MULTINOMIAL LOGISTIC REGRESSION

Logistic Regression with SAS

Sahil Palsaniya (110093)

This work is carried out as part of a Logistic Regression course at Warsaw School of Economics in the academic year 2022/2023 under care of PhD Adam Korczyński

MASTER'S DEGREE

Advanced Analytics – Big Data

Evaluating the Impact of Perception of national infrastructure on Income Type

1. Introduction

Traditionally, income had been associated with the labour work and wages earned in the labour market. However, the world has changed and there are so many diverse sources of income nowadays. This transformation has piqued my curiosity, specifically about the factors that drive this shift. So, we started to become curious what kind of factors may influence the people' main source of income. Among those possible factors, we wanted to see if 'People' satisfaction with their country's national infrastructure has any effect on their source of income. By national infrastructure we mean societal (e.g., education, healthcare) and economic (e.g., state of the economy) aspects of a country. We believe these seemingly distinct aspects eventually translate to overall satisfaction with countries' present state for living and thus influence their sources of income.
The main hypothesis we aim to investigate is: "Individuals with a negative perception of their country's national infrastructure are more likely to rely on Social Grants/Benefits Income as their main source of income compared to Labour or Capital income."

We will use SAS studio to proceed with descriptive analysis, create a multinomial logistic regression model and interpret the results in detail.

2. Descriptive Statistics

2.1. Frequency distribution and missing data

The ESS dataset is a survey data. Which means we've got elements like 'Not Applicable (66)', 'Refusal (77)', 'Don't know (88)', and 'No answer (99)' (sometimes even presented as 7 or 9) mixed in the data. In essence, these are missing values, and they require our attention before we proceed with any further analysis. Before we decide to address these gaps with techniques

such as imputation or consider removing them entirely, we will perform a thorough assessment to investigate the proportion of missing data in each variable, and do they form any recognizable pattern?

For checking missing data patterns, we used the PROC MI statement. The most significant missing data pattern in our dataset is the absence of the 'inc' variable, which accounts for over 18% of the data (Group 17). This is significant because it constitutes a considerable portion of our data.

Other notable patterns include missing data in '_stfeco', '_stfedu', and some other combinations. However, each of these groups individually constitutes less than 3% of your data, which may not be as impactful as the missing 'inc' values. Here, we only show the part of our result table since most of the rows show a very low percent which is almost 0. You can find the original whole table in the Appendix.
Based on these observations, we decided to drop the missing values and also the ‘inc’ (hinctnta) variable since it could heavily impact our analysis and is also not very important for our topic.

Missing Data Patterns

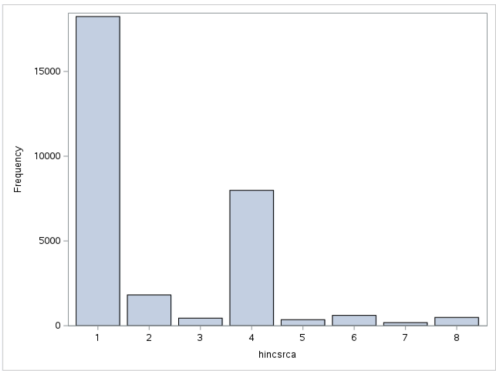
										Group Means						
Group	Y	inc	_stfhlth	_stfedu	_stfeco	_agea	_stfgov	Freq	Percent	Y	inc	_stfhlth	_stfedu	_stfeco	_agea	_stfgov
1	X	X	X	X	X	X	X	24221	72.62	1.764006	1.986004	2.227860	2.243466	2.016060	2.364684	1.939309
2	X	X	X	X	X	X	.	247	0.74	1.862348	1.801619	2.178138	2.251012	1.979757	2.295547	.
3	X	X	X	X	.	X	X	207	0.62	1.932367	1.734300	2.246377	2.207729	.	2.323671	2.009662
4	X	X	X	X	.	X	.	73	0.22	1.630137	1.739726	2.246575	2.219178	.	2.205479	.
5	X	X	X	.	X	X	X	866	2.60	2.274827	1.608545	2.038106	.	1.811778	2.728637	1.792148
6	X	X	X	.	X	X	.	50	0.15	1.880000	1.580000	2.020000	.	1.880000	2.480000	.
7	X	X	X	.	.	X	X	66	0.20	2.757576	1.272727	2.136364	.	.	2.939394	1.818182
8	X	X	X	.	.	X	.	35	0.10	2.514286	1.142857	2.314286	.	.	2.771429	.
9	X	X	.	X	X	X	X	41	0.12	1.878049	1.829268	.	2.268293	1.926829	2.219512	1.804878
10	X	X	.	X	X	X	.	2	0.01	2.000000	1.500000	.	1.500000	1.500000	3.000000	.
11	X	X	.	X	.	X	X	7	0.02	2.285714	1.857143	.	2.428571	.	2.285714	2.000000
12	X	X	.	X	.	X	.	3	0.01	1.666667	2.333333	.	2.000000	.	1.333333	.
13	X	X	.	.	X	X	X	61	0.18	2.278689	1.524590	.	.	1.885246	2.573770	1.688525
14	X	X	.	.	X	X	.	7	0.02	1.285714	1.714286	.	.	1.428571	2.000000	.
15	X	X	.	.	.	X	X	6	0.02	3.000000	1.000000	.	.	.	3.000000	1.666667
16	X	X	.	.	.	X	.	27	0.08	2.296296	1.370370	.	.	.	2.444444	.
17	X	.	X	X	X	X	X	6100	18.29	1.709016	.	2.089508	2.097377	1.814098	2.244098	1.809672
18	X	.	X	X	X	X	.	147	0.44	1.707483	.	2.204082	2.238095	1.931973	2.081633	.
19	X	.	X	X	.	X	X	99	0.30	1.707071	.	2.444444	2.313131	.	1.969697	2.030303
20	X	.	X	X	.	X	.	58	0.17	1.603448	.	2.327586	2.310345	.	1.844828	.

