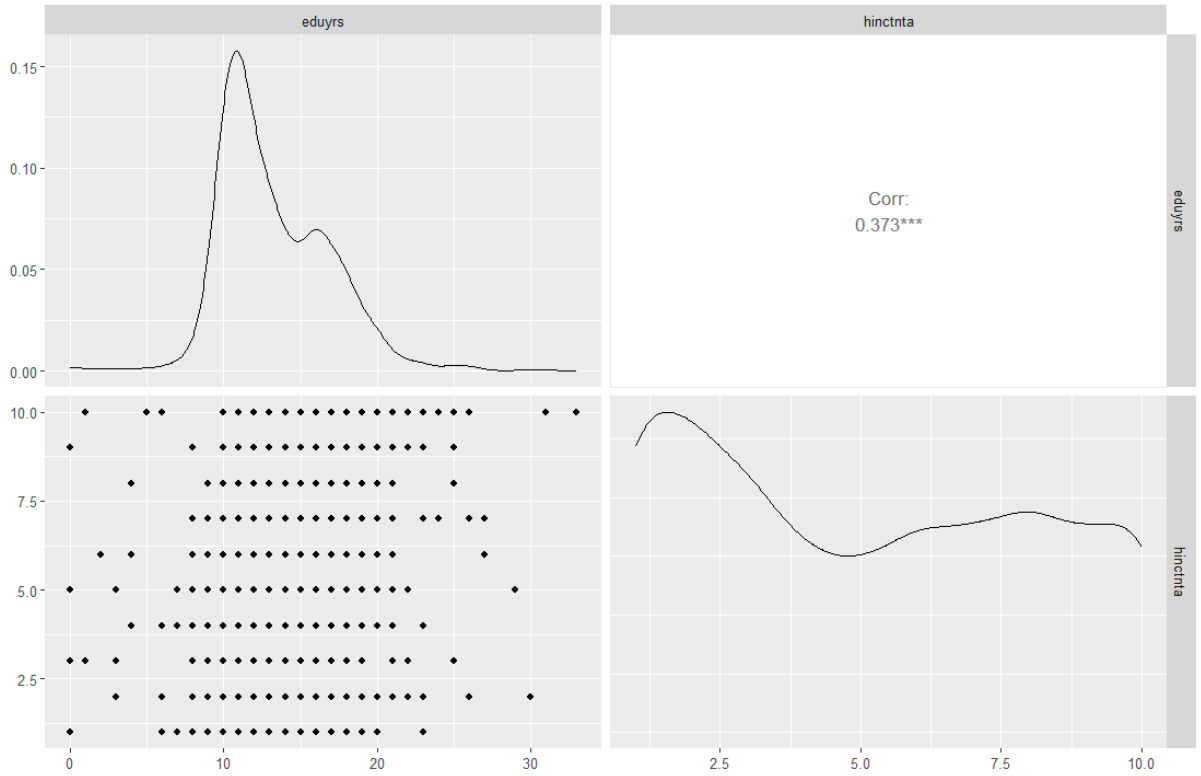
Appendix E: Boxplot of Variables II



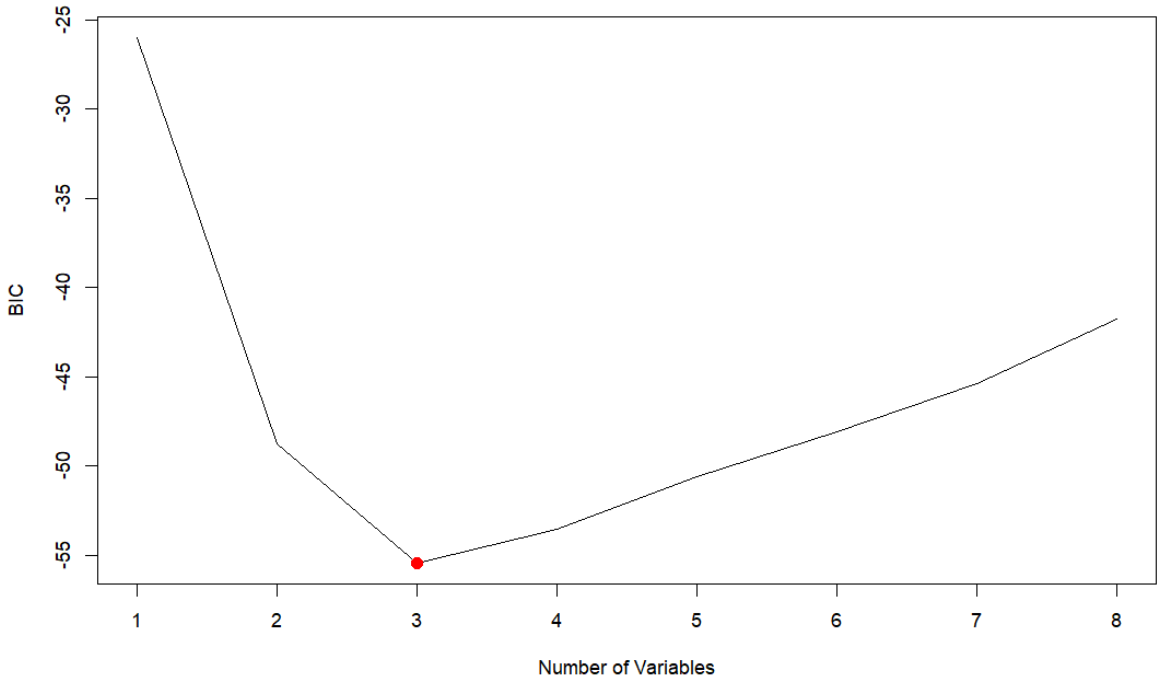
Appendix F: Analysis of Interaction Term I



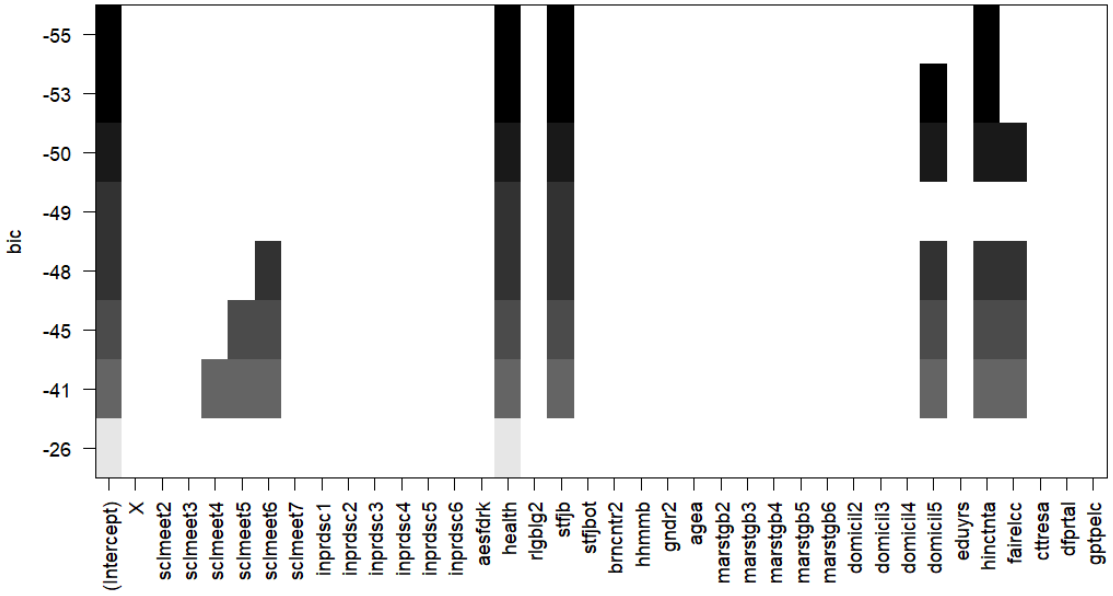
Appendix G: Analysis of Interaction Term II



Appendix H: Analysis of Interaction Term III



Appendix I: Lowest BIC Value from 3 Variable Models



Appending J: Various Model Propositions from BIC