

# **Exploring the Impact of Demographic and Socioeconomic Factors on Individual Life Satisfaction**

IOE 591 Team 10: Szu-Tung Chen, Tzu-Hsuan Chuang, Kaixian Mao, Lucy Lin

## **1. Introduction**

The World Values Survey (WVS) [1] is a global research initiative focused on assessing the impact of the stability of people's social, political, economic, religious, and cultural values worldwide. WVS has been conducted globally every five years since 1981 as the largest non-commercial academic social survey program.

This report aims to utilize the WVS wave 7 (2017-2022) dataset to explore how demographic and socioeconomic factors shape life satisfaction, stratified by respondents' sexes. It includes an exploratory data analysis, delving into the descriptive analysis among 31 predictors representative of demographic and socioeconomic dimensions. Subsequently, we employ linear regression models to show the connections between these predictors and individual life satisfaction. To ensure the robustness of our model, diagnostics, testing-based model selections, and interaction identifications are performed to assess the assumptions and variables. Finally, we conclude the analysis by evaluating the accuracy of our model with root mean squared error.

Two models were proposed with selected predictors including occupational group, family savings, and scale of incomes for females, and highest educational level and family savings for males. Both models incorporate significant interaction terms. The robustness of our models is evident, with three out of four achieving R-squared values exceeding 0.6. Considering potential variance in reporting Likert scale data, differences within 1.5 may not be significant for the response variable. Plotting recorded against predicted response variables, most data points fall within the  $\pm 1.5$  region, reinforcing the models' effectiveness.

The evaluation results of our proposed models demonstrate high robustness. It is also noteworthy that all models incorporate the addition of significant interaction terms, and the reduced models also include predictors that emerge in the interaction terms but were not originally selected. This approach enhances the models' capacities to capture intricate relationships within the data.

## 2. Analysis Procedure

Referring to the WVS Wave 7 Variable Report, the initial set of demographic and socioeconomic variables including a total of 31 (Table I) have been incorporated into our analysis as predictors.

Table I. Demographic and Socioeconomic Variables

Q 260	Sex
Q 261	Birth Year
Q 262	Age
Q 263	Respondent Immigrant Status
Q 264	Mother Immigrant Status
Q 265	Father Immigrant Status
Q 266 - Q 268	Country of Birth: Self, Mom, Dad
Q 269	Citizenship
Q 270	Number of People in Household
Q 271	Live with Parents or not
Q 272	Language at Home
Q 273	Marital Status
Q 274	Number of Children
Q 275 - Q 278	Highest Educational Level: Self, Spouse, Mother, Father
Q 279 - Q 280	Employment Status: Self, Spouse
Q 281 - Q 283	Respondent Occupational group: Self, Spouse, Father
Q 284	Employment: public, private, private non-profit
Q 285	If the chief wage earner: Yes/No
Q 286	Family Savings During Past Year
Q 287	Social Class
Q 288	Scale of Income
Q 289	Religious Denomination
Q 290	Racial Belonging / Ethnic Group

Originally intending to explore life satisfaction across five continents using the dataset collected from 120 countries and societies, we encountered challenges with the dataset's predominantly categorical predictors. Linear regression models were not accurate at this broader continental level. However, upon decomposing the data to country levels, we found that linear regression could provide meaningful insights, yielding a fair R square value.

With our group's base in the United States, we subsequently shifted our project scope to delve into how demographic and socioeconomic factors influence life satisfaction within the US. Due to the uneven distribution of data for other ethnicities in the dataset, we focused on the white ethnic group. This strategy allows a more nuanced analysis tailored to the dataset's characteristics.

Building upon the foundational insights gained from data pre-processing and exploratory data analysis, we proceeded with model diagnostics, model selection, identification of significant interactions, and levels aggregation to construct models informed by our discoveries.

## 2.1 Data Pre-processing

Data preprocessing was conducted to enhance the quality and suitability of raw data for analysis. We translated category codes into their actual meanings to improve code readability. Rows containing responses labeled as 'No answer,' 'Not asked,' 'Missing,' or 'NA' were excluded from the dataset. This refinement has resulted in our analysis now centering on a dataset encompassing a total of 462 data points.

## 2.2 Exploratory Data Analysis

An exploratory data analysis (EDA) was conducted to gain insights into the WVS wave 7 dataset. We first identified and removed missing or unusable responses, and categorical predictors were transformed into factors. A total of 462 data points remain for our analysis with 191 females and 271 males. The distribution of their responses to life satisfaction (Q49) is illustrated in the boxplot below (Fig. 1). Our subsequent analysis will involve building separate models for each sex group.

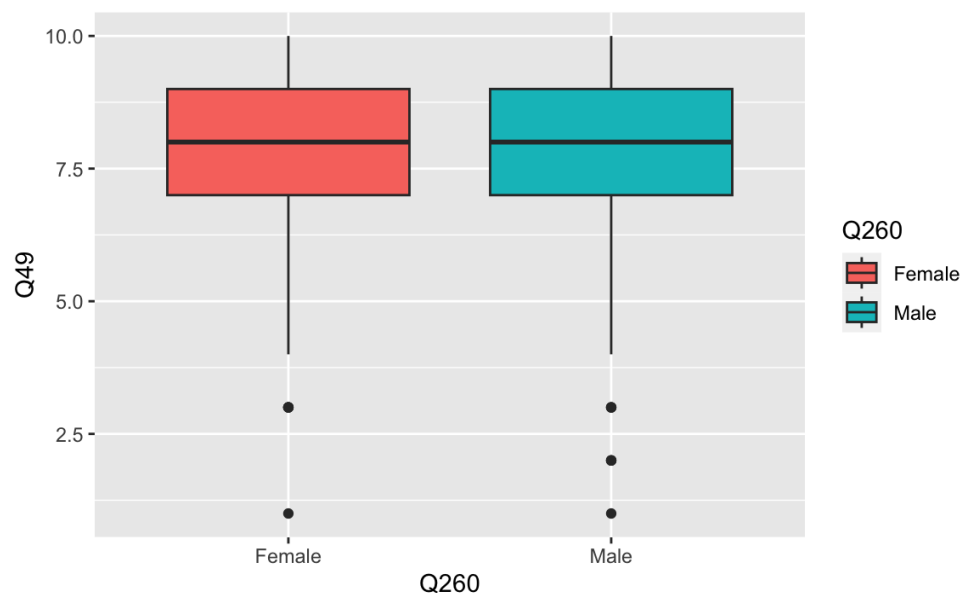
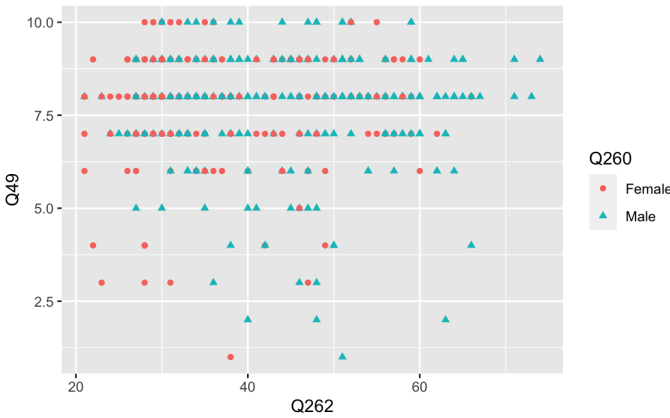
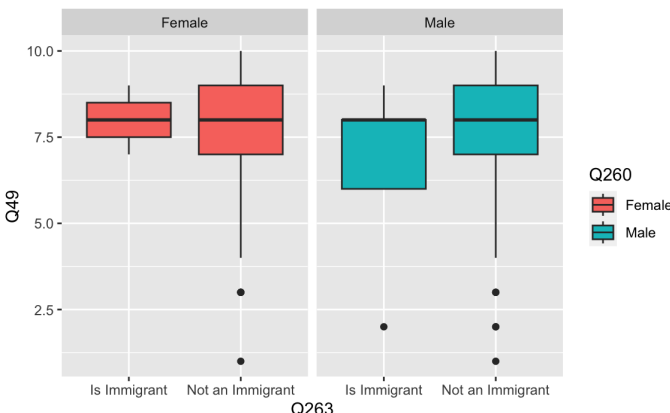


Fig. 1 The boxplot for the responses to life satisfaction among two sexes

Below, we present visualizations that illustrate the distribution of responses for the remaining predictors (Table II), enhancing our understanding of the dataset.

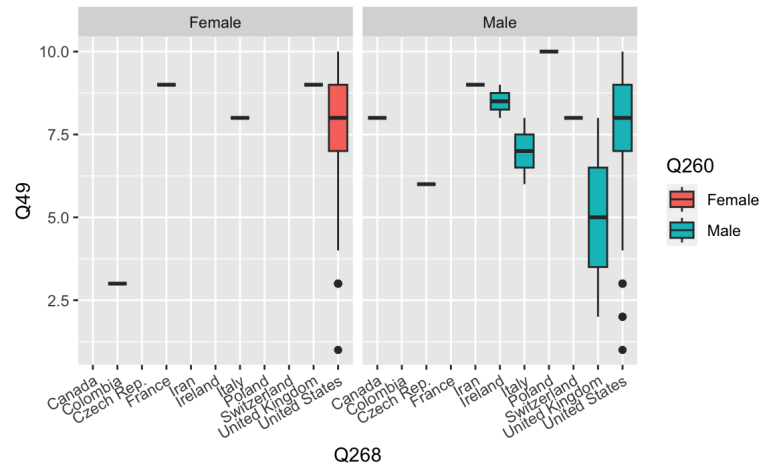
Table II. The distributions of responses for each level of the predictors

	Female	Male	Graph
Age (Q262)			
Min.	21.00	21.00	
1st Qu.	29.00	34.00	
Median	34.00	44.00	
Mean	37.24	44.42	
3rd Qu.	46.00	53.00	
Max.	66.00	74.00	
Immigrant Status (Q263 - Q265, Q269)			
Is Immigrant	3	8	
Not Immigrant	188	263	

Is Immigrant (Mom)	11	15	<p>Q264</p>
Not an immigrant (Mom)	180	256	
Is Immigrant (Dad)	4	15	<p>Q265</p>
Not an immigrant (Dad)	187	260	
Is citizen	190	269	<p>Q269</p>
Not citizen	1	2	