2.2. Target Variable hincsrca - Main source of household income

The question respondents were asked is as follows: “Please consider the income of all household members and any income which may be received by the household as a whole. What is the main source of income in your household?”. You can see possible answers to this question in the table below.

Levels	Categories
1	Wages or salaries
2	Income from self-employment (excluding farming)
3	Income from farming
4	Pensions
5	Unemployment/redundancy benefit
6	Any other social benefits or grants
7	Income from investments, savings etc.
8	Income from other sources
77	Refusal
88	Don't know
99	No answer

You can see the distribution of our target variable on the plot below. For the majority of respondents the main source of income is either “Wages or salaries” (60.56%) or “Pensions” (26.52%). The distribution between all other main income sources except two mentioned before is relatively the same.



The FREQ Procedure

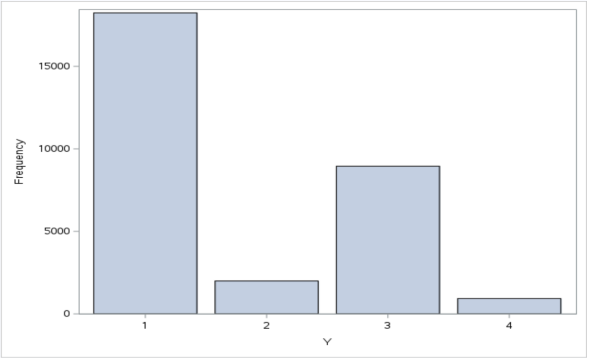
hincsrca	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	18239	60.56	18239	60.56
2	1813	6.02	20052	66.58
3	444	1.47	20496	68.06
4	7987	26.52	28483	94.58
5	355	1.18	28838	95.76
6	608	2.02	29446	97.78
7	182	0.60	29628	98.38
8	487	1.62	30115	100.00

For an effective Multinomial Logistic Regression analysis, we grouped these into four categories. The 'Income from self-employment (excluding farming) (2)' was treated as 'Capital Income', given its nature as wealth-generated income. Although 'Income from

farming (3)' could be considered 'Labour Income', we placed it in 'Other Income' due to its low representation (1.47%).

Post-regrouping, our target variable 'Y' included 'Labour Income (1)' at 60.56%, 'Capital Income (2)' at 6.62%, 'Social Grants/Benefit Income (3)' at 29.72%, and 'Other Income (4)' at 3.09%. This regrouping aids in clearer analysis and interpretation.

Previous Levels	New Levels	Categories
1	1	Labour Income
2, 7	2	Capital Income
4, 5, 6	3	Social Grants/ Benefit Income
3, 8	4	Other Income

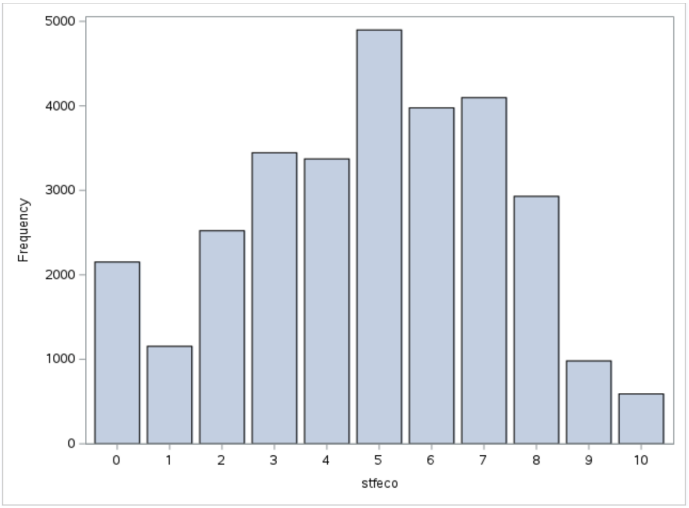


2.3. Explanatory Variables

Now we will describe 5 explanatory variables that we choose for our model. Those 5 variables are *stfeco, stfgov, stfhlth, stfedu,* and *agea.*

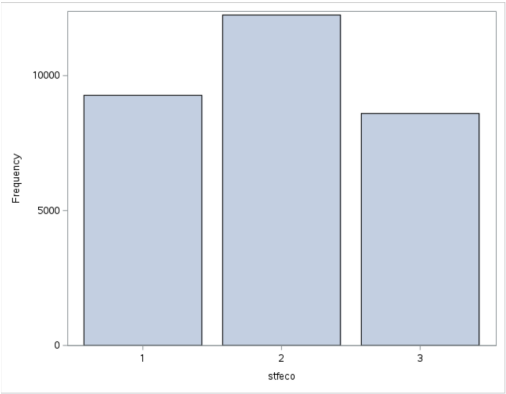
2.3.1. Variable stfeco: How satisfied with present state of economy in country

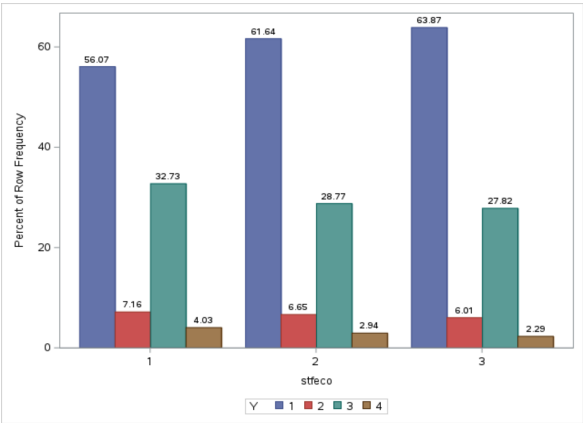
Possible answers for this variable are from 0 to 10 with "0" standing for "Extremely dissatisfied" and "10" standing for "Extremely satisfied". You can see on the left plot below that distribution of variable *stfeco*. To have a better interpretation of modelling results later, we decided to group the levels of this variable into 3 following categories: "Low Satisfaction (1)", "Mid Satisfaction (2)", "High Satisfaction (3)". As per the table below, levels 0, 1, 2, and 3 are now equal to level 1 (low satisfaction); levels 4, 5, and 6 are now equal to level 2 (mid satisfaction); and levels 7, 8, 9, and 10 are now equal to level 3 (high satisfaction).



