Statistics and Methodology

Group Project Report

Group #20

Members:

Alvaro Melendez Gutierrez
(SNR: 2047635 | ANR: u203902)

Fernando Catala Resquin
(SNR: 2048042 | ANR: u270800)

Francesco Lauria
(SNR: 2041074 | ANR: u819639)

Konstantinos Soiledis
(SNR: 2037001 | ANR: u226393)

Miftahul Ridwan
(SNR: 2040778 | ANR: u989303)

1 Data Cleaning

1.1 List of variable to be used

In this section, we report the name of the variable, the human-readable description of it, and the question regarding specific variable.

1.1.1 Variables' name

Variable codes, names, types, role and which exercise it is going to be used are reported as follow:

Code	Name	Type	Role	Exercise
V2	Country code	Categorical	Predictor	3 & 4
V10	Feeling of happiness	Continuous	Outcome	3
V11	State of health	Continuous	Predictor	3 & 4
V23	Satisfaction with your life	Continuous	Predictor	4
V57	Marital status	Categorical	Predictor	4
V58	How many children do you have	Discreet	Predictor	4
V59	Satisfaction with financial situation of household	Continuous	Outcome	4
V240	Sex	Categorical	Predictor	4
V242	Age	Continuous	Predictor	4
V248	Highest educational level attained	Continuous	Predictor	4

Table 1: Variable to be used

1.1.2 Variables' Description

1. Country (V2)

   Defines the country of origin of the participant. The countries are coded in a specific numeric form (e.g. China = 156).

2. Feeling of Happiness (V10)

   Measures the participant's perception of own happiness based on a Likert scale from 1 to 4, with 1 being "Very happy" and 4 being "Not at all happy".

3. State of Health (V11)

   The question asked the participant about his conception of his own health and the answers were given based on a Likert scale from 1 to 4, with 1 being "Very Good" and 4 being stated as "Poor".

4. Satisfaction with your life (V23)

   The question explores the amount of satisfaction of the participant. The variable is determined by a Likert scale of 1 to 10 with 1 being completely dissatisfied and 10 being completely satisfied.

5. Marital status (V57)

   The question is measuring if the participant belongs in some form of relationship, or being single. The coding does not follow any specific order regarding the assignment in the status groups. The values are 1) Married; 2) Living together as married; 3) Divorced; 4) Separated; 5) Widowed; and 6) Single..

6. How many children do you have? (V58)

   The variable 'Children' is exploring if the participant has any children. The variable starts from 0 up to possible number of children of the participant, with 0 being used as no children.

7. Satisfaction with financial situation of household (V59)

   Measures the participant's perception of satisfaction with own financial situation of household, on a scale from 1 to 10, with 1 being "Completely dissatisfied" and 10 being "Completely satisfied".

8. Sex (V240)

   Codes respondent's sex by observation. Values are 1="Male", and 2="Female".

9. Age (V242)

   This numeric variable is specifying the age of the participant.

10. Highest educational level attained (V248)

    Specifies the level of education of each participant. The scale used is from 1 to 9 with 1 being the state of no formal education and 9 being the possession of university level degree.

1.1.3 List of questions for each variable (*predictors*)

1. Country (V2):

   (a) Will people in the USA be the happiest compared to other countries in the sample?

   (b) Will people in the USA have higher financial satisfaction than other countries in the sample?

2. State of Health (V11):

   Will people who perceive themselves as healthy feel happier than those who perceive unhealthy?

3. Satisfaction with your life (V23)

   Will people who claim to be satisfied with their lives have higher financial satisfaction than those who claim to be not satisfied?

4. Marital status (V57)

   (a) Will people who are married have higher financial satisfaction than those that remain single?

   (b) Will people who had been married but then became alone (e.g. divorced, separated, widowed) have less financial satisfaction than those that remain married?

5. How many children do you have? (V58)

   Will people who have more children have less financial satisfaction than those who have few children?

6. Sex (V240)

   Will men have higher financial satisfaction than women?

7. Age (V242)

   Will older people have higher financial satisfaction than younger people?