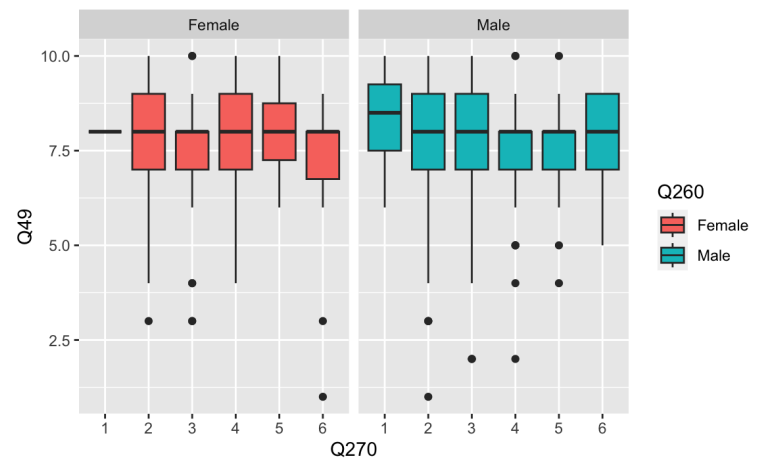
Country of Birth (Q266 - Q268)			
Canada	1	2	<p>Q266</p>
Germany	1	0	
Guinea	0	1	
Italy	0	2	
Spain	1	0	
United Kingdom	0	2	
United States	188	264	
Virgin Island (British) (Mom)	1	0	<p>Q267</p>
Canada (Mom)	0	1	
El Salvador (Mom)	1	0	
Germany (Mom)	3	0	
Iran (Mom)	0	1	
Israel (Mom)	1	0	
Italy (Mom)	1	4	
Mexico (Mom)	0	1	
Russia (Mom)	0	1	
Switzerland (Mom)	0	1	
Turkey (Mom)	0	1	
United Kingdom (Mom)	3	5	
United States (Mom)	181	256	

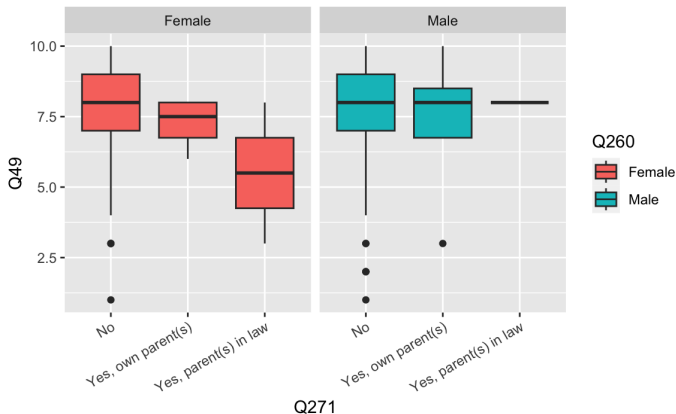
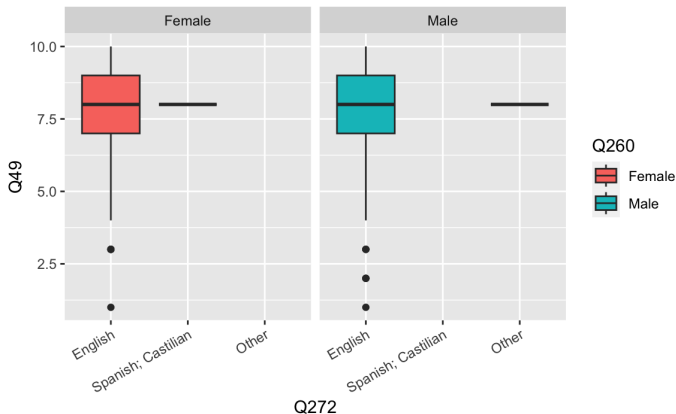
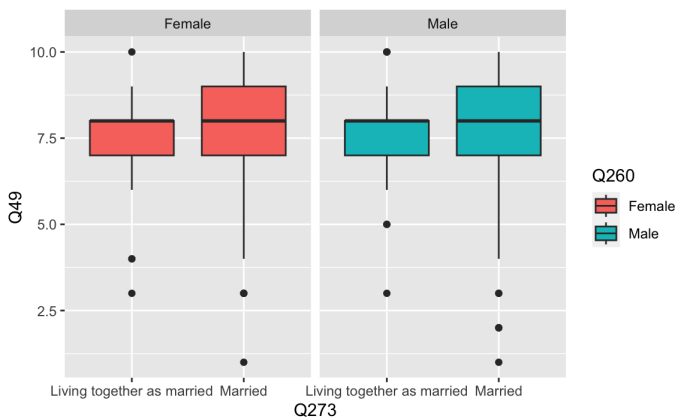
Canada (Dad)	0	1
Columbia (Dad)	1	0
Czech Rep. (Dad)	0	1
France (Dad)	1	0
Iran (Dad)	0	1
Ireland (Dad)	0	2
Italy (Dad)	1	2
Poland (Dad)	0	1
Switzerland (Dad)	0	1
United Kingdom (Dad)	1	2
United States (Dad)	187	260



### Number of People in Household (Q270)

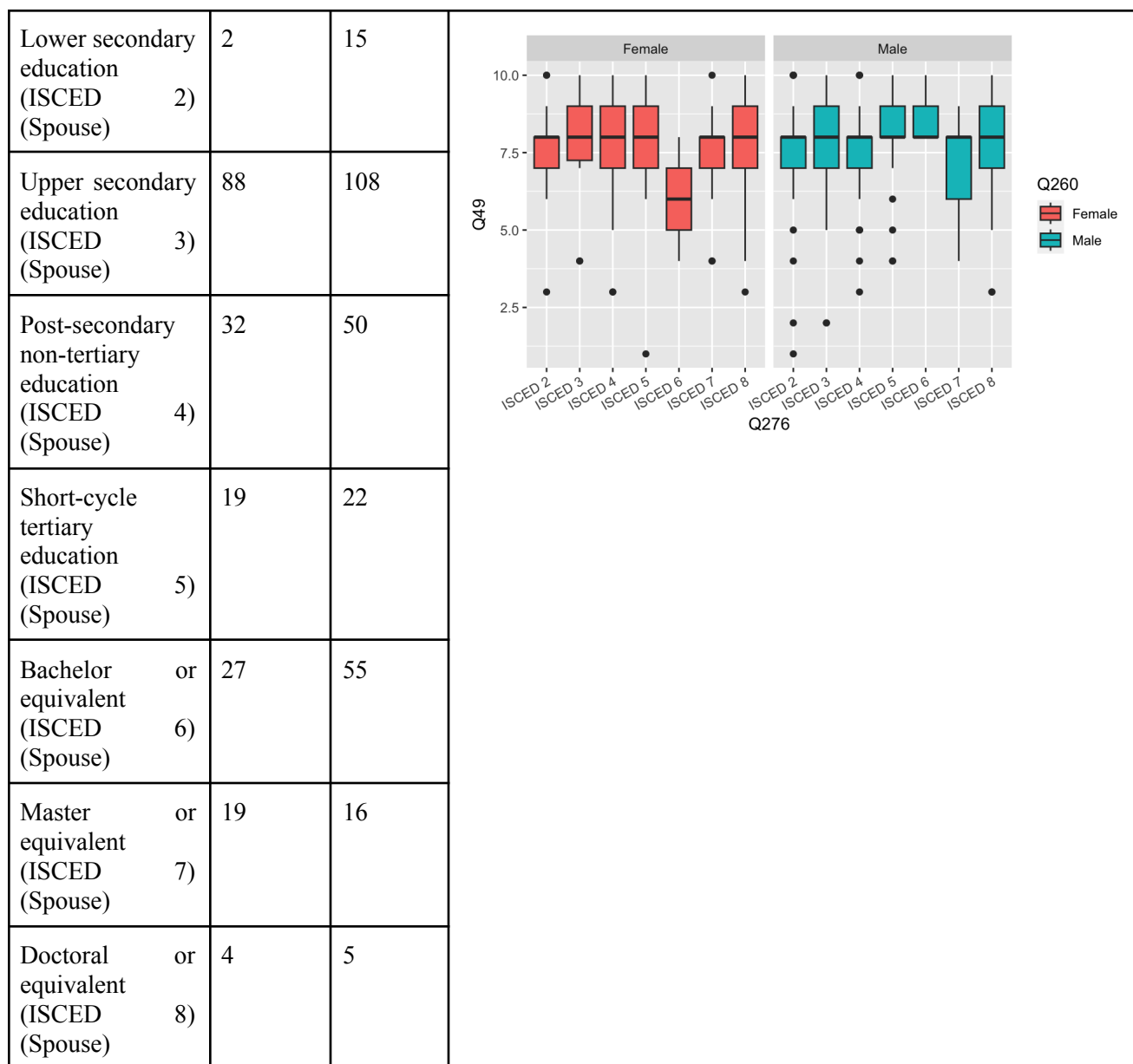
1	2	4
2	78	106
3	38	56
4	43	70
5	18	22
6	12	13