Previous Levels	New Levels	Categories
0~3	1	Low Satisfaction
4~6	2	Mid Satisfaction
7~10	3	High Satisfaction

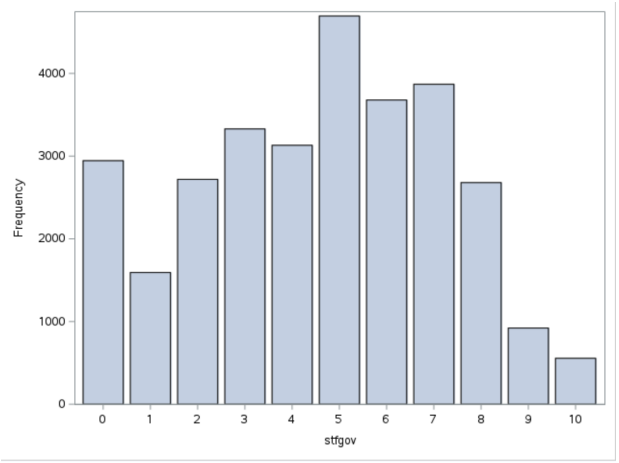
The distribution of our variable 'stfeco', after grouping, reveals that the highest proportion of respondents, around 40.66%, demonstrate medium satisfaction. In contrast, the group expressing low satisfaction contains the smallest proportion of respondents at 30.79%.However, upon analysis of our target variable 'Y' across the 'stfeco' categories shows that the group indicating low satisfaction with the state of the economy exhibits the highest proportion of 'Social Grants/Benefit Income' as their income source. To elaborate, about 33.91% of those reporting low economic satisfaction state that their primary income source is social grants or benefits.This indicates that individuals with lower satisfaction regarding their country's economy are more likely to depend on social grants or benefit income as their main income source.

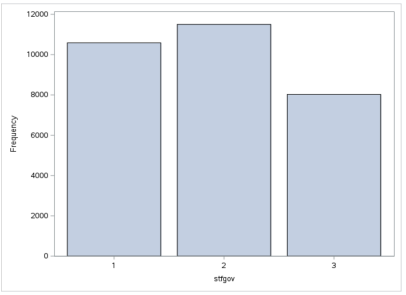




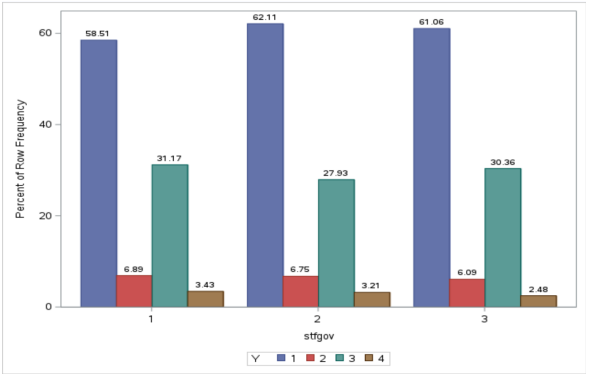
2.3.2. Variable stfgov: How satisfied with the national government

You can see on the plot below the distribution of variable *stfgov* with "0" standing for "Extremely dissatisfied" and "10" standing for "Extremely satisfied". As in the previous case, after checking the distribution of variable *stfgov*, we decided to group the levels of this variable into the same 3 categories: "Low Satisfaction", "Mid Satisfaction", "High Satisfaction". Levels 0 - 3 are equal to level 1 (low satisfaction); levels 4 - 6 are equal to level 2 (mid satisfaction); and levels 7 - 10 are equal to level 3 (high satisfaction).

