NATIONAL RESEARCH UNIVERSITY
HIGHER SCHOOL OF ECONOMICS

Faculty of Computer Science
Bachelor's Programme "Data Science and Business Analytics"

Life Satisfaction Analysis

Elizaveta Anikina

Group БПАД212

The Problem and Its Significance

Life satisfaction is a crucial indicator of overall well-being and happiness. Understanding the factors that influence life satisfaction can provide valuable insights for individuals, policymakers, and researchers. This research focuses on identifying key variables that affect individual life satisfaction. The primary variable of interest is the self-reported life satisfaction score. The goal is to analyze how various demographic, socioeconomic, and personal characteristics impact life satisfaction, enabling more accurate predictions and informed decision-making.

The dataset used offers a comprehensive collection of individual-level data, including variables such as age, gender, education level, marital status, occupation, income, health status, and lifestyle factors. The diversity of the dataset provides a rich basis for exploring the complexities of life satisfaction, allowing us to uncover patterns and relationships that might not be evident in more aggregated data.

Data Description

Variables

life_satis: Self-reported life satisfaction score (target variable).

age: Age of the respondent.

sex: Gender of the respondent.

educ: Education level.

mar_st: Marital status.

occupation: Type of occupation.

vacation: Paid leave in the last 12 months.

money: Income level.

fin_state_change: Change in financial status.

fin_resp: Financial behavior.

change_job: Desire for new job.

social_media: Use of social media.

preg: Pregnancy status.

physical_activity: Level of physical activity.

alcohol: Alcohol consumption.

smoker: Smoking status.

depression: Depression status.

health: Health status.

level_of_trust: Level of trust in others.

num_of_children: Number of children.

is_religious: Religious status.

Data Type and Structure

Our data is cross-sectional, facilitating ease of analysis and manipulation using various statistical tools and software. The dataset includes a substantial number of records, providing a robust sample for statistical analysis and ensuring the reliability of the research findings.

Data Preprocessing

We began by translating our categorical variables. This step was essential to ensure the data was consistent and usable for analysis. We focused on the '*Occupation*' and '*Mar_st*' (marital status) variables. Initially, the 'Occupation' variable included a wide range of job titles and classifications and the '*Mar_st*' variable captured various marital statuses, including Single, Married, Divorced, Widowed, etc.

By examining their distributions, we identified classes that were either similar or had very few observations and consolidated them. Additionally, categories with very few observations were merged into an "Other" category to ensure they did not disproportionately affect the model due to their sparsity.

There are benefits of consolidation such as improved model performance, increased robustness, and enhanced interpretability. Consolidating categories helped to reduce the complexity of the model by decreasing the number of dummy variables created for categorical data. This reduction in complexity helps to prevent overfitting and improves the generalizability of the model. By ensuring that each category has a

sufficient number of observations, the statistical power of the analysis is enhanced, making the results more robust and reliable.

Exploratory Data Analysis (EDA)

Next, we analyzed the distribution of our target variable, '*life_satis*'. This step helped us understand the baseline distribution and guided us in considering different scenarios for our analysis.

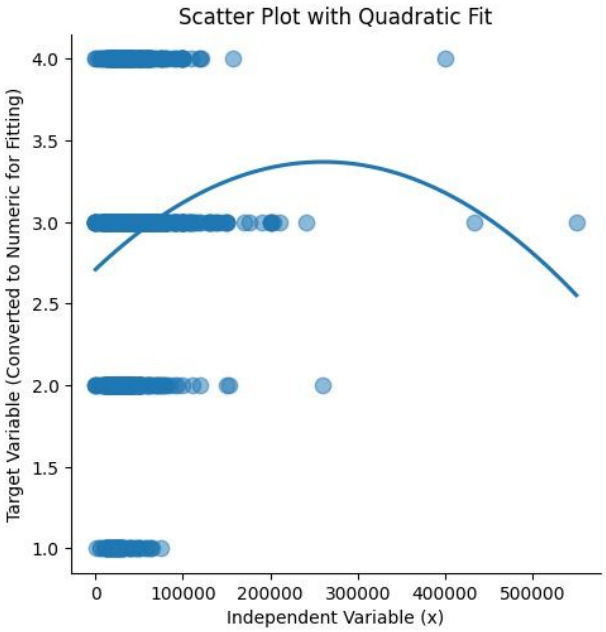


Figure 1
Money

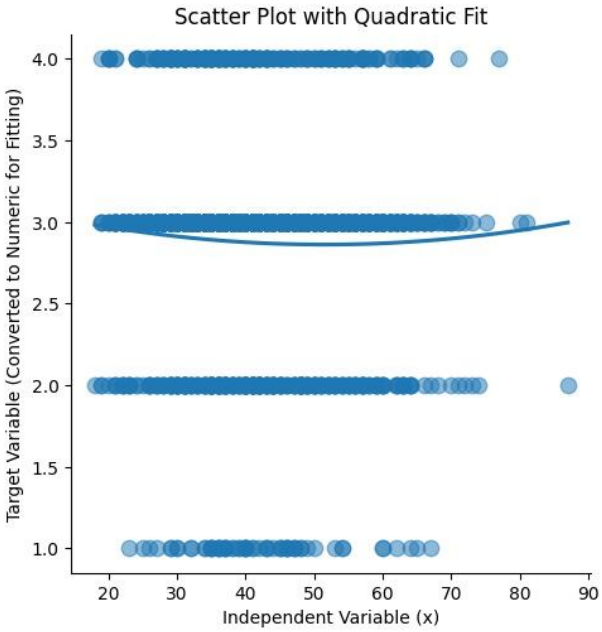


Figure 2
Age

Based on the scatter plots (Figure 1, Figure 2), here are some interpretations for the relationships between **life_satis** (life satisfaction) and the other variables:

Age: There doesn't appear to be a clear trend in life satisfaction with respect to age. The points are quite scattered, indicating that age may not have a strong direct impact on life satisfaction.

