Statistics and Methodology

Project Report

Professor: Dr. Kyle Lang

Group number: 6

Name	Analytical Role	SNR	ANR
Aayushi Pandey	Data Preparation	2038981	u848144
Shreshtha Sharma	Predictive Modelling	2037590	u434770
Nasharetty Wiesken	Inferential Modelling	2011820	u876834
Carolyn Landbrug	Documentation	2045788	u901160

Data Preparation

After preprocessing the data according to the instructions given, we created a subset of data including variables relevant for the inference and prediction tasks. The data provided had five categories of non-response (-1,-2,-3,-4,-5). Out of these, the -3 category denoted "Not Applicable". It wouldn't have made sense for us to impute the missing values for this category, as it didn't exist in the true population in the first place. First, we tried to modify the scale of the variables which had -3 in the data. We replaced -3 with 5 for variables V181(Worries: losing/not finding a job) and V182(Worries: not being able to give one's children a good education) as 5 would indicate absolutely no worry on that scale. It wasn't possible to modify the scale for other variables so we deleted all the rows left with -3. This left us with 12243 observations. The remaining values (-1,-2,-4,-5) were labeled as NA. After this, we performed a univariate outlier analysis using the median absolute deviation method with cut off value 2.5. This method was chosen because it is not sensitive to outliers like the studentized residuals. The analysis was performed only for the age variable (V242) as all the other variables ranged on a fixed scale (eg- 1 to 4 or 1 to 10). We found 6 outliers from the age column and they were labeled as NA. After doing this we evaluated the extent of the missing data problem.

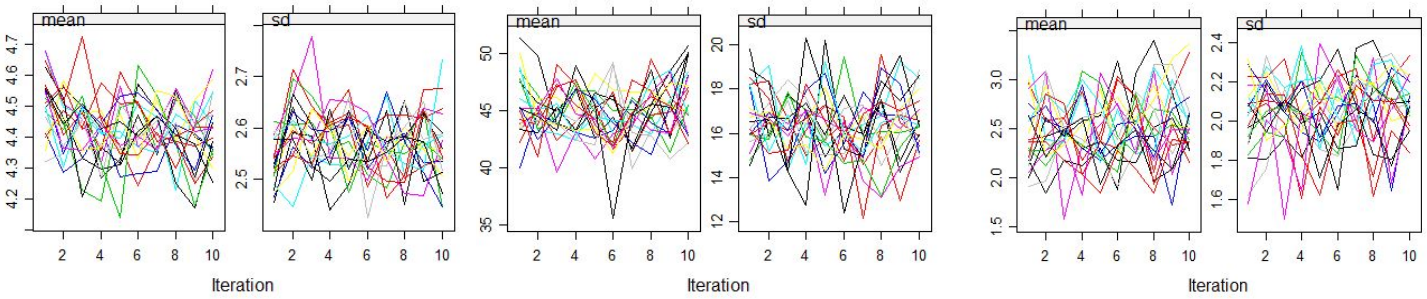
a) Percentage missing from each column

Range	**Mean**	**Median**
0.0245038 to 14.2122029 (Minimum for V240 - Sex, Maximum for V186- Worries: Govt tapping my phone)	4.545813	4.141142

b) Covariance coverages

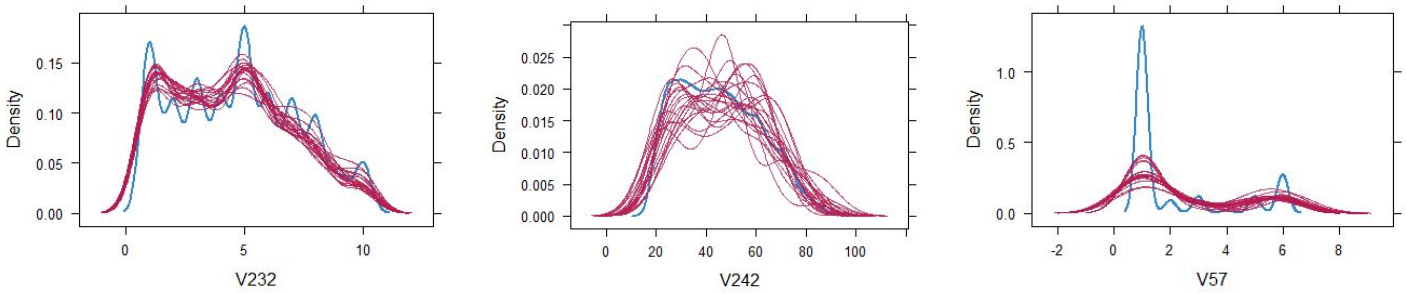
Range	**<=0.8**
0.7797109 to 0.9997550	(V153, V186), (V186, V203) V153: When science and religion conflict, religion is right,V186: Worries: Govt. tapping my phone, V203: Justifiable: Homosexuality