8. Highest educational level attained (V248)

   Will people with higher levels of education attained (e.g. a university degree) have higher financial satisfaction than people with lower levels of education?

1.2 Proportion of missing data for each variable

We ran a column-wise missing data analysis and found no missing data for the subset of the full dataset containing all variables that we will use to answer sections 3 and 4.

1.3 Check for univariate outliers

In this section, we ran univariate outliers analysis, and report the indicies of variable with univariate outliers

1.3.1 Justification of the outlier test chosen

We used the Boxplot method (Tukey, 1977) for all non-categorical variables, since we considered it had the most benefits towards our objective:

- It admits not normally distributed variables;
- It is not sensitive to outliers, unlike other methods, since it does not use the mean nor the dispersion;
- It provides an additional layer of distinction of outlier types: possible vs. probable; and
- It is graphically intuitive to analyze

On the other hand, a key criterion to select an outlier test is a high break-down point. The Boxplot method has a break-down point of 25%, which is not as high as in the Median Absolute Deviation method (50%), however its benefits outweigh this point.

1.3.2 Row indices of outliers

The Boxplot method has an output both possible and probable outliers. However, the ones we are considering and treating are probable outliers. We found possible outliers for V10 (Happiness) and V58 (Children) and probable outliers only for V58 (Children). For the probable outliers, there are 278 rows flagged as probable outliers (*full list of rows reported in Annex 1*).

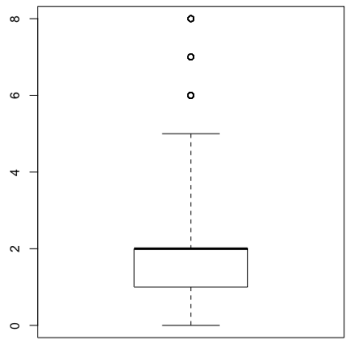

Figure 1: Boxplot for Children (V58)

1.4 Treatment of univariate outliers

After we discovered the variable with univariate outliers, we use Winsorizing treatment to treat it.

1.4.1 Treatment method chosen

The method used for treating univariate outliers is Winsorizing. This method deletes the values that are considered outliers and replaces them with the most extreme legal value that was not considered as an outlier.

In our analysis, values for V58 (Children) above 5, were replaced with 5.

1.4.2 Justification of the outlier test chosen

The reason for using Winsorizing is that we wanted to avoid deleting the values, since this would lead to discarding approximately 278 observations. Even though the sample is large (approximately 3000 rows), we suspected that deleting so many rows would weaken the results in the rest of the project.

1.5 Check for multivariate outliers

In this section, we ran multivariate outliers analysis, report row indicies flagged to be potential mutlivariate outliers, and remove them from the dataset.

1.5.1 Justification of the outlier test chosen

The multivariate outlier test chose is Mahalanobis distance. This test uses a mean distribution of feature values and calculates the distance to it of particular unlikely combinations. It is a generalization of the internally studentized residuals.

1.5.2 Row indices of outliers

We ran robust Mahalanobis Distance analysis and flagged 513 rows of potential multivariate outliers (full list of rows reported in Annex 2).

1.5.3 Exclusion of outlying observations from the data

We remove the rows that flagged as potential multivariate outliers. After removal, the number of rows in the dataset decreased to 12.463 rows.

2 Exploratory Data Analysis

In this section, we ran Exploratory Data Analysis to the dataset to better understand the data we are working with.

2.1 Countries represented in the data

We found out that there are 5 countries in our dataset, namely China (code 156), Germany (code 276), Russia (code 643), United States (code 840), and India (code 356).