Sex: Life satisfaction seems to be fairly similar between sexes, with no significant difference noticeable from the plot.

Education (educ): Higher levels of education might show a slight increase in life satisfaction, but the trend isn't very strong.

Marital Status (mar_st): There may be some differences in life satisfaction based on marital status, but the plot doesn't show a clear pattern.

Occupation: Different occupations do not show a strong correlation with life satisfaction. Most points are clustered without a clear trend.

Vacation: There is a slight indication that taking vacations may be associated with higher life satisfaction.

Money: Higher income generally appears to correlate with higher life satisfaction, although there is considerable variability.

Financial State Change (fin_state_change): Positive changes in financial state seem to be associated with higher life satisfaction.

Financial Responsibility (fin_resp): The plot doesn't show a clear trend between financial responsibility and life satisfaction.

Change Job (change_job): This variable doesn't show a clear trend in relation to life satisfaction.

Social Media Usage (social_media): The relationship between social media usage and life satisfaction is not clear from the plot.

Pregnancy (preg): This variable has many zeros (indicating non-pregnant) and a few ones (indicating pregnant), with no clear trend in life satisfaction.

Physical Activity (physical_activity): There is some indication that higher physical activity levels may be associated with higher life satisfaction.

Alcohol Consumption (alcohol): There doesn't appear to be a strong trend between alcohol consumption and life satisfaction.

Smoking (smoker): Smokers and non-smokers don't show a clear difference in life satisfaction.

Depression: Higher levels of depression are generally associated with lower life satisfaction, indicating a strong negative correlation.

Health: Better health is associated with higher life satisfaction, showing a positive correlation.

Level of Trust (level_of_trust): Higher levels of trust seem to correlate with higher life satisfaction.