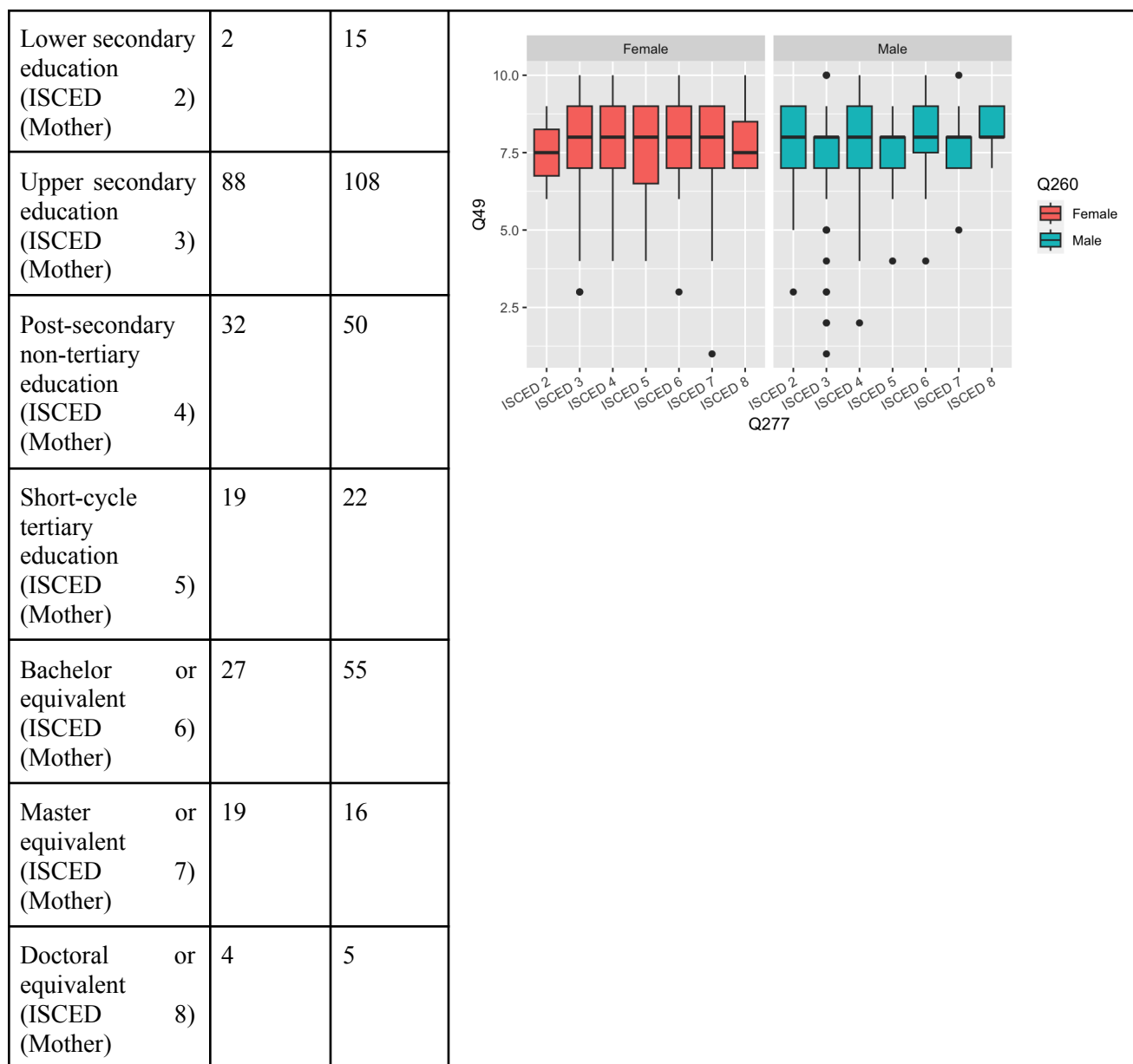


Do you live with your parents (Q271)			
No	185	266	 <p>Q271</p>
Yes, own parent(s)	4	4	
Yes, parent(s) in law	2	1	
Language at Home (Q272)			
English	190	270	 <p>Q272</p>
Spanish; Castilian	1	0	
Other	0	1	
Marital status (Q273)			
Married	161	246	 <p>Q273</p>
Living together as married	30	25	

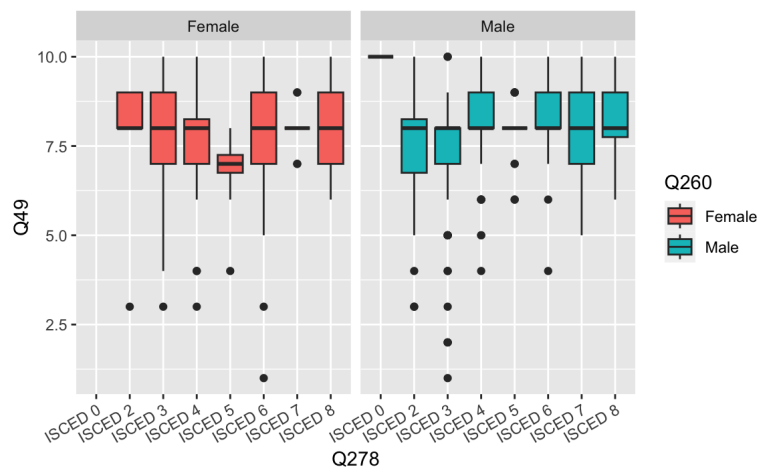


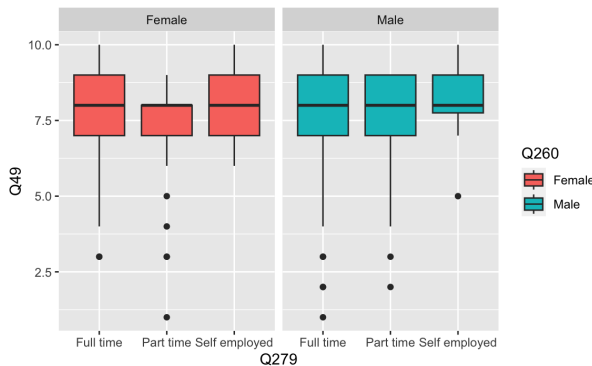
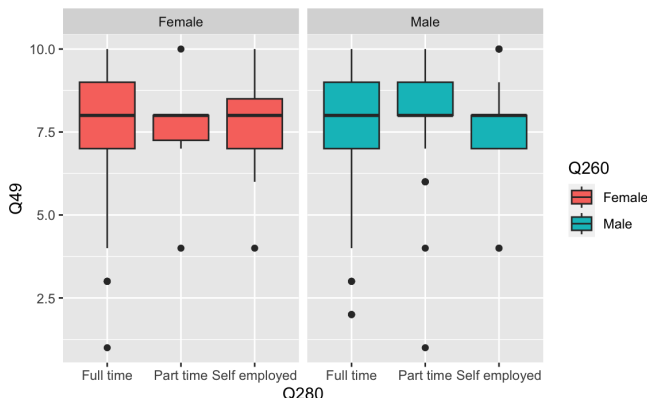
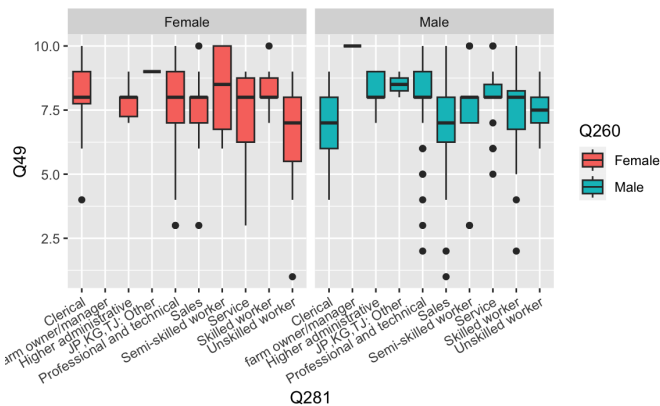
How many children do you have (Q274)			
0	56	57	
1	33	51	
2	65	91	
3	19	47	
4	10	18	
5	8	4	
6	0	3	
Highest educational level: Respondent [ISCED 2011] (Self) (Q275-Q278)			
Lower secondary education (ISCED 2)	0	1	
Upper secondary education (ISCED 3)	39	22	
Post-secondary non-tertiary education (ISCED 4)	36	63	
Short-cycle tertiary education (ISCED 5)	26	24	
Bachelor or equivalent (ISCED 6)	59	96	
Master or equivalent (ISCED 7)	26	38	
Doctoral or equivalent (ISCED 8)	5	27	

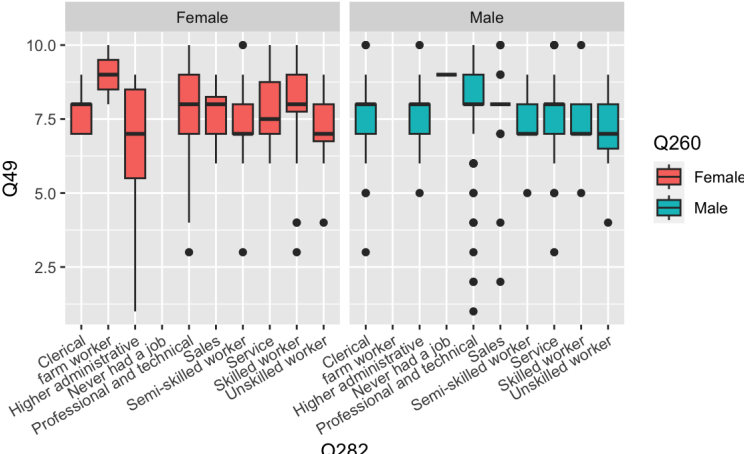
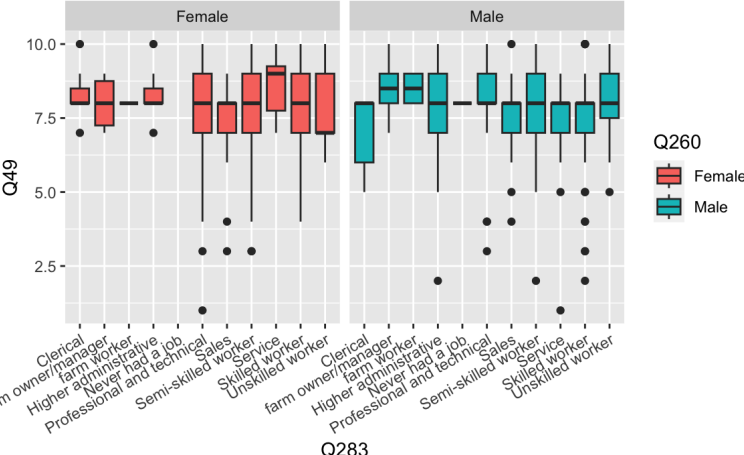




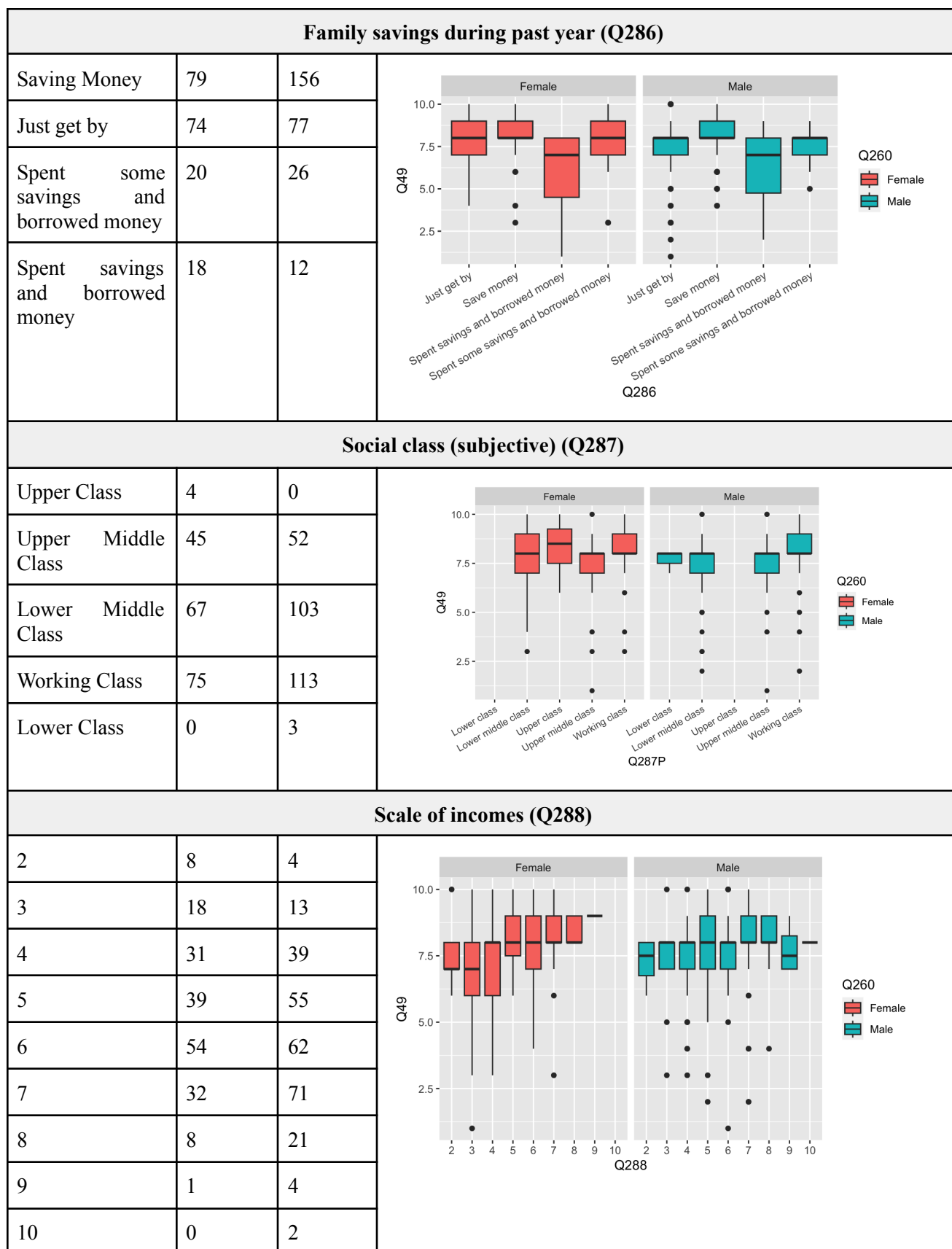
Early childhood education (ISCED 0) / no education (Father)	0	1
Lower secondary education (ISCED 2) (Father)	7	24
Upper secondary education (ISCED 3) (Father)	74	89
Post-secondary non-tertiary education (ISCED 4) (Father)	48	48
Short-cycle tertiary education (ISCED 5) (Father)	8	17
Bachelor or equivalent (ISCED 6) (Father)	30	51
Master or equivalent (ISCED 7) (Father)	9	21
Doctoral or equivalent (ISCED 8) (Father)	15	20