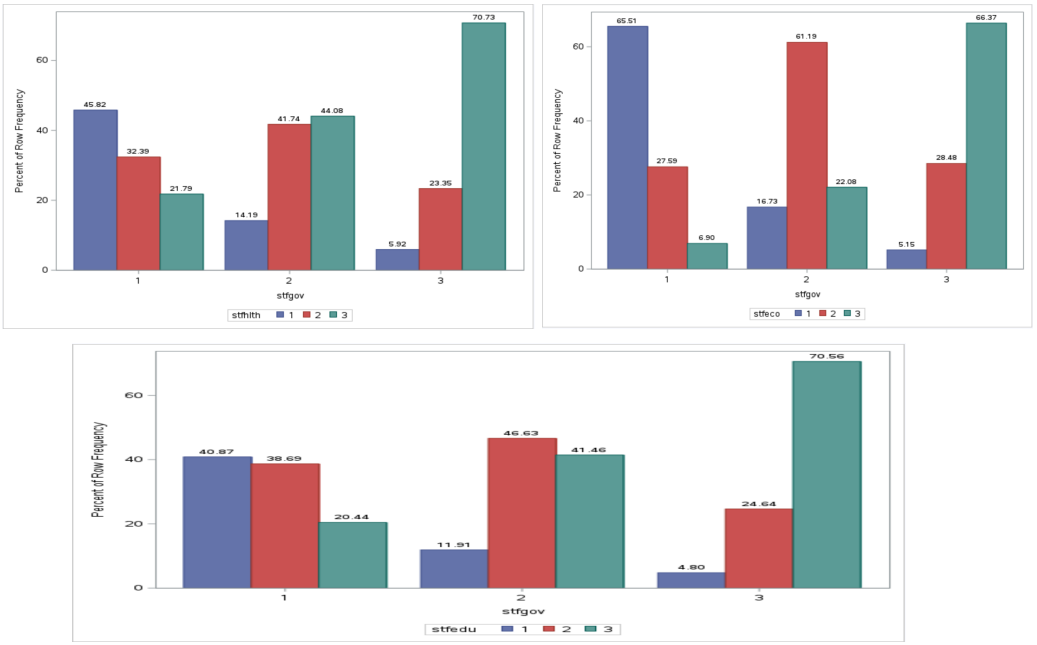




The largest share of respondents fell into the 'Mid Satisfaction (2)' category at 38.20%, closely followed by 'Low Satisfaction (1)' at 35.15%. The 'High Satisfaction (3)' category was less represented, holding 26.65%. The distribution shows little variation in proportional representation of income sources across the satisfaction categories, suggesting a limited discriminatory performance of 'stfgov' for predicting income source. These trends highlight that while there may be some influence, satisfaction with government does not distinctly differentiate between income sources.

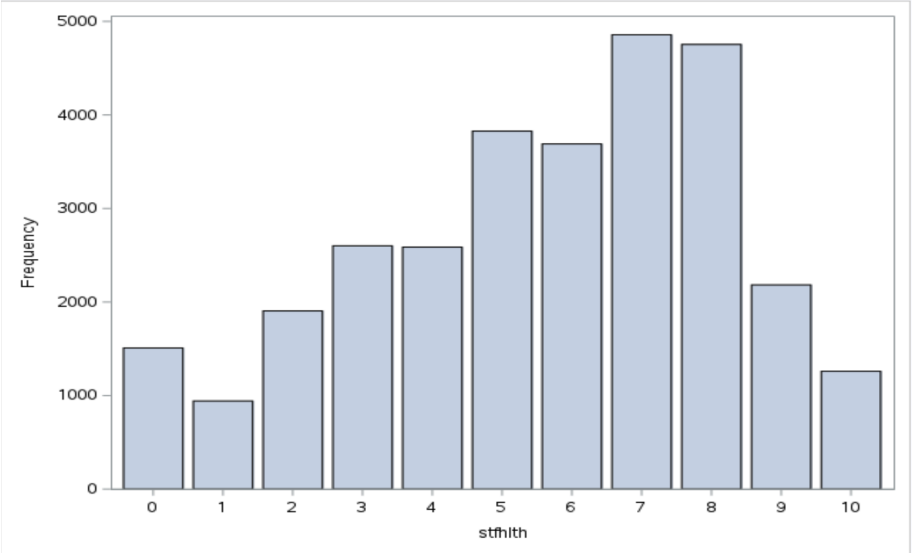


Although not the primary focus of our study, upon comparing stfgov with our other main explanatory variables, it is interesting to observe that people who have low satisfaction with current government have low satisfaction level in health services, education and current state of economy. This points towards collinearity between our variables. However, we will analyse collinearity in detail at the later stages of our project.



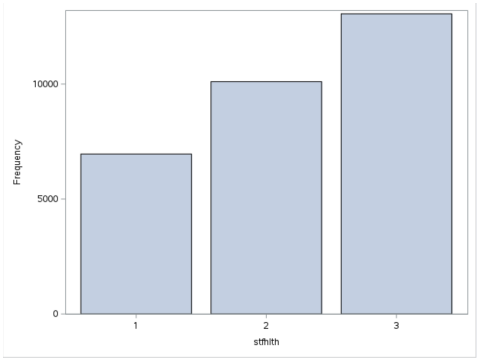
2.3.2. Variable stfhlth: How satisfied with the health services

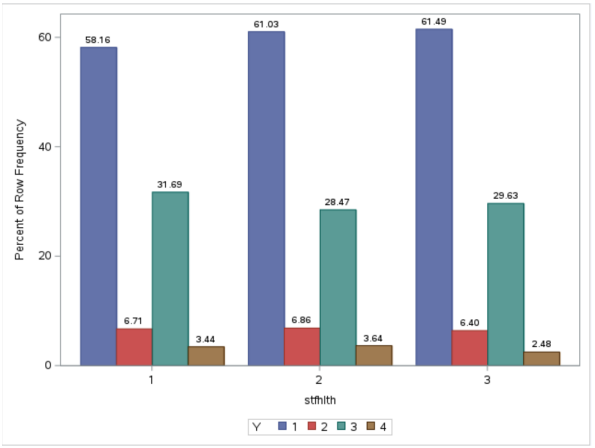
Possible answers for our 'stfhlth' variable, which refers to the level of satisfaction with health services, are same as our other explanatory variables with "0" standing for "Extremely dissatisfied" and "10" standing for "Extremely satisfied".



we see a substantial proportion of respondents grouped into the 'High Satisfaction' category, representing 43.35% of responses. Following this, the 'Mid Satisfaction', 33.55%, and the 'Low Satisfaction' group made up 23.01% of the responses.