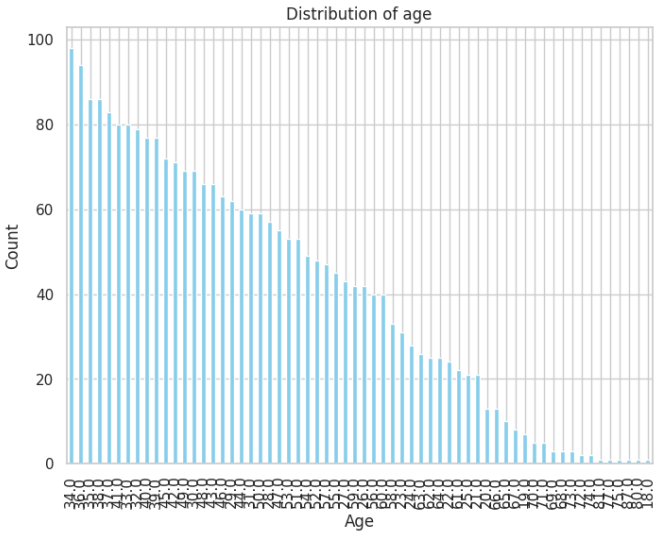


Figure 3

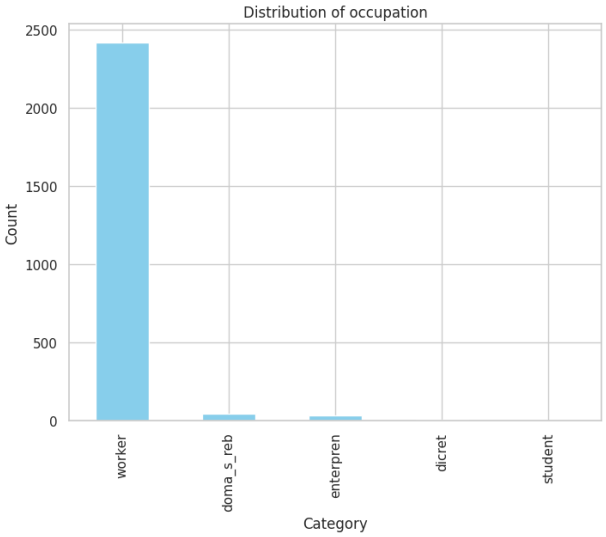


Figure 4

Number of Children (num_of_children): There is no clear trend between the

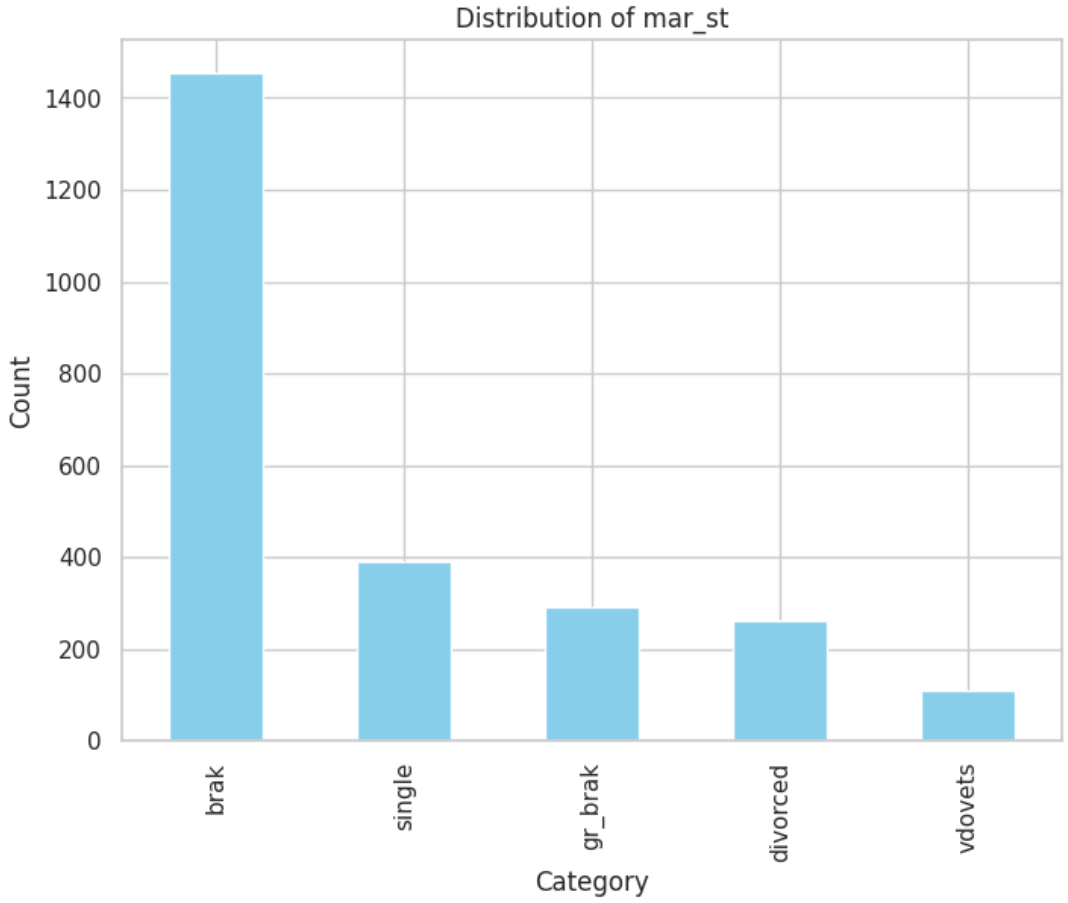


Figure 5

number of children and life satisfaction.

Religiosity (is_religious): Religious individuals might have slightly higher life satisfaction, but the trend isn't strong.

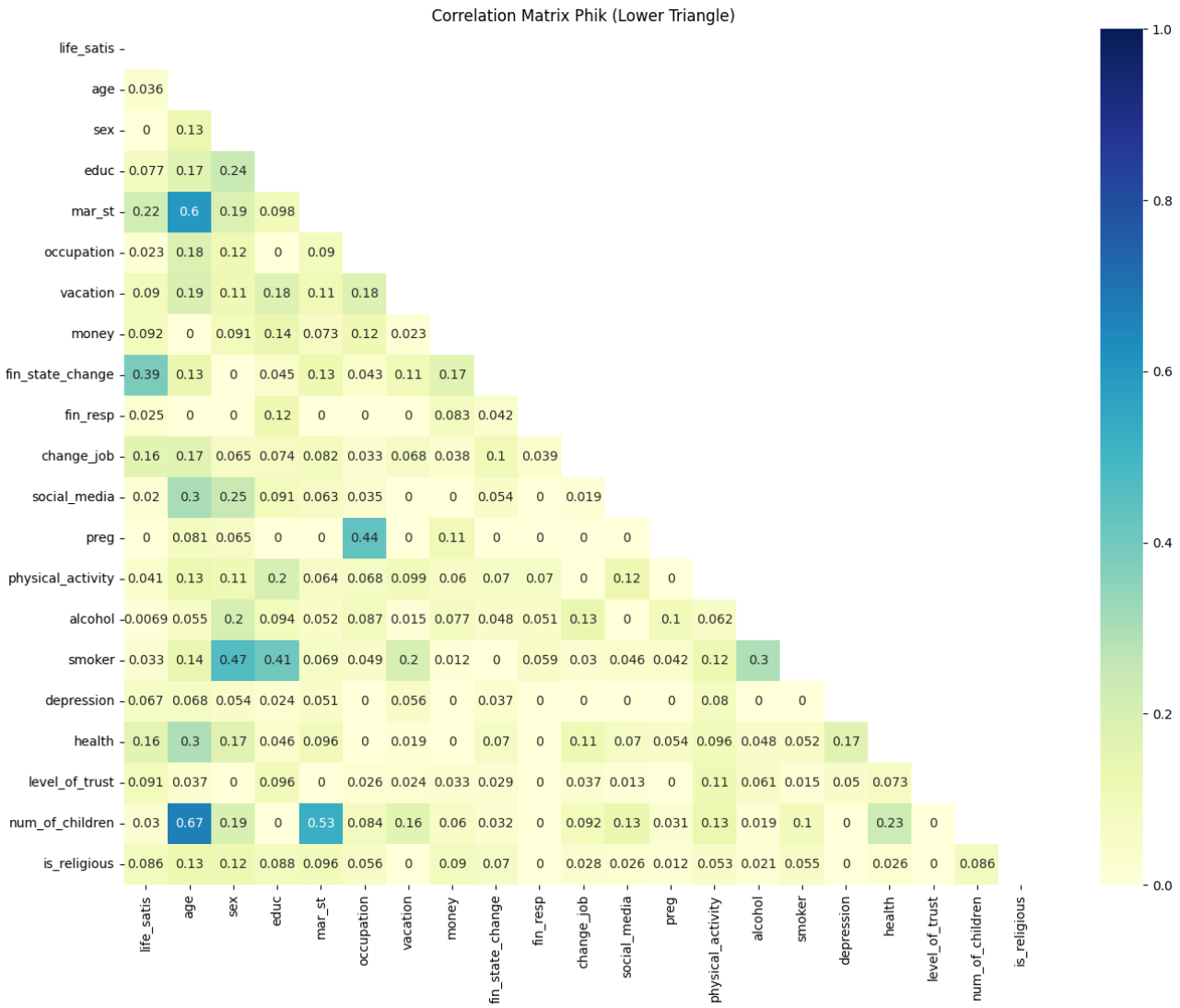


Figure 6

Summary Statistics Insights

Life Satisfaction (life_satis): Ranges from 1 to 5 with a mean of 3.51.