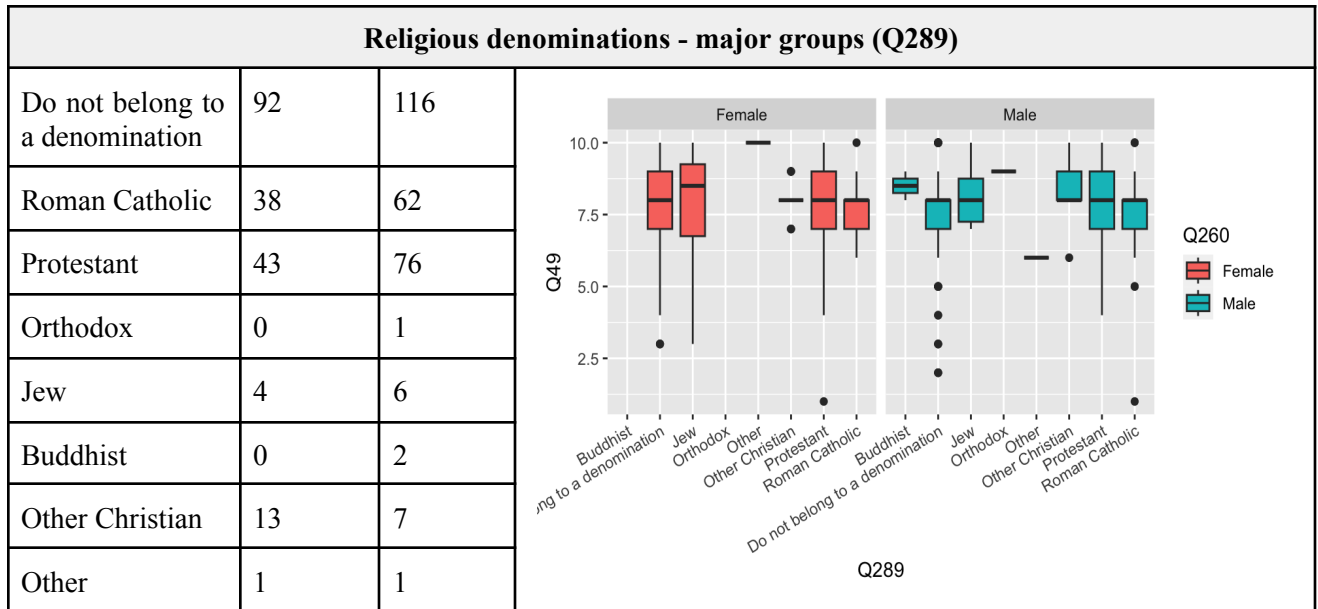
Employment status (Self) (Q279-Q280)			
Full time (30+)	139	231	
Part time (<30)	34	20	
Self employed	18	20	
Full time (30+) (Spouse)	170	204	
Part time (<30) (Spouse)	6	46	
Self employed (Spouse)	15	21	
Respondent - Occupational group (Q281-Q283)			
Professional and technical	84	117	
Higher administrative	10	18	
Clerical	32	13	
Sales	21	30	
Service	26	27	
Skilled worker	6	40	
Semi-skilled worker	4	13	
Unskilled worker	7	10	
Farm	0	1	

owner/manager			 <p>Q260</p> <p>Female</p> <p>Male</p> <p>Q282</p>
JP, KG, TJ: Other	1	2	
Clerical (spouse)	9	45	
farm worker (spouse)	2	0	
Higher administrative (spouse)	15	24	
Never had a job (spouse)	0	1	
Professional and technical (spouse)	70	125	
Sales (spouse)	20	26	
Semi-skilled worker (spouse)	9	5	
Service (spouse)	14	29	
Skilled worker (spouse)	44	9	
Unskilled worker (spouse)	8	7	
Clerical (Father)	7	9	 <p>Q283</p> <p>Female</p> <p>Male</p>
farm owner/manager (Father)	6	8	
farm worker (Father)	1	4	
Higher administrative (Father)	7	29	
Never had a job (Father)	0	1	
Professional and technical (Father)	49	62	
Sales (Father)	22	25	
Semi-skilled worker (Father)	16	27	

Service (Father)	8	21	
Skilled worker (Father)	58	70	
Unskilled worker (Father)	17	15	
Sector of employment (Q284)			
Government or public institution	41	73	<p>Q284</p>
Private business or industry	125	175	
Private non-profit organization	25	23	
Are you the chief wage earner in your house (Q285)			
Yes	42	194	<p>Q285</p>
No	149	77	







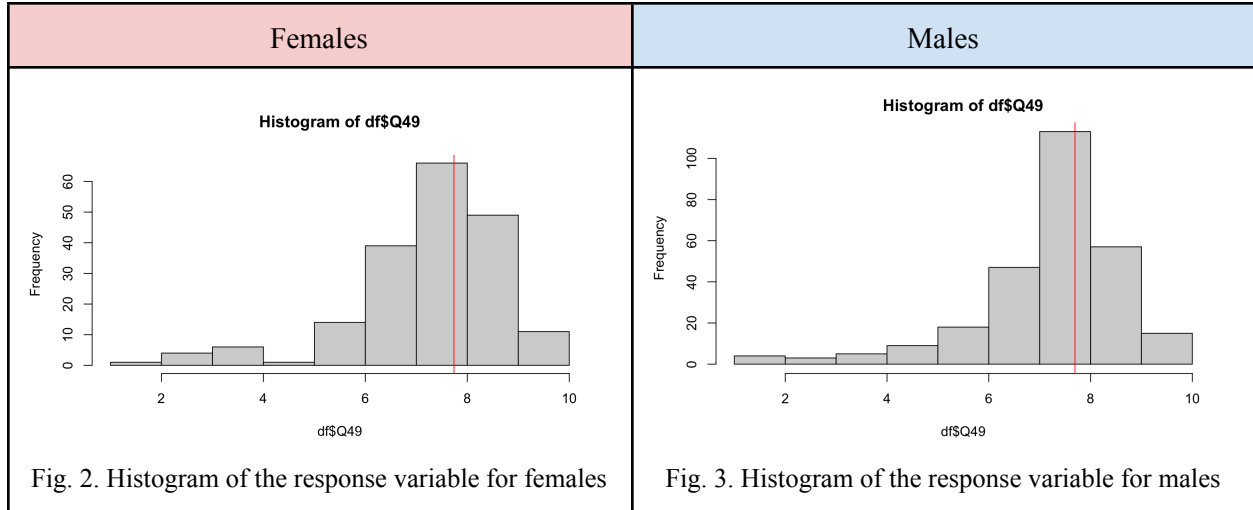
Following a thorough examination, we decided to remove predictors (Table III) characterized by an excess of levels or instances where data points were overly concentrated within a limited range of levels.

Table III. Predictors Removed by EDA

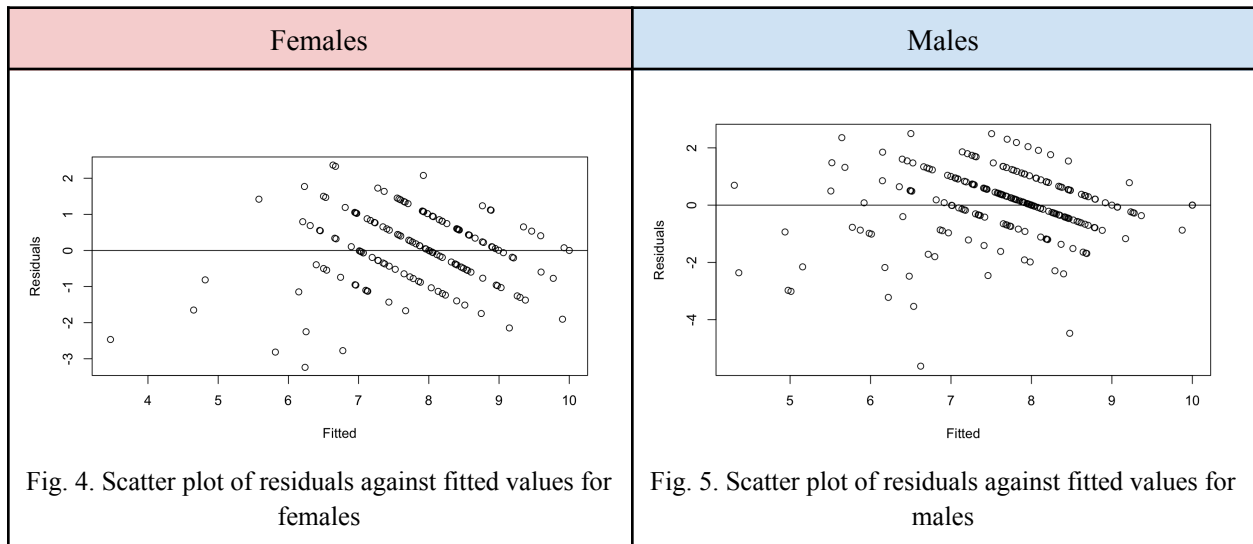
Q 261	Birth Year
Q 263	Born inside or outside the country
Q 264	Mother Immigrant Status
Q 265	Father Immigrant Status
Q 266 - Q 268	Country of Birth: Self, Mom, Dad
Q 269	Citizenship
Q 271	Live with Parents or not
Q 272	Language at Home
Q 290	Racial Belonging / Ethnic Group