2.2 Sample sizes for each country

Code	Country	Sample Size
156	China	2,265
276	Germany	1,998
643	Russia	2,394
840	United States	2,164
356	India	3,822

Table 2: Sample Size of Each Country

2.3 Report of means, medians, standard deviations, and ranges of each continuous variable

Code	Variable Name	Mean	Median	Standard Deviation	Range
V10	Feeling of Happiness	1.88	2	0.64	1 - 4
V11	State of Health	2.17	2	0.84	1 - 4
V23	Satisfaction with your life	6.87	7	2.04	1 - 10
V59	Satisfaction with financial situation of household	5.98	6	2.31	1 - 10
V242	Age	45.01	44	16.26	17 - 94
V248	Highest educational level attained	5.6	5	2.53	1 - 9

Table 3: Descriptive Statistics of each continuous variable

2.4 Frequency tables of each categorical variable

	Country (V2)	Freq
1	China	2,265
2	Germany	1,998
3	India	3,822
4	Russia	2,394
5	United States	2,163

	Marital Status (V57)	Freq
1	Married	8,611
2	Living Together as Married	573
3	Divorced	730
4	Separated	131
5	Widowed	758
6	Single	1,839

	Sex (V240)	Freq
1	Male	6,373
2	Female	6,269

Table 4: Frequency table for categorical variables

2.5 Histograms of each variable

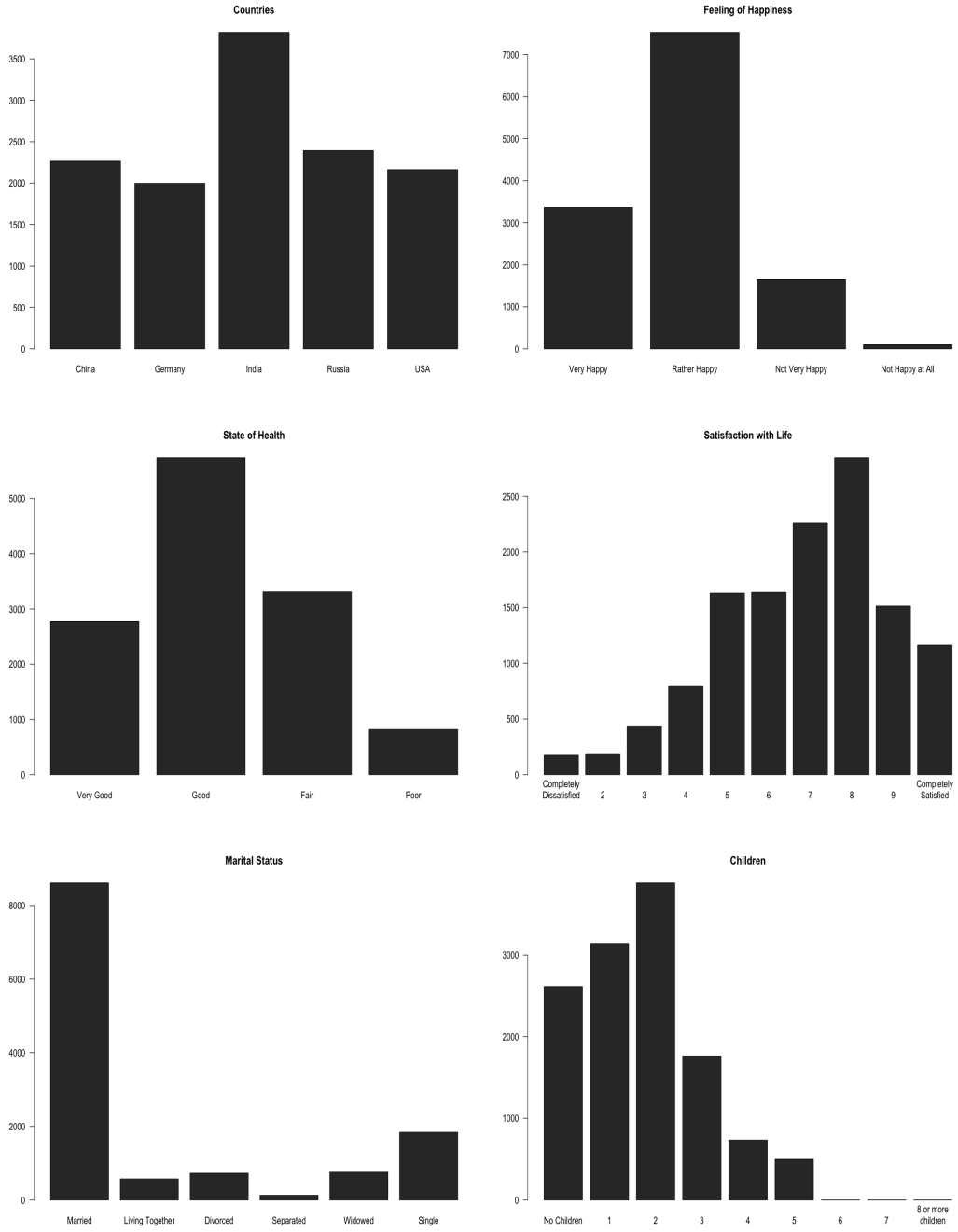

Figure 2: Histogram of Each Variable

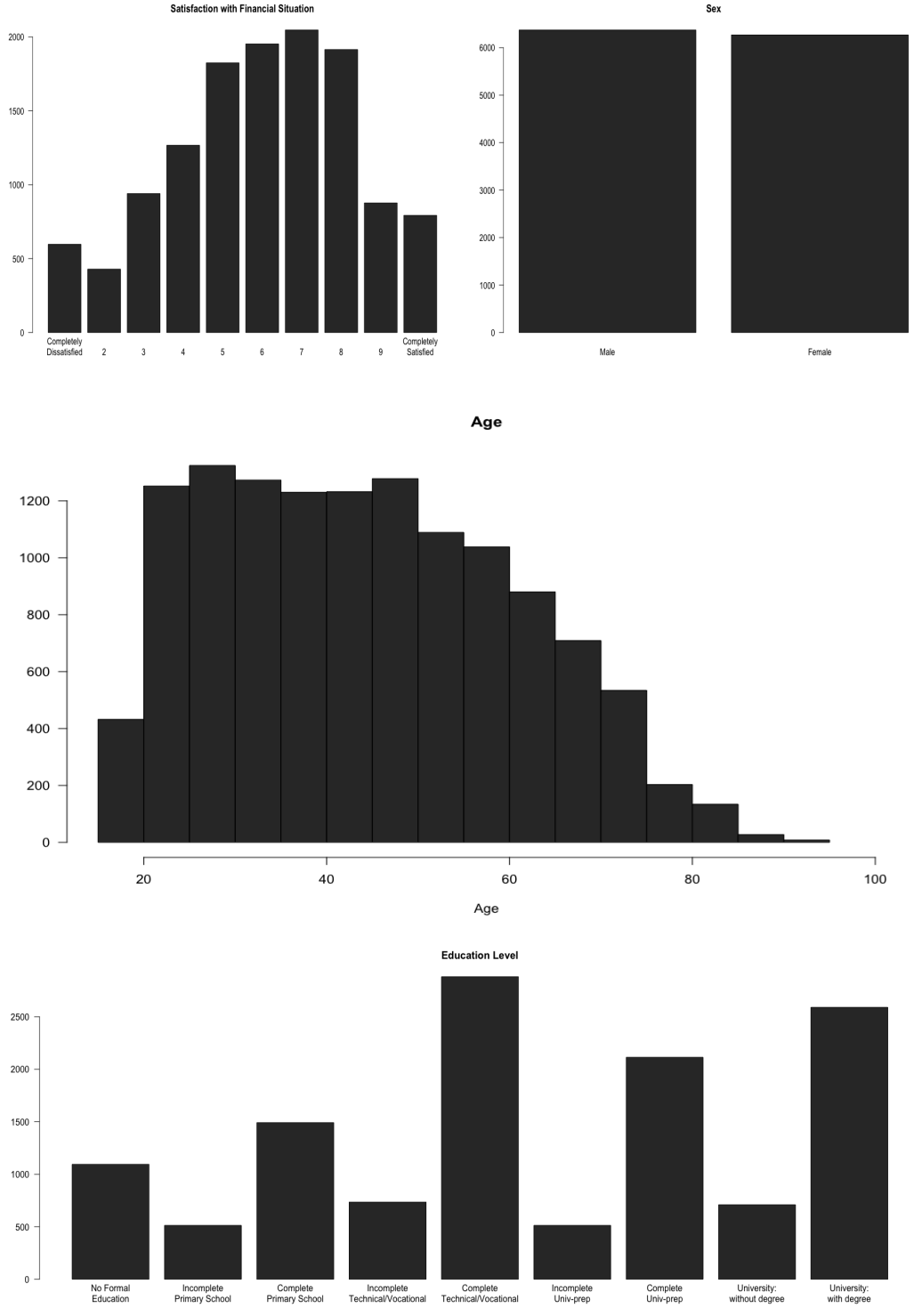

Figure 2: Histogram of Each Variable (*Cont.*)