Age: Ranges from 18 to 87 with a mean age of about 42 years.

Sex: Binary variable (0 for male and 1 for female) with nearly equal distribution.

Education (educ): Ranges from 1 to 6.

Vacation: Binary variable (0 or 1) indicating whether the individual took a vacation.

Money: Wide range from 0 to 550,000 with an average of about 37,875.

Financial State Change (fin_state_change): Ranges from 1 to 5.

Financial Responsibility (fin_resp): Mostly 0 with very few 1s.

Change Job (change_job): Binary variable indicating job change.

Social Media (social_media): Majority use social media (0 or 1).

Pregnancy (preg): Mostly 0s.

Physical Activity (physical_activity): Varied from 0 to 6.

Alcohol Consumption (alcohol): Binary variable indicating alcohol consumption.

Smoker: Binary variable.

Depression: Mostly 0s.

Health: Ranges from 2 to 5.

Level of Trust (level_of_trust): Ranges from 1 to 3.

Number of Children (num_of_children): Mostly 0 or 1.

Religious (is_religious): Mostly 1s, indicating religious individuals.

Distribution of features Insights (Figures 3, 4, and 5)

Life Satisfaction (life_satis): Normally distributed with a peak around 4.

Age: Skewed towards younger individuals.

Vacation: Binary distribution with more 0s than 1s.

Money: Right-skewed with a few high-income outliers.

Financial State Change (fin_state_change): Mostly around 3.

Financial Responsibility (fin_resp): Predominantly 0.

Change Job (change_job): Predominantly 0.

Social Media (social_media): Most people use social media (value 1).

Pregnancy (preg): Almost all 0s.

Physical Activity (physical_activity): Right-skewed with many 0s.

Alcohol Consumption (alcohol): More people drink alcohol (value 1).

Smoker: More non-smokers (value 0).

Depression: Very few instances (mostly 0s).

Health: Normally distributed around 3.

Level of Trust (level_of_trust): Peaks around 2.

Number of Children (num_of_children): Mostly 1.

Religious (is_religious): Mostly 1s, indicating a religious population.

Correlation Matrix Insights (figure 6)

Life Satisfaction has moderate positive correlations with **health** and **level of trust**.

Age shows a positive correlation with **money** and **number of children**, and a negative correlation with **social media** usage.

Vacation is positively correlated with **life satisfaction** and **money**.

Money is moderately correlated with **life satisfaction**, indicating higher income may contribute to higher satisfaction.

Physical Activity shows weak correlations with other features, indicating it's somewhat independent.

Health has a moderate positive correlation with **life satisfaction**, suggesting healthier individuals are more satisfied.

Depression has a negative correlation with **life satisfaction** and **health**.

Categorical Features Insights

Marital Status (mar_st): The most common marital status is *'brak'*, followed by *'gr_brak'* and *'single*'. Other categories are less frequent.

Occupation: The majority of individuals are categorized as *'worker*', with fewer in other categories.

Model Building

We chose two models — OLS and Ordered Logistic Model as a baseline. We use them to try various ways to manipulate 'life satisfaction' variable, such as, using binary target variable, using multiple classes — specifically configurations with three, four, and five classes, modifying the classification by collapsing some categories or removing third of the data as per different strategies outlined.

OLS

The initial step in our regression analysis was fitting an Ordinary Least Squares (OLS) model with robust errors with a set of predictors to assess their effect on the dependent variable. A key outcome from this step was a low R-squared value. A low R-squared indicates that the model does not account for much of the variance in the dependent variable.

Ordered Logistic Model

Then we used the Ordered Logistic Regression Model for various datasets, including binary (and simple Logit for this case) and multi-class target variables (with three, four, and five classes).

This approach serves as a control scenario where no data manipulation or additional techniques are applied beyond the basic model fitting. It allows to establish a performance benchmark for the ordered logistic model on the raw, unmodified dataset.

We applied smoothing techniques to reduce noise and variability in the data, leading to more stable and generalizable model estimates. Smoothing is particularly useful when dealing with ordinal data that may have irregular distributions across categories.

Then we considered random undersampling technique to address class imbalance in the dataset by reducing the number of instances in the overrepresented classes. While undersampling can help in balancing the classes, it also leads to a loss of potentially valuable data. This reduction in dataset size led to underfitting or biased estimates, thus, we came to the conclusion that we won’t be using it.

Ramsey Test

We performed the Ramsey test to verify the model's specification, ensuring that no significant variables were omitted. The results indicated that the model was correctly specified.

Results

- The null hypothesis (H0) of the Ramsey RESET test states that the model is correctly specified, meaning that no significant omitted variable exists.

- If the p-value from the F-test is small (typically less than 0.05), we reject the null hypothesis, suggesting that the model may be misspecified.
- If the p-value is large, we fail to reject the null hypothesis, indicating that the model is likely correctly specified.