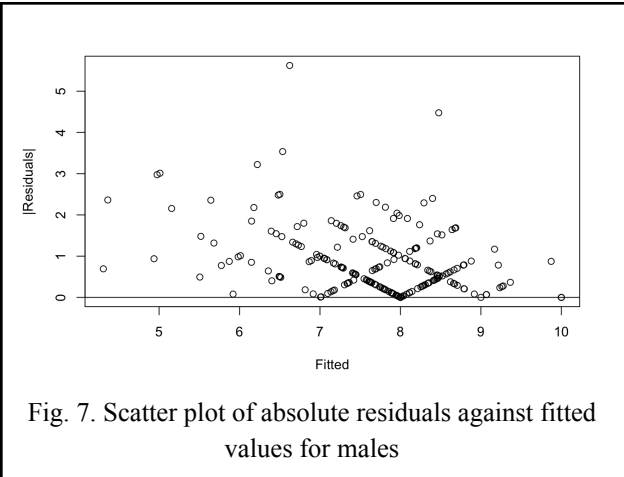
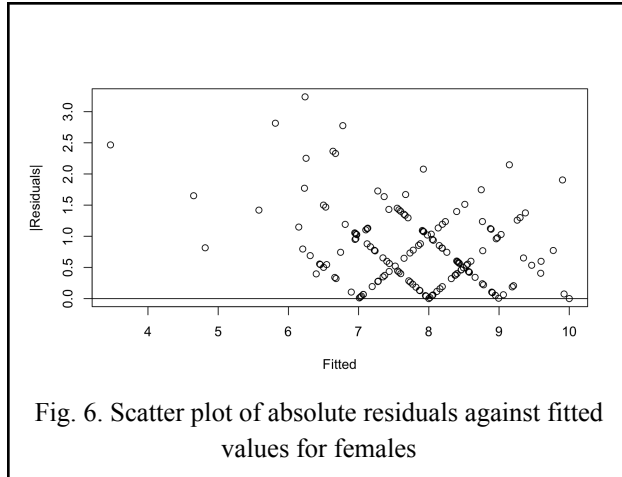
## 2.3 Model Diagnostics

Model diagnostics were conducted to ensure the robustness of our model. Histograms depicting the distribution of response variable — life satisfaction (Q49) for both females (Fig. 2) and males (Fig. 3) were plotted to investigate its skewness.

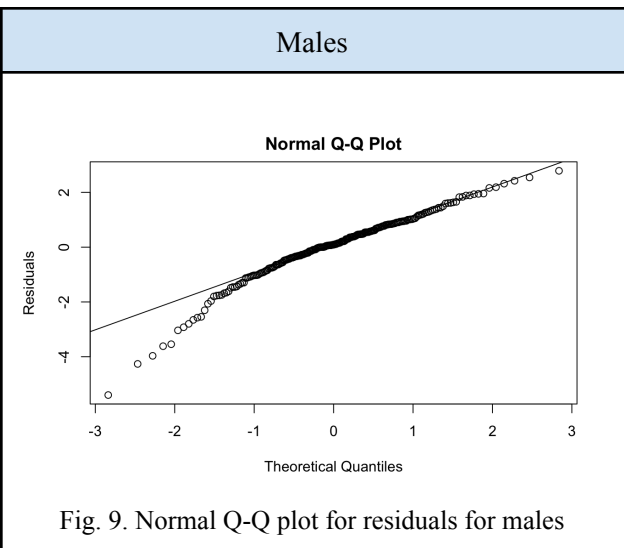
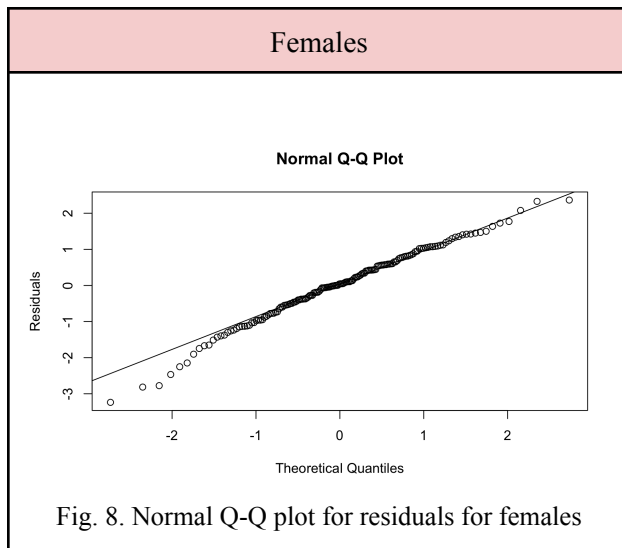


After fitting the full model, we plotted residuals against fitted values (Fig. 4, Fig. 5) and absolute residuals against fitted values (Fig. 6, Fig. 7) to help identify heteroscedasticity. The appearance of multiple bands in the plot can be attributed to the ordinal nature of the model's response variable, which ranges on a scale from 1 to 10 with scores restricted to whole numbers [2]. As a result, the fitted values conform to this constraint, creating noticeable bands in the plot. The heteroscedasticity is not observed, suggesting that y transformation is not needed.



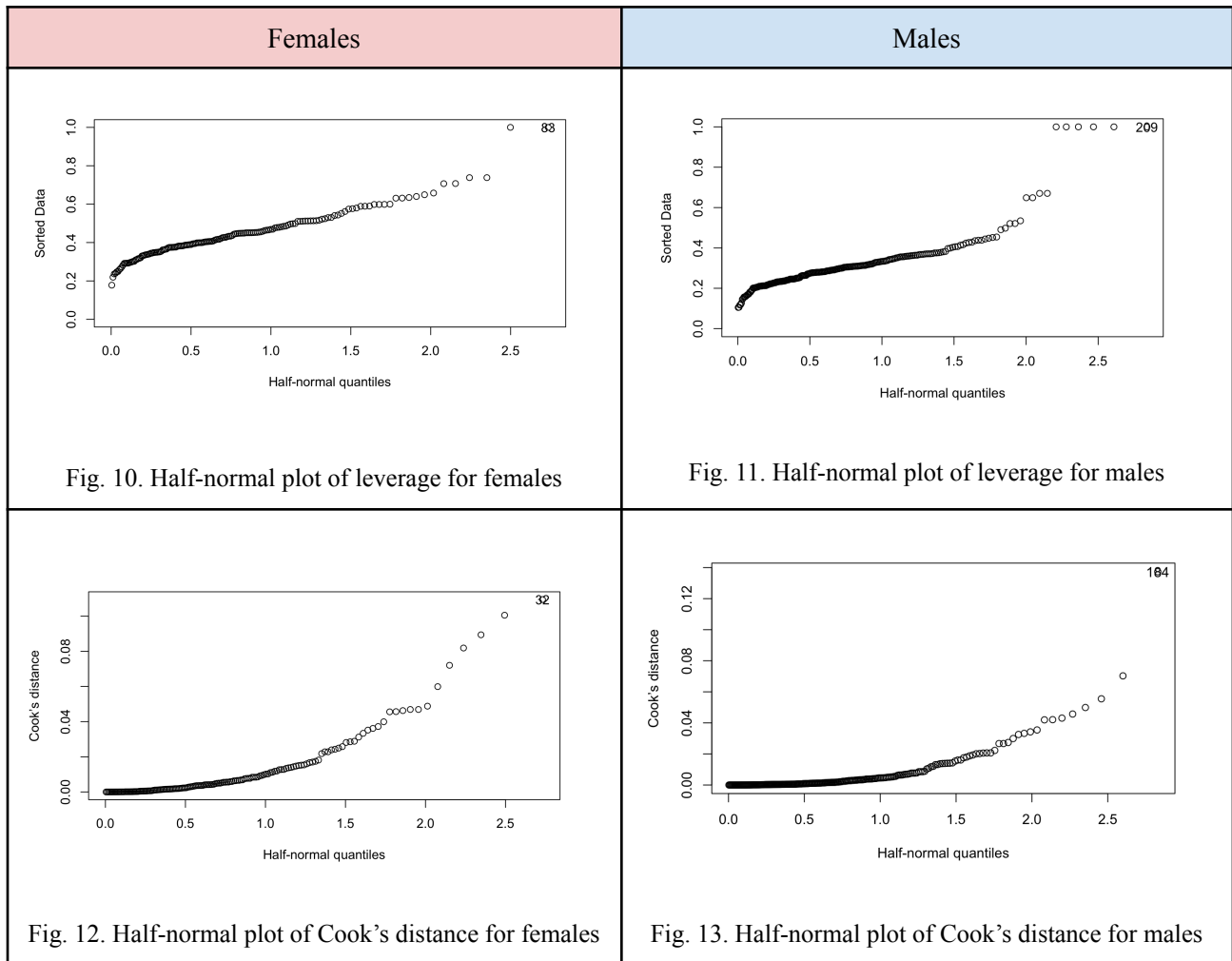


To further validate our model, we check the normality assumption using a Q-Q plot (Fig. 8, Fig. 9), comparing sorted residuals against ideal normal observations. While short-tailed errors are observed on the left side, the non-normality is not severe and can be reasonably ignored.



Complementing these assessments, a Shapiro-Wilk test was employed to statistically evaluate the normality of the data. The obtained p-value is 0.115 for Female. The obtained p-value for male is 0.5012 after eliminating unusual points using Cook's distance, indicating that the training data follows a normal distribution. The observations above affirm the satisfaction of model assumptions. Consequently, transformation is not necessary for either the response variable or predictors.

Lastly, we pinpointed unusual points with large leverage with a half-normal plot (Fig. 10, Fig. 11) and proceeded to remove influential points with Cook's distance larger than 0.05 for females or 0.025 for males (Fig. 12, Fig. 13).



## 2.4 Model Selection

Given the huge number of predictors, we performed test-based model selection to discern the most suitable ones. Backward, forward, and stepwise selection techniques yielded consistent results (see Appendix). Specifically, for females, occupational group (Q282), family savings (Q286), and scale of incomes (Q288), were selected, while for males, the selected predictors were highest educational level (Q275) and family savings (Q286).

## 2.5 Identifying significant interactions between predictors

By pairing each predictor and constructing linear models, we identified significant interaction terms for both females and males (Table III) with p-values smaller than 0.0001. We determined the significance level at 0.0001 to prevent the inclusion of an excessive number of interaction terms that would result in overfitting issues.