2.6 Kernel density plots of each continuous variable

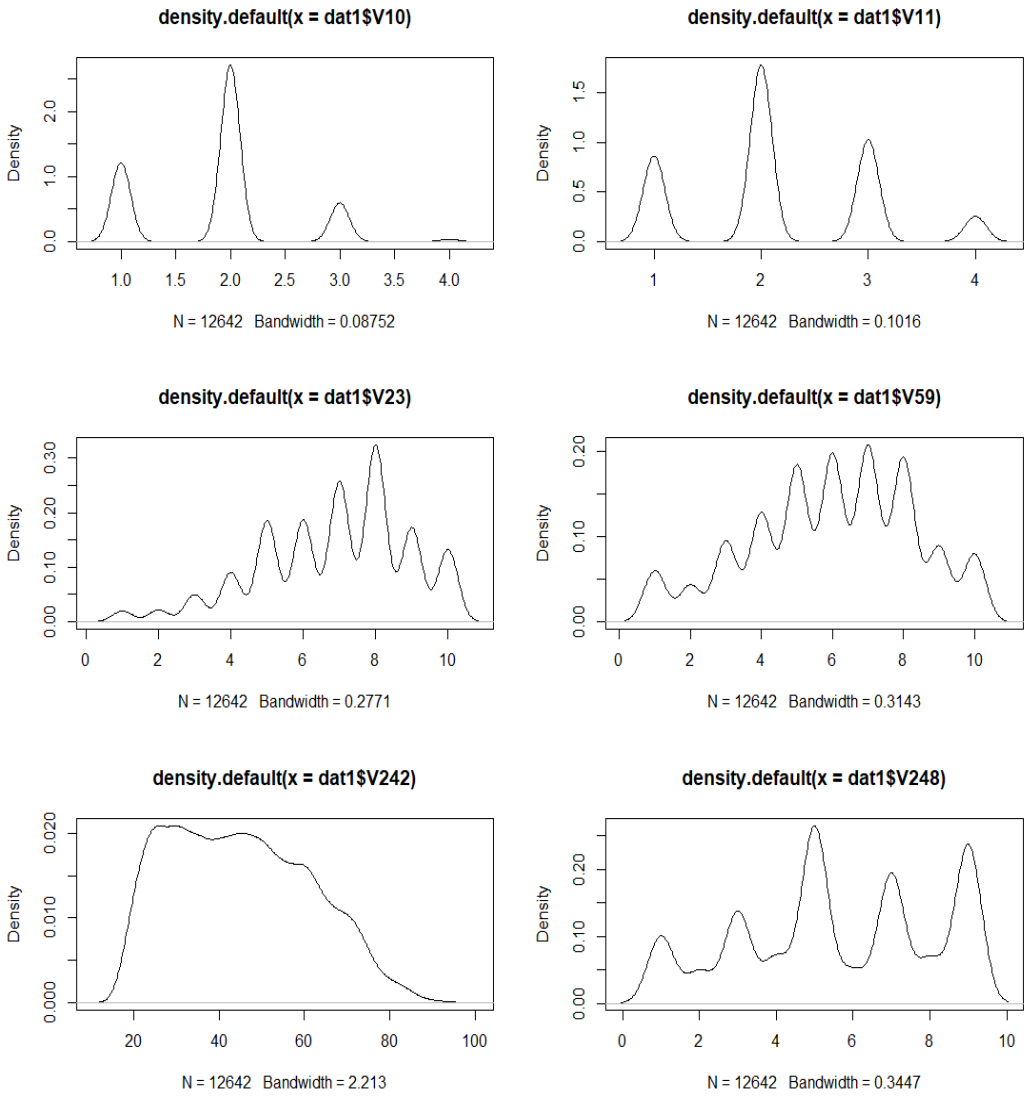

Figure 3: Kernel density plots of continuous variables

2.7 Comments about EDA and potential implications for inferential/predictive analysis

For some variables (V2: Country, V57: Marital status) there is a disproportionately large frequency of observations for one value ('India', 'Married' respectively). Therefore, we expect conclusions to be biased towards respondents to those values.

Also, for all continuous variables we find that the mean value and the median value are quite similar, which by looking into histograms, indicate a rather symmetrical distribution. Regarding V242 (Age), it stands out for its high standard error when comparing to other variables. However, this is explained by its rather even distribution across possible values, which ranges from 17 to 84. This means, conclusions would be rather unbiased regarding age.

In all variables that involved a self-assessment with a scoring scale (V10, V11, V23, V59), we notice that most answers are above the middle-point of the scale. Therefore, overall conclusions would be somewhat biased towards respondents who consider themselves in a good situation respecting those variables.

Furthermore, since the sample of country' variable has very few options (5), we are able to draw conclusions out of them and between them. But unfortunately, we would not feel comfortable with extrapolating results to broader categories (e.g. West vs. East, developed vs. developing, continents).

3 Multiple Linear Regression

In this section, we ran multiple linear regression of Country (V2) and State of Health (V11) effect on Feeling of Happiness (V10).

3.1 Is there a significant effect of country on feelings of happiness (happiness)?

The effect is significant but very small ($R^2 = 0.04951$, F-statistic = 164.6, p-value = < 2.2e-16). On top of that, these are the coefficients:

	β	SE	t-value	p-value
USA	1.729	0.013	128.3	<2e-16
Germany	1.889	0.014	134.6	<2e-16
Russia	2.103	0.012	164.1	<2e-16
China	1.996	0.013	151.5	<2e-16
India	1.753	0.01	172.9	<2e-16

Table 5: Regression output of Country (V2) on Feeling of Happiness (V10)

3.2 The most happy country

The most happy country is United States of America (USA) with $\beta = 1.729$, $SE = 0.013$, $t =$ 128.3, and $p =$ <2e-16.

3.3 The least happy country

The least happy country is Russia with $\beta = 2.103$, $SE = 0.012$, $t =$ 164.1, and $p =$ <2e-16.

3.4 Is there a significant effect of subjective state of health (V11) on happiness after controlling for country?

Yes ($R^2 = 0.206$, F-statistic = 655.7, p-value = < 2.2e-16). On top of that, these are the coefficients:

	β	SE	t-value	p-value
USA	1.125	0.017	65.1	<2e-16
Germany	1.226	0.018	66.38	<2e-16
Russia	1.284	0.02	63.73	<2e-16
China	1.322	0.01	73.08	<2e-16
India	1.106	0.015	69.4	<2e-16
Health (V11)	0.312	0.006	49.86	<2e-16

Table 6: Regression output of Country (V2) and State of Health (V11) on Feeling of Happiness (V10)

3.5 The most happy country after controlling for health

The most happy country, after controling for State of Health is India with $\beta = 1.106$, $SE = 0.015$, $t = 69.4$, and $p = <$2e-16.

3.6 The least happy country after controlling for health

The least happy country, after controling for State of Health is China with $\beta = 1.322$, $SE = 0.02$, $t = 73.08$, and $p = <$2e-16.

3.7 How country-specific levels of happiness change after controlling for health?

We observe that the effect of health is significant. Therefore, it has an effect in country-specific levels of happiness. After controlling for health, all levels of happiness appear to be higher. In particular, now we notice that India replaced the USA in having the highest levels of happiness, and that China replaced Russia in having the lowest levels.