Likelihood ratio

We implemented the likelihood ratio test to compare two models: a full model with

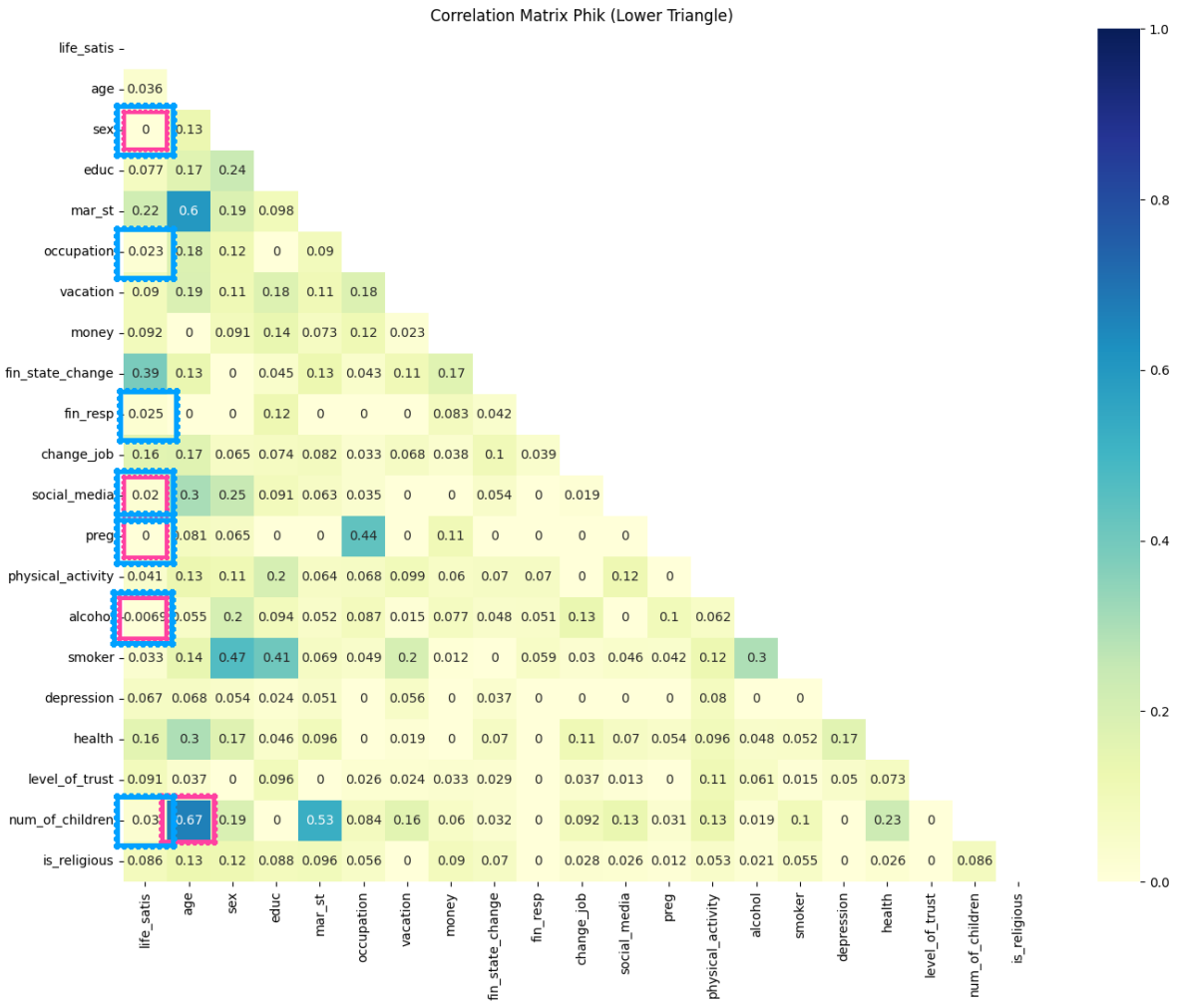


Figure 7

all predictors and a reduced model with exclusion of predictors with low correlation with target variable and those, which may lead to multicollinearity.

We chose 2 thresholds for correlation: 0.02 and 0.03, so we obtained the following lists of variables (Image below):

- thresholds <= 0.02

1. sex, alcohol, preg, socia_media
2. num_of_children due to multicollinearity

- thresholds <= 0.03
  1. sex, alcohol, preg, socia_media, fin_resp, occupation, num_of_children

We defined the models:
• Full Model: An ordered logistic regression model (**OrderedModel**) is defined with all predictors using the logit link function. The model is fit to the training data using the BFGS optimization method.
• Reduced Model: Another ordered logistic regression model is defined similarly but without chosen predictors). This model is also fit using the BFGS method.

Then we computed the test statistic. The likelihood ratio test statistic is calculated as **-2 * (reduced_model.llf - full_model.llf)**. This involves the difference in the log likelihoods of the reduced model and the full model, multiplied by -2.

Based on the p-value, we can draw a conclusion that:
• For threshold 0.02 p-value: 0.60 , so we accept H0 -> restricted model is better
• For threshold 0.03 p-value: p-value: 0.016, so we reject H0 -> full model is better
AIC values also enhance our results, so we decided to exclude sex, alcohol, preg, socia_medis and num_of_children parameters.

Variance Inflation Factor (VIF)

After identifying a low R-squared, we conducted a VIF analysis.
VIF values greater than 10 indicate serious multicollinearity, suggesting that the predictor variables are highly correlated. This can cause issues in determining the individual effect of each predictor on the dependent variable because it becomes difficult to isolate the influence of one predictor from another.
Despite a high VIF, we decided not to exclude the variable "age" from the model since age is a crucial predictor of life satisfaction.

Based on VIF we removed "Education," “Fin_state_change,” and "Health" to reduce multicollinearity. This way we simplified the model, making it more interpretable while potentially improving the accuracy of the coefficient estimates for the remaining variables.

Chow test

We conducted a chow test on the variables of age and money to identify structural gaps in the sample.

Previously, we divided the sample into: low income, middle income, high income. The boundaries were chosen by quantiles: the split between low income and average income was 22,000, and between average and high 45,000.

We also divided the sample by age: 18-30 years old, 30-50 and 50+.

The results of the chow money test showed that there is a structural gap between people with different incomes. That is, the relationship between happiness and money varies significantly at the sample boundaries - it differs among people with low, medium and high incomes.

As for the chow age test: this suggests that there is no structural gap. This means that the relationship between age and standard of living is relatively stable in different age groups.