Table III. Significant Interaction Terms

	Female		Male	
Term 1	Q275	Q278	Q275	Q281
	The highest education level of the respondent	The highest education level of the respondent's father	The highest education level of the respondent	Occupational group
Term 2	Q276	Q278	Q276	Q282
	The highest education level of the respondent's spouse	The highest education level of the respondent's father	The highest education level of the respondent's spouse	Spouse occupational group
Term 3			Q282	Q286
			Spouse occupational group	Family savings during the past year

For females, Q275, Q276, and Q278, being part of the significant interaction terms, are additionally incorporated into the model with model selection, alongside previously selected Q285, Q286, and Q288. As for males, Q276, Q281, and Q282 are part of the significant interaction terms, so they are additionally incorporated into model selection, along with terms Q275 and Q286 kept after model selection.

## 2.6 Levels Aggregation

Given that many factors have numerous unbalanced levels, we performed levels aggregation to factors that have interaction terms to efficiently eliminate dummy variables. We took into account the significance level, the number of data points, and the nature of levels while aggregating the levels. This contributes to more reliable and interpretable models. For females, 3 predictors were aggregated (Table IV), while for males, 5 predictors were aggregated (Table V).

Table IV. Aggregated Terms for Females

	Level	Aggregate or not		Level	Aggregate or not		Level	Aggregate or not
Q275	ISCED 0		Q276	ISCED 0		Q278	ISCED 0	
	ISCED 1			ISCED 1			ISCED 1	
	ISCED 2			ISCED 2			ISCED 2	Aggregated
	ISCED 3			ISCED 3			ISCED 3	
	ISCED 4			ISCED 4			ISCED 4	Aggregated
	ISCED 5			ISCED 5			ISCED 5	

	ISCED 6	Aggregated		ISCED 6	Aggregated		ISCED 6	Aggregated
	ISCED 7			ISCED 7			ISCED 7	
	ISCED 8			ISCED 8			ISCED 8	

Table V. Aggregated Terms for Male

	Level	Aggregate or not		Level	Aggregate or not
Q275	ISCED 0		Q276	ISCED 0	
	ISCED 1			ISCED 1	
	ISCED 2	Aggregated		ISCED 2	Aggregated
	ISCED 3			ISCED 3	
	ISCED 4			ISCED 4	Aggregated
	ISCED 5			ISCED 5	
	ISCED 6			ISCED 6	
	ISCED 7			ISCED 7	Aggregated
	ISCED 8			ISCED 8	
Q281	Never had a job	Aggregated	Q282	Never had a job	Aggregated
	Professional and technical			Professional and technical	
	Higher administrative			Higher administrative	
	Clerical			Clerical	
	Sales			Sales	
	Service			Service	
	Skilled Worker			Skilled Worker	
	Semi-skilled worker	Aggregated		Semi-skilled worker	Aggregated
	Unskilled worker			Unskilled worker	
	Farm worker			Farm worker	
	Farm owner, farm manager			Farm owner, farm manager	

	JP,KG,TJ: Other			JP,KG,TJ: Other	
Q286	Save money				
	Just get by				
	Spent some savings and borrowed money	Tried but failed to aggregate it			
	Spent savings and borrowed money				

## **2.7 Final Models and Validation**

To validate the model, we divided the data into an 80:20 split, allocating 80% for the training set. Additionally, we implemented a while loop to ensure that our training set covers all levels for each factor.

To balance the size of the predictor and the accuracy of the model, we proposed the two following methods to select the model. We consider the full model with selected interaction terms and the model selected by test-based methods with interaction terms. If a factor that is included in interaction terms is not selected by the test-based methods, we will add it to the model. We then compared the two models using the F test to determine whether the selected model was acceptable.



### 3. Result

Two models were proposed for each sex group with selected predictors including occupational group, family savings, and scale of incomes for females, and highest educational level and family savings for males. Both models incorporate significant interaction terms.

#### 3.1 Two Proposed Models for Two Sex Groups

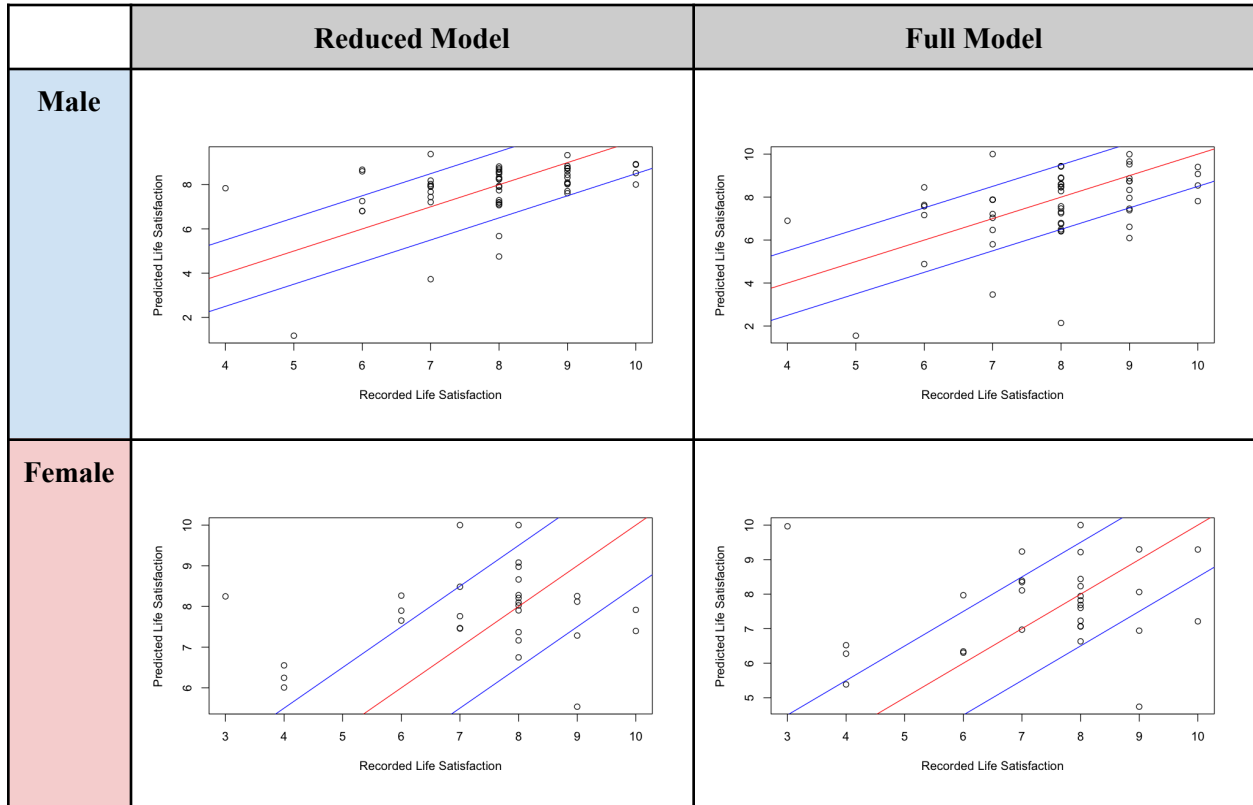
The statistics and model details are shown in the following table (Table VI).

Table VI. Test Statistics and Other Details for the Linear Regression Models

		R Square	Adjusted R Square	F Statistic (p-value)	Training RMSE	Testing RMSE
Male	Reduced Model	0.6077	0.4194	3.229 (<0.001)	0.9630	1.3789
		Q49 ~ Q275 + Q276 + <b>Q281</b> + <b>Q282</b> + Q286 + <b>Q288</b> + <b>Q275:Q281</b> + <b>Q276:Q282</b> + <b>Q282:Q286</b>				
	Full Model	0.7041	0.4214	2.491 (<0.001)	0.8363	1.6759
		Q49 ~ Q270 + <b>Q275</b> + Q276 + Q277 + Q278 + Q279 + Q280 + Q281 + <b>Q282</b> + Q283 + Q284 + <b>Q285</b> + Q286 + <b>Q287P</b> + <b>Q288</b> + <b>Q275:Q281</b> + <b>Q276:Q282</b> + Q282:Q286				
	F-statistic: 1.0141 (p-value: 0.461) → <b>Reject H0</b> : Reduced Model is Acceptable					
Female	Reduced Model	0.387	0.2155	2.257 (0.0008)	0.927975	1.857864
		Q49 ~ <b>Q275</b> + Q276 + Q278 + <b>Q285</b> + <b>Q286</b> + <b>Q288</b> + <b>Q275:Q278</b> + Q276:Q278				
	Full Model	0.6224	0.2082	3.06 (0.0404)	0.7282625	1.982029
		Q49 ~ Q262 + Q270 + Q273 + Q274 + Q275 + Q276 + Q277 + Q278 + Q279 + <b>Q280</b> + <b>Q281</b> + <b>Q282</b> + Q283 + Q284 + <b>Q285</b> + <b>Q286</b> + Q287P + <b>Q288</b> + Q289 + Q275:Q278 + Q276:Q278				
	F-statistic: 1.5255 (p-value: 0.1457) → <b>Reject H0</b> : Reduced Model is Acceptable However, considering R square and testing result, the full model is more satisfied.					

Given that there could be some variance between subjects for reporting likert scale data, the differences within 1.5 might not be significantly different, we plotted two blue auxiliary lines indicating the  $\pm 1.5$  region (Table VI). The red line shows the perfect fitting line.

Table VII. Plots of Recorded Against Predicted Response Variable



Based on the F-test ANOVA table and validation using the testing data the reduced models are not significantly different from the full model.

### 3.2 Significant Main Effects

- Q275: Self highest educational level (Male Full Model and Female Reduced Model)
- Q280: Spouse employment status (Female Full Model)
- Q281: Self occupational group (Male Reduced Model and Female Full Model)
- Q282: Mother occupational group (Male Reduced Model, Male Full Model, and Female Full Model)
- Q285: Are you the chief wage earner in your house (Male Full Model, Female Full Model, and Female Reduced Model)
- Q286: Family savings during past year (Female Reduced and Full Model)
- Q287: Social class (Male Full Model)
- Q288: Scale of incomes (All Models)

The predictors that are bold in Table indicated that it is significant.

Scale of incomes (Q288) is the most influential factor across genders and models, significantly impacting life satisfaction (Q49). Both genders are affected by their highest educational level (Q275) and occupational group (Q281), although the significance varies across models. Females appear more impacted by factors like spouse employment status (Q280) and family savings in the past year (Q286)—which especially holds significance in both full and reduced models. Conversely, males seem more influenced by social class (Q287). Mother's occupational group (Q282) and being the primary wage earner in the household (Q285) are both significant for both genders, but males are notably more affected

by the former, while females have greater sensitivity to the latter. In summary, factors related to the family have a stronger influence on females compared to males.

### 3.3 Significant Interaction Terms For Females

Interaction plots are constructed for significant interaction terms to visualize their interplay.