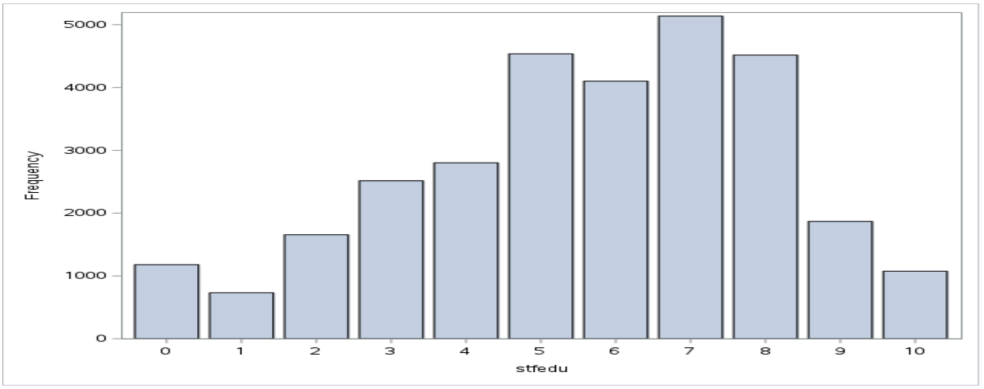
In observing the intersection between 'stfhlth' and our target variable 'Y', we notice a relatively consistent pattern of distributions across the three satisfaction categories. The 'Labour Income (1)' option was predominant across all satisfaction categories, followed by 'Social Grants/Benefit Income (3)', and then 'Capital Income (2)' and 'Other Income (4)'. The representation of income sources within each satisfaction level remained quite steady, suggesting that 'stfhlth' may not be a robust predictor of income source. This analysis proposes that while satisfaction with health might have some influence, it does not clearly differentiate between different income sources.



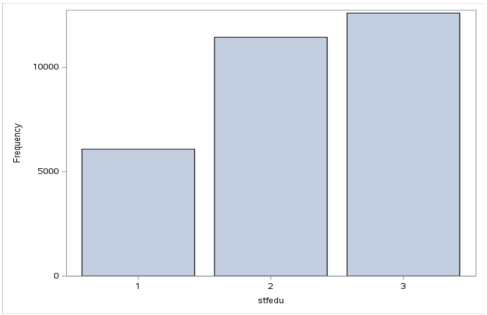


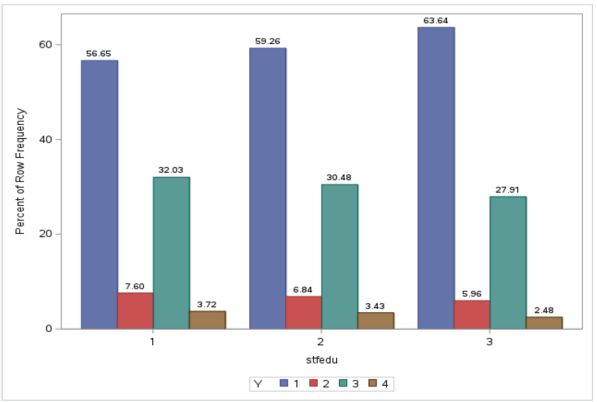
2.3.2. Variable stfedu: How satisfied with current state of education

'stfedu', indicates respondents' satisfaction with the current state of education. It has a similar distribution as stf health. High Satisfaction (3)' category holds the largest proportion of responses, comprising 41.83% of the total. The 'Medium Satisfaction (2)' category comes next with 37.98%, while 'Low Satisfaction (1)' accounts for 20.19% of responses.



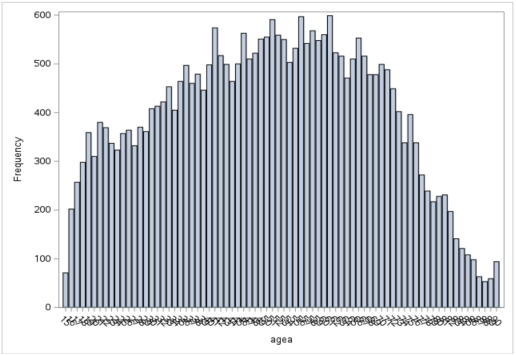
When viewing the association between 'stfedu' and our target variable 'Y', we see a pattern similar to 'stfhlth'. The most common income source across all education satisfaction levels remains 'Labour Income (1)', followed by 'Social Grants/Benefit Income (3)', then 'Capital Income (2)', and finally 'Other Income (4)'.
The consistency in these distributions suggests that, like 'stfhlth', the 'stfedu' variable may not be a particularly influential predictor for income source. Although the satisfaction level with education might have some bearing on income source, the distinction among the different sources of income is not clear-cut within each category of 'stfedu'.

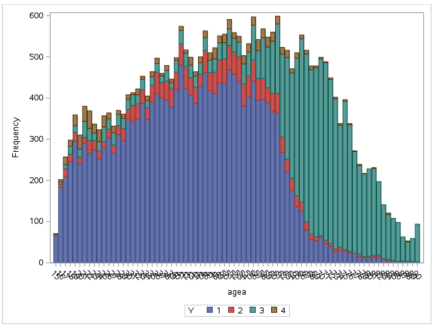




2.3.2. Variable agea: How satisfied with current state of education

The distribution of age can be seen below. Percentage of income from social grants or benefits starts slowly increasing for respondents in their fifties and continues to increase even more drastically in their sixties and above. For respondents that are seventy years old and older the main source of income is almost solely income from social grants or benefits.