We decided to treat the missing data using multiple imputation because it models both the uncertainty in predicted values and the imputation model itself. This leads to more accurate estimates of standard errors, confidence intervals and prediction intervals. Pmm (predictive mean matching) was chosen as the method of imputation for all variables except categorical ones: sex (logistic regression, V240), do you live with your parents (logistic regression, V250), and marital status (Bayesian polytomous regression, V57). This is because Pmm produces more plausible estimates than the normal regression method in case the normality assumption is violated (Horton and Lipsitz 2001, p. 246), which is the case here as except the categorical i.e nominal variables all are ordinal variables which are assumed continuous. For the categorical variables, the methods are chosen according to the number of categories. The imputation was carried out using the mice package with 10 iterations and 20 datasets. To get better imputations, the predictor matrix was specified such that for each variable, only the variables having a correlation greater or equal to 0.2 are included as predictors. Since correlation doesn't make sense for categorical variables, gender was also separately specified as a predictor for other variables because it can influence many variables for instance: a university education is more important for a boy than for a girl (V52). After performing the imputations, convergence and sanity checks were performed. For some of the variables, the plots are presented in Figures 1 and 2.



(a) (b) (c)

Figure 1: Convergence plot for (a) V232 (Nature of tasks: routine vs creative) (b) V242 (Age) (c) V57 (Marital Status)



(a) (b) (c)

Figure 2: Density plot for (a) V232 (Nature of tasks: routine vs creative) (b) V242 (Age) (c) V57 (Marital Status)

After obtaining the multiply imputed datasets, we checked for multivariate outliers using Mahalanobis distance as a metric (quantile - 99.9%). The categorical variables were excluded during testing. 1642 observations were labeled as multivariate outliers in at least one dataset. But it does not make sense to treat those observations as outliers that are flagged in fewer datasets. Hence, in the subsequent sections we have reported our analyses with and without removing the observations which are marked as multivariate outliers in at least 15 out of 20 datasets (there were 964 such observations in total).

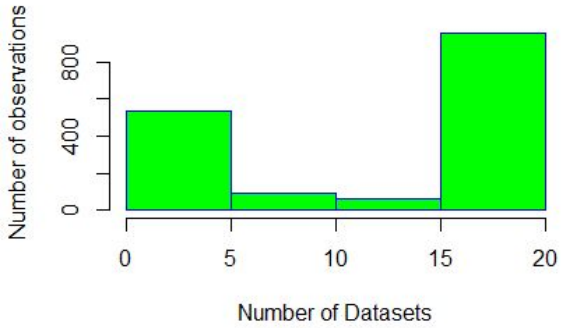


Figure 3: Frequency distribution of the number of datasets in which an observation was flagged outlier

Inferential Modelling

We chose the question: *Are conservative attitudes good or bad for your psychological well-being?* We used the variable V10 (feeling of happiness) as the dependent variable which simulates the psychological well-being. Conservatives attitudes have a wide range of themes such as political, religion, social issues, etc. To narrow down the analysis we only covered social conservatism. The following variables were used as social conservative predictors: V45, V47, V203-210. Moreover, to make a model that does take into account other confounding variables, sex V240, age V242, education V248 and V11 state of health were included to control for the sample characteristics. Model comparison was done to elicit the difference between the two models and finally answer the research question.

EDA was performed to facilitate variable selection. Significant correlation was found between happiness and education (cor: -0.07947108, t = -8.7474, df = 12039, p-value < 2.2e-16), happiness and health status (cor:0.4183085, t = 50.535, df = 12041, p-value: <2.2e-16), happiness and age (cor: 0.09920828 , t = 10.937, df = 12034, p-value < 2.2e-16). These variables were included in the full model. Some social conservative attitudes appear to be related to sex (figure 4). Hence, the interactions term V240*V47 and V240*V45 were also included in the full model.

(a) (b)

Figure 4: Barplot (a) V240 and V45, (b) V240 and V47

Model comparison

We report results of multiple linear regression using multiply imputed datasets with multivariate outliers removed. Coefficients obtained from model 1 tell us what is the effect of the social predictor(i) on the level of happiness after controlling for other social predictors. This model explains 2.7% of the variation in the level of happiness. A test was performed using the intercept-only model to ascertain whether the R2 value of model 1 is significantly greater than 0. We found that predictors explained a significantly greater proportion of variability in levels of happiness (F[df1= 10, df2= 34179.61]= 19.98763, p < 0.001) compared to an intercept only model.