1. Between the highest education level of the respondent and that of their father

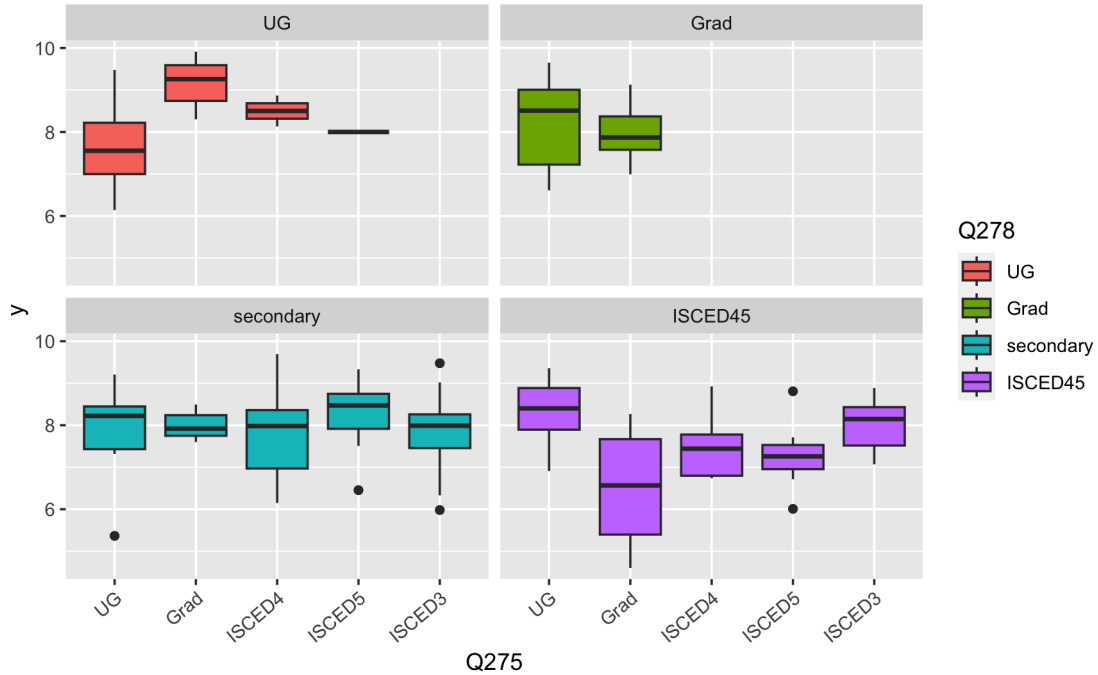


Fig. 14. Interaction Terms of the highest education level of respondents (Q275) and their fathers (Q278)

Stability is observed when the respondent's father has a lower educational level (secondary, ISCED 4&5, as shown in the two bottom boxplots). However, despite generally stable expectations, there are outliers, and the spread of scores within the sample is not significantly wide. There appears to be a noteworthy contrast in score distribution when comparing respondents at the graduate level with fathers who have educational backgrounds categorized under ISCED levels 4 and 5. This discrepancy in scores could potentially arise due to communication barriers that affect satisfaction scores.

However, when the respondent's father holds a higher degree, such as undergraduate and graduate levels, there are significant differences in scores. This difference could potentially be attributed to heightened expectations from fathers with advanced educational backgrounds, leading respondents to report lower life satisfaction scores.

2. Between the highest education level of the respondent's spouse and the respondent's father.

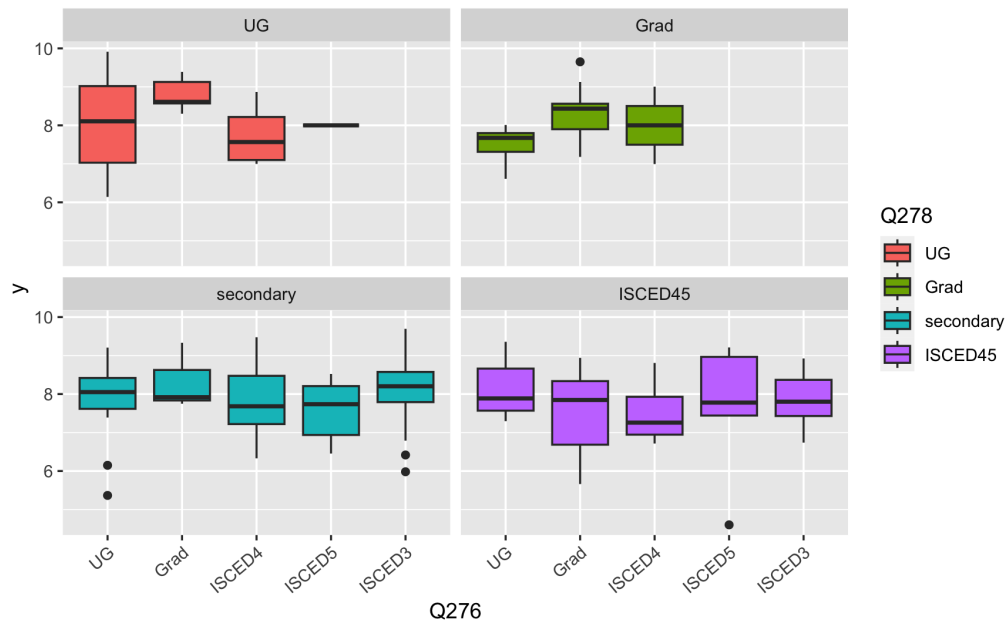


Fig. 15. Interaction Terms of the highest education level of their spouses (Q276) and fathers (Q278)

When considering scenarios where both the respondent's spouse and father possess lower educational qualifications, there is an increased likelihood of reporting lower satisfaction scores. Notably, instances where spouses hold ISCED 5 degrees and fathers possess ISCED 4 & 5 degrees might result in significantly low scores.

### 3.4 Significant Interaction Terms For Males

1. Between the highest education level of the respondent and their occupational group

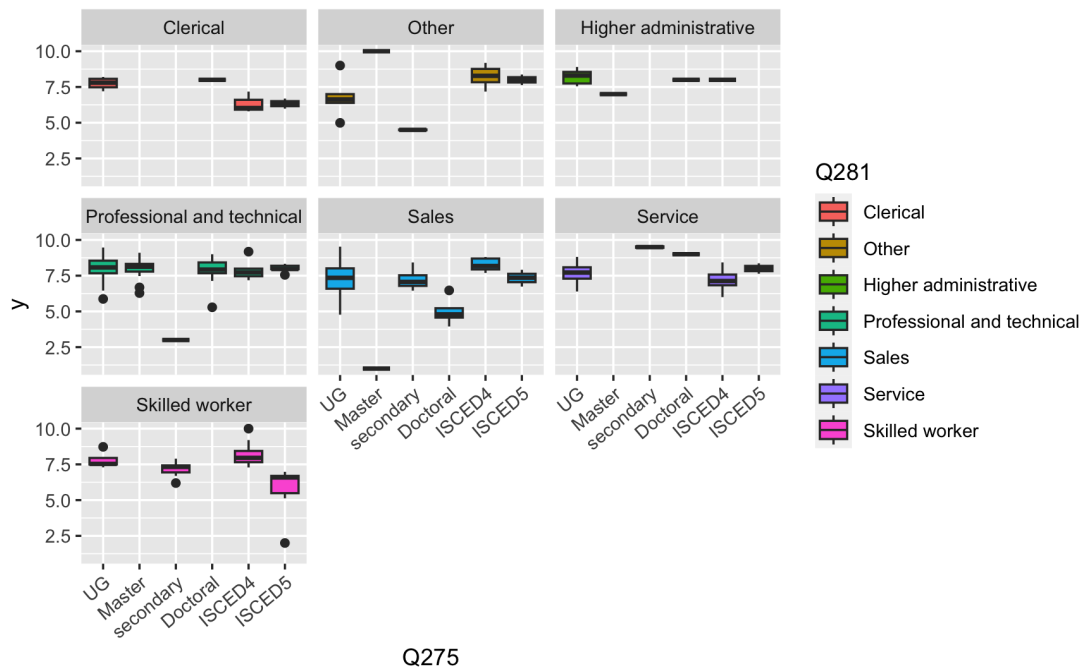


Fig. 16. Interaction Terms of the highest education level of respondents (Q275) and their occupational group (Q281)

Fig 16. indicates that life satisfaction does not differ significantly among individuals with varying education levels within certain occupational groups. These groups include clerical, higher administrative, professional and technical, service, and skilled workers. However, individuals with higher education levels, such as Master's and Doctoral degrees, tend to have lower life satisfaction when working in sales.

2. Between the highest education level of the respondent's spouse and their occupational group

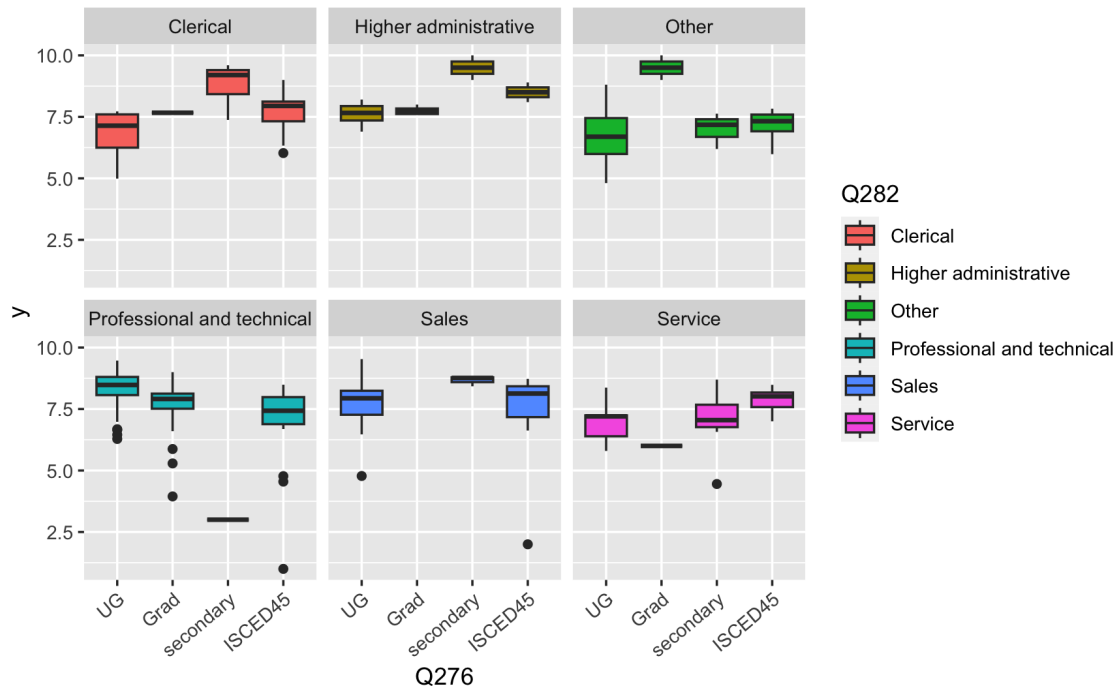


Fig. 17. Interaction Terms of the highest education level of respondents' spouses (Q276) and their occupational group (Q282)

Respondent's spouses with secondary education engaged in professional and technical roles exhibit significantly lower scores compared to all other degree levels. In addition, within the professional and technical groups, individuals whose spouses hold ISCED 4 & 5 qualifications have lower life satisfaction scores compared to those with spouses having other educational levels.