On top of that, after adding some intuitive questions to further understand these results could be: Do well-known high pollution levels in China have triggered these results? Do high levels of income in the USA come at the cost of an unhealthy lifestyle? This type of questions remains pendant for a future study.

4 Predictive Modeling

In this section, we are trying to build a linear regression models to predict satisfaction with the financial situation of their household (*FinSat*).

4.1 Selection of three non-nested sets of predictors to use in predicting Financial Satisfaction

First of all, we set a baseline model which we would compare the others' performance against. This baseline model included the prediction of our target variable V59 (Financial Satisfaction) with V242 (Sex). Then, for each of our models, we would include the pairs of variables, updating our baseline model adding the subsequent pair sets.

We selected three sets of paired values, that we would use as predictors together with Sex. These sets are:

$$\hat{Y}_{FinStat} = \beta_1\, Satisfaction\, with\, Life + \beta_2\, Country + \beta_3\, Sex + \epsilon \tag{1}$$

$$\hat{Y}_{FinStat} = \beta_1\, Children + \beta_2\, Marital\, Status + \beta_3\, Sex + \epsilon \tag{2}$$

$$\hat{Y}_{FinStat} = \beta_1\, Age + \beta_2\, Education\, Level + \beta_3\, Sex + \epsilon \tag{3}$$

As we have multiple predictors, the algorithm used for prediction is Multiple Linear Regression.

4.2 Rationale for choosing the three sets of predictors

Model	Rationale
1	Men and women have different perceptions of financial satisfaction according to their country and to their perceived satisfaction with life. (e.g. unsatisfied woman in Germany vs. satisfied man in India)
2	Men and women have different perceptions of financial satisfaction at the presence of children and given their marital status. (e.g. single man with no children vs. divorced woman with several children)
3	Men and women have different perceptions of financial satisfaction according to their age and to their education level. (e.g. older highly educated man vs. younger uneducated woman)

4.3 Report of the cross-validation error (CVE) from each model

Model	CVE
1	3.536
2	5.213
3	5.262

4.4 Best performing model

According to the results of the 10-fold cross-validation, the model that performed the best is Model (1) with MSE: 3.536

4.5 Estimated prediction error of the best performing model

The estimated prediction error of the best performing model is 3.580.

Annex 1

Row indices flagged to be **Univariate Outliers**

67	374	588	974	1116	1554	1576	1577	1606
1638	2402	2410	2545	2610	2795	2859	2911	2922
2974	3236	3339	3630	3943	3998	4579	4596	5030
5385	5423	5447	5670	6383	6447	6567	6619	6623
6984	7018	7036	7095	7096	7169	7185	7234	7254
7263	7303	7309	7336	7340	7342	7410	7442	7470
7645	7697	7699	7757	7770	7884	7913	8006	8025
8028	8039	8068	8110	8124	8171	8180	8214	8219
8250	8299	8312	8349	8400	8494	8564	8567	8585
8615	8634	8731	8777	8810	8818	8823	8836	8837
8906	8932	8944	8946	9049	9073	9141	9150	9191
9194	9259	9390	9478	9574	9590	9593	9596	9598
9600	9602	9604	9606	9624	9629	9632	9634	9635
9644	9645	9647	9652	9667	9690	9708	9710	9712
9716	9734	9832	9841	9852	9895	9901	9905	9936
9955	10032	10040	10043	10044	10045	10048	10049	10051
10054	10055	10060	10063	10064	10065	10066	10068	10078
10082	10084	10089	10090	10091	10092	10132	10140	10271
10308	10334	10549	10576	10580	10581	10583	10590	10591
10604	10616	10669	10675	10676	10686	10703	10715	10717
10799	10800	10907	10913	10918	11031	11043	11045	11096
11101	11188	11196	11262	11270	11339	11433	11436	11465
11474	11488	11491	11495	11553	11565	11566	11594	11629
11652	11690	11711	11713	11717	11722	11730	11733	11737
11746	11749	11750	11759	11760	11764	11773	11776	11778
11849	11858	11860	11878	11887	11907	11951	12024	12032
12078	12093	12100	12138	12176	12188	12220	12235	12238
12256	12257	12259	12267	12273	12277	12279	12320	12342

