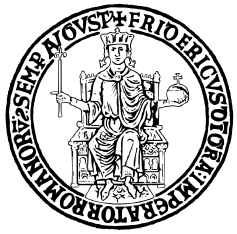
Univeristy of Naples Federico II



Political Sciences Department

Master's Degree Program in
Statistical Sciences for Decision Making

Research Project for Internship

Financial Crime in Romania: An Analysis of Social Perception

Supervisor:
Professor Maria Iannario

Student:
Antonio Cola
Student ID: M10/399

Academic Year 2022/2023

Contents

1 Introduction

In the current socio-economic landscape of Romania, understanding and analyzing phenomena related to financial crime play a fundamental role in safeguarding the integrity of the financial system and formulating targeted and effective public policies. The present research project aims to undertake an in-depth analysis of the underlying dynamics of financial crime perception within the Romanian community, using as its foundation the text "*Financial Crime in Romania: A Community Pulse Survey.*" Through the adoption of a rigorous methodological approach, the project seeks to reanalyze the data derived from the survey conducted between May 27 and June 6, 2022, on a sample of 1856 participants. In particular, attention will be focused on examining the variables of interest expressed in the original work, which include the degree of tax compliance, tax morality, perception of the level of institutional corruption, competence in detecting money laundering risk, and attitudes toward sharing banking information.

The first phase of the project entails a detailed analysis of each variable, aiming to highlight their characteristics, interrelationships, and potential emerging trends within the Romanian context. Subsequently, advanced statistical analysis techniques will be applied to construct predictive models that can provide a deeper understanding of the examined dynamics.

The ultimate goal of the project is to provide a comprehensive and in-depth analytical framework of the dynamics related to financial crime perception in Romania. The results obtained will not only contribute to academic knowledge in this field but may also have significant implications for the formulation of public policies aimed at countering and preventing financial crime in the Romanian context.

2 Exploratory Analysis

2.1 Description of the Dataset

The dataset consists of 1856 observations and includes 10 variables, each providing insight into various aspects related to taxation, financial behavior, and socio-demographic characteristics of the participants:

- **taxpay**: it refers to the degree of tax compliance;
- **receipt**: it refers to tax morality;
- **corruption**: it refers to the perception of the level of institutional corruption;
- **antimoneylaundering**: it refers to the competence in detecting the risk of money laundering;
- **infopay**: it refers to the attitude towards sharing banking information;
- **age**: it refers to the age of the interviewee;
- **gender**: it refers to the gender of the interviewee;
- **region**: it refers to the region of residence;
- **work**: it refers to the potential type of occupation;
- **education**: it refers to the level of education.

In summary, this dataset provides a comprehensive view of respondents' tax compliance, financial behavior, perceptions, and socio-demographic characteristics. The variations in responses across the variables present opportunities for further exploration and insights into the factors influencing these behaviors and attitudes.

2.2 Variables Distributions

Understanding the distribution of variables in a dataset is important because it enables us to grasp the data's structure and assess the presence of anomalies or patterns. This, in turn, can aid in making decisions regarding the analysis techniques to employ and formulating reliable conclusions about the dataset.
The first distributions to be examined are those concerning the *dependent variables*.

The variable **taxpay** refers to the degree of tax compliance. This variable was assessed through the following question:
Regarding paying your tax obligations, which one of the following behaviours is common for you and the people around you:

Original Categories	Encoding
Long before the deadline, regardless of the discounts	1
Long before the deadline to benefit from discounts	2
Very close to deadline	3
Over the deadline	4

After modifying the modes of the variable for easier interpretation and representation, its distribution was examined:

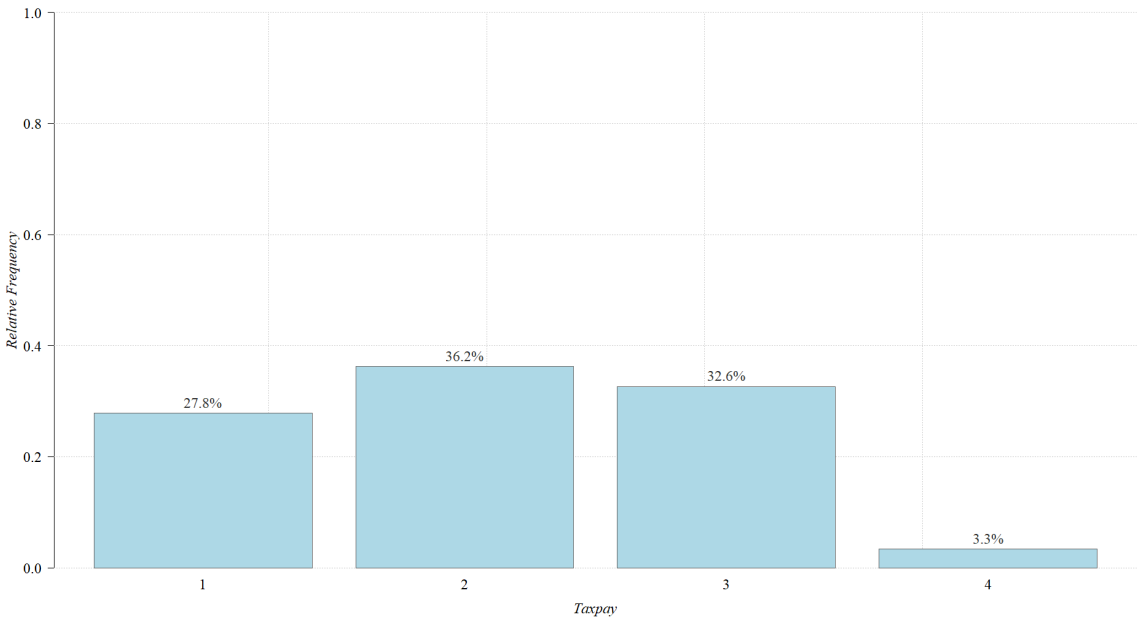


Figure 1: Taxpay Distribution

The bar chart presented highlights that 27.8% of the respondents declare to pay taxes well in advance regardless of discounts, 36.2% declare to pay taxes well before the deadline to benefit from discounts, 32.6% declare to pay taxes close to the deadline, and 3.3% declare to pay taxes after the deadline.
A first observation that can be made is that the relative majority of participants, amounting to 36.2%, prefer to pay taxes before the deadline to obtain economic benefits in the form of discounts. This could suggest that financial incentives are effective in motivating people to make early payments.
On the other hand, the 27.8% who pay taxes well in advance without considering the discounts appear to adopt a more responsible and diligent approach to their tax obligations. This might reflect a good financial culture or a specific attention to personal financial planning.
The 32.6% who declare to pay taxes close to the deadline might indicate a tendency towards procrastination or some negligence in managing tax deadlines. This behavior could lead to penalties or additional stress in the final stages.
Finally, the 3.3% who pay taxes after the deadline represents a relatively low percentage, but it could still indicate a possible lack of awareness of the negative consequences associated with payment delays.
In general, the chart reflects various approaches to tax payment among participants, each of which could be influenced by different factors that will be analyzed later on.

The variable **receipt** refers to tax morality. This variable was assessed through the following question:
When you buy goods and services from shops, restaurants, hotels, salons, etc, please choose one or more of the following that best applies to you:

Original Categories	Encoding
Always receiving the receipt	1
Usually received the receipt	2
Always asking for receipt	3
Receiving receipt but let it there	4
Not bothered by not receiving the receipt	5

After modifying the modes of the variable for easier interpretation and representation, its distribution was examined:

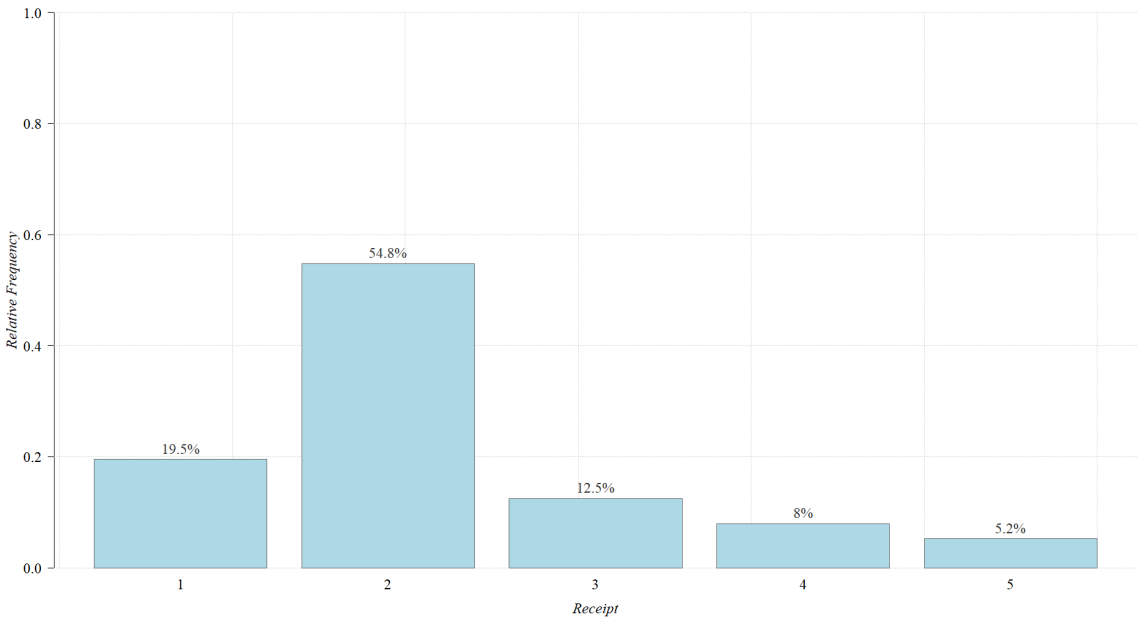


Figure 2: Receipt Distribution

The bar chart presented highlights that 19.5% of the respondents declare to always receive the receipt, 54.8% declare to usually receive the receipt, 12.5% declare to always ask for the receipt, 8% declare to receive the receipt and leave it there, and 5.2% declare to not be bothered by not receiving the receipt.
A primary observation is that the majority of respondents, amounting to 54.8%, tend to usually receive the receipt. This could indicate that many individuals have the habit of requesting or receiving a receipt, although not necessarily consistently.
Additionally, it's interesting to note that 12.5% of the interviewed individuals claim to always ask for the receipt. This can be interpreted as a good practice for financial control and expense documentation.
On the other hand, the 8% who receive the receipt but leave it there might suggest a certain negligence in storing important expense-related documents. This behavior could lead to challenges in tracking expenses in the long term.
The relatively low percentage of 5.2% who are not bothered by not receiving the receipt could indicate a lack of awareness regarding the necessity of retaining documentary evidence for potential transactions or returns.
In general, the chart reflects diverse habits regarding receipt reception and retention, showcasing a variety of attitudes towards expense management and financial responsibility.

The variable **corruption** refers to the perception of the level of institutional corruption. This variable was assessed through the following question:
How do you perceive the level of corruption in Romanian public institutions, on a scale from 1 (very low) to 5 (very high)? Please choose one of the following:

Original Categories	Encoding
Very low level of corruption	1
Low level of corruption	2
Medium level of corruption	3
High level of corruption	4
Very high level of corruption	5

After modifying the modes of the variable for easier interpretation and representation, its distribution was examined:

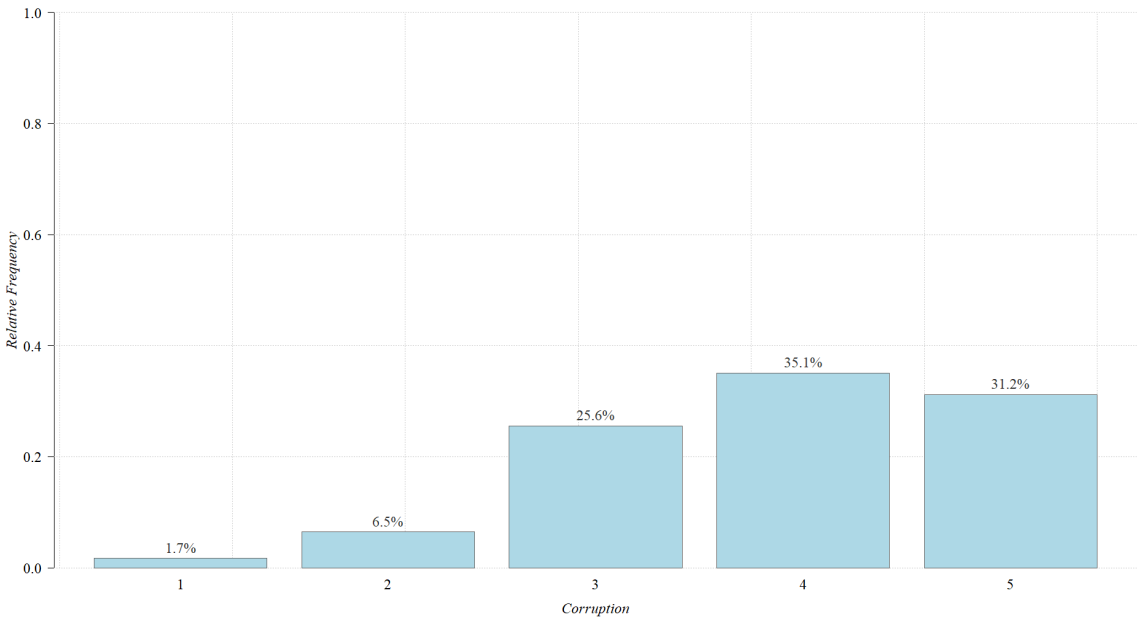


Figure 3: Corruption Distribution

The bar chart presented highlights that 1.7% of the respondents declare to have a very low perception of institutional corruption, 6.5% declare to have a low perception of institutional corruption, 25.6% declare to have a medium perception of institutional corruption, 35.1% declare to have a high perception of institutional corruption, and 31.2% declare to have a very high perception of institutional corruption.
A primary observation is that a significant percentage, amounting to 66.3% (sum of the percentages of high and very high perception), of the respondents perceive a high or very high level of institutional corruption. This suggests that the majority of participants holds a critical view regarding the integrity of institutions, indicating widespread distrust in their functioning.
Furthermore, it's interesting to note that 25.6% of the participants have a medium perception of institutional corruption. This might indicate that some individuals consider corruption to be an existing issue but do not perceive it as particularly widespread or pervasive.
On the other hand, the 1.7% who declare to have a very low perception could represent a group of individuals who tend not to consider corruption a relevant or prevalent issue within institutions.
The distribution of responses reflects a range of viewpoints concerning institutional corruption, underscoring the need for further investigation and analysis to understand the reasons behind these diverse perceptions.

The variable **antimoneylaundering** refers to the competence in detecting the risk of money laundering. This variable was assessed through the following question:
Regarding the risk of money laundering, do you think that people have suitable knowledge to be able to recognize a suspicious transaction in a business? Please choose one of the following:

Original Categories	Encoding
Very low skills	5
Low skills	4
Medium skills	3
High skills	2
Very high skills	1

After modifying the modes of the variable for easier interpretation and representation, its distribution was examined:

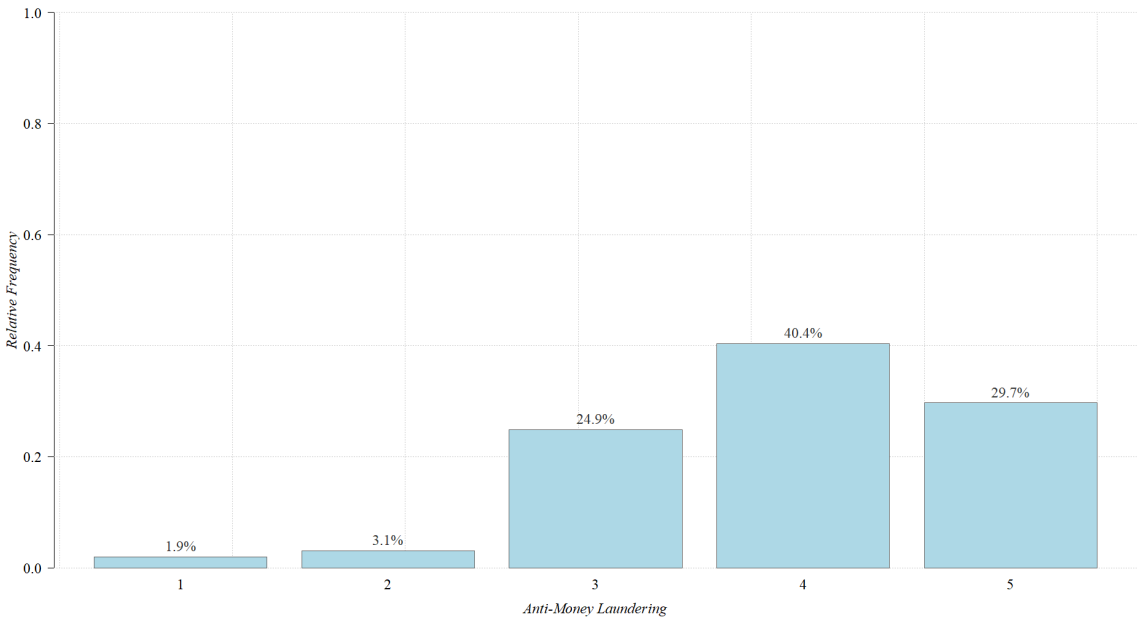


Figure 4: Anti-Money Laundering Distribution

The bar chart presented highlights that 1.9% of the respondents declare to have a very high level of anti-money laundering skills, 3.1% declare to have a high level of anti-money laundering skills, 24.9% declare to have a medium level of anti-money laundering skills, 40.4% declare to have a low level of anti-money laundering skills, and 29.7% declare to have a very low level of anti-money laundering skills.
A primary observation is that a significant percentage, amounting to 70.1% (sum of the percentages of very low and low levels), of the participants declare to have relatively low anti-money laundering skills. This could suggest a potential lack of knowledge and awareness regarding anti-money laundering practices.
Furthermore, the 1.9% who declare to have a very high level of skills could represent a group of individuals with specific experience or training in the field of anti-money laundering. On the other hand, the 3.1% with high skills might reflect those who possess above-average knowledge but not extremely advanced.
The 24.9% with medium skills could represent individuals with basic knowledge who could still benefit from training or greater familiarity with anti-money laundering practices.
The distribution of responses highlights a range of competency levels in anti-money laundering, with a concerning percentage of participants declaring relatively low skills. This could indicate the need for training and awareness efforts to improve understanding and implementation of anti-money laundering measures.