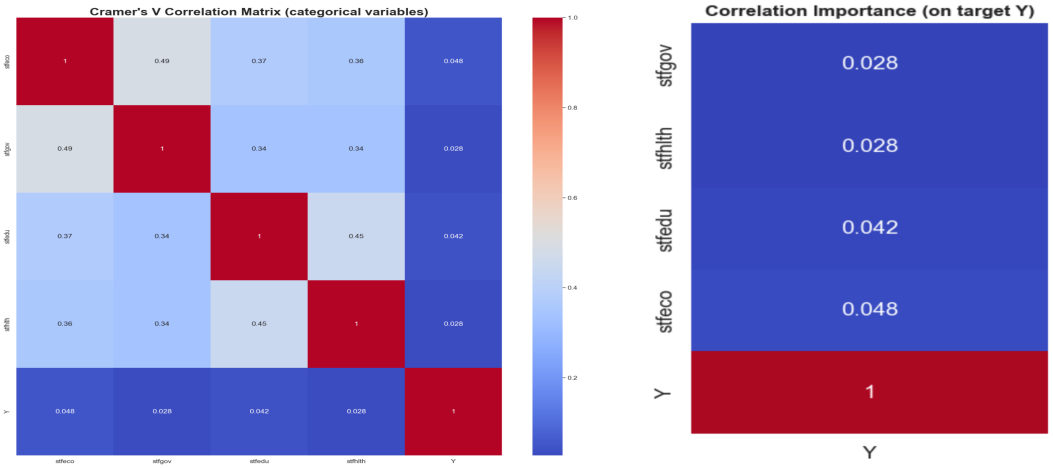




3. Substantive Analysis

3.1. Collinearity assessment

Before we build our model, we need to check if there is any possible correlation between our X explanatory variables. It is important since collinearity in logistic regression may lead to overestimation of standard errors and regression coefficients. Since we only have 1 continuous variable, we will only check collinearity between categorical variables in our dataset, we will use Cramer's V coefficients test. For this Collinearity assessment part, we utilised 'Python' to see a more intuitive correlation matrix as below (code can be found in the appendix).



As we can see in the plot above, there is no severe correlation. Even though it's moderate correlation with 0.49, stfgov and stfeco also show a bit of correlation. It is natural that they

are correlated since a national government literally influences a lot of the economy of the country based on their policy(and vice versa). There is also correlation of 0.45 between stfhlth and stfedu and thus can be combined, but we would still like to interpret their individual effects on sources of income because they each give us information about different infrastructure in an economy and since it is not a very strong correlation.

3.2. Stepwise Variable Selection

The summary of stepwise selection shows the variables that were entered into the model in order of their significance, with agea being the most significant, followed by stfeco, stfedu, and stfhlth. Each of these variables was found to significantly improve the model at the time they were entered, hence they were included in the final model. ‘stfgov was not included in the final model because it did not meet the 0.05 significance level for entry into the model. This would suggest that, given the other variables in the model, stfgov does not provide additional significant information for predicting the outcome variable Y in our dataset.

Summary of Stepwise Selection

Step	Effect Entered	Effect Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq
1	agea		3	1	11616.9012		<.0001
2	stfeco		6	2	153.3388		<.0001
3	stfedu		6	3	29.8671		<.0001
4	stfhlth		6	4	17.4014		0.0079

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
stfhlth	6	17.3762	0.0080
stfedu	6	33.9599	<.0001
stfeco	6	77.1309	<.0001
agea	3	6934.1560	<.0001

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	14637.6390	3	<.0001
Score	11616.9012	3	<.0001
Wald	6951.4902	3	<.0001

3.3. Model estimates

For Maximum Likelihood Estimate, there are some variables with higher p-value than 0.05, which means they are statistically insignificant. In the result table below, the variables with the red box are insignificant ones. Although most of the categories required to answer our hypothesis survived, stfhlth variable's second category with respect to labour income (Y=3) turns out to be insignificant. However, no clear conclusion can be made here, let's look at odd's ratio for better interpretation.

Analysis of Maximum Likelihood Estimates

Parameter		Y	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Exp(Est)
Intercept		2	1	-2.9667	0.1033	824.7550	<.0001	0.051
Intercept		3	1	-8.0132	0.1067	5642.9973	<.0001	0.000
Intercept		4	1	-2.9851	0.1412	447.0894	<.0001	0.051
stfhlth	2	2	1	0.1601	0.0713	5.0485	0.0246	1.174
stfhlth	2	3	1	0.0310	0.0511	0.3671	0.5446	1.031
stfhlth	2	4	1	0.2276	0.0964	5.5724	0.0182	1.256
stfhlth	3	2	1	0.2245	0.0772	8.4497	0.0037	1.252
stfhlth	3	3	1	0.0682	0.0553	1.5204	0.2176	1.071
stfhlth	3	4	1	0.0355	0.1093	0.1056	0.7452	1.036
stfedu	1	2	1	0.1725	0.0703	6.0198	0.0141	1.188
stfedu	1	3	1	0.0862	0.0520	2.7508	0.0972	1.090
stfedu	1	4	1	0.0392	0.0966	0.1647	0.6849	1.040
stfedu	3	2	1	-0.1943	0.0605	10.3282	0.0013	0.823
stfedu	3	3	1	-0.1070	0.0437	6.0093	0.0142	0.899
stfedu	3	4	1	-0.2164	0.0871	6.1777	0.0129	0.805
stfeco	1	2	1	0.1214	0.0607	4.0000	0.0455	1.129
stfeco	1	3	1	0.1701	0.0442	14.7780	0.0001	1.185
stfeco	1	4	1	0.3953	0.0829	22.7505	<.0001	1.485
stfeco	3	2	1	-0.1040	0.0628	2.7448	0.0976	0.901
stfeco	3	3	1	-0.1911	0.0453	17.7595	<.0001	0.826
stfeco	3	4	1	-0.1688	0.0955	3.1239	0.0772	0.845
agea		2	1	0.0144	0.00156	85.0572	<.0001	1.014
agea		3	1	0.1293	0.00156	6886.6657	<.0001	1.138
agea		4	1	-0.00245	0.00226	1.1839	0.2766	0.998

3.4. Odds ratios with CI & Interpretations

Since we are considering 3 variables to associate national infrastructure ‘stfhlth’, ‘stfedu’ and ‘stfeco’, we will interpret each of them separately.