Annex 2

Row indices flagged to be **Multivariate Outliere**

14	19	20	23	175	324	416	427	432	440	453	455	466
485	518	522	587	839	845	891	993	1003	1174	1248	1625	1628
1639	1696	1745	1846	1905	1907	2030	2039	2273	2327	2341	2345	2346
2379	2423	2458	2574	2578	2585	2599	2602	2688	2737	2756	2764	2797
2805	2822	2898	2943	3091	3163	3223	3237	3241	3274	3287	3295	3349
3458	3464	3494	3531	3559	3584	3621	3723	3755	3760	3998	4003	4160
4255	4269	4270	4272	4319	4398	4452	4536	4588	4595	4621	4624	4627
4634	4704	4712	4825	4842	4874	4917	4945	4980	4993	5014	5016	5026
5030	5098	5149	5164	5192	5308	5327	5347	5359	5385	5423	5431	5432
5433	5513	5586	5592	5593	5596	5644	5645	5649	5660	5670	5690	5692
5705	5765	5770	5779	5783	5786	5801	5842	5843	5871	5874	5875	5901
5916	5923	5934	5988	6051	6057	6069	6120	6132	6193	6194	6214	6251
6272	6274	6315	6393	6447	6449	6452	6457	6458	6488	6496	6525	6537
6544	6553	6556	6557	6595	6599	6603	6619	6637	6667	6670	6710	6729
6742	6765	6782	6808	6823	6830	6843	6849	6860	6899	7062	7084	7113
7209	7214	7268	7317	7325	7340	7346	7374	7399	7423	7557	7651	7682
7714	7727	7728	7775	7798	7867	7922	7996	8040	8042	8054	8077	8092
8119	8130	8155	8175	8180	8223	8260	8280	8308	8348	8364	8419	8480
8530	8543	8547	8552	8573	8574	8610	8615	8677	8706	8712	8726	8752
8818	8827	8834	8837	8873	8991	8994	9019	9044	9059	9080	9091	9116
9139	9146	9190	9194	9244	9248	9259	9300	9355	9375	9394	9406	9449
9450	9451	9452	9454	9456	9476	9512	9524	9572	9583	9590	9596	9602
9609	9610	9624	9638	9644	9647	9649	9652	9693	9709	9722	9734	9753
9761	9762	9794	9834	9844	9849	9850	9858	9865	9883	9885	9888	9889
9890	9895	9896	9897	9900	9901	9907	9909	9947	9948	9952	9975	9978
9984	9986	9994	9995	9998	9999	10010	10017	10022	10023	10028	10031	10040
10041	10049	10057	10060	10062	10068	10084	10086	10091	10093	10094	10095	10097
10168	10252	10264	10271	10275	10292	10392	10420	10604	10616	10632	10642	10697
10700	10708	10714	10719	10730	10732	10735	10738	10749	10775	10783	10784	10788
10809	10893	10901	10903	10918	10934	10950	10963	10964	10981	10994	11005	11015
11018	11023	11028	11031	11032	11041	11043	11056	11163	11183	11188	11193	11194
11196	11197	11205	11207	11213	11227	11308	11312	11317	11326	11330	11331	11381
11391	11392	11393	11394	11395	11428	11431	11435	11459	11464	11471	11496	11553
11567	11571	11583	11593	11601	11617	11621	11625	11637	11646	11659	11669	11726
11747	11764	11771	11860	11881	11903	11908	11918	11926	11928	11951	11963	11965
11967	11978	12010	12012	12047	12054	12055	12064	12065	12068	12078	12082	12105
12108	12192	12207	12235	12254	12275	12278	12283	12308	12316	12324	12327	12335
12341	12346	12347	12401	12444	12455	12460	12461	12469	12479	12501	12517	12536
12553	12596	12602	12692	12733	12773	12820	12833	12834	12868	12896	13006	13031
13093	13096	13102	13118	13144	13151							