The variable **infopay** refers to the level of education. This variable was assessed through the following question:
When you or the people around you ask for bank transactions (making a bank deposit, withdrawing cash from the account etc.), and the bank officer asks you to complete some details regarding the source of the money, or where the amount goes, then you or the people around you generally have different reactions. To help us analyse your responses, please indicate which of the following categories represents you best:

Original Categories	Encoding
Offer any asked details	3
Bothered by asked details but finally provided	2
Refuse to offer any asked details	1

After modifying the modes of the variable for easier interpretation and representation, its distribution was examined:

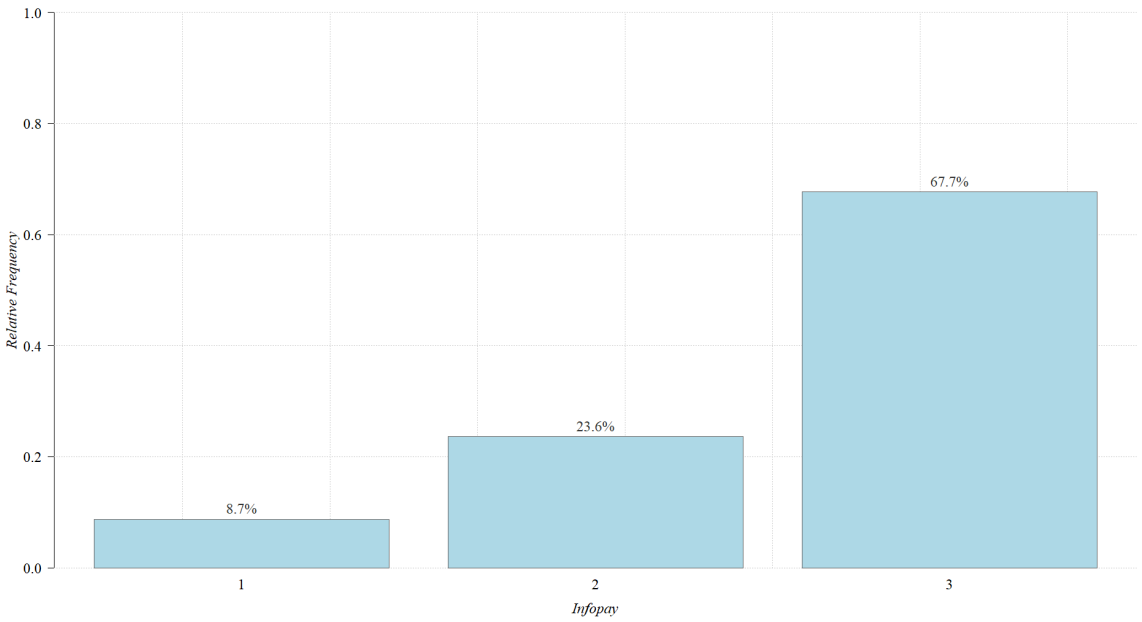


Figure 5: Infopay Distribution

The bar chart presented highlights that 8.7% of the respondents declare to refuse to provide the requested information, 23.6% declare to provide the requested information despite being bothered by the question, and 67.7% declare to provide all the requested information.
A primary observation is that the overwhelming majority of participants, amounting to 67.7%, are willing to provide all the requested information. This might reflect a inclination towards collaboration and data sharing.
On the other hand, the 23.6% who declare to provide the information despite being bothered by the question suggests that some individuals might feel uncomfortable or embarrassed about the request but still choose to respond. This could indicate a certain reluctance mixed with a willingness to cooperate.
The 8.7% who refuse to provide the requested information represent a relatively low percentage, but it could reflect a concern for privacy or a desire to keep certain information confidential.
In general, the chart highlights various responses to the act of providing requested information, showing a mix of reactions that could be influenced by factors such as the sensitivity of the questions, trust in the requesting entity, and one's propensity to share personal information.

After analyzing the distribution of the dependent variables, the distribution of the *socio-demographic variables* was examined.

The variable **age** refers to the age of the interviewee:

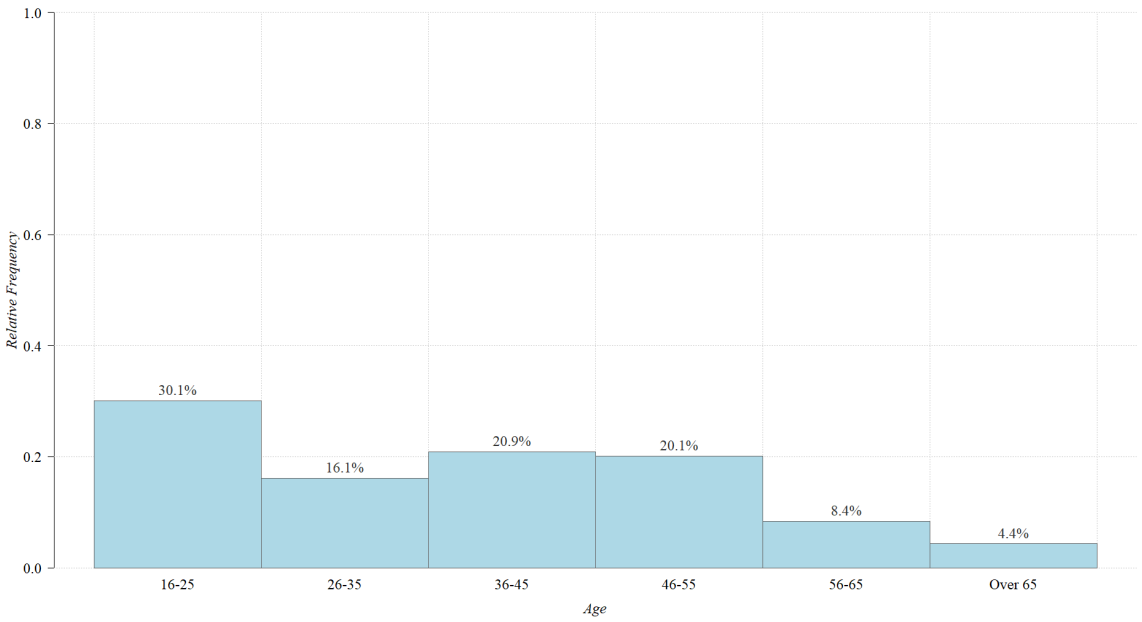


Figure 6: Age Distribution

The presented histogram highlights the distribution of participants' ages and provides interesting insights into the demographic composition of the sample:
Young Age Bracket (16-25 years): The youngest age bracket represents 30.1% of the participants. This might indicate substantial involvement of the youth, reflecting their interest and attention to relevant topics in today's world.
Young Adult Age Bracket (26-35 years): The age bracket between 26 and 35 years constitutes 16.1%. This group could include individuals in the early stages of their careers or making important decisions in both work and personal spheres.
Adult Age Bracket (36-45 years): The age bracket between 36 and 45 years makes up 20.9%. This range might encompass people in the middle of their careers, dealing with work and family responsibilities.
Middle-Aged Bracket (46-55 years): The age bracket between 46 and 55 years accounts for 20.1%. This group could involve individuals approaching the later stages of their careers, planning for retirement, and transitioning to post-work life.
Elderly Age Bracket (56-65 years): The age bracket between 56 and 65 years is represented by 8.4%. This might consist of people nearing retirement or already enjoying post-work life.
Elderly Age Bracket (Above 65 years): The percentage of 4.4% represents those above the age of 65. This group could include retirees and seniors who have gained extensive life and societal experience.
In summary, the chart reflects a balanced demographic distribution with significant participation from both younger and adult age groups. This could suggest a cross-sectional interest in the surveyed topic among individuals of various ages and life stages.
Usually, the age variable is collected and used as a continuous variable. In this case, the variable *age* is grouped into age brackets. For this reason, and for easier subsequent use, it has been decided to dichotomize it as follows:

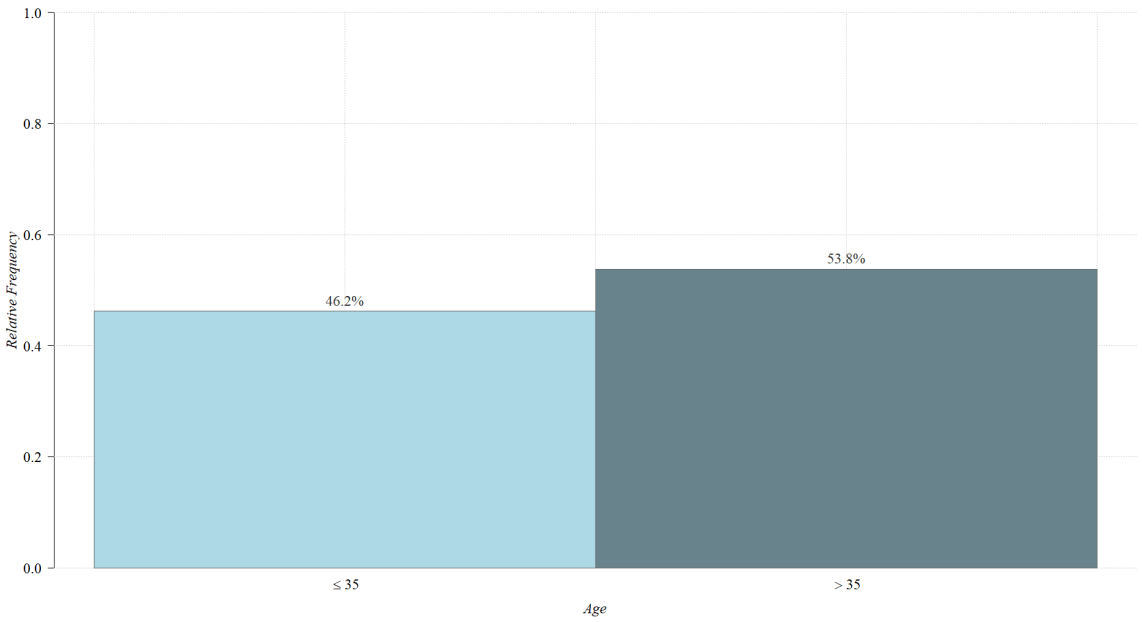


Figure 7: Age Distribution

The dichotomization of age into age groups of 35 or younger and over 35 brings to light an interesting division among the participants based on age. The majority of participants, at 53.8%, fall into the over 35 age group, while 46.2% are aged 35 or younger.
This division can hold significant implications in terms of data analysis and interpretation. For instance, it could be intriguing to assess whether there are differences in responses to questions between these two age categories. Younger participants might have distinct perspectives compared to older ones, thereby influencing the overall data trends. Furthermore, this breakdown can be useful in identifying any patterns or trends that might emerge within specific age groups.
Dichotomizing age in this manner simplifies data analysis and interpretation, focusing on two distinct groups that could provide valuable insights into how different age brackets react to the topics covered in the survey.

The variable **gender** refers to the gender of the interviewee:

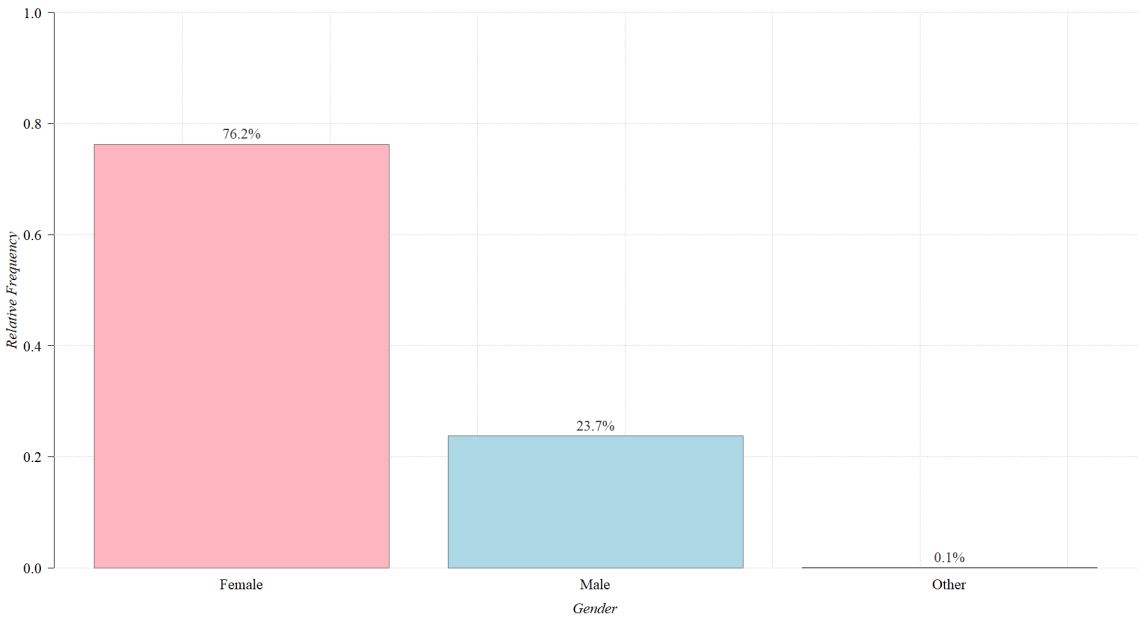


Figure 8: Gender Distribution

The presented bar chart highlights that 76.2% of the respondents are female, 23.7% are male, and 0.1% belong to another gender. This last category, which corresponds to a single observation, has been removed, reducing the number of observations from 1856 to 1855. Below is the updated distribution:

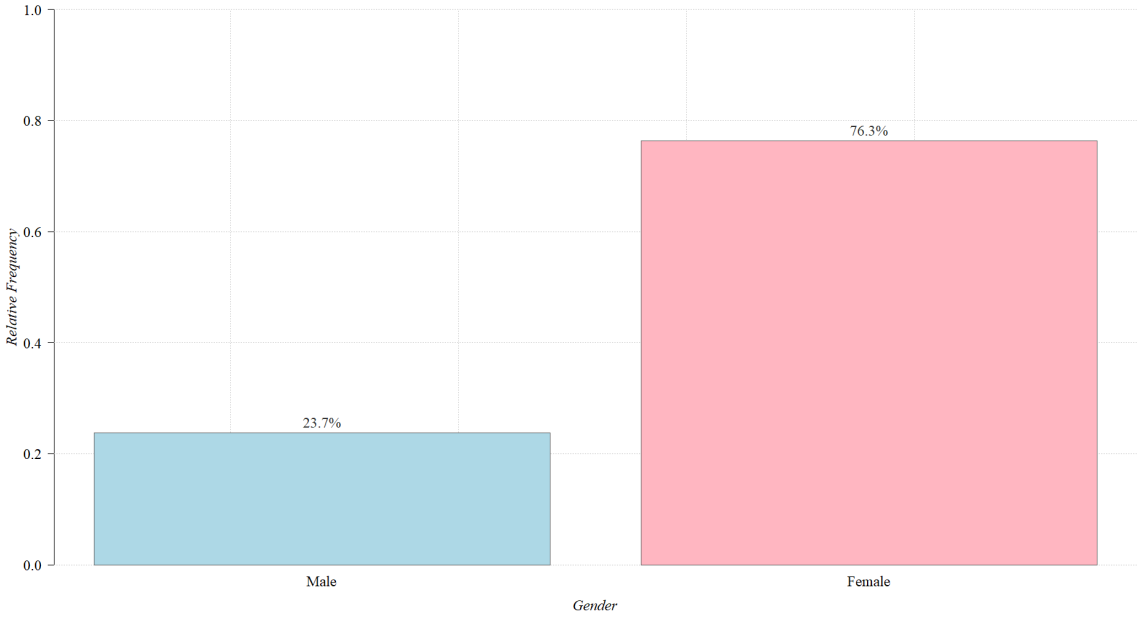


Figure 9: Gender Distribution

The presented bar chart highlights that 23.7% of the respondents belong to the male gender, and 76.3% belong to the female gender.
The distribution of genders reflects a noticeable asymmetry towards the female gender among the participants. This could be attributed to various factors, including the sample composition, the survey topic, and the methodology employed for participant recruitment. Understanding the differing levels of engagement between genders can contribute to a more in-depth analysis of the survey results.

The variable **region** refers to the region of residence:

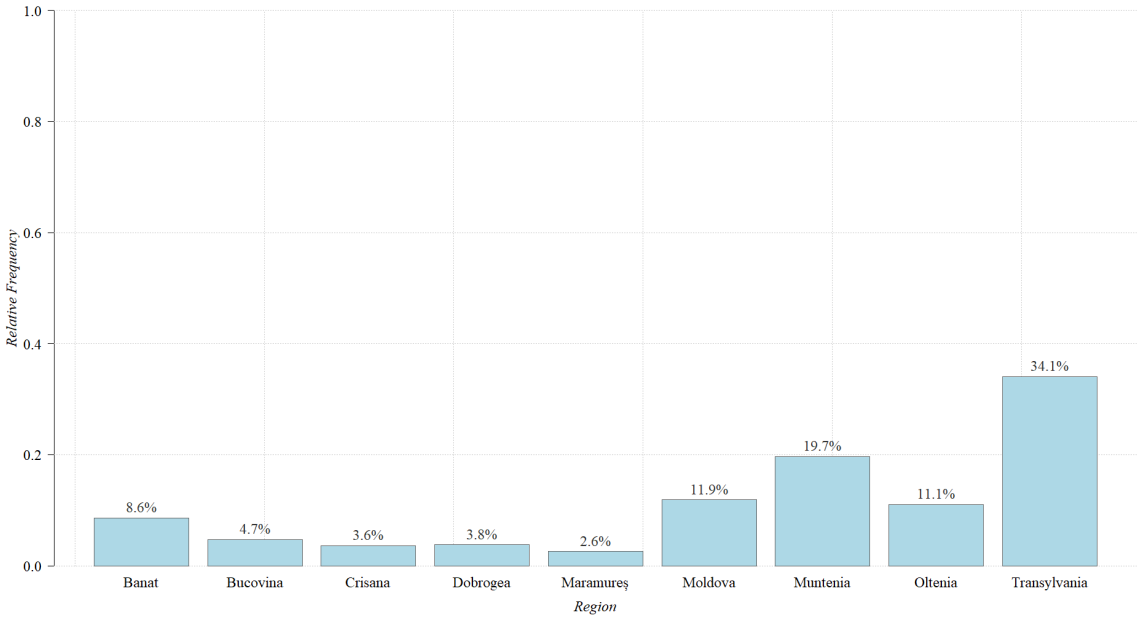


Figure 10: Region Distribution

The presented bar chart provides an interesting overview of the geographical distribution of participants across different regions. The chart reflects a diverse distribution of participants among the various regions. While Transylvania represents the region with the highest percentage of residents (34.1%), followed by Muntenia (19.7%), other regions such as Banat (8.6%), Moldova (11.9%), and Oltenia (11.1%) also exhibit significant presence. Regions like Bucovina (4.7%), Crisana (3.6%), Dobrogea (3.8%), and Maramures (2.6%) show lower percentages, but they remain relevant considering the sample size. Transylvania and Muntenia, which include the cities of Cluj-Napoca and Bucharest respectively, emerge as the regions with the highest percentages of residents. This could be attributed to the presence of important cultural, economic, and institutional centers in these areas.
Subsequently, for easier subsequent use, it was decided to dichotomize the variable based on the Human Development Index.

The **Human Development Index (HDI)** is a composite measure developed by the United Nations to assess human well-being and the level of development in a country or region. It is a multidimensional indicator that takes into account three main dimensions of human well-being: *life expectancy, education, and income.*
Life Expectancy at Birth: This represents the average number of years a person can expect to live at birth. It reflects the quality and access to healthcare, nutrition, and general health conditions of a population.
Years of Education: This measures the average years of education received by individuals of a certain age. It includes both primary and secondary education and reflects access to education and attention to training.
Gross National Income (GNI) Per Capita: This component measures the average income earned by an individual in a year. Per capita income is adjusted based on purchasing power, taking into account differences in cost of living between countries.
The HDI combines these three components into a single index that ranges from 0 to 1, where 0 represents lower human development and 1 represents the highest. The HDI is a fundamental tool for understanding the well-being of populations and assessing progress towards sustainable development goals and improving quality of life.