- **Satisfaction with the economy (stfeco)**: Compared to people who are less satisfied with the economy (stfeco = 1), people who are moderately satisfied (stfeco = 2) have 0.844 times the odds of relying on Social Grants/Benefit Income as their main income source. Those who are highly satisfied (stfeco = 3) have even lower odds, with 0.697 times the odds of relying on Social Grants/Benefit Income. This aligns with our hypothesis, indicating that people less satisfied with the economy are more likely to rely on social benefits/grants income.
- **Satisfaction with education (stfedu)**: Similarly, people who are less satisfied with education (stfedu = 1) are comparatively less likely (0.917 times the odds) to rely on Social Grants/Benefit Income than people who are moderately satisfied (stfedu = 2) or highly satisfied (stfedu = 3).
- **Satisfaction with healthcare (stfhlth)**: Compared to people who are less satisfied with healthcare (stfhlth = 1), people who are moderately satisfied (stfhlth = 2) have about the same odds (1.031 times) of relying on Social Grants/Benefit Income. However, people who are highly satisfied (stfhlth = 3) have slightly higher odds (1.071 times) of relying on Social Grants/Benefit Income. This does not directly support your hypothesis. However, the observed trend for stfhlth reflects a complex dynamic in which individuals who rely more heavily on Social Grants/Benefit Income might actually be more satisfied with healthcare services. This could be the case if those individuals are also the ones who benefit most directly from social healthcare services or public health provisions, thus increasing their satisfaction with the healthcare system.

In conclusion, lower satisfaction with the state of the economy and education, key aspects of national infrastructure, are associated with a greater likelihood of individuals relying on Social Grants/Benefits Income. The relationship between satisfaction with healthcare and

income source is less clear, with similar odds across satisfaction levels. But as argued above, It indicates that in fact stfhlth is probably not a good indicator to prove or disprove our particular hypothesis.

Odds Ratio Estimates and Profile-Likelihood Confidence Intervals					
Effect	Y	Unit	Estimate	95% Confidence Limits	
stfhlth 2 vs 1	2	1.0000	1.174	1.021	1.350
stfhlth 2 vs 1	3	1.0000	1.031	0.933	1.140
stfhlth 2 vs 1	4	1.0000	1.256	1.040	1.518
stfhlth 3 vs 1	2	1.0000	1.252	1.076	1.457
stfhlth 3 vs 1	3	1.0000	1.071	0.961	1.193
stfhlth 3 vs 1	4	1.0000	1.036	0.837	1.284
stfedu 1 vs 2	2	1.0000	1.188	1.035	1.363
stfedu 1 vs 2	3	1.0000	1.090	0.984	1.207
stfedu 1 vs 2	4	1.0000	1.040	0.860	1.256
stfedu 3 vs 2	2	1.0000	0.823	0.731	0.927
stfedu 3 vs 2	3	1.0000	0.899	0.825	0.979
stfedu 3 vs 2	4	1.0000	0.805	0.679	0.955
stfeco 1 vs 2	2	1.0000	1.129	1.002	1.271
stfeco 1 vs 2	3	1.0000	1.185	1.087	1.293
stfeco 1 vs 2	4	1.0000	1.485	1.262	1.747
stfeco 3 vs 2	2	1.0000	0.901	0.797	1.019
stfeco 3 vs 2	3	1.0000	0.826	0.756	0.903
stfeco 3 vs 2	4	1.0000	0.845	0.700	1.017
agea	2	1.0000	1.014	1.011	1.018
agea	3	1.0000	1.138	1.135	1.142
agea	4	1.0000	0.998	0.993	1.002

3.5 Model diagnostics

The metrics such as AIC, SC, -2 Log L will be useful when we compare this model to the other models. Usually, the lower these values are, the better the model is. For R-Square, the higher this value is, the better the model is (in terms of prediction power). In our project, all these metrics are not utilized that much as we didn't build other models to compare, but they were actually used in the process of Variable Selection with Step wise method to compare models by adding and deleting explanatory variables.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	57321.276	42526.020
SC	57346.214	42725.527
-2 Log L	57315.276	42478.020

R-Square	0.3890	Max-rescaled R-Square	0.4572

Deviance and Pearson Goodness-of-Fit Statistics				
Criterion	Value	DF	Value/DF	Pr > ChiSq
Deviance	7816.4313	5769	1.3549	<.0001
Pearson	15373.4955	5769	2.6648	<.0001

For our model, we cannot rely entirely on the Deviance and Pearson Goodness-of-Fit Statistics given the complexity of our model which contains multiple explanatory variables with on average 3 categories per variable, as well as two continuous variables (age and number of child). These statistics show a significant deviation from a good fit with p-values less than 0.0001. Hence, we would normally check the Hosmer and Lemeshow Goodness-of-Fit test for a more definitive understanding of the model's performance. However, a noteworthy limitation is the fact that these tests often reject the hypothesis of perfect fit with very low p-value in large data samples. Our data sample is quite large with 38053 observations. The influence of sample size might lead to a rejection of the model due to the sheer volume of the data, as suggested in academic research (Nattino, Pennell and Lemeshow, 2020).

To address this issue, we could consider resampling a smaller subset of our data and retesting or utilising a parameter which isn't dependent on the sample size. Although we haven't implemented these steps in this case due to constraints on skills and time, it's important to be aware of these limitations and possible ways to address them.

Meanwhile, it's encouraging to observe that our model's Akaike Information Criterion (AIC) and Schwarz Criterion (SC) decreased significantly when the covariates were included. This suggests that the model with the covariates is a better fit compared to the intercept-only model. The R-Square value of 0.3890 and the Max-rescaled R-Square value of 0.4572 further validate that our model explains a reasonable proportion of the variation in the outcome variable, lending credibility to our findings despite the limitations.