This test showed that we should check our model on different intervals of financial statuses of our sample.

According to the results of the Chow test, we divided our dataset into three parts according to the level of financial income according to the variable “money”:

1. low income up to 22,000
2. Average income from 22,000 to 45,000
3. High income: more than 45,000

(We divided the sample by quartiles: 25% - 22000, 75% - 45000)

Next, we ran our final model** separately on these three subsamples.

OrderedModel Results

Dep. Variable:	life_satis	**Log-Likelihood:**	-2844.0
Model:	OrderedModel	**AIC:**	5734.
Method:	Maximum Likelihood	**BIC:**	5872.
Date:	Fri, 17 May 2024		
Time:	19:44:20		
No. Observations:	3004		
Df Residuals:	2981		
Df Model:	21		

	coef	std err	z	P>|z|	[0.025	0.975]
age	0.0040	0.004	1.043	0.297	-0.003	0.011
educ	0.2156	0.040	5.445	0.000	0.138	0.293
vacation	0.7096	0.079	9.036	0.000	0.556	0.864
money	0.0291	0.006	4.608	0.000	0.017	0.041
fin_state_change	0.8166	0.050	16.298	0.000	0.718	0.915
fin_resp	-0.7946	0.418	-1.900	0.057	-1.614	0.025
change_job	-0.1302	0.116	-1.120	0.263	-0.358	0.098
physical_activity	-0.0074	0.024	-0.301	0.764	-0.055	0.041
smoker	0.3595	0.083	4.326	0.000	0.197	0.522
depression	0.1326	0.176	0.752	0.452	-0.213	0.478
health	0.7230	0.072	10.071	0.000	0.582	0.864
level_of_trust	0.2710	0.052	5.185	0.000	0.169	0.373
is_religious	0.1372	0.071	1.938	0.053	-0.002	0.276
money_sq	-0.0002	6.09e-05	-3.258	0.001	-0.000	-7.91e-05
occupation_enterpren	0.3716	0.396	0.939	0.348	-0.404	1.147
occupation_student	0.8052	1.210	0.666	0.506	-1.566	3.177
occupation_worker	-0.1094	0.209	-0.523	0.601	-0.519	0.300
mar_st_divorced	0.2039	0.139	1.467	0.142	-0.068	0.476
mar_st_gr_brak	0.1660	0.133	1.246	0.213	-0.095	0.427
mar_st_single	0.0430	0.119	0.362	0.717	-0.190	0.276
mar_st_vdovets	-0.0489	0.216	-0.226	0.821	-0.473	0.375
1/2	6.6495	0.421	15.797	0.000	5.824	7.474
2/3	0.5648	0.029	19.609	0.000	0.508	0.621

Figure 8

**Our final model is an Ordered Logit model for three subsamples with classes thrown out after the Likelihood Ratio test and with the addition of Money Squared Term.

The following results were obtained.

	coef	std err	z	P>|z|	[0.025	0.975]
age	-0.0193	0.010	-1.898	0.058	-0.039	0.001
educ	0.0339	0.109	0.312	0.755	-0.179	0.247
vacation	0.5068	0.233	2.179	0.029	0.051	0.963
money	0.0646	0.018	3.569	0.000	0.029	0.100
fin_state_change	1.0257	0.144	7.148	0.000	0.744	1.307
fin_resp	-1.1280	0.897	-1.257	0.209	-2.886	0.630
change_job	-1.4352	0.285	-5.030	0.000	-1.994	-0.876
physical_activity	-0.0848	0.059	-1.435	0.151	-0.201	0.031
smoker	-0.0078	0.223	-0.035	0.972	-0.444	0.429
depression	-0.2898	0.406	-0.714	0.475	-1.085	0.505
health	0.4779	0.187	2.556	0.011	0.111	0.844
level_of_trust	0.3208	0.143	2.250	0.024	0.041	0.600
is_religious	-0.5010	0.189	-2.653	0.008	-0.871	-0.131
money_sq	-0.0004	0.000	-2.147	0.032	-0.001	-3.21e-05
occupation_enterpren	-1.6552	1.052	-1.573	0.116	-3.718	0.407
occupation_student	0	5.64e+07	0	1.000	-1.11e+08	1.11e+08
occupation_worker	-1.6266	0.727	-2.238	0.025	-3.051	-0.202
mar_st_divorced	-0.9503	0.321	-2.961	0.003	-1.579	-0.321
mar_st_gr_brak	-0.6832	0.334	-2.045	0.041	-1.338	-0.029
mar_st_single	-0.8154	0.298	-2.732	0.006	-1.400	-0.230
mar_st_vdovets	-0.8191	0.460	-1.781	0.075	-1.721	0.083
1/2	3.3907	1.292	2.625	0.009	0.859	5.922

Figure 9

We see that the accuracy of the forecast has increased in comparison with the result of the same model, where we did not divide the sample into three parts. For low, middle and high income, the accuracy of the model was 75.91%, 83.87%, and 79.43%, respectively.