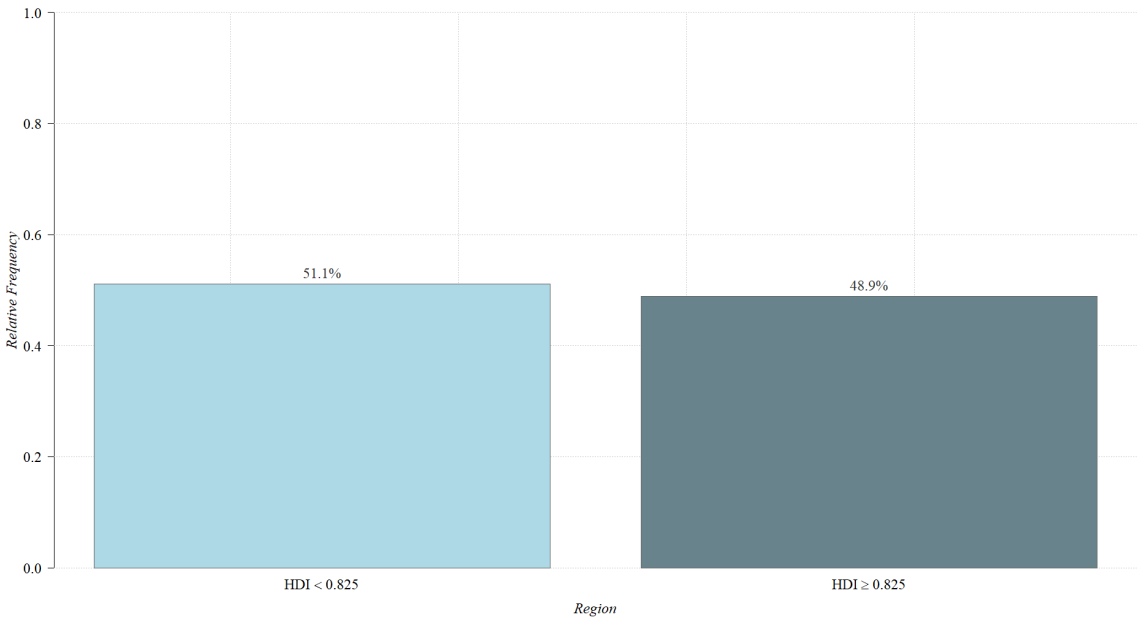


Figure 11: Region Distribution

The presented bar chart provides relevant information about the distribution of participants based on the Human Development Index (HDI) of their residential locations in Romania.
The division of participants into the two HDI categories highlights a disparity in development conditions among different locations. While 51.1% reside in places with an HDI lower than 0.825 (Bucovina, Dobrogea, Moldova, Muntenia, Oltenia), the remaining 48.9% live in places with an HDI equal to or greater than this value (Banat, Crisana, Maramures, Transylvania). This suggests a variety of levels of human development within the country.
The difference in HDI may be reflected in the behaviors, attitudes, and perceptions of participants. For instance, those living in areas with lower HDI might face different socioeconomic challenges compared to those residing in more developed areas.
HDI is often associated with access to services and resources such as education, healthcare, and employment. Participants from higher HDI locations might have better access to these resources, thereby influencing their financial behavior and perceptions about taxation and corruption.

The variable **work** refers to the potential type of occupation:

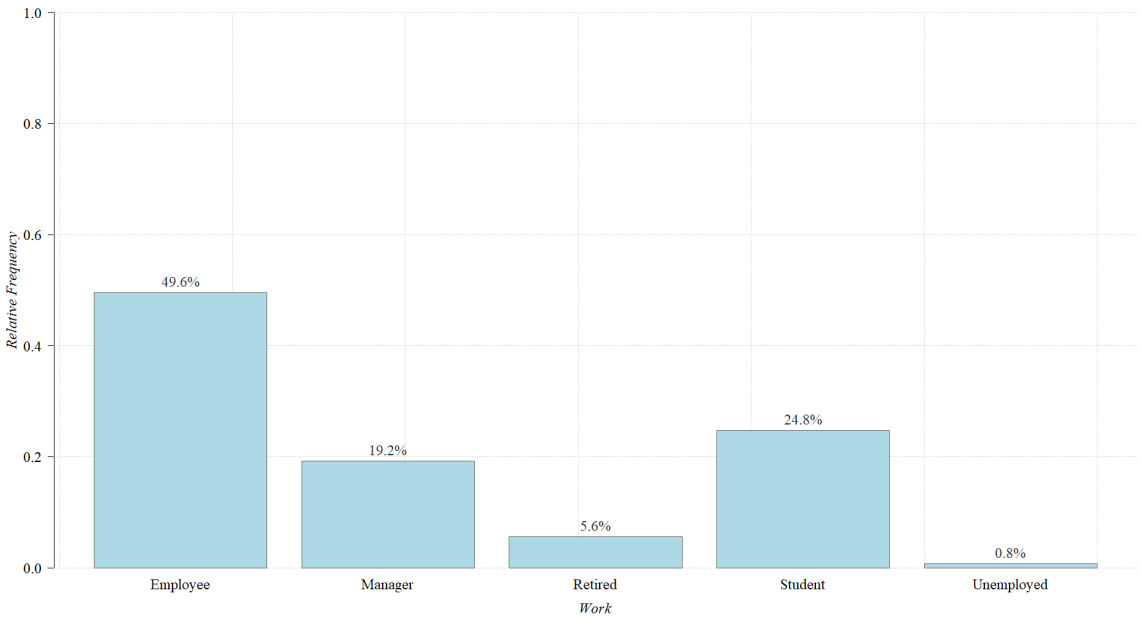


Figure 12: Work Distribution

The presented bar chart illustrates a diverse distribution of occupations among the interview participants.
The majority of participants, at 49.6%, are employees. This could indicate a significant presence of individuals employed across various industries and sectors, reflecting the occupational structure of the sample.
The manager occupation accounts for 19.2% of the participants. This percentage suggests that a substantial portion of the sample holds leadership and management roles, which could impact their financial perceptions and economic decisions.
Retirees make up 5.6% of the participants, indicating a small yet significant portion of individuals in retirement age. This category might have different financial experiences and attitudes compared to those still in the workforce.
Students constitute 24.8% of the sample. This percentage reflects the presence of individuals in the process of completing their education, and it's important to consider how their financial situation often relies on family resources.
Finally, the 0.8% of unemployed participants represent a relatively small group. However, it's interesting to note how even unemployment can influence financial attitudes and perceptions among participants.
The diversity of occupations within the sample could lead to a range of financial perspectives, attitudes towards saving and investing, as well as perceptions regarding fiscal and financial matters in general.
Subsequently, for easier subsequent use, it was decided to dichotomize the variable as follows:

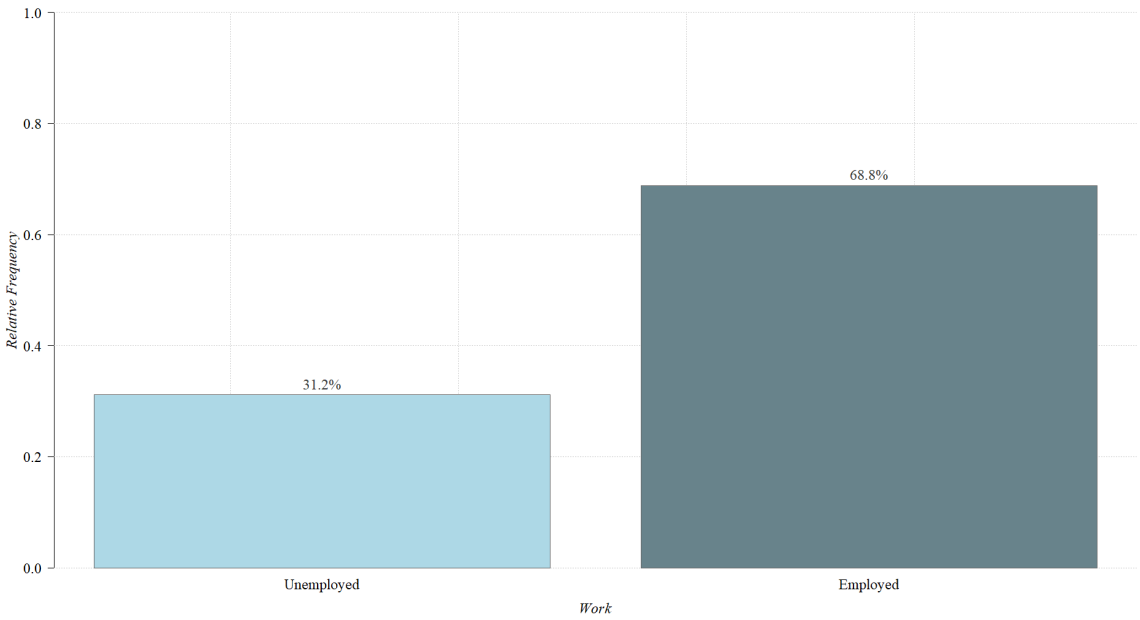


Figure 13: Work Distribution

The division between the unemployed (Retired, Student, Unemployed) and employed (Employee, Manager) provides an interesting view of the sample's composition in terms of employment status and occupation.
The relatively high percentage (31.2%) of respondents belonging to the category of unemployed suggests a diversification in terms of occupation or employment situation. This could reflect complex economic dynamics or a variety of age groups within the sample.
On the other hand, the majority (68.8%) of respondents are employed. This might indicate a strong presence of individuals actively engaged in the workforce, bringing with them experiences and perceptions tied to their employment situation.
The difference in financial experiences and perspectives between the unemployed and employed groups could be significant. Employed participants might have greater financial stability and a different outlook on financial and tax management, whereas those who are unemployed, retired, or students might face distinct challenges and perspectives.
Further analysis of participants' responses could provide insights into the underlying reasons for their employment status and how it influences their financial and tax-related perceptions.

The variable **education** refers to the level of education:

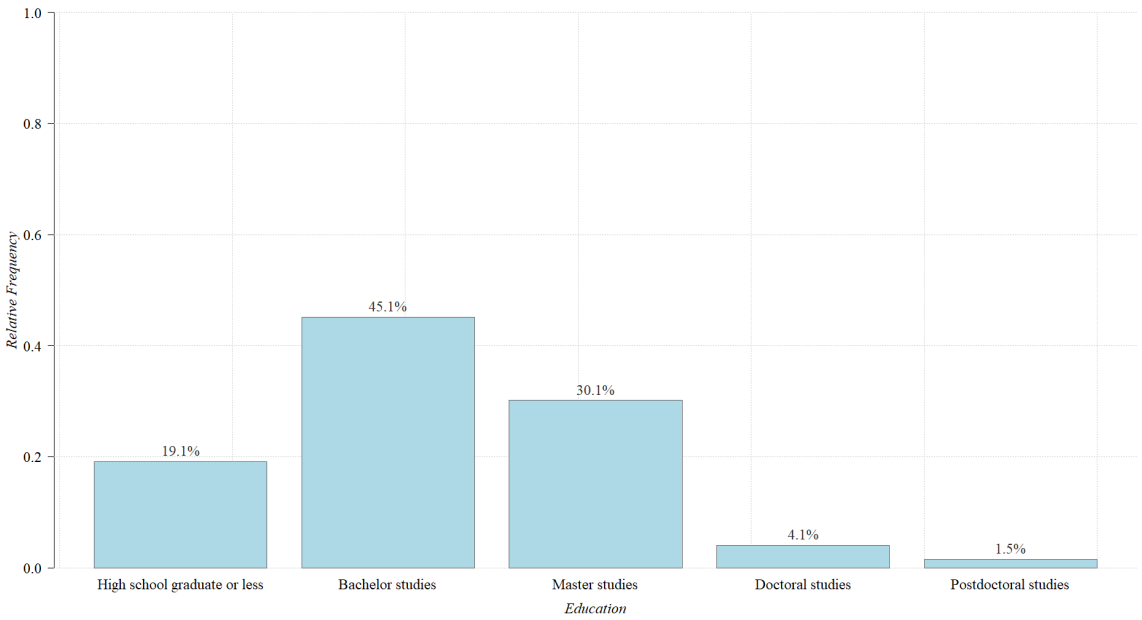


Figure 14: Education Distribution

The analysis of the distribution of education levels within the sample provides interesting insights into the participants' educational background.
The 19.1% of the respondents have at most a high school diploma, indicating a substantial portion of participants with a basic education level. This group might encompass individuals with diverse backgrounds and experiences, potentially contributing to a range of financial perspectives. Their relatively lower educational attainment might influence their financial decision-making, emphasizing the importance of accessible financial literacy initiatives.
A notable finding is that 45.1% of the participants hold a bachelor's degree, indicating a significant proportion of individuals with basic university education. This might reflect increased accessibility to higher education or a general trend towards pursuing university education among the population.
The presence of 30.1% with a master's degree suggests that a substantial number of participants pursued further education at an advanced level. This could indicate an interest in deepening their education or a desire to acquire specialized skills.
The 4.1% with a doctoral degree represents a minority that has achieved the highest academic level of education. This could comprise individuals highly specialized and committed to research or advanced professional fields.
The inclusion of 1.5% who have continued their studies beyond a doctoral degree underscores an even greater commitment to academic research or the pursuit of further specializations.
The diversity in education levels within the sample can lead to a variety of perspectives and knowledge among participants. Differences in education might influence financial attitudes, understanding of tax-related matters, and participation in complex financial decisions.
Subsequently, for easier subsequent use, it was decided to dichotomize the variable as follows:

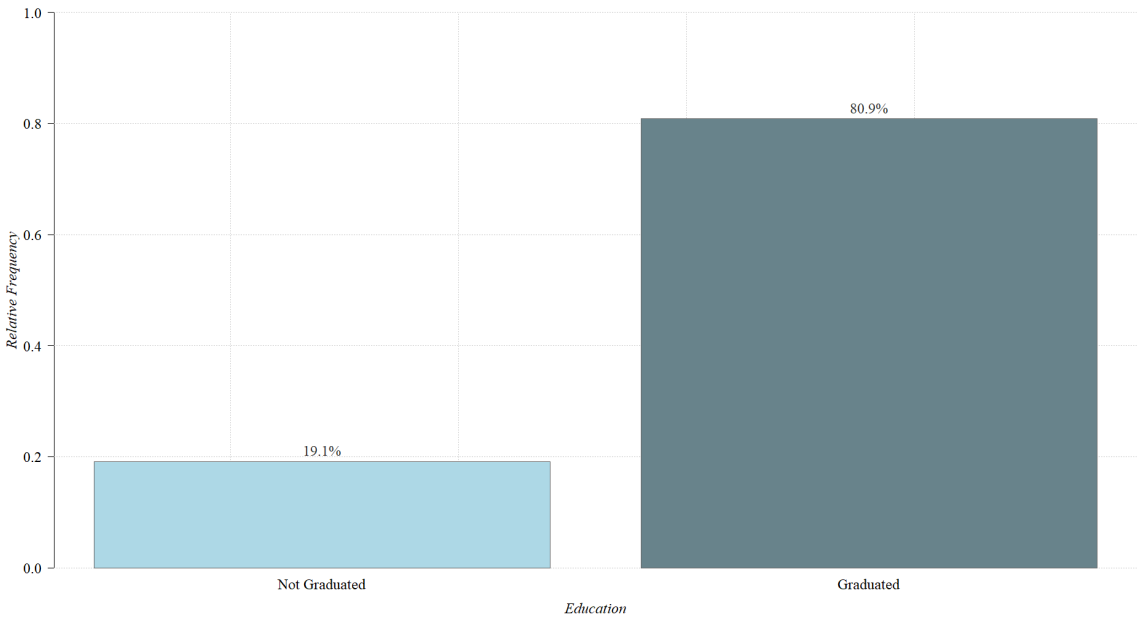


Figure 15: Education Distribution

The highlighted bar chart shows that a noteworthy percentage of 19.1% of the participants who have not obtained a degree indicates a significant portion of the sample with a basic educational level. This group may include individuals with diverse backgrounds and experiences, potentially contributing to a variety of financial perspectives. Their relatively lower education could influence their financial decisions, underscoring the importance of accessible financial literacy initiatives. Understanding the financial behaviors and attitudes of this subgroup could provide valuable insights into the impact of education on financial choices and the potential need for tailored educational programs. On the other hand, the substantial majority (80.9%) of participants who have obtained at least a bachelor's degree might indicate a solid educational foundation, potentially influencing their financial habits and investment decisions differently from those with a lower educational level.

2.3 Joint Distribution

Analyzing the joint distribution of variables in data analysis is crucial for understanding the relationships and interactions among different variables. This process provides a more comprehensive and detailed picture of the data, revealing hidden connections and trends that may not emerge from the analysis of individual variables. Examining the joint distribution can uncover patterns, dependencies, and associations between variables, allowing for the discovery of how changes in one variable can influence another. This analysis is crucial for building accurate predictive models and drawing robust conclusions from the data. Another advantage of joint analysis is the ability to identify subsets of data that share common characteristics or exhibit significant differences. This can lead to important discoveries and guide informed decisions.

Below are the joint distributions with respect to the variable **age**:

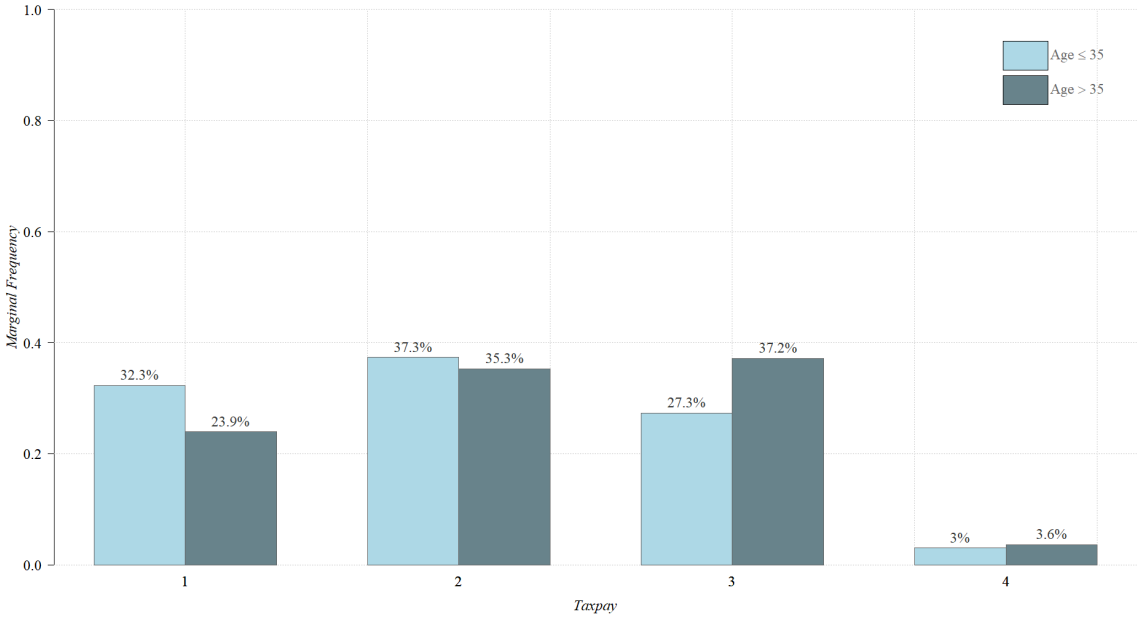


Figure 16: Taxpay Distribution by Age

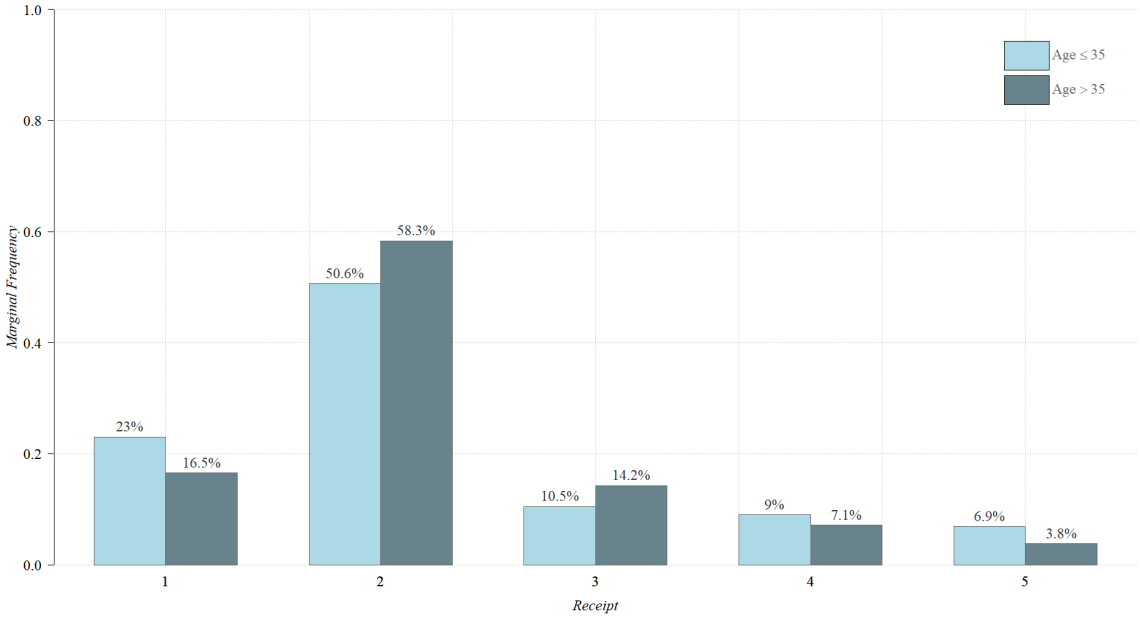


Figure 17: Receipt Distribution by Age

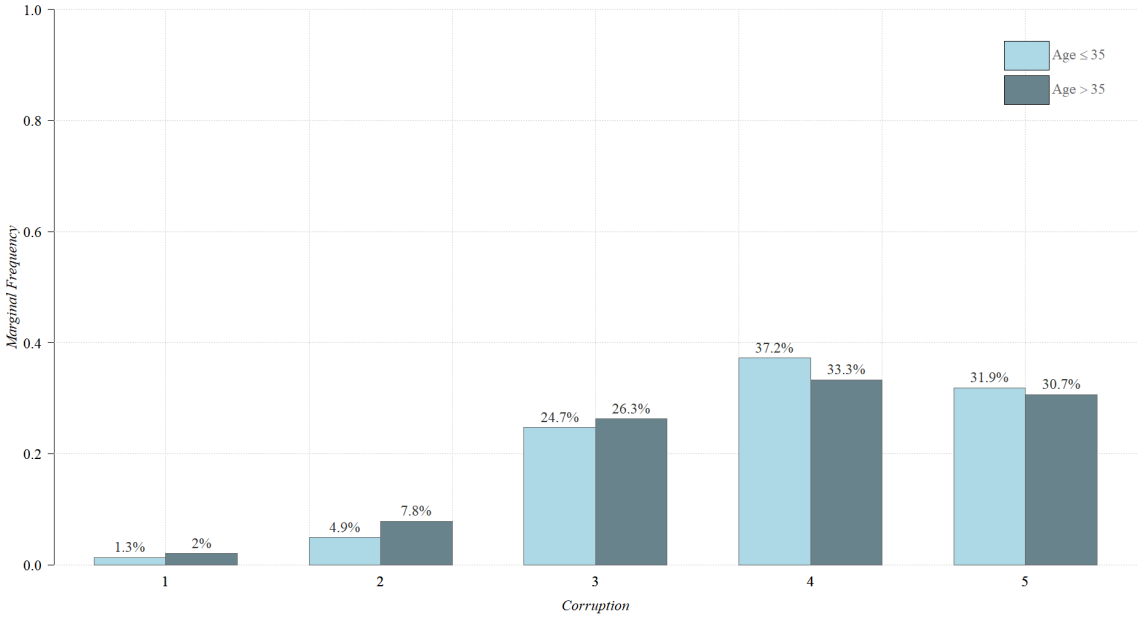


Figure 18: Corruption Distribution by Age

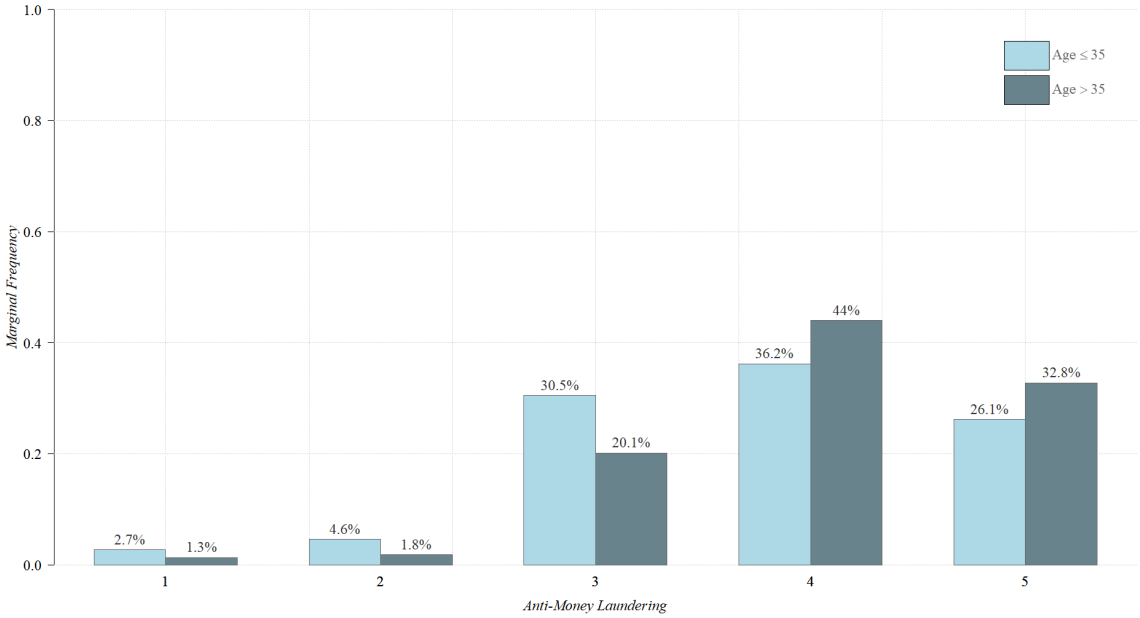


Figure 19: Anti-Money Laundering Distribution by Age

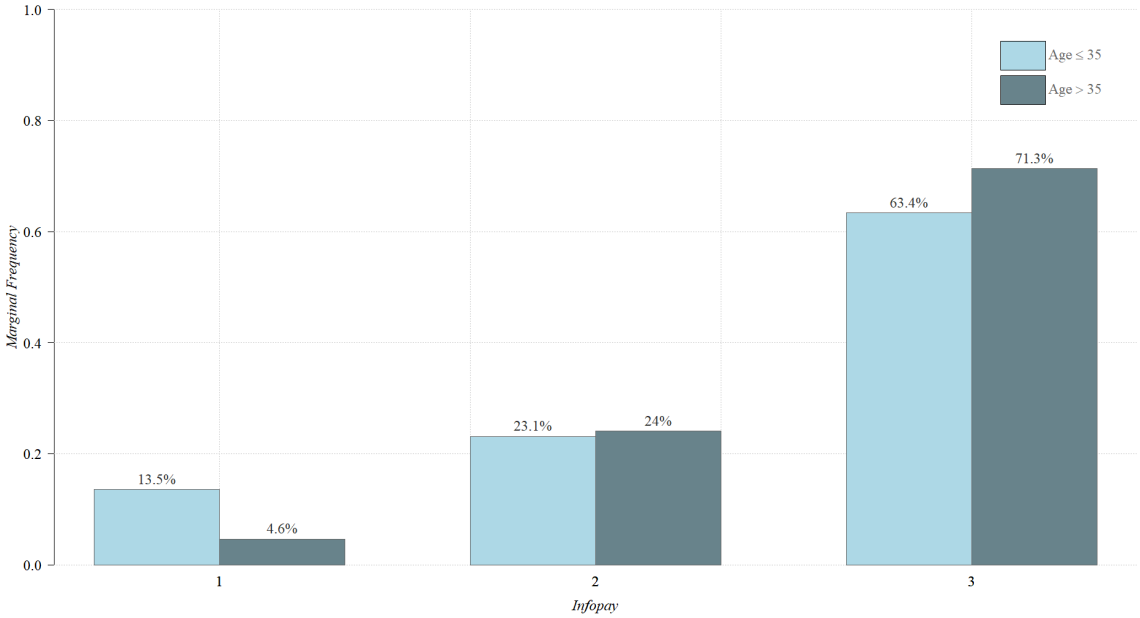


Figure 20: Infopay Distribution by Age

From the previous bar charts, a visual difference in the frequencies of the *age* variable modes can be observed, which has been tested using proportion tests (Table 2). The test results on each of the aforementioned distributions confirm that the relative frequencies are not the same in the reference populations, with high significance indicated by a $p\text{-}value < 0.01$ value in all cases. Therefore, it can be asserted that individuals under 35, compared to those over 35, are more likely to pay taxes in advance, are less concerned about receipts, have a higher perception of corruption, possess a higher level of anti-money laundering skills, and are more inclined to provide requested information. This outcome suggests that the *age* variable could prove to be significant in a subsequent model.

Below are the joint distributions with respect to the variable **gender**:

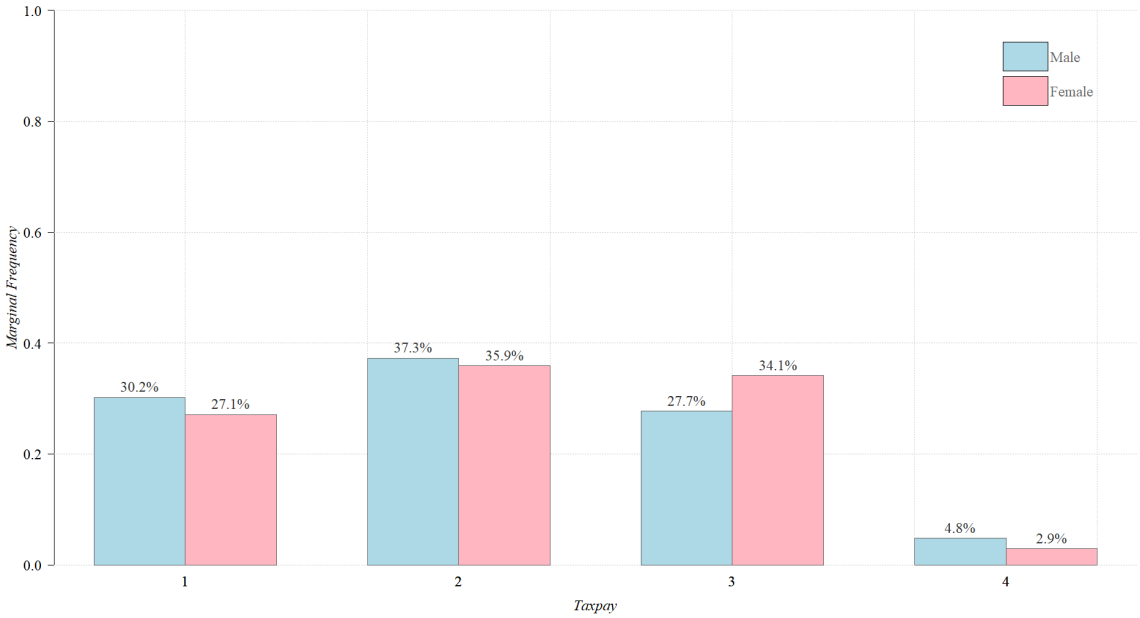


Figure 21: Taxpay Distribution by Gender

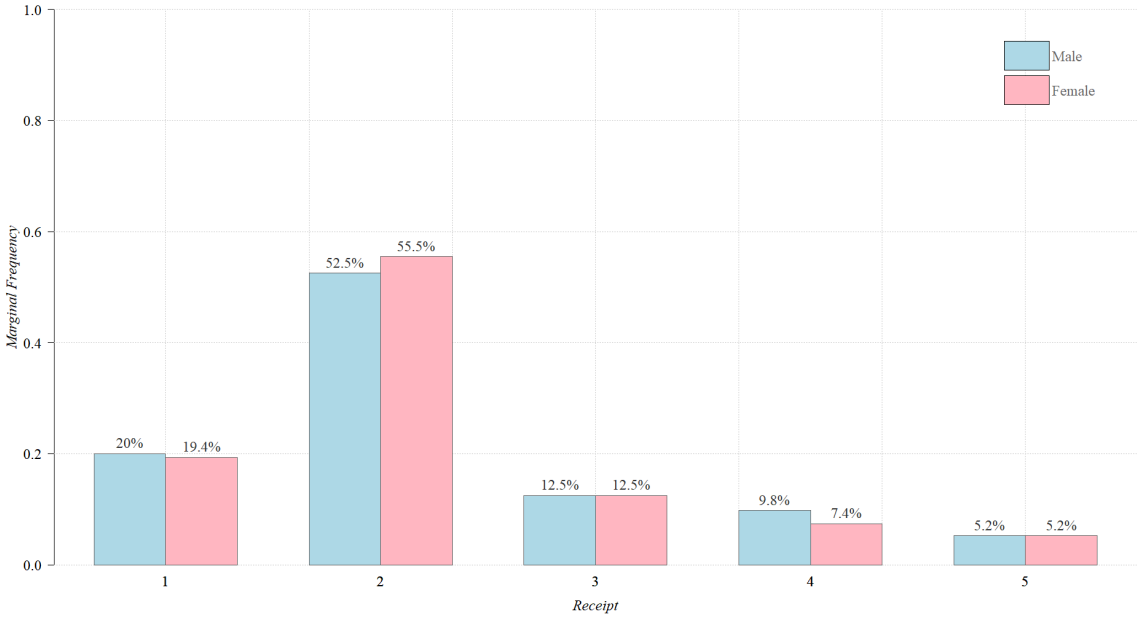


Figure 22: Receipt Distribution by Gender

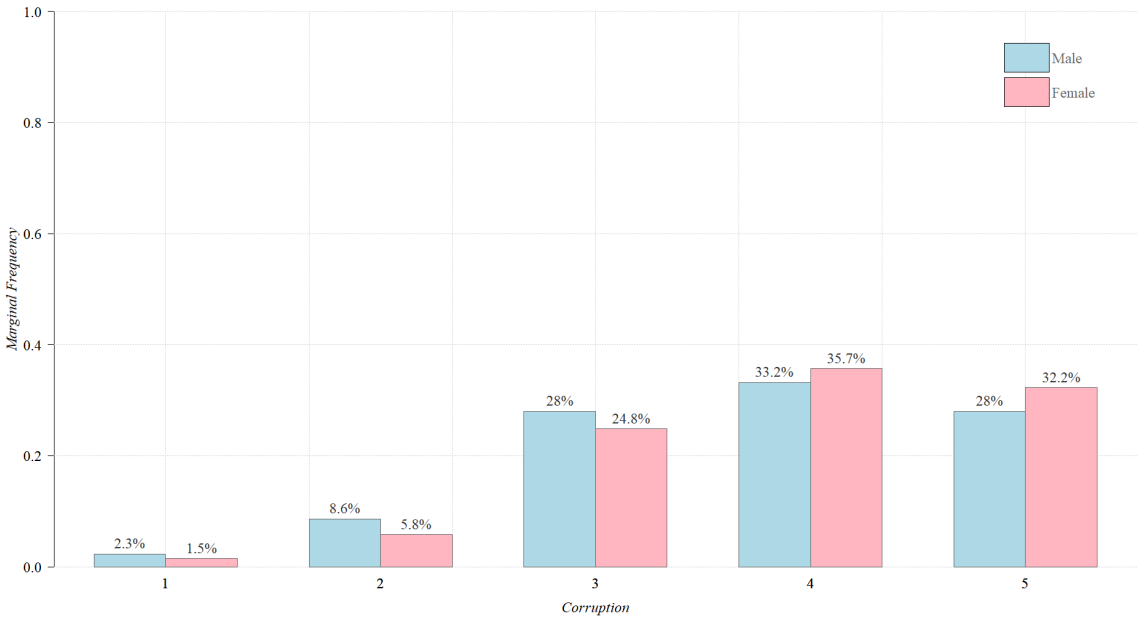


Figure 23: Corruption Distribution by Gender

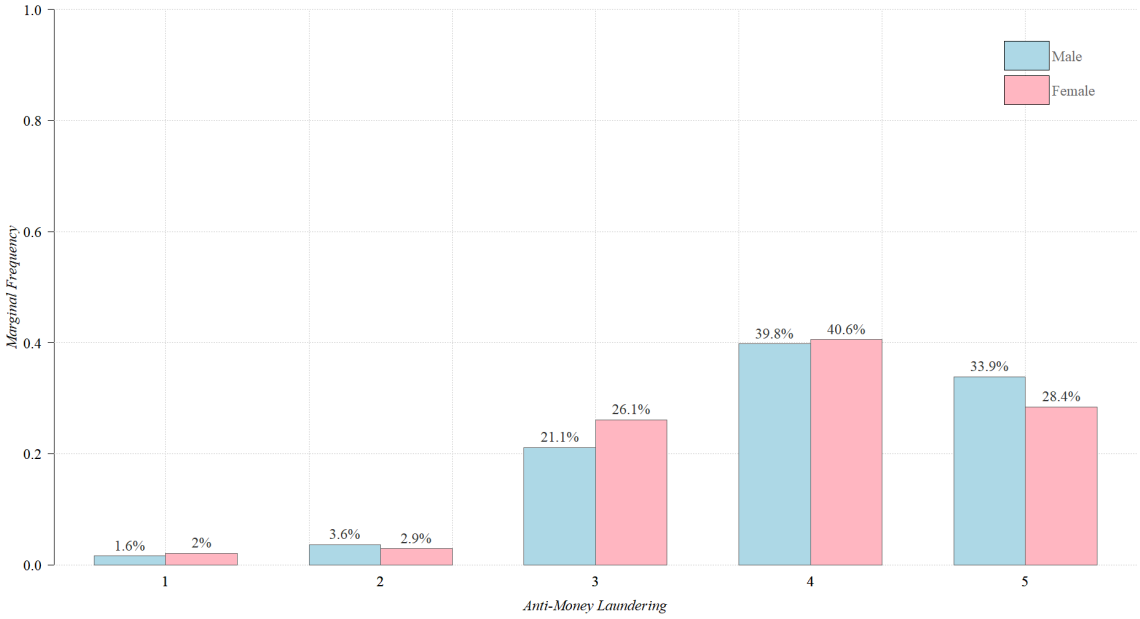


Figure 24: Anti-Money Laundering Distribution by Gender

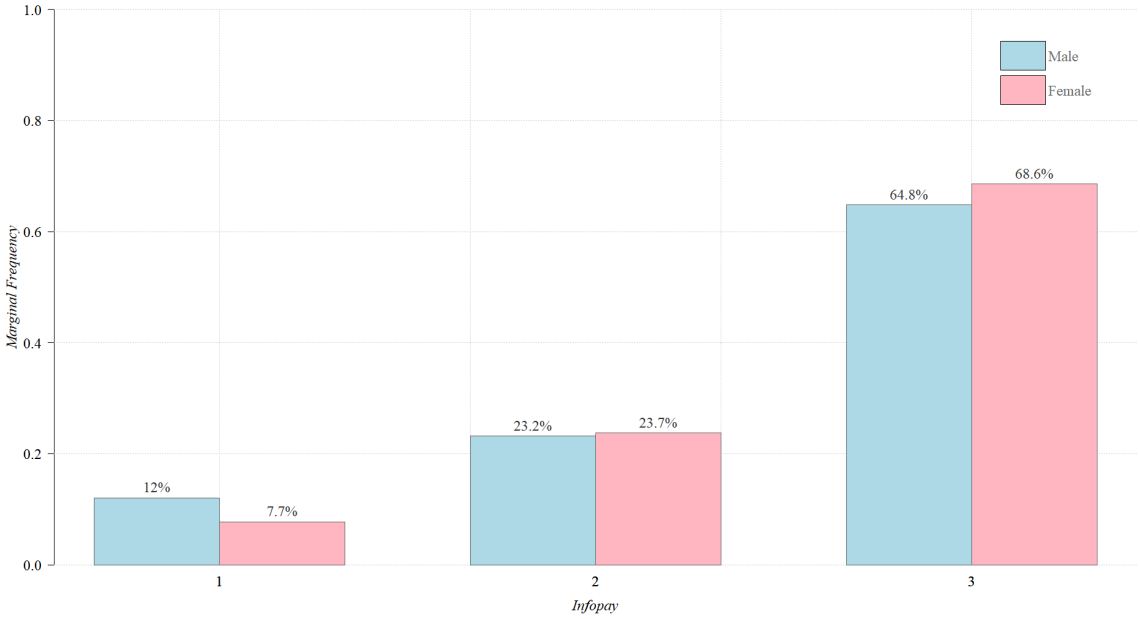


Figure 25: Infopay Distribution by Gender

From some of the previous bar charts (Figure 21 and 25), a visual difference in the frequencies of the *gender* variable modes can be observed, which has been tested using proportion tests (Table 2). The test results on these distributions confirm that the relative frequencies are not the same in the reference populations, with high significance indicated by a $p\text{-}value<0.01$ value in all cases. Therefore, it can be asserted that males, compared to females, are more likely to pay taxes in advance and provide requested information. This outcome suggests that the *gender* variable could prove to be significant in a subsequent model.