Model 1 is nested in model 2 (which includes additional variables and interaction terms). The full model explained 19.1% of variation in the level of happiness. The predictors explain a significantly greater proportion of variability in levels of happiness (F[df1= 16 , df2= 68889.98]= 125.1444, p < 0.001) compared to an intercept only model.

The additional variation explained by model 2 relative to model 1 is equal to $\Delta R2 = 0.164$. We tested if $\Delta R2=0.164$ represents a significantly greater degree of explained variation. The increase in R2 turned out to be significant (F[df1=6,df2=42430.45] = 325.221, p<0.001).

Model	Predictors
Model 1	**Social conservative attitudes:** V45(When jobs are scarce, men should have more right to a job than women), V47(If a woman earns more money than her husband, it's almost certain to cause problems), V203(Justifiable: Homosexuality), V204(Justifiable: Abortion), V205(Justifiable: Divorce), V206(Justifiable: Sex before marriage), V207(Justifiable: Abortion), V208(Justifiable: For a man to beat his wife), V209(Justifiable: Parents beating children), V210(Justifiable: Violence against other people)
Model 2	Model 1 + **Demographic variables:** V240(Age), V242(Sex), V248(Highest educational level attained) **Individual characteristic:** V11(State of health) **Interaction terms:** V240*V47(Men have more right to jobs when scarce conditional on male=0), V240*V45 (women earning more cause household issues conditional on male =0)

Table 1: Models used for inferential modelling

	Estimate	Std.Error	t Statistic	Df	P value
Intercept	1.2564103815	0.0354631858	35.4285819	4455.4176	0.000000e+00
V45	-0.0130401133	0.0102686690	-1.2698932	4510.7182	2.041882e-01
V47	-0.0265716475	0.0107883597	-2.4629924	2141.3082	1.385645e-02
V203	-0.0125765629	0.0026514178	-4.7433350	3984.8763	2.175877e-06
V204	0.0037802319	0.0034071181	1.1095101	954.9437	2.674895e-01
V205	0.0152950354	0.0035043748	4.3645547	2111.9367	1.335260e-05
V206	0.0086638700	0.0029786933	2.9086143	1305.3364	3.691794e-03
V207	0.0027589604	0.0034348405	0.8032281	4395.5450	4.218863e-01
V208	0.0245839070	0.0059594761	4.1251792	2260.9532	3.838466e-05
V209	-0.0008305122	0.0032513956	-0.2554325	5037.9274	7.983995e-01
V210	0.0118042762	0.0051285910	2.3016607	5427.2624	2.139201e-02
V242	-0.0014772698	0.0003694932	-3.9980973	7221.2478	6.448999e-05
V248	-0.0119814775	0.0024443389	-4.9017252	9783.0329	9.652466e-07
V2402	-0.1109586601	0.0386846702	-2.8682850	3481.6805	4.151864e-03
V11	0.3190791390	0.0072170629	44.2117719	8755.4286	0.000000e+00
V47:V2402	0.0158916823	0.0151836295	1.0466326	2265.8393	2.953807e-01
V45:V2402	0.0350845490	0.0144154136	2.4338219	1993.1494	1.502767e-02

Table 2: Statistical Results for model 2

Conclusion

From model 2 output (table 2), we can infer that the effect of social conservative predictors V205, V206, V208, V210 is statistically significant at significance level 0.05 and affects the psychological well being in a positive way. The effect of social conservative predictors V47, V203 is statistically significant at significance level 0.05 and affects the psychological well being in a negative way. The social conservative predictors V45, V204, V207, V209 don't have a statistically significant effect on psychological well being at significance level 0.05.

[Note - In our submission, we have included code that can be used to generate the results for multiply imputed data without multivariate outliers removed. The test statistics change but the high-level results remain the same (except that V210 does not have a statistically significant effect). We don't state them here due to word limit constraint.]

Predictive Modelling

For predictive modelling we chose Satisfaction with life (V23) as the outcome variable and used multiple linear regression for building the model. The dependent variable ranges on a scale of 1 (completely dissatisfied) to 10 (completely satisfied). EDA was conducted to facilitate variable selection. We found out that the majority of the survey participants were married, hence it made sense to include some predictors reflecting family life satisfaction. We also included other potential satisfaction indicators and demographic variables as predictors (Table 3).