Final models

Ordered Logit

$$\text{life\_satis} = \Lambda(0.0040 \pm 0.004 \times \text{age} + 0.2156 \pm 0.040 \times \text{educ} + 0.7096 \pm 0.079 \times \text{vacation} + \ldots) \text{ (Logit) (Figure 8)}$$

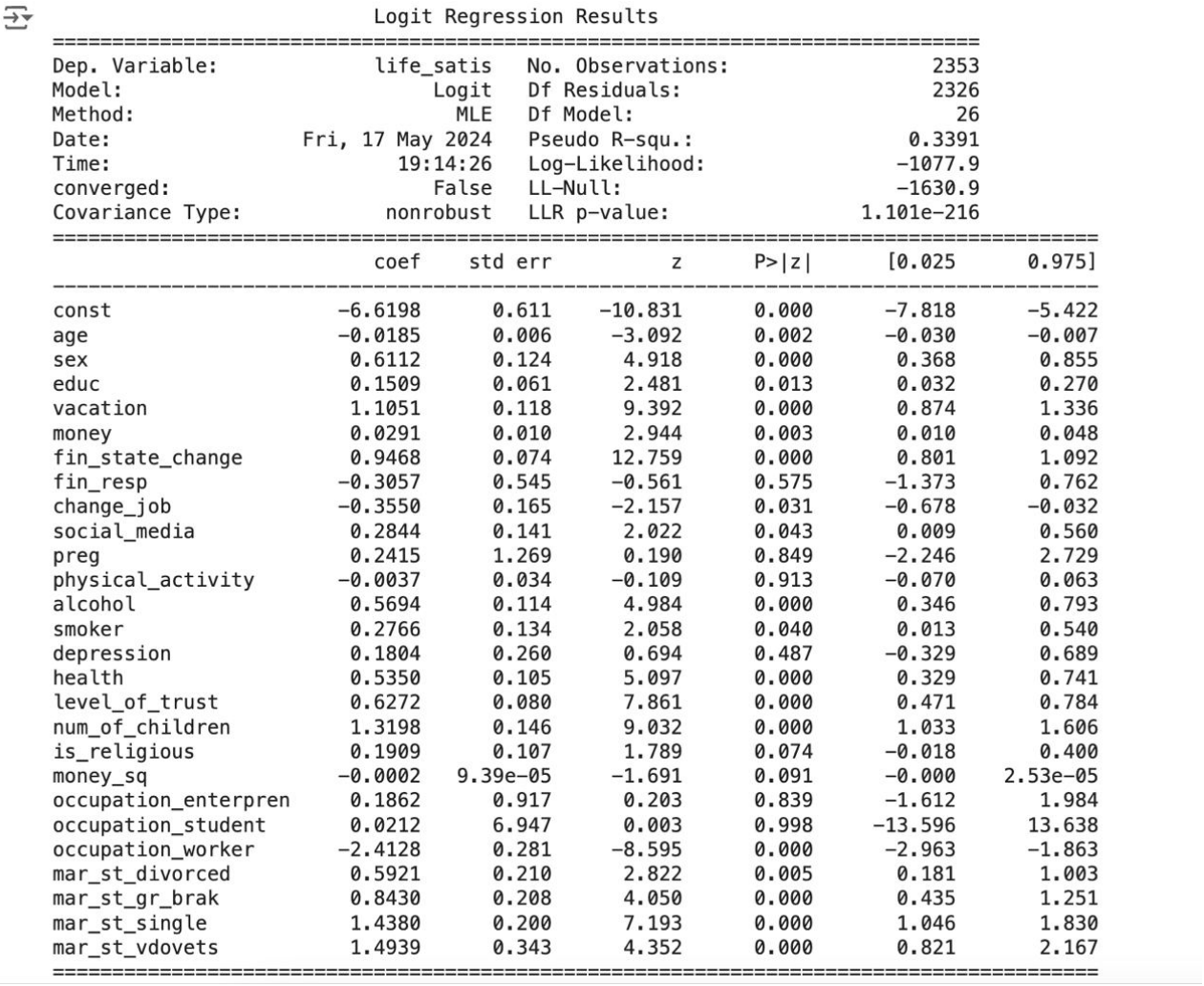
The ordered logit model explains the variable *'life_satis'* (life satisfaction), which is treated as a binary variable. The coefficients in the model represent the log-odds of being in a higher category of life satisfaction as the predictor variables increase. We used the following distribution of life satisfaction : 1+2, 3, 4+5 and achieved 53% accuracy.

Ordered model

$$\text{life\_satis} = \Lambda((-0.0193 \pm 0.010 \times \text{age}) + (0.0339 \pm 0.010 \times \text{educ}) + (0.5068 \pm 0.233 \times \text{vacation}) + \ldots) \text{ (Logit) (Figure 9)}$$

We achieved accuracy of 80.49%

Logit model



Logit Regression Results

Dep. Variable:	life_satis	No. Observations:	2353
Model:	Logit	Df Residuals:	2326
Method:	MLE	Df Model:	26
Date:	Fri, 17 May 2024	Pseudo R-squ.:	0.3391
Time:	19:14:26	Log-Likelihood:	-1077.9
converged:	False	LL-Null:	-1630.9
Covariance Type:	nonrobust	LLR p-value:	1.101e-216

	coef	std err	z	P>|z|	[0.025	0.975]
const	-6.6198	0.611	-10.831	0.000	-7.818	-5.422
age	-0.0185	0.006	-3.092	0.002	-0.030	-0.007
sex	0.6112	0.124	4.918	0.000	0.368	0.855
educ	0.1509	0.061	2.481	0.013	0.032	0.270
vacation	1.1051	0.118	9.392	0.000	0.874	1.336
money	0.0291	0.010	2.944	0.003	0.010	0.048
fin_state_change	0.9468	0.074	12.759	0.000	0.801	1.092
fin_resp	-0.3057	0.545	-0.561	0.575	-1.373	0.762
change_job	-0.3550	0.165	-2.157	0.031	-0.678	-0.032
social_media	0.2844	0.141	2.022	0.043	0.009	0.560
preg	0.2415	1.269	0.190	0.849	-2.246	2.729
physical_activity	-0.0037	0.034	-0.109	0.913	-0.070	0.063
alcohol	0.5694	0.114	4.984	0.000	0.346	0.793
smoker	0.2766	0.134	2.058	0.040	0.013	0.540
depression	0.1804	0.260	0.694	0.487	-0.329	0.689
health	0.5350	0.105	5.097	0.000	0.329	0.741
level_of_trust	0.6272	0.080	7.861	0.000	0.471	0.784
num_of_children	1.3198	0.146	9.032	0.000	1.033	1.606
is_religious	0.1909	0.107	1.789	0.074	-0.018	0.400
money_sq	-0.0002	9.39e-05	-1.691	0.091	-0.000	2.53e-05
occupation_enterpren	0.1862	0.917	0.203	0.839	-1.612	1.984
occupation_student	0.0212	6.947	0.003	0.998	-13.596	13.638
occupation_worker	-2.4128	0.281	-8.595	0.000	-2.963	-1.863
mar_st_divorced	0.5921	0.210	2.822	0.005	0.181	1.003
mar_st_gr_brak	0.8430	0.208	4.050	0.000	0.435	1.251
mar_st_single	1.4380	0.200	7.193	0.000	1.046	1.830
mar_st_vdovets	1.4939	0.343	4.352	0.000	0.821	2.167