Below are the joint distributions with respect to the variable **region**:

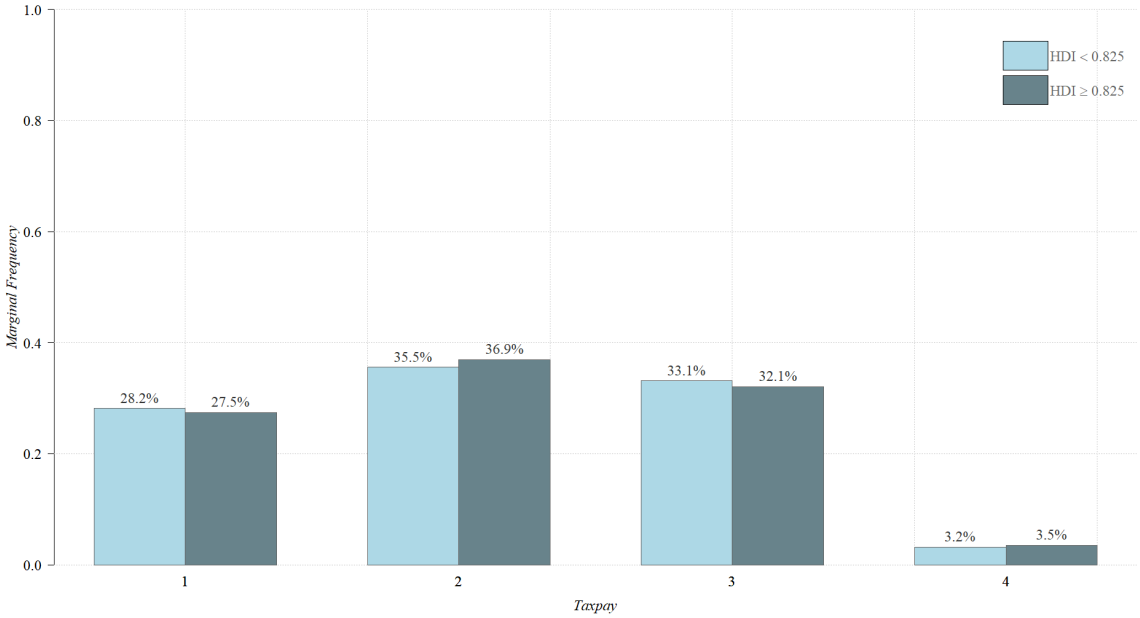


Figure 26: Taxpay Distribution by Region

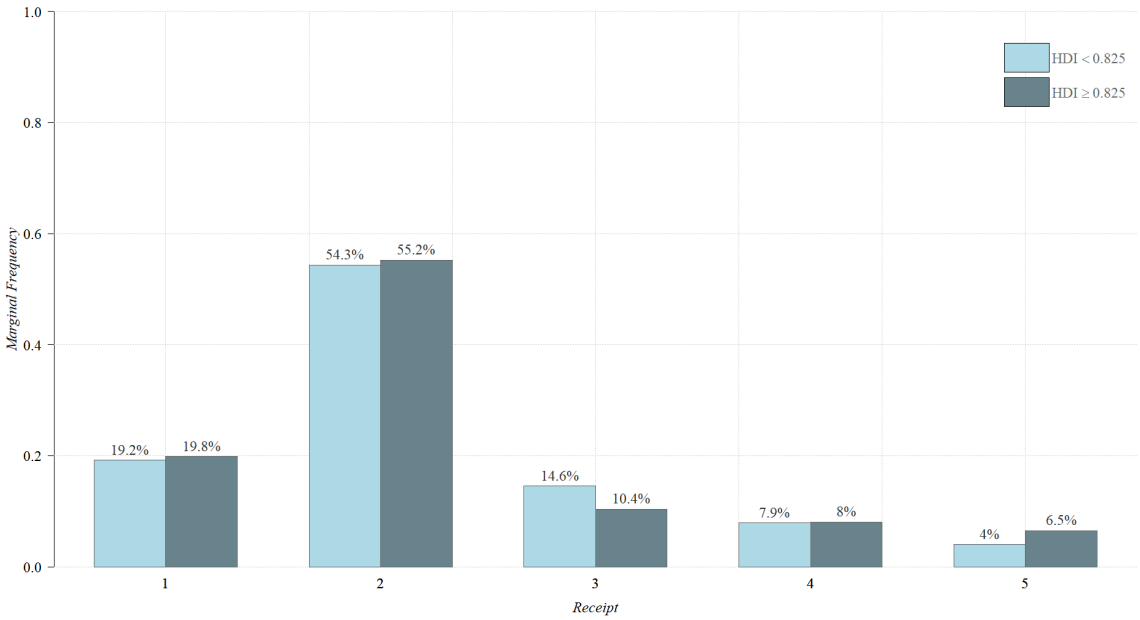


Figure 27: Receipt Distribution by Region

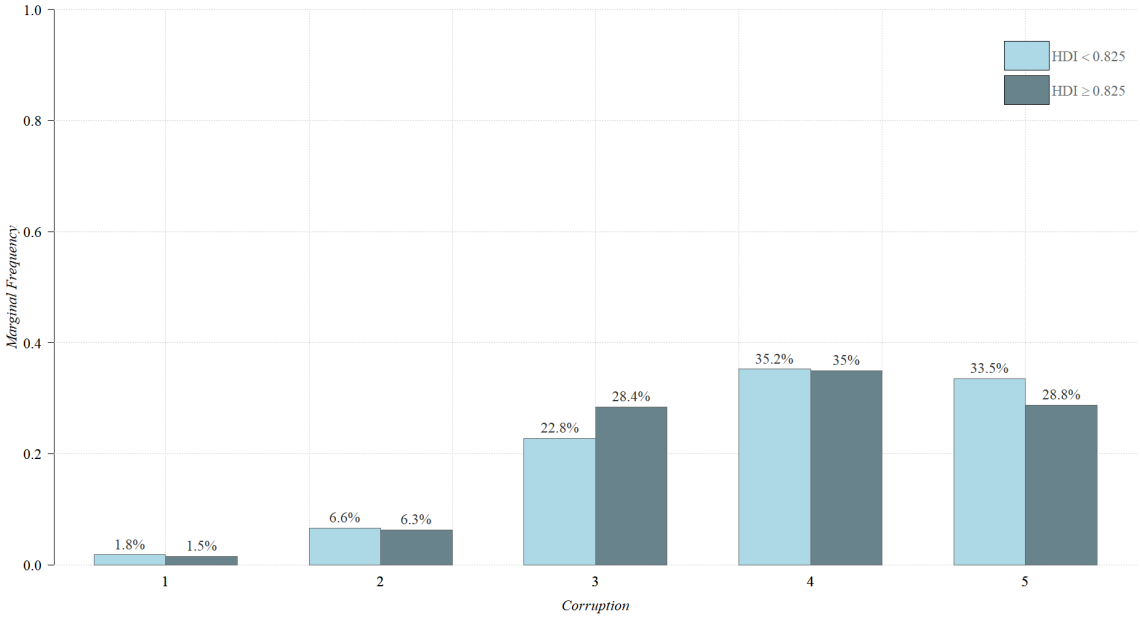


Figure 28: Corruption Distribution by Region

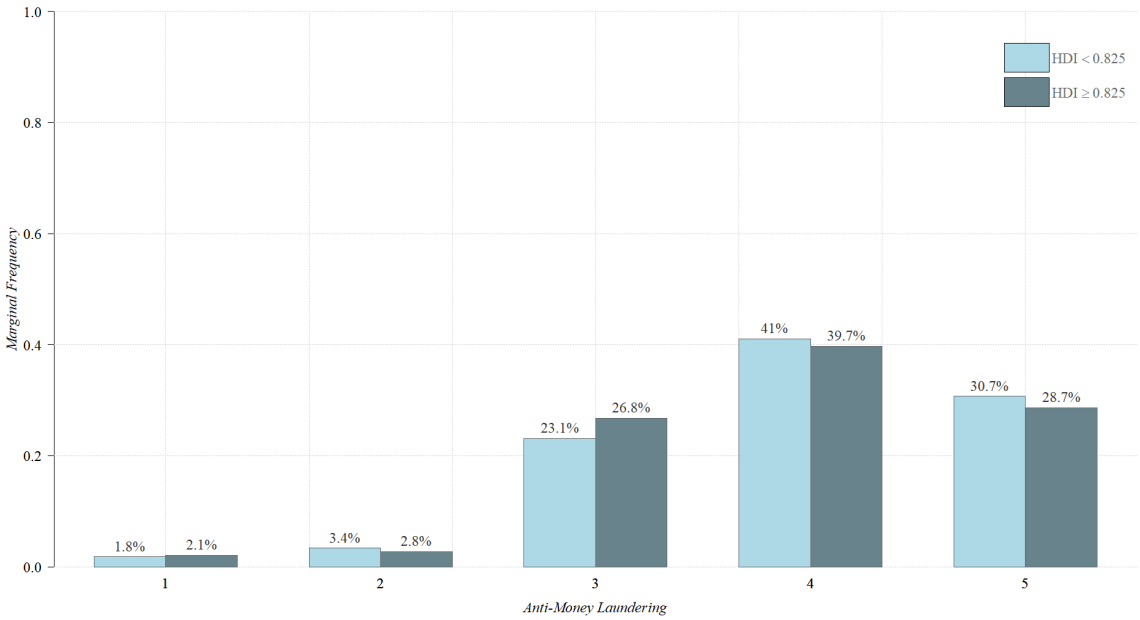


Figure 29: Anti-Money Laundering Distribution by Region

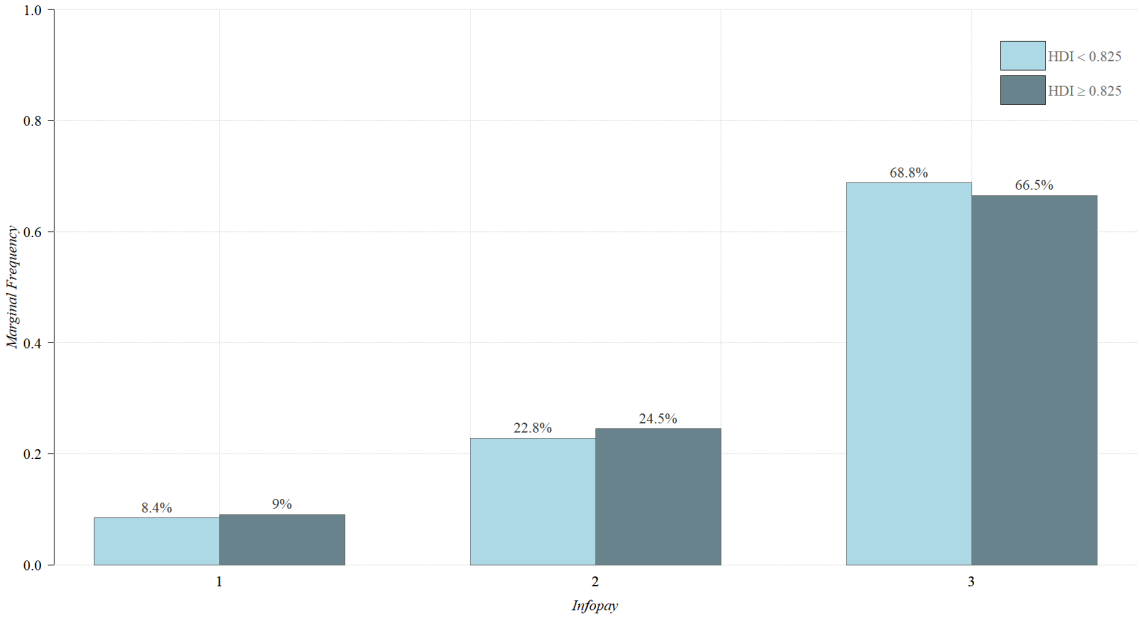


Figure 30: Infopay Distribution by Region

From one of the previous bar charts (Figure 27), a visual difference in the frequencies of the *region* variable modes can be observed, which has been tested using a proportion test (Table 2). The test result on this distribution confirms that the relative frequencies are not the same in the reference populations, with high significance indicated by a $p\text{-}value < 0.01$. Therefore, it can be stated that individuals residing in a location with an HDI lower than 0.825, compared to individuals residing in a location with an HDI equal to or greater than 0.825, are more interested in receipts. This outcome suggests that the *region* variable could prove to be significant in a subsequent model.

Below are the joint distributions with respect to the variable **work**:

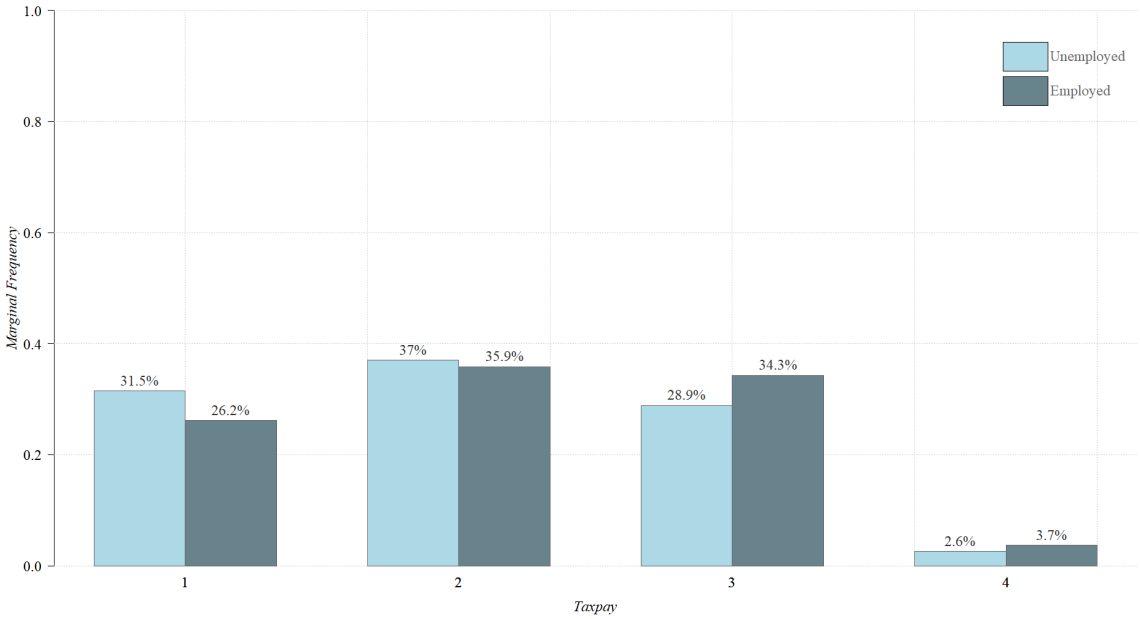


Figure 31: Taxpay Distribution by Work

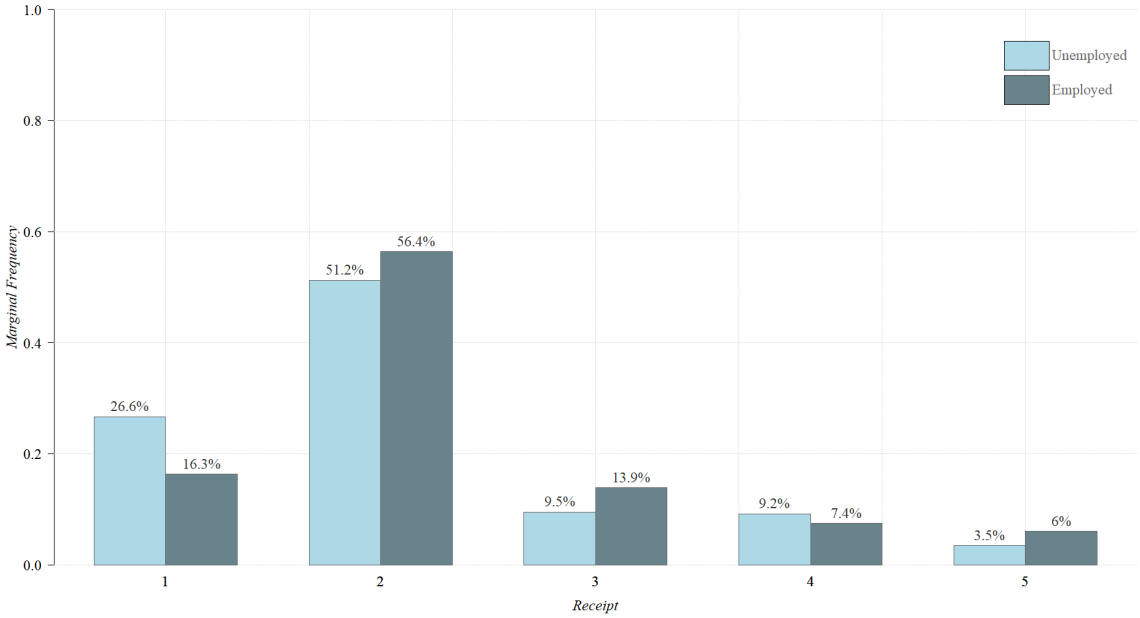


Figure 32: Receipt Distribution by Work

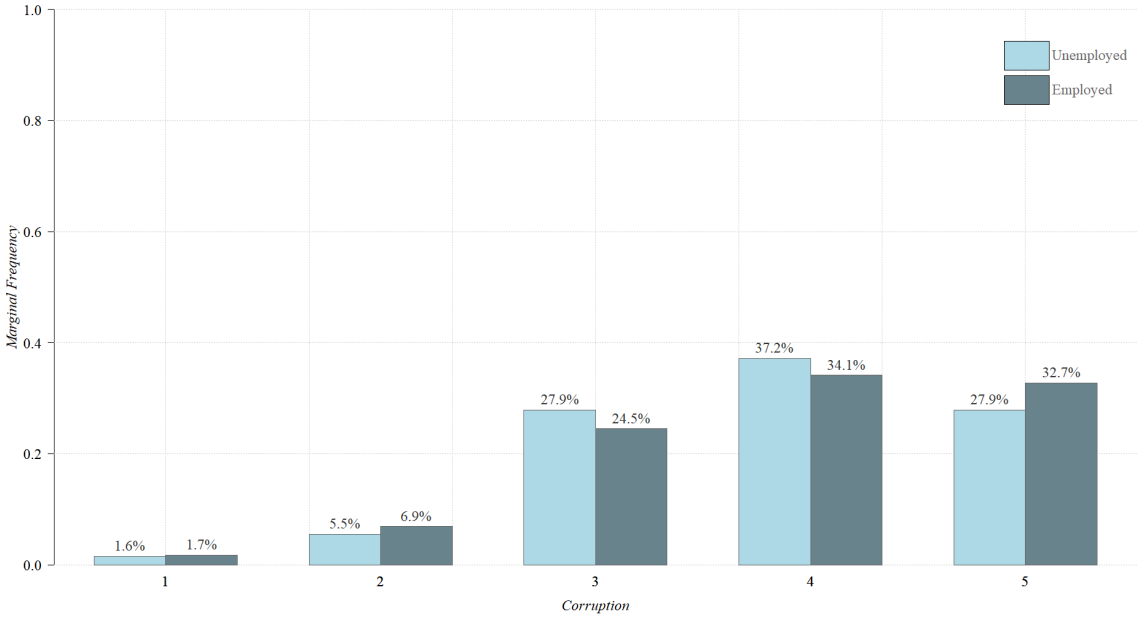


Figure 33: Corruption Distribution by Work

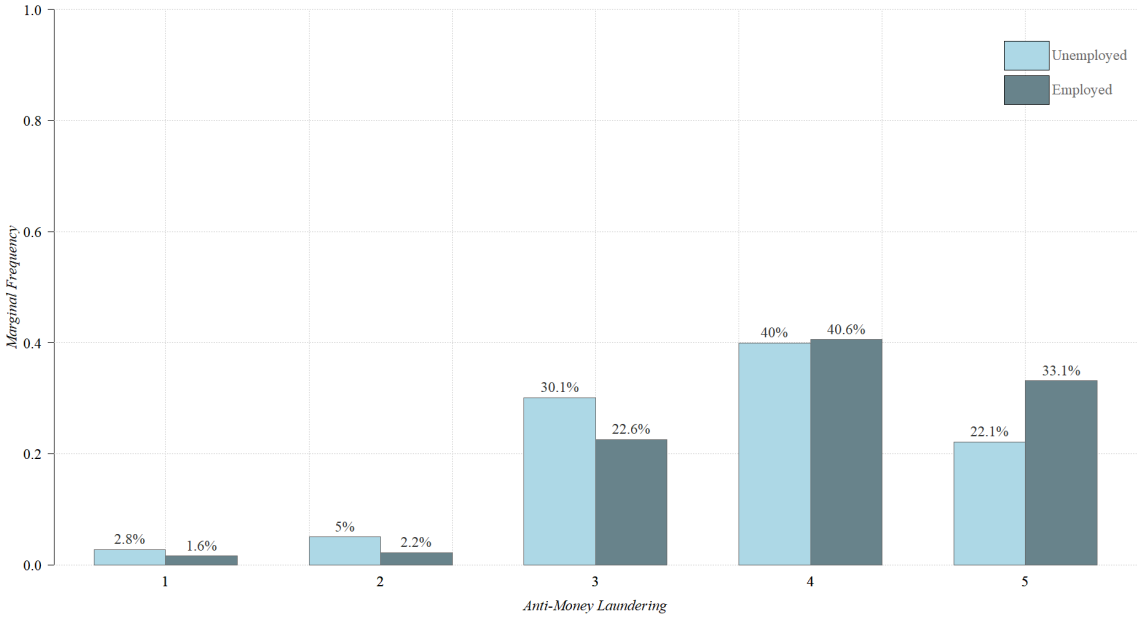


Figure 34: Anti-Money Laundering Distribution by Work

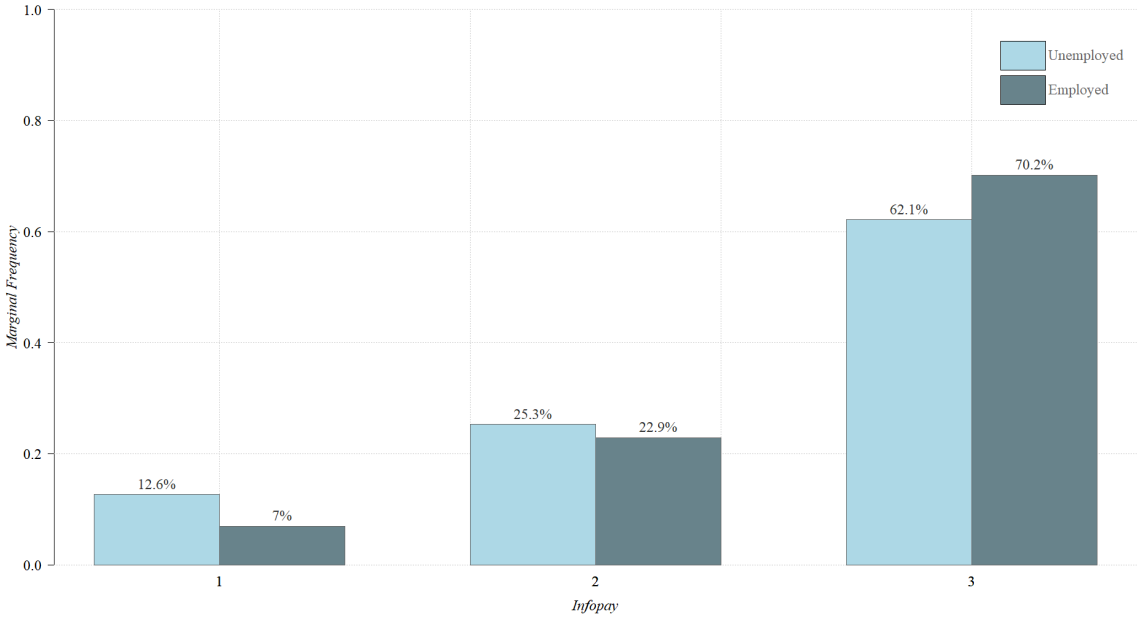


Figure 35: Infopay Distribution by Work

A noticeable visual difference in the frequencies of the *work* variable modes can be observed from the preceding bar charts, which has been tested using proportion tests (Table 2). The test results on four of the previous distributions (Figure 31, 32, 34, 35) confirm that the relative frequencies are not the same in the reference populations, with high significance indicated by a $p\text{-}value<0.01$ in all cases. Therefore, it can be stated that the unemployed, compared to the employed, are more likely to pay taxes in advance, show more interest in receipts, have a higher level of anti-money laundering skills, and are more likely to provide requested information. This outcome suggests that the *work* variable could potentially be significant in a subsequent model.

Below are the joint distributions with respect to the variable **education**:

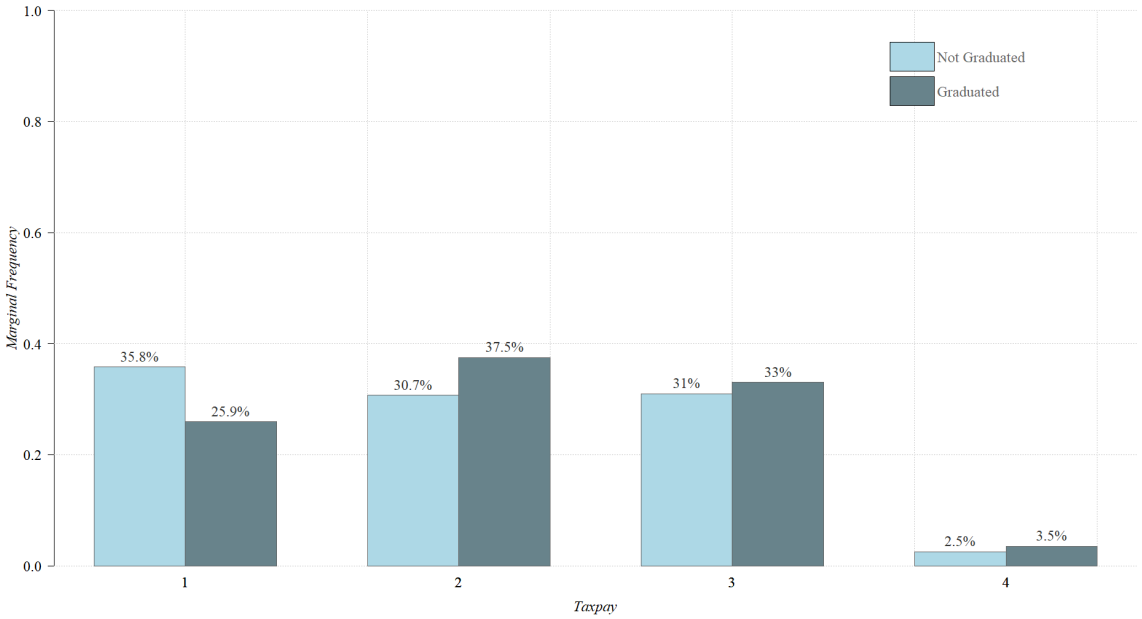


Figure 36: Taxpay Distribution by Education

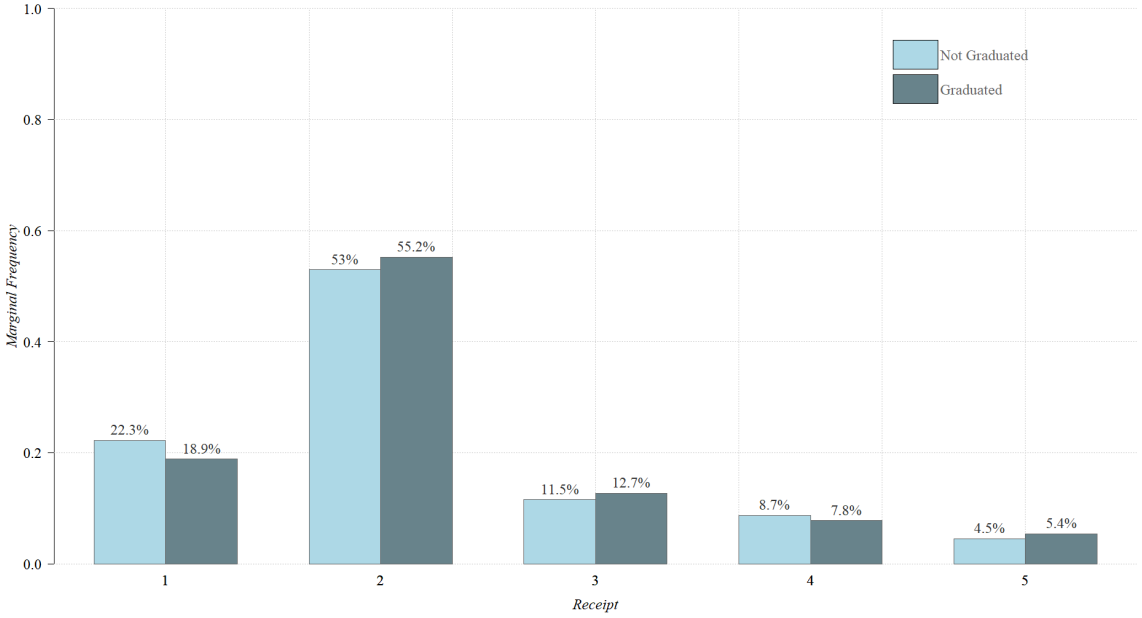


Figure 37: Receipt Distribution by Education

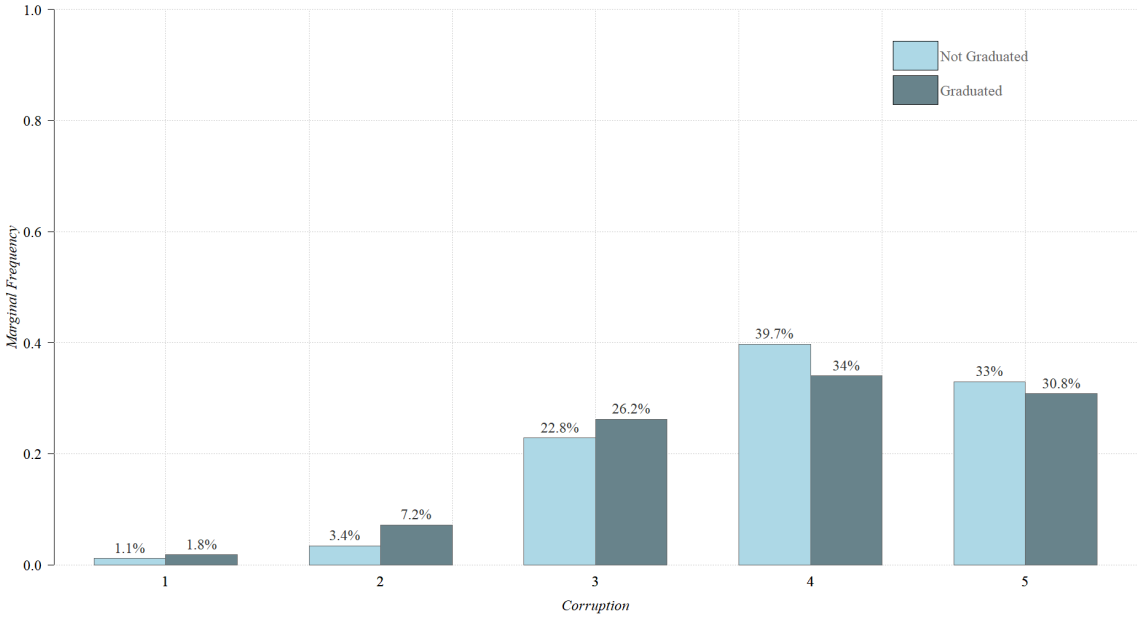


Figure 38: Corruption Distribution by Education

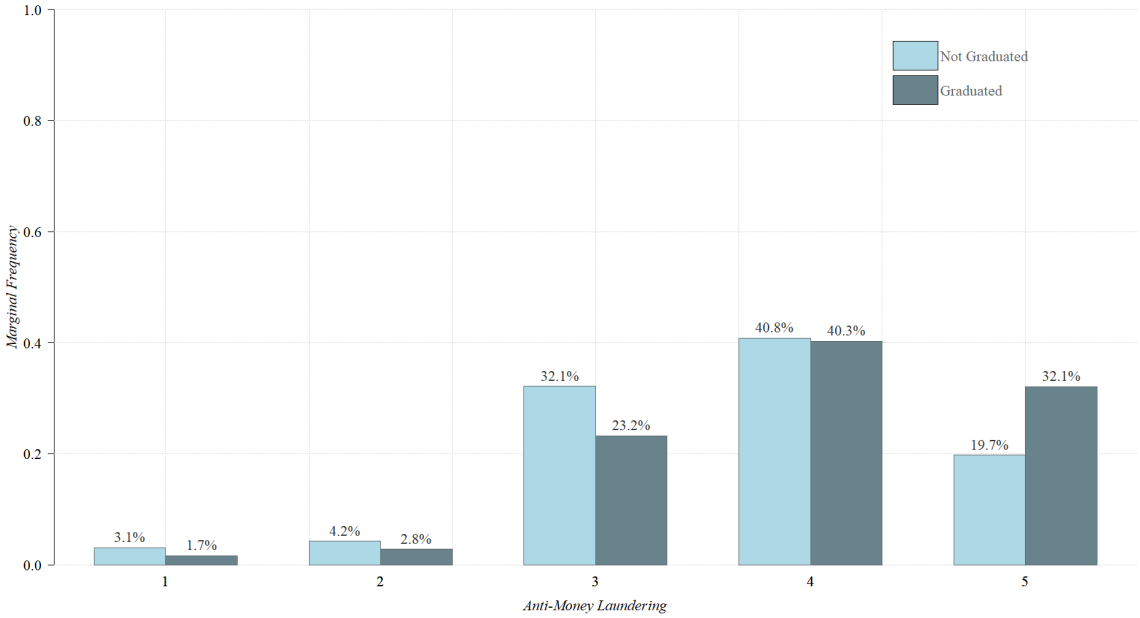


Figure 39: Anti-Money Laundering Distribution by Education

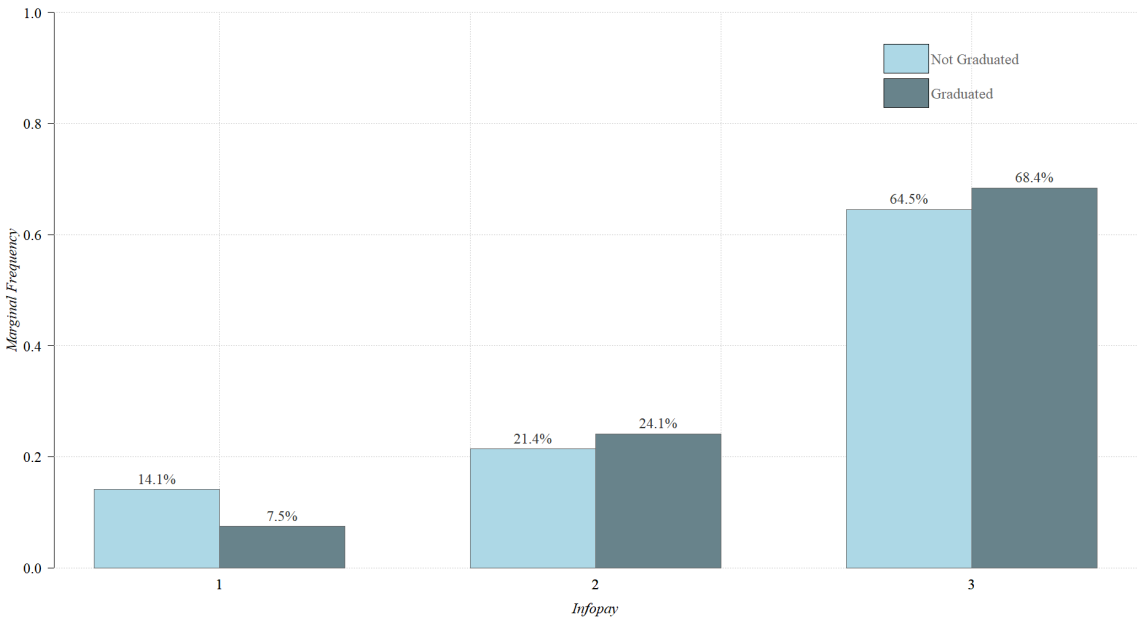


Figure 40: Infopay Distribution by Education

A noticeable visual difference in the frequencies of the *education* variable modes can be observed from the preceding bar charts, which has been tested using proportion tests (Table 2). The test results on four of the previous distributions (Figure 36, 38, 39, 40) confirm that the relative frequencies are not the same in the reference populations, with high significance indicated by a $p\text{-}value<0.01$ in all cases. Therefore, it can be stated that individuals without a degree, compared to those with a degree, are more likely to pay taxes in advance, have a higher perception of corruption, possess a higher level of anti-money laundering skills, and are more likely to provide requested information. This outcome suggests that the *education* variable could potentially be significant in a subsequent model.

2.4 Correlation

The knowledge of the **correlation** between variables is important because it allows to identify the relationship between two or more quantitative variables. This information is useful for understanding data behavior and can be used to predict the behavior of one variable based on the knowledge of another. In this case, correlations between only the dependent variables were analyzed. The results are reported in the following **correlation matrix**:

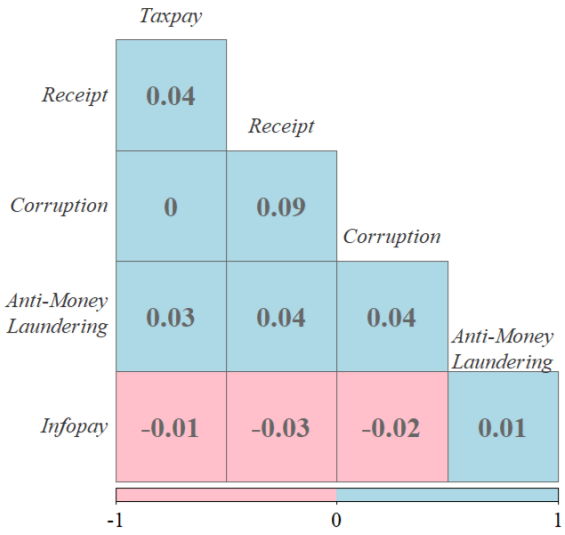


Figure 41: Correlation Matrix

Typically, the correlation matrix is a *symmetric matrix*, meaning it is a square matrix that is equal to its transpose. This implies that element *(i, j)* in the matrix is equal to element *(j, i)* in the same matrix. In this case, to avoid redundant data, the correlation matrix has been reduced to a lower triangular form. Subsequently, data with a unitary value on the main diagonal have been removed. The correlation values for the dependent variables are all very low, with the only negative value associated with the *infopay* variable. Therefore, it can be concluded that there is no significant linear dependence among the dependent variables. It's important to note that the absence of linear dependence doesn't exclude the presence of other complex or nonlinear relationships among the dependent variables, that may emerge later on.