4. Extended research on the topic (Innovative aspect)

In a first-of-its-kind study, we dissected the socio-economic situation in Eastern and Southern European countries by assessing the impact of economic and health satisfaction, as well as education level, on people's happiness. We innovatively divided the sample into four distinct age groups: young (18-30), middle-aged (31-50), senior (51-65), and old (above 65). This granular age categorization helped in unveiling the varying effects of these parameters on happiness across different life stages.

We utilized the PROC LOGISTIC function in SAS software to implement logistic regression and derive maximum likelihood estimates, helping us understand the relationships between the predictor variables and happiness level. An innovative step in our study was the use of multinomial logistic regression, which can handle dependent variables with more than two categories.

In the first part of the study involving the young age group, the relationship between happiness and economic satisfaction (stfeco), health satisfaction (stfhlth), and education level (stfedu) was explored. The findings showed significant variations in the odds of being happy based on economic and health satisfaction, as well as education level. For example, when comparing different levels of economic satisfaction, we found that the odds of being happy were reduced for lower levels of satisfaction, indicating a strong positive relationship between economic satisfaction and happiness.

For the middle-aged group, a similar pattern emerged, but the strength of the relationship between economic satisfaction and happiness appeared to be stronger than in the young group. Furthermore, the study innovatively highlighted the diminishing role of education level in affecting happiness as people age.

Finally, in the senior and old age groups, the effects of health satisfaction became more prominent, further emphasizing the changing dynamics of happiness predictors across age groups.

Through the integration of detailed age segmentation and the application of multinomial logistic regression, we provided a more nuanced understanding of the predictors of happiness across different life stages in Eastern and Southern Europe. This innovative approach offers valuable insights for policymakers and social workers striving to enhance happiness and well-being in these regions. By targeting policies to specific age groups and focusing on the

most influential factors within each group, our research can help pave the way towards more effective strategies for improving public happiness and satisfaction.

Odds Ratio Estimates and Profile-Likelihood Confidence Intervals					
Effect	Y	Unit	Estimate	95% Confidence Limits	
stfeco 2 vs 1	2	1.0000	0.987	0.831	1.174
stfeco 2 vs 1	3	1.0000	0.634	0.516	0.777
stfeco 2 vs 1	4	1.0000	0.589	0.438	0.790
stfeco 3 vs 1	2	1.0000	0.836	0.676	1.034
stfeco 3 vs 1	3	1.0000	0.398	0.301	0.524
stfeco 3 vs 1	4	1.0000	0.451	0.301	0.664
stfedu 2 vs 1	2	1.0000	0.781	0.647	0.942
stfedu 2 vs 1	3	1.0000	1.084	0.863	1.367
stfedu 2 vs 1	4	1.0000	1.117	0.816	1.539
stfedu 3 vs 1	2	1.0000	0.650	0.529	0.798
stfedu 3 vs 1	3	1.0000	0.885	0.683	1.146
stfedu 3 vs 1	4	1.0000	0.660	0.450	0.966

Odds Ratio Estimates and Profile-Likelihood Confidence Intervals					
Effect	Y	Unit	Estimate	95% Confidence Limits	
stfeco 2 vs 1	2	1.0000	0.797	0.663	0.959
stfeco 2 vs 1	3	1.0000	0.728	0.645	0.822
stfeco 2 vs 1	4	1.0000	0.562	0.430	0.732
stfeco 3 vs 1	2	1.0000	0.622	0.504	0.765
stfeco 3 vs 1	3	1.0000	0.513	0.447	0.589
stfeco 3 vs 1	4	1.0000	0.319	0.225	0.446

Odds Ratio Estimates and Profile-Likelihood Confidence Intervals					
Effect	Y	Unit	Estimate	95% Confidence Limits	
stfedu 2 vs 1	2	1.0000	1.059	0.601	1.938
stfedu 2 vs 1	3	1.0000	0.696	0.549	0.877
stfedu 2 vs 1	4	1.0000	0.725	0.392	1.384
stfedu 3 vs 1	2	1.0000	1.200	0.693	2.170
stfedu 3 vs 1	3	1.0000	0.680	0.538	0.854
stfedu 3 vs 1	4	1.0000	0.397	0.201	0.794

5. Conclusion

In conclusion, we have discussed the influence factors on main source of household income throughout EDA, checked multicollinearity Cramer's V for categorical variables, using Python. Based on all those steps, we created the multinomial logistic regression model that supported our hypothesis (true) about the influence of 'satisfaction on one's own country's present economy' on 'type of main source of household income'. Innovation came in the form of multinomial logistic regression, allowing us to handle our happiness variable that had more than two categories effectively. This enabled us to discern the nuanced influences of the predictor variables on happiness across varying age groups. The effect of economic satisfaction on happiness was found to be significant across all age groups, with its magnitude varying according to age

6. Bibliography

Nattino, G., Pennell, M. and Lemeshow, S., 2020. Assessing the goodness of fit of logistic regression models in large samples: A modification of the Hosmer‐Lemeshow test. Biometrics, 76(2), pp.549-560.

Analyticsvidhya. 2015. What to do when Hosmer lemeshow test fails during Logistic regression?. [online] Available at: <https://discuss.analyticsvidhya.com/t/what-to-do-when-hosmer-lemeshow-test-fails-during-logistic-regression/2304/2> [Accessed 10 June 2021].

Statalist (The STATA forum). 2021. Hosmer Lemeshow test for large data. [online] Available at: <https://www.statalist.org/forums/forum/general-stata-discussion/general/1596105-hosmer- lemeshow-test-for-large-data> [Accessed 10 June 2021].

6. Appendix

https://github.com/sahilsingh1997/Multinomial-Logistic-Regression-