Figure 10

$$\text{logit}(\Pr(\text{life\_satis}=1|X))=$$

$$
\begin{aligned}
&-6.6198 \\
&-0.0185 \times \text{age} \\
&+0.6112 \times \text{sex} \\
&+0.1509 \times \text{educ} \\
&+1.1051 \times \text{vacation} \\
&+0.0291 \times \text{money} \\
&-0.9468 \times \text{fin\_state\_change} \\
&+0.3067 \times \text{fin\_resp} \\
&-0.3550 \times \text{change\_job} \\
&+0.2844 \times \text{social\_media} \\
&+2.4415 \times \text{preg} \\
&-0.0037 \times \text{physical\_activity} \\
&+0.5964 \times \text{alcohol} \\
&+0.2766 \times \text{smoker} \\
&+0.1800 \times \text{depression} \\
&+0.5350 \times \text{health} \\
&+0.3272 \times \text{level\_of\_trust} \\
&+1.6118 \times \text{num\_children} \\
&-0.1909 \times \text{is\_religious} \\
&+9.000e{-}05 \times \text{money\_sq} \\
&-0.1862 \times \text{occupation\_entrepren} \\
&+0.0212 \times \text{occupation\_student} \\
&-2.4218 \times \text{occupation\_worker} \\
&+0.5921 \times \text{mar\_st\_divorced} \\
&+0.8430 \times \text{mar\_st\_gr\_brak} \\
&-1.4839 \times \text{mar\_st\_single} \\
&+1.4380 \times \text{mar\_st\_vdovets}
\end{aligned}
$$

The logit model (Figure 10) is a standard binary logit model, which is used to predict a binary outcome from predictor variables. We used it to predict the likelihood of an individual achieving a certain level of life satisfaction ('satisfied' vs. 'not satisfied'). We achieved accuracy of 79.18%.

Conclusion

- Ordered Logit model provide better performance compared to OLS and simple Logit model.
- SMOTE smoothing technique enhance models' accuracy.
- Money variable has non-linear relation with target variable that is why inclusion of squared term yields better performance.
- Likelihood Ratio test helped to identify excessive variables and simplify our model.
- Chow test provide insights for division data into separate social classes for avoiding data shifts. And according to accuracy results this insights we consider useful.
- The AIC and BIC values of 654.2 and 750.5, respectively, for the ordered logit model suggest that the model fits the data reasonably well, with the lower AIC indicating a potentially more favorable balance of model complexity and fit.

The following conclusions will be regarding obtained results from model compared to our initial assumptions.
This variables effect coincides with our assumptions:

- **Education (educ)**: With a coefficient of -0.0339 and a p-value of 0.001, education negatively impacts the target variable. The negative sign indicates that as education increases, the probability of achieving a lower category in the target variable increases

- **Vacation**: This variable has a positive coefficient of 0.5068 with a p-value of 0.001, suggesting that higher vacation scores lead to a higher category in the target variable. This indicates a beneficial effect of vacation on the target variable.

- **Financial Situation Change (fin_state_change)**: This has a significant positive coefficient of 1.0257 ($p < 0.001$), indicating that improvements in financial situation strongly push the target variable towards higher categories.

- **Health**: This variable's positive coefficient of 0.4779 ($p = 0.011$) suggests that better health status is associated with higher scores on the target variable.

- **Level of Trust (level_of_trust)**: The positive coefficient of 0.3208 (p = 0.024) suggests that a higher level of trust is associated with higher categories of the target variable.

- **Marital Status**: All categories of marital status listed (divorced, break, single, and widowed) have significant negative effects on the target variable compared to being married, suggesting that marital stability might positively influence the target variable.

- M**oney:** The coefficient for the term "money" output is 0.0646 with a standard error of 0.018 and a p-value of 0.001. This is statistically significant, indicating a reliable effect in the model.

And this differs from our initial assumptions:

- **Occupational Category**: Being an entrepreneur (occupation_entrepren) has a significant negative coefficient of -1.6552, suggesting a negative impact on the target variable compared to other professions. Similarly, being a worker (occupation_worker) with a coefficient of -1.6266 (p = 0.025) also negatively impacts the target variable.

- **Religiousness (is_religious)**: Having a significant negative coefficient of -0.5010 (p = 0.008), being more religious correlates with lower scores on the target variable.

- **Money Squared (money_sq)**: The extremely small but significant coefficient of -0.0004 (p = 0.032) indicates a slight negative effect as the squared money value increases.

So we assume the following possible reasons of such interesting effects:

- **Occupational Category (Entrepreneurs and Workers)**: Entrepreneurs and workers may face negative impacts due to stress and long hours which could detract from well-being.

- **Religiousness**: Higher religiousness might result in lower scores on the target variable due to differing life values or priorities.

- **Money Squared (money_sq)**: The negative coefficient suggests that excessive wealth could lead to stress or social isolation, indicating diminishing returns on well-being.