3 Test

3.1 Chi-Squared Test (Independence Test)

The verification of independence between two or more phenomena constitutes a preliminary phase prior to the establishment of more explicit and significant relationships, especially in the context of constructing a statistical model.
The procedure of the test can be summarized as follows:

Null Hypothesis	Alternative Hypothesis	Critical Region
H_0 : X and Y are Indipendent	H_0 : X and Y are not Indipendent	$RC(\alpha) : X^2 > \chi^2_{(\alpha,g=(k-1)(h-1))}$

The test statistic, also known as Chi-Squared Test, can be calculated as follow:

$$X^2 = n\left(\sum_{i=1}^{k}\sum_{j=1}^{h}\frac{(n_{ij})^2}{n_{i.}n_{.j}} - 1\right)$$

Below is the table related to the independence tests conducted on the data:

Variable 1	Variable 2	Chi-Squared	p-value	Signif.
taxpay	receipt	29.2954	3.561e-03	**
taxpay	corruption	17.0637	1.472e-01	
taxpay	antimoneylaundering	10.9590	5.324e-01	
taxpay	infopay	29.4255	5.053e-05	***
taxpay	age	26.3933	7.89e-06	***
taxpay	gender	9.1348	2.755e-02	*
taxpay	region	0.6669	8.81e-01	
taxpay	work	9.1839	2.694e-02	*
taxpay	education	14.9031	1.901e-03	**
receipt	corruption	44.0509	1.938e-04	***
receipt	antimoneylaundering	53.6137	5.995e-06	***
receipt	infopay	25.6150	1.222e-03	**
receipt	age	30.2901	4.272e-06	***
receipt	gender	2.9403	5.679e-01	
receipt	region	12.2220	1.577e-02	*
receipt	work	36.3311	2.474e-07	***
receipt	education	3.0232	5.54e-01	
corruption	antimoneylaundering	60.0888	5.056e-07	***
corruption	infopay	17.7553	2.314e-02	*
corruption	age	10.1688	3.768e-02	*
corruption	gender	9.3033	5.395e-02	.
corruption	region	9.4656	5.046e-02	.
corruption	work	7.0275	1.344e-01	
corruption	education	11.6417	2.022e-02	*
antimoneylaundering	infopay	10.3975	2.382e-01	
antimoneylaundering	age	49.3461	4.944e-10	***
antimoneylaundering	gender	7.6635	1.047e-01	
antimoneylaundering	region	4.1803	3.822e-01	
antimoneylaundering	work	38.5861	8.482e-08	***
antimoneylaundering	education	28.8466	8.399e-06	***
infopay	age	46.5837	7.664e-11	***
infopay	gender	8.0000	1.832e-02	*
infopay	region	1.1144	5.728e-01	
infopay	work	19.3777	6.197e-05	***
infopay	education	15.9416	3.454e-04	***

Table 1: Chi-Squared Test Table

3.2 Proportions Test

The proportions test is a statistical test used to verify whether the difference between the relative frequencies of categories in a dichotomous variable is significant or not.
The procedure of the test can be summarized as follows:

Null Hypothesis	Alternative Hypothesis	Critical Region
$H_0 : \Theta_1 - \Theta_2 = 0$	$H_0 : \Theta_1 - \Theta_2 \neq 0$	$RC(\alpha) : Z_{(n_1,n_2)} \leq z_\alpha$

The test statistic, also known as z-test, can be calculated as follow:

$$Z_{n_1,n_2} = \frac{p_1 - p_2}{\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Where p_1 is the proportion observed in group one with size n_1, $p_2$is the proportion observed in group two with size n_1, p and q are the overall proportions.

Below is the table related to the proportion tests conducted on the data:

Variable 1	Variable 2	Chi-Squared	df	p-value	Signif.
taxpay	age	26.3933	3	7.89e-06	***
taxpay	gender	9.1348	3	2.755e-02	*
taxpay	region	0.6669	3	8.81e-01	
taxpay	work	9.1839	3	2.694e-02	*
taxpay	education	14.9031	3	1.901e-03	**
receipt	age	30.2901	4	4.272e-06	***
receipt	gender	2.9403	4	5.679e-01	
receipt	region	12.2220	4	1.577e-02	*
receipt	work	36.3311	4	2.474e-07	***
receipt	education	3.0232	4	5.54e-01	
corruption	age	10.1688	4	3.768e-02	*
corruption	gender	9.3033	4	5.395e-02	.
corruption	region	9.4656	4	5.046e-02	.
corruption	work	7.0275	4	1.344e-01	
corruption	education	11.6417	4	2.022e-02	*
antimoneylaundering	age	49.3461	4	4.944e-10	***
antimoneylaundering	gender	7.6635	4	1.047e-01	
antimoneylaundering	region	4.1803	4	3.822e-01	
antimoneylaundering	work	38.5861	4	8.482e-08	***
antimoneylaundering	education	28.8466	4	8.399e-06	***
infopay	age	46.5837	2	7.664e-11	***
infopay	gender	8.0000	2	1.832e-02	*
infopay	region	1.1144	2	5.728e-01	
infopay	work	19.3777	2	6.197e-05	***
infopay	education	15.9416	2	3.454e-04	***

Table 2: Proportions Test Table

Below is the table to interpret the level of significance of the tests:

Signif.	%
***	0.1%
**	1%
*	5%
.	10%
	Not Signif.

4 Generalized Linear Model

4.1 Proportional Odds Model (POM)

The **Proportional Odds Model (POM)** is a regression model for the ordinal classification of dependent variables that uses logistic regression to estimate proportional odds ratios. The main difference compared to ordinary logistic regression is that POM takes into account the order among the categories of the dependent variable, assigning different weights to each category and modeling the probability of belonging to one category compared to the others. This model is particularly useful in situations where the dependent variable has more than two categories, and the order among them is important. The first step involved adopting a sequential strategy (*stepwise regression*) for the selection of explanatory variables. In particular, the *backward elimination* strategy was chosen, which is a *step-down* procedure. It starts with the most complete regression model, including all explanatory variables, progressively eliminating variables that are not significant. Then, a new model is estimated at each step with the significant variables from the previous step. Finally, the procedure stops when all the variables present in the model have been found significant at a certain level, meaning that all calculated p-values fall below a predetermined critical threshold ($p - value < 0.05$).

The first model is related to the variable **taxpay**:

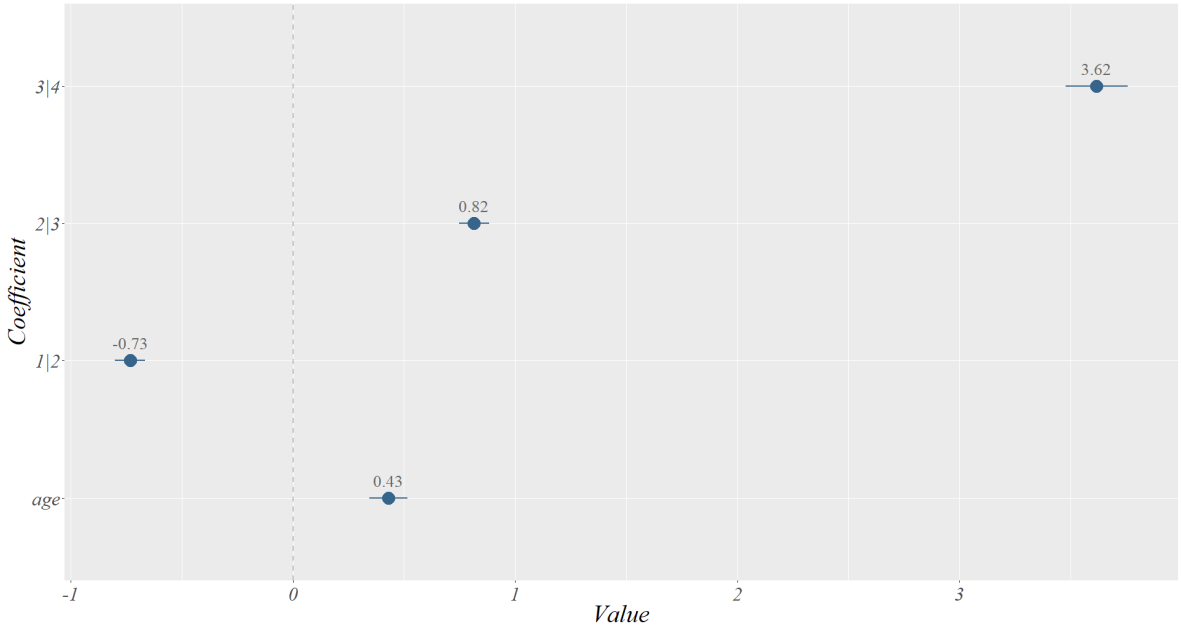


Figure 42: Coefficients Plot POM Taxpay

The POM is not straightforward to interpret because it doesn't specify a relationship between observed variables but rather between the probability of an event and one or more explanatory variables. One possible way to interpret this model is through *odds ratios*, which are the ratio of the probability of an event to the probability of its negation:

$$odds(E) = \frac{Pr(E)}{Pr(\overline{E})} = \frac{Pr(E)}{1 - Pr(E)}$$

Applying the logarithm yields the *log-odds:*

$$log - odds(E) = \log(odds(E)) = \log\left[\frac{Pr(E)}{1 - Pr(E)}\right] = logit(Pr(E))$$

The obtained model takes the following forms:

$$logit(Pr(taxpay \leq 1)) = -0.73 + 0.43\, age$$
$$logit(Pr(taxpay \leq 2)) = 0.82 + 0.43\, age$$
$$logit(Pr(taxpay \leq 3)) = 3.62 + 0.43\, age$$

Using odds to interpret the model with respect to the explanatory variables yields that:

$$e^{\beta_1} = e^{0.43} = 1.54$$

For individuals over 35, the probability of paying taxes in advance is over 54% higher than for those under 35. Here are some indices to compare the full model (M_c) with the final model (M_f):

	Full Model	*Final Model*
AIC	4449.75	4445.04
BIC	4493.96	4467.14
log-Lik (ℓ)	-2216.88	-2218.52

Having used the backward elimination method for selecting explanatory variables, we are dealing with nested models. The table presents three evaluation criteria: **Akaike Information Criterion (AIC)**, **Bayesian Information Criterion (BIC)**, and **log-likelihood (log-Lik)**. We prefer the model with lower values for these indices. Furthermore:

$$BIC = AIC + (p+1)[\log n - 2]$$

It will always be:

$$BIC > AIC \quad if \quad n > 8$$

Therefore, the BIC index is more "stringent" in the choice among multiple models. The AIC index tends to overparameterize the models it selects, so BIC is preferred when using a parsimonious model is important. The AIC index is biased, while the BIC index for common models is consistent and as n increases, it specifies the correct model with probability 1. However, the BIC and AIC indices are used to compare non-nested models.
Another method for comparing nested models is the **likelihood ratio test (LRT)**:

$$LRT = -2(\ell(M_f) - \ell(M_c)) = -2(-2218.52 - 2216.88) = 3.29 \rightarrow \chi^2_{g=4} \qquad p - value = 0.51$$

Given that the $p - value > 0.05$, we reject the null hypothesis. This means that the full model and the final model fit the data equally well. Therefore, we should use the final model because the additional predictor variables in the full model do not offer a significant improvement in fit.
Subsequently, the Brant-Wald test was performed:

	χ^2_g	g	*p-value*
Omnibus	1.31	2	0.52
age	1.31	2	0.52

The **Brant-Wald test** is conducted to compare proportional odds and generalized models. A Brant-Wald test is a hypothesis test on the significance of the difference in model coefficients, producing a chi-squared statistic. A low p-value in a Brant-Wald test indicates that the coefficient does not satisfy the proportional odds assumption.
From the results obtained, we can't reject the proportional odds assumption.

The second model is related to the variable **recipt**:

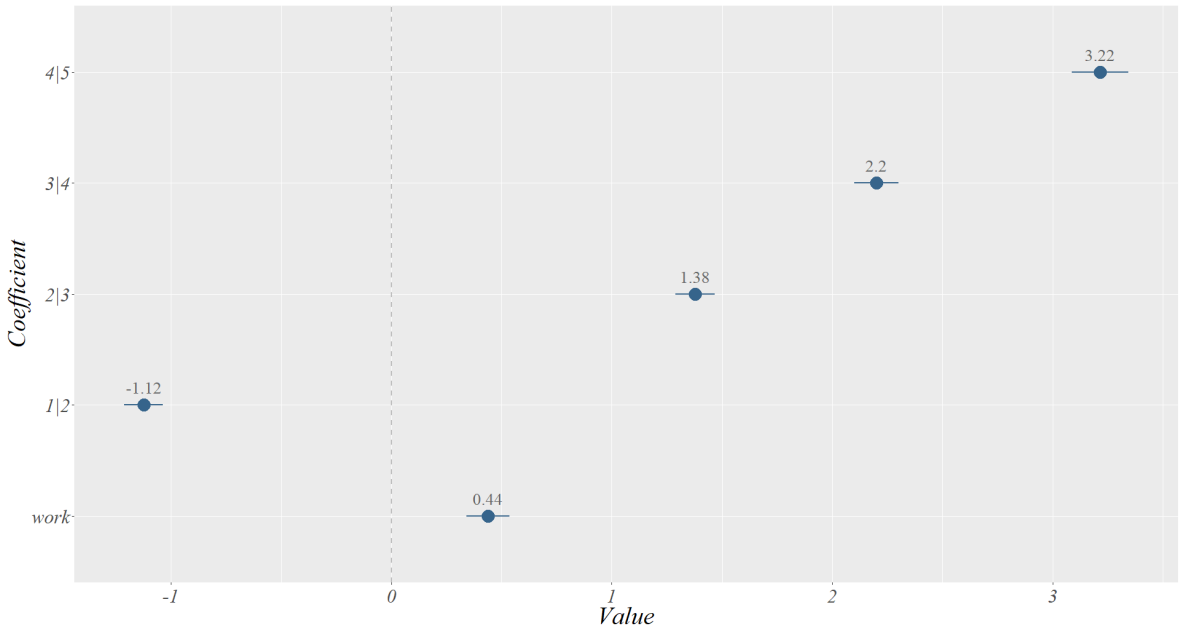


Figure 43: Coefficients Plot POM Receipt

The obtained model takes the following forms:

$$logit(Pr(receipt \leq 1)) = -1.12 + 0.44\, work$$
$$logit(Pr(receipt \leq 2)) = 1.38 + 0.44\, work$$
$$logit(Pr(receipt \leq 3)) = 2.20 + 0.44\, work$$
$$logit(Pr(receipt \leq 4)) = 3.22 + 0.44\, work$$

Using odds to interpret the model with respect to the explanatory variables yields that:

$$e^{\beta_1} = e^{0.44} = 1.55$$

For employed individuals, the probability of paying more attention to the receipt is over 55% higher than for the unemployed.
Here are some indices to compare the full model (M_c) with the final model (M_f):

	Full Model	*Final Model*
AIC	4687.24	4681.54
BIC	4736.97	4709.17
log-Lik (ℓ)	-2334.62	-2335.77

$$LRT = -2(\ell(M_f) - \ell(M_c)) = -2(-2335.77 - 2334.62) = 2.30 \rightarrow \chi^2_{g=4} \qquad p - value = 0.68$$

Given that the $p - value > 0.05$, we reject the null hypothesis. This means that the full model and the final model fit the data equally well. Therefore, we should use the final model because the additional predictor variables in the full model do not offer a significant improvement in fit.
Subsequently, the Brant-Wald test was performed:

	χ^2_g	g	*p-value*
Omnibus	15.11	3	0.00
age	15.11	3	0.00

From the results obtained, we can reject the proportional odds assumption.

The third model is related to the variable **corruption**:

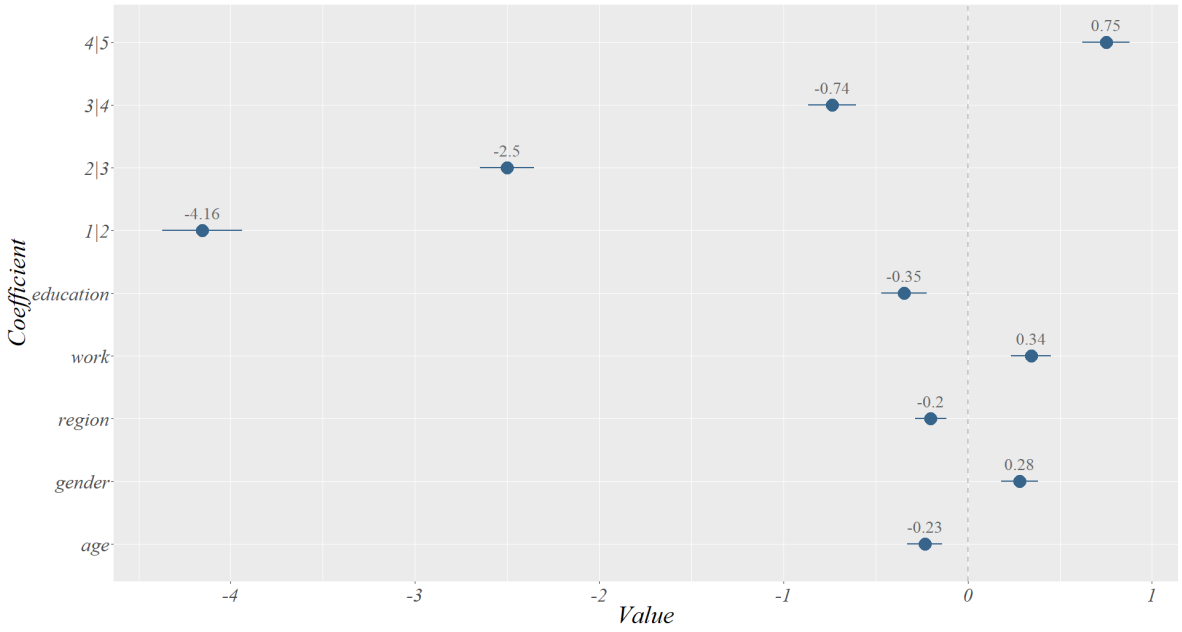


Figure 44: Coefficients Plot POM Corruption

The obtained model takes the following forms:

$$logit(Pr(receipt \leq 1)) = -4.16 - 0.23\, age + 0.28\, gender - 0.20\, region + 0.34\, work - 0.35\, education$$

$$logit(Pr(receipt \leq 2)) = -2.50 - 0.23\, age + 0.28\, gender - 0.20\, region + 0.34\, work - 0.35\, education$$

$$logit(Pr(receipt \leq 3)) = -0.74 - 0.23\, age + 0.28\, gender - 0.20\, region + 0.34\, work - 0.35\, education$$

$$logit(Pr(receipt \leq 4)) = 0.75 - 0.23\, age + 0.28\, gender - 0.20\, region + 0.34\, work - 0.35\, education$$

Using odds to interpret the model with respect to the explanatory variables yields that:

$$e^{\beta_1} = e^{-0.23} = 0.79; \qquad e^{\beta_2} = e^{0.28} = 1.32; \qquad e^{\beta_3} = e^{-0.20} = 0.82; \qquad e^{\beta_4} = e^{0.34} = 1.40; \qquad e^{\beta_5} = e^{-0.35} = 0.70;$$

For individuals over 35, the probability of having a low level of corruption perception is 21% lower than for those under 35. For men, the probability of having a low level of corruption perception is 32% higher than for women. For individuals residing in a place with an HDI greater than or equal to 0.825, the probability of having a low level of corruption perception is 18% lower than for individuals residing in a place with an HDI less than 0.825. For employed individuals, the probability of having a low level of corruption perception is 40% higher than for the unemployed. For graduates, the probability of having a low level of corruption perception is 30% lower than for non-graduates.
In this case, none of the variables were found to be significant, so we stick with the full model.
Here are some indices:

	Full Model	*Final Model*
AIC	4902.20	-
BIC	4951.93	-
log-Lik (ℓ)	-2442.10	-

Subsequently, the Brant-Wald test was performed:

	χ^2_g	g	*p-value*
Omnibus	16.81	15	0.33
age	1.75	3	0.62
gender	2.39	3	0.50
region	3.88	3	0.27
work	1.16	3	0.76
education	2.77	3	0.43

From the results obtained, we can't reject the proportional odds assumption.