All the variables considered were included in the full model. For the restricted model, we only included the variables having a significant (at 0.01 level of significance) and higher (greater than 0.2) correlation with the dependent variable. In model 3, some interaction terms were added to the full model. The choice was made based on the results from EDA (see table 4 and figure 5). In model 4, the interaction terms were added to the restricted model.

Hold out cross-validation can be sensitive to the training-validation split. To overcome this, we chose 10-fold cross-validation for evaluating the performance of our candidate models. We first evaluated our results on the multiply imputed data after removing multivariate outliers. The cross-validation errors for the four models were 2.387269, 2.410522, 2.369497, 2.383834 respectively. We selected the third model (full model + interactions) as our final model as it had the lowest cross-validation error. The test set prediction error for our final model was MSE = 2.351383. We concluded that adding interaction terms to the full set of variables boosted prediction performance.

For completeness, we also computed the results on multiply imputed datasets before removing multivariate outliers. The cross-validation errors for the four models were 2.646790, 2.663177, 2.622420, 2.629105. The third model had the lowest error for this case too. The test set prediction error for this model was MSE = 2.511924.

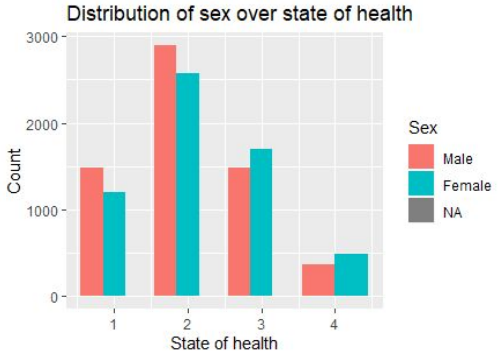


Figure 5: Men appear to be more on the healthy side (1 represents most healthy)

Model	Predictors
1. Full Model	**Family life:** V59 (Satisfaction with financial situation of household, V102(How much do you trust: your family), V182(Worries: not being able to give one's children a good education), V188(Family gone without enough food to eat), V190(Family gone without needed medicine/medical treatment), V191(Family gone without a cash income), V189(Family felt unsafe from crime in own house) **Individual characteristics:** V10 (Feeling of happiness), V11 (State of health), V55 (Freedom of choice and control over life), V75(Being very successful is important to this person), V104(How much do you trust: people you know personally), V143(Thinking about meaning and purpose of life), V233(Nature of tasks: independence), V232(Nature of tasks: routine/creative) **Society:** V238(Social class), V170(Secure in neighbourhood) **Demographic variables:** V240 (sex), V242(age), V248(education), V57(marital status) **Other:** V181(Worries: losing/not finding a job)
2. Restricted model (Linear correlation > 0.2)	**Family life:** V59 (Satisfaction with financial situation of household, V188(Family gone without enough food to eat), V190(Family gone without needed medicine/medical treatment), V191(Family gone without a cash income) **Individual characteristics:** V10 (Feeling of happiness), V11 (State of health), V55 (Freedom of choice and control over life, V233 (Nature of tasks: independence) **Society and country:** V238(Social class), V170(Secure in neighbourhood) **Demographic variables:** V240 (sex), V57(marital status)
3. Full Model + Interactions 4. Restricted Model+ Interactions	**Interactions included:** a. State of health (V11) and age(V242) b. State of health (V11) and gender (V240) c. Financial Satisfaction of household (V59) and education (V248) d. Financial Satisfaction of household (V59) and social class (V238) e. Financial Satisfaction of household (V59) with V188 (Family gone without enough food to eat), V190 (Family gone without needed medicine/medical treatment), V191(Family gone without a cash income)

Table 3: Models used for predictive modeling

Interaction pairs	Correlation	t statistic, degrees of freedom	P value
V11,V242	0.325301	t = 37.912, df = 12145	< 2.2e-16
V59, V248	0.09078382	t = 10.027, df = 12099	<2.2e-16
V59, V238	-0.3128789	t = -35.551, df = 11647	<2.2e-16
V59, V188	0.2207018	t = 24.539, df = 11760	< 2.2e-16
V59, V190	0.2306912	t = 25.661, df = 11715	< 2.2e-16
V59, V191	0.3491218	t = 40.256, df = 11675	< 2.2e-16

Table 4: Correlation test for candidate interaction pairs

References

1. Horton, N. J., & Lipsitz, S. R. (2001). Multiple imputation in practice: comparison of software packages for regression models with missing variables. *The American Statistician*, *55*(3), 244-254.