3. Between the respondent's spouse's occupational group and the family savings during past year

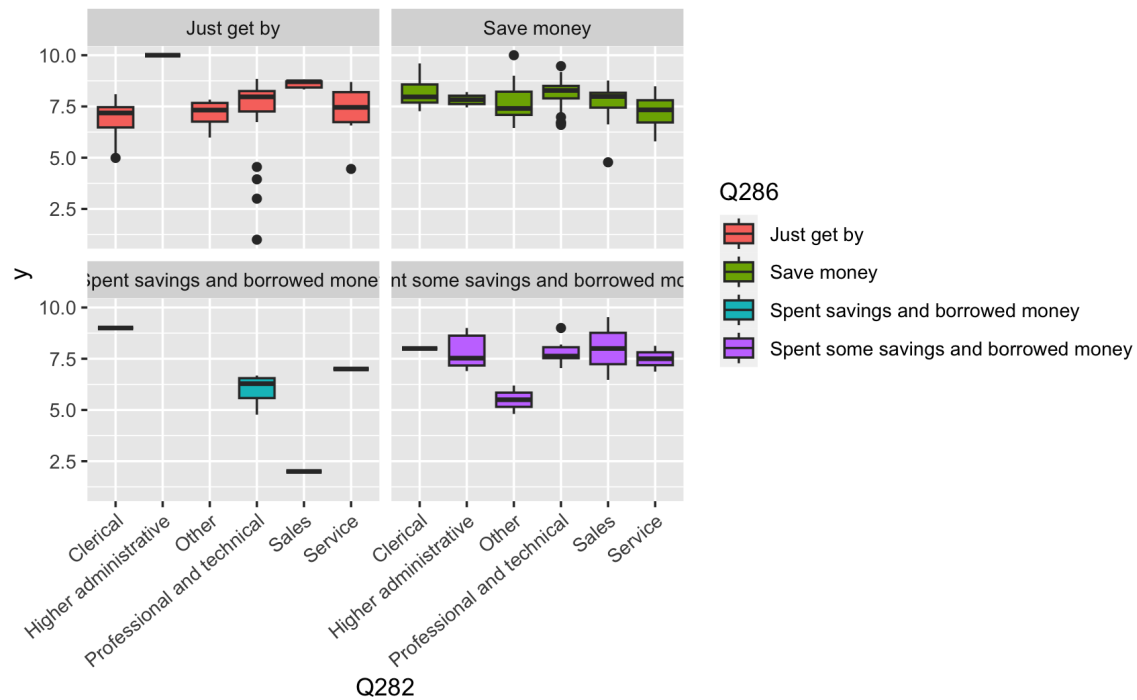


Fig. 18. Interaction Terms of the highest education level of respondents' spouses' occupational group (Q282) and their family savings during past year (Q286)

Our model shows that people who work in professional and technical or sales who spent some saving and borrowing money have significantly higher life satisfaction. In the interaction plot, we can also observe that people have similar satisfaction when they save money for the past year regardless of their occupation. However, people with different occupations show different life satisfaction when they have some savings while also borrowing money for the past year.

## **4. Conclusion and Discussion**

### **4.1 Model Results**

Two distinct models have been proposed to explore the influence of demographic and socioeconomic factors on individual life satisfaction within each gender group. In each model, predictors are categorized into two components: reduced or full predictors, and significant interaction terms. The process involved testing-based model selections to ensure consistent predictor choices for both genders. For females, the selected partial predictors include occupational group (Q282), family savings (Q286), and scale of incomes (Q288). In contrast, for males, the chosen predictors include the highest educational level (Q275) and family savings (Q286). Both models incorporate significant interaction terms. Notably, the reduced model incorporates predictors not initially chosen through the selection process but appeared as significant in interaction terms. This approach seeks to capture the nuanced impact of various factors on life satisfaction for both males and females.

The robustness of our models is evident as three out of four yield R-squared values exceeding 0.6. Furthermore, considering potential variance between subjects in reporting Likert scale response variables, differences within 1.5 may not be significantly different. By plotting records against predicted response variables along with two auxiliary lines indicating the  $\pm 1.5$  region, it reveals that the majority of data points fall within these lines. This visualization reinforces the effectiveness of our models.

### **4.2 Similarity and Difference in Variable Selection**

Family savings during the past year (Q286) is the only predictor selected for both females and males. This commonality suggests that the status of family savings holds substantial influence over an individual's life satisfaction regardless of gender.

For females, life satisfaction is not only impacted by family savings but also by their occupation (Q282) and the scale of their incomes (Q288). The nature of their work and the level of their earnings play significant roles in shaping their overall satisfaction with life. As for males, the only predictor selected in addition to family savings is the highest educational level (Q275). The level of education they have achieved emerges as a distinct factor shaping their satisfaction with life.

### **4.3 Similarity and Difference in Significant Interactions**

There are notable differences in the significant interactions identified within each sex group. Specifically, for females, both the respondent's educational level (Q275) and the respondent's spouse's educational level (Q276) show an interplay with the respondent's father's educational level (Q278). All the predictors contributing to significant interactions are associated with the education levels of the respondents themselves or their family members. This demonstrates the significance of interactions involving family education levels in impacting females' life satisfaction.

Similarly, for males, the respondent's educational level (Q275) and the respondent's spouse's educational level (Q276) show significant interaction with other variables. Specifically, the respondent's educational level (Q275) interacts with their occupation group (Q281), while the respondent's spouse's educational level (Q276) interacts with the respondent's spouse's occupation group (Q282).



#### **4.4 Discussion**

While the World Values Survey (WVS) Wave 7 provides a comprehensive dataset, we observed an imbalance in the distribution of levels within variables. Our aim to sample a balanced dataset for training the linear regression models faced challenges in obtaining a sufficiently large sample size.

Furthermore, given the ordinal nature of the response variable (life satisfaction score) in the dataset, we tried to employ ordinal regression, but our attempts were unsuccessful. Despite linear regression not being the most accurate method for modeling ordinal response variables, our models produced statistically acceptable results. However, it's crucial to interpret these outcomes with caution, considering the inherent limitations of linear regression in capturing the nuanced relationships within ordinal data.

## Reference

[1] Haerpfer, C., Inglehart, R., Moreno, A., Welzel, C., Kizilova, K., Diez-Medrano J., M. Lagos, P. Norris, E. Ponarin & B. Puranen et al. (eds.). 2020. World Values Survey: Round Seven – Country-Pooled Datafile. Madrid, Spain & Vienna, Austria: JD Systems Institute & WVSA Secretariat. doi.org/10.14281/18241.1

[2] McAllister, "Why do your residual vs fitted plots look so weird?" Prof McAllister Blog. [Online]. Available: <https://profmmcallister.com/2020/03/23/why-do-you-residual-vs-fitted-plots-look-so-weird/> (accessed Nov. 26, 2023)

## Appendix. Model Selection Results

### Females

#### 1. Forward Selection

Selection Summary						
Step	Variable Entered	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	Q286	0.1233	0.1088	13.7853	664.3120	1.4339
2	Q288	0.1701	0.1516	5.3876	656.1633	1.3990
3	Q282	0.2471	0.1946	-9.7215	654.1480	1.3631

#### 2. Backward Selection

Elimination Summary						
Step	Variable Removed	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	Q277	0.4813	0.198	-31.7588	691.2109	1.3602
2	Q284	0.4797	0.2088	-33.3960	687.7924	1.3511
3	Q278	0.4663	0.2268	-32.4277	680.4827	1.3356
4	Q262	0.4663	0.2327	-34.4139	678.5042	1.3304
5	Q283	0.4329	0.2383	-29.0026	671.7192	1.3256
6	Q275	0.4141	0.2408	-26.8285	667.7487	1.3234
7	Q270	0.4126	0.2442	-28.4880	666.2320	1.3205
8	Q281	0.3838	0.2492	-24.1015	659.0702	1.3161
9	Q273	0.3816	0.2514	-25.5976	657.7499	1.3142
10	Q280	0.3731	0.2509	-25.7093	656.2751	1.3146
11	Q289	0.3501	0.248	-22.6137	652.9220	1.3172
12	Q287P	0.3335	0.243	-20.9119	651.6053	1.3216
13	Q276	0.2915	0.2241	-13.5936	650.8946	1.3379
14	Q285	0.2823	0.2186	-13.5383	651.2946	1.3427
15	Q274	0.2719	0.212	-13.2429	651.9386	1.3483
16	Q279	0.2471	0.1946	-9.7215	654.1480	1.3631

### 3. Stepwise

### Selection

Stepwise Selection Summary							
Step	Variable	Added/ Removed	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	Q286	addition	0.123	0.109	13.7850	664.3120	1.4339
2	Q288	addition	0.170	0.152	5.3880	656.1633	1.3990
3	Q282	addition	0.247	0.195	-9.7220	654.1480	1.3631

### Males

#### 1. Forward Selection

Selection Summary						
Step	Variable Entered	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	Q286	0.0936	0.0810	251.2251	801.9735	1.4772
2	Q275	0.1570	0.1209	220.5321	798.0144	1.4448

#### 2. Backward Selection

Elimination Summary						
Step	Variable Removed	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	Q282	0.3914	0.1511	129.7183	832.3441	1.4198
2	Q283	0.3688	0.1722	139.3989	820.3864	1.4021
3	Q274	0.3686	0.177	137.4607	818.4282	1.3980
4	Q277	0.3556	0.1889	142.1980	810.9359	1.3878
5	Q280	0.3529	0.1947	141.5943	807.8587	1.3828
6	Q276	0.3303	0.1941	151.2385	803.4068	1.3834
7	Q287P	0.3195	0.1945	154.7617	800.8986	1.3830
8	Q279	0.3123	0.1946	156.5105	799.2374	1.3830
9	Q262	0.3088	0.1949	156.2851	798.3359	1.3827
10	Q284	0.2994	0.1925	159.1546	797.3224	1.3848
11	Q285	0.295	0.1917	159.3978	796.6846	1.3854
12	Q270	0.2873	0.1871	161.3852	797.0854	1.3894
13	Q273	0.2762	0.1787	165.1145	798.4898	1.3966
14	Q278	0.2267	0.1532	188.6243	799.0393	1.4180
15	Q281	0.1679	0.1281	216.9244	797.1568	1.4389
16	Q288	0.157	0.1209	220.5321	798.0144	1.4448

#### 3. Stepwise Selection

# Stepwise Selection Summary

Step	Variable	Added/ Removed	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	Q286	addition	0.094	0.081	251.2250	801.9735	1.4772
2	Q275	addition	0.157	0.121	220.5320	798.0144	1.4448