The fourth model is related to the variable **antimoneylaundering**:

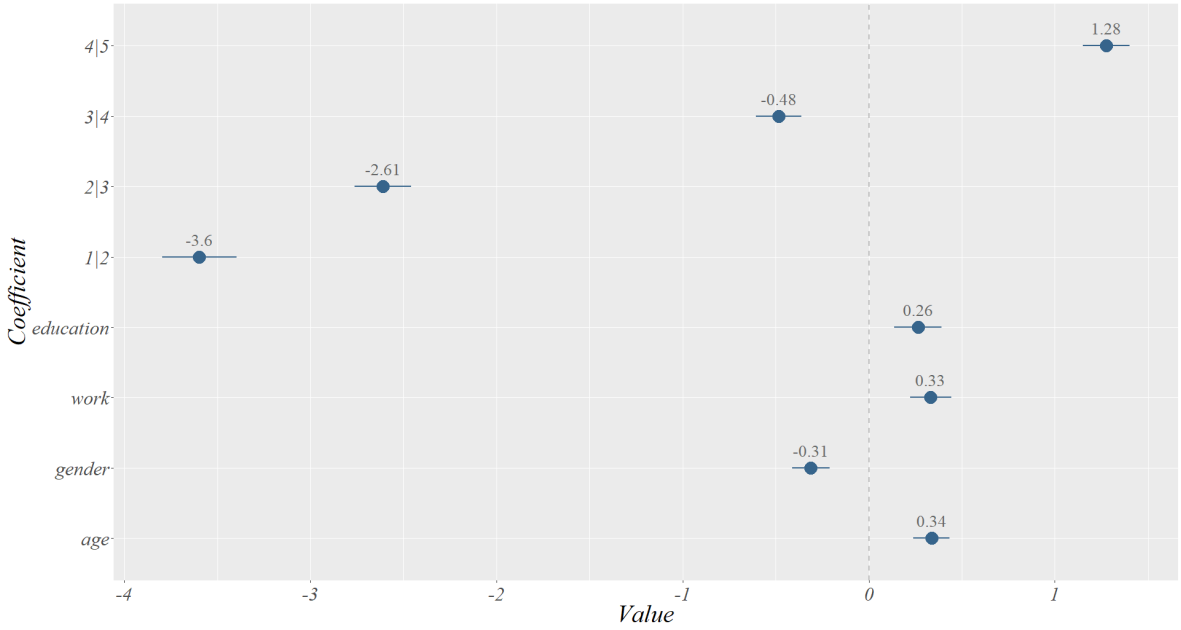


Figure 45: Coefficients Plot POM Anti-Money Laundering

The obtained model takes the following forms:

$$logit(Pr(antimoneylaundering \leq 1)) = -3.60 + 0.34\, age - 0.31\, gender + 0.33\, work + 0.26\, education$$

$$logit(Pr(antimoneylaundering \leq 2)) = -2.61 + 0.34\, age - 0.31\, gender + 0.33\, work + 0.26\, education$$

$$logit(Pr(antimoneylaundering \leq 3)) = -0.48 + 0.34\, age - 0.31\, gender + 0.33\, work + 0.26\, education$$

$$logit(Pr(antimoneylaundering \leq 4)) = 1.28 + 0.34\, age - 0.31\, gender + 0.33\, work + 0.26\, education$$

Using odds to interpret the model with respect to the explanatory variables yields that:

$$e^{\beta_1} = e^{0.34} = 1.40; \qquad e^{\beta_2} = e^{-0.31} = 0.73; \qquad e^{\beta_3} = e^{0.33} = 1.39; \qquad e^{\beta_4} = e^{0.26} = 1.30;$$

For individuals over 35, the probability of having a low level of anti-money laundering skills is more than 40% higher than for those under 35. For women, the probability of having a low level of anti-money laundering skills is 27% lower than for men. For the employed, the probability of having a low level of anti-money laundering skills is 39% higher than for the unemployed. For graduates, the probability of having a low level of anti-money laundering skills is 30% higher than for non-graduates. Here are some indices to compare the full model (M_c) with the final model (M_f):

	Full Model	*Final Model*
AIC	4312.18	4612.06
BIC	4661.91	4656.26
log-Lik (ℓ)	-2297.09	-2298.03

$$LRT = -2(\ell(M_f) - \ell(M_c)) = -2(-2298.03 - 2297.09) = 1.87 \rightarrow \chi^2_{g=1} \qquad p - value = 0.17$$

Given that the $p - value > 0.05$, we reject the null hypothesis. This means that the full model and the final model fit the data equally well. Therefore, we should use the final model because the additional predictor variables in the full model do not offer a significant improvement in fit.
Subsequently, the Brant-Wald test was performed:

	χ^2_g	g	*p-value*
Omnibus	21.81	12	0.04
age	15.78	3	0.00
gender	2.28	3	0.52
work	2.38	3	0.50
education	4.12	3	0.25

From the results obtained, we can reject the proportional odds assumption.

The fifth model is related to the variable **corruption**:

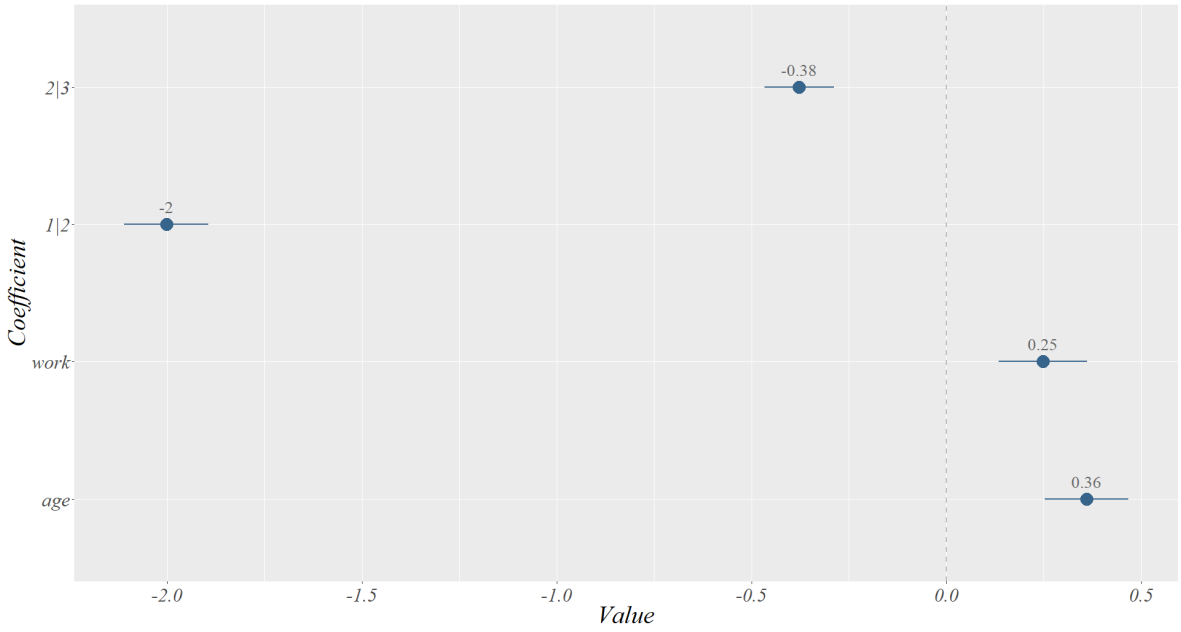


Figure 46: Coefficients Plot POM Infopay

The obtained model takes the following forms:

$$logit(Pr(antimoneylaundering \leq 1)) = -2.00 + 0.36\, age + 0.25\, work$$
$$logit(Pr(antimoneylaundering \leq 2)) = -0.38 + 0.36\, age + 0.25\, work$$

Using odds to interpret the model with respect to the explanatory variables yields that:

$$e^{\beta_1} = e^{0.36} = 1.43; \qquad e^{\beta_2} = e^{0.25} = 1.28$$

For individuals over 35, the probability of not wanting to provide the requested information increases by 43% compared to those under 35. For the employed, the probability of not wanting to

provide the requested information increases by 43% compared to the unemployed.
Here are some indices to compare the full model (M_c) with the final model (M_f):

	Full Model	*Final Model*
AIC	3020.12	3017.14
BIC	3058.80	3039.24
log-Lik (ℓ)	-1503.06	-1504.57

$$LRT = -2(\ell(M_f) - \ell(M_c)) = -2(-1504.57 - 1503.06) = 3.01 \rightarrow \chi^2_{g=3} \qquad p - value = 0.39$$

Given that the $p - value > 0.05$, we reject the null hypothesis. This means that the full model and the final model fit the data equally well. Therefore, we should use the final model because the additional predictor variables in the full model do not offer a significant improvement in fit.
Subsequently, the Brant-Wald test was performed:

	χ^2_g	g	*p-value*
Omnibus	23.39	2	0.00
age	20.72	1	0.00
work	0.00	1	0.97

From the results obtained, we can reject the proportional odds assumption.

4.2 Graded Response Model

The **Graded Response Model**, also known as the *Samejima Model* after its developer *Hiroshi Samejima* in 1969, is a statistical model used in the analysis of responses to test items or questionnaires. This model is primarily employed in the field of *Item Response Theory (IRT)*, which is a theory aimed at assessing individuals' abilities or latent traits based on their responses to questions or statements. The graded response model is a variant within the broader framework of Item Response Models and is specifically designed to handle ordinal or categorical data. This model is often used in situations where individuals' responses to a set of questions can be ordered in a graded manner or into categories. In essence, the graded response model assesses the probability that an individual will give a particular response or fall into a specific category based on their abilities or latent traits. This model takes into account that questions may have more than two response options and that these options can be ordered based on the level of difficulty or ability required to respond correctly.
The model is widely used in the fields of education, psychology, and other areas where it is necessary to assess individuals' abilities or latent traits through questionnaires or tests.

Model formulation:

$$\log \frac{p(Y_{ij} \geq y|\theta_i)}{p(Y_{ij} < y|\theta_i)} = \lambda_j(\theta_i - \beta_{iy}) \quad with \quad j = 1, ..., J \quad and \quad y = 1, ..., I_j - 1$$

Probability of scoring y on item j:

$$p_{iy}(\theta_i) = \frac{e^{\lambda_j(\theta_i - \beta_{jy})}}{1 + e^{\lambda_j(\theta_i - \beta_{jy})}} - \frac{e^{\lambda_j(\theta_i - \beta_{j,y+1})}}{1 + e^{\lambda_j(\theta_i - \beta_{j,y+1})}} = p^*_{jy}(\theta_i) - p^*_{j,y+1}(\theta_i)$$

where:

$$p^*_{jy}(\theta_i) = \frac{e^{\lambda_j(\theta_i - \beta_{jy})}}{1 + e^{\lambda_j(\theta_i - \beta_{jy})}} (2PL\ Model) \qquad p^*_{jI_j}(\theta_i) \equiv 0 \qquad p^*_{j0}(\theta_i) \equiv 1.$$

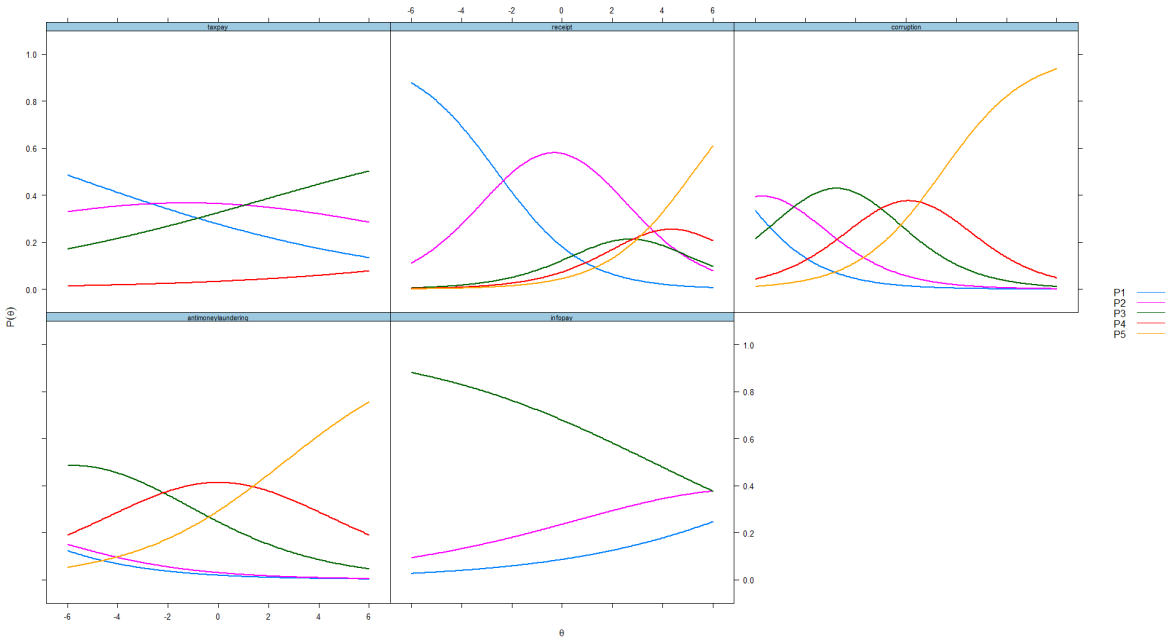


Figure 47: Item Probability Functions

It is interesting to examine the probabilities of responding to specific categories in an item's response scale. These probabilities are graphically displayed in the category response curves (CRCs). Each symmetrical curve represents the probability of endorsing a response category. These curves have a functional relationship with theta, as theta increases, the probability of endorsing a category increases and then decreases as responses transition to the next higher category. The CRCs indicate that the response categories cover a wide range of theta.
To assess the model fit, the *M2* index, designed specifically for evaluating the fit of item response models for ordinal data, was used. Additionally, the M2-based root mean square error of approximation served as the primary fit index, along with the *standardized root mean square residual (SRMSR)* and the *comparative fit index (CFI)*:

	M2	**df**	**p**	**RMSEA**	**RMSEA_5**	**RMSEA_95**	**SRMSR**	**TLI**	**CFI**
Stats	3.20	5	0.67	0	0	0.03	0.02	1.20	1

The obtained $RMSEA = 0$ (95% CI[0.00, 0.03]) and $SRMSR = 0.02$ suggest that data fit the model reasonably well using suggested cutoff values of $RMSEA \leq 0.06$ and $SRMSR \leq 0.08$ as suggested guidelines for assessing fit. The $CFI = 1$ was above a recommended 0.95 threshold.

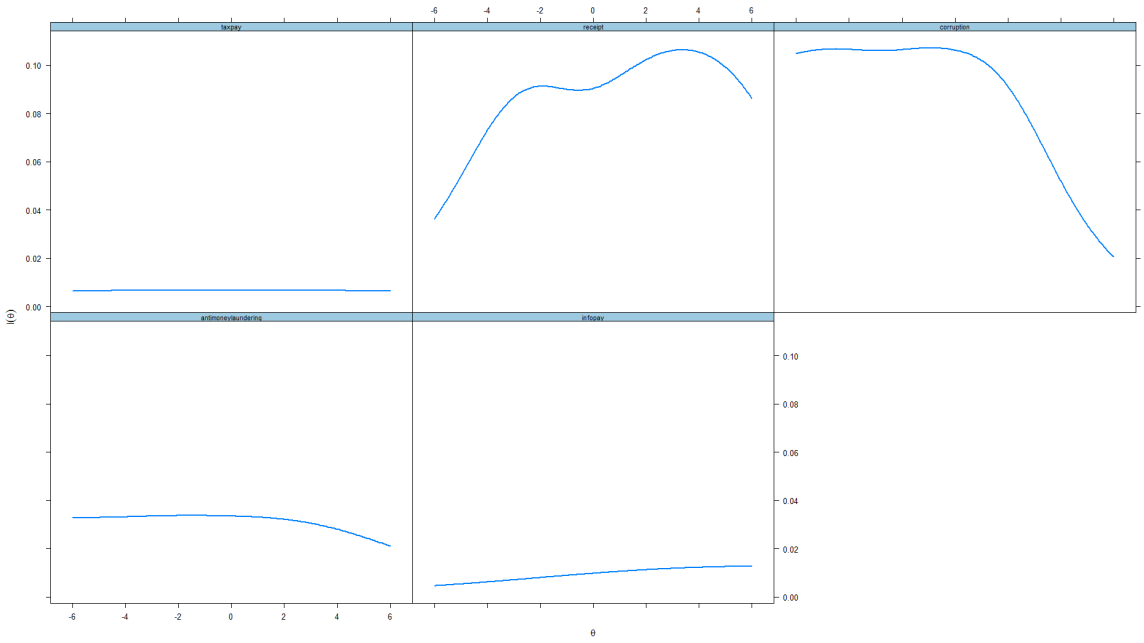


Figure 48: Item Information Curves

In polytomous models, the amount of information an item contributes depends on its slope parameter: the larger the parameter, the more information the item provides. Further, the farther apart the location parameters (b1, b2, b3, b4), the more information the item provides. Typically, an optimally informative polytomous item will have a large location and broad category coverage (as indicated by location parameters) over theta.
Information functions are best illustrated by the item information curves for each item as displayed above. These curves show that item information is not a static quantity, rather, it is conditional on levels of theta. The relationship between slopes and information is illustrated here. Item 1 had the lowest slope and is, therefore, the least informative item.On the other hand, Item 2 had the highest slope and provides the highest amount of statistical information. Items tended to provide the most information between [-2, 2] theta range. The wavy form of the curves reflects the fact that item information is a composite of category information, that is, each category has an information function which is then combined to form the item information function.

	a	**b1**	**b2**	**b3**	**b4**
taxpay	0.15	-6.37	3.83	22.38	-
receipt	0.58	-2.60	1.95	3.44	5.23
corruption	0.59	-7.14	-4.32	-1.23	1.43
antimoneylaundering	0.34	-11.84	-8.91	-2.61	2.62
infopay	-0.21	11.35	3.58	-	-

References

[1] Agresti A. *Analysus of Ordinal Categorical Data*. Wiley, 2 edition, 2010.

[2] Agresti A. *Categorical Data Analysis*. Wiley, 3 edition, 2013.

[3] Dobson A. *An Introduction to Generalized Linear Model*. Springer, 2 edition, 1990.

[4] Monica Violeta Achim and Robert W McGee. *Financial Crime in Romania: A Community Pulse Survey*. Springer Nature, 2023.

[5] Frank B Baker, Seock-Ho Kim, et al. *The basics of item response theory using R*. Springer, 2017.

[6] Blangiardo G. C. *Elementi di Demografia*. Il Mulino, 1 edition, 2006.

[7] Li Cai and Mark Hansen. Limited-information goodness-of-fit testing of hierarchical item factor models. *British Journal of Mathematical and Statistical Psychology*, 66(2), 2013.

[8] R Philip Chalmers. mirt: A multidimensional item response theory package for the r environment. *Journal of statistical Software*, 48, 2012.

[9] Piccolo D. *Statistica*. Il Mulino, 3 edition, 2010.

[10] Murphy K. P. *Machine Learning: A Probabilistic Perspective*. MIT press, 1 edition, 2012.

[11] Fumiko Samejima. Graded response model. In *Handbook of modern item response theory*, pages 85–100. Springer, 1997.