Buongiorno

An exploratory study on Household Finance and Consumption of Italian households

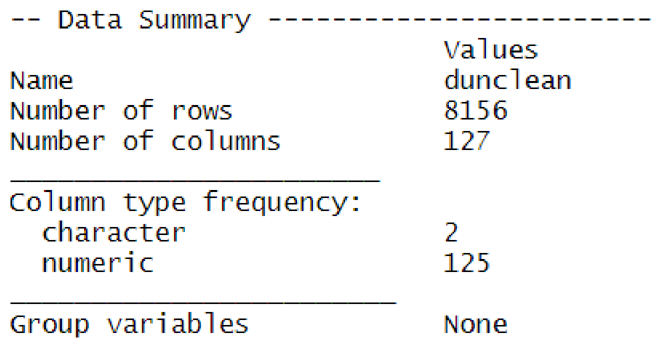
A group Project for the course of Statistical Data Analysis

The Household Finance and Consumption Survey dataset from European survey data of Banca D’Italia is a comprehensive source of information on household balance sheets and related economic and demographic variables.

It provides information on household size, demographics, financial and non-financial assets, liabilities, income, and consumption patterns of households in Italy. The dataset was collected by sampling households from all regions of the country and can be used for exploratory study of household finances in Italy.

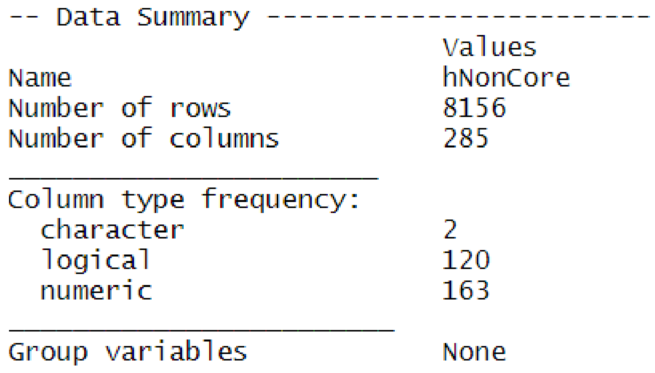
Dataset

D1



```
-- Data Summary ------------------------
                           Values
Name                       dunclean
Number of rows             8156
Number of columns          127
_______________________
Column type frequency:
  character                2
  numeric                  125
________________________
Group variables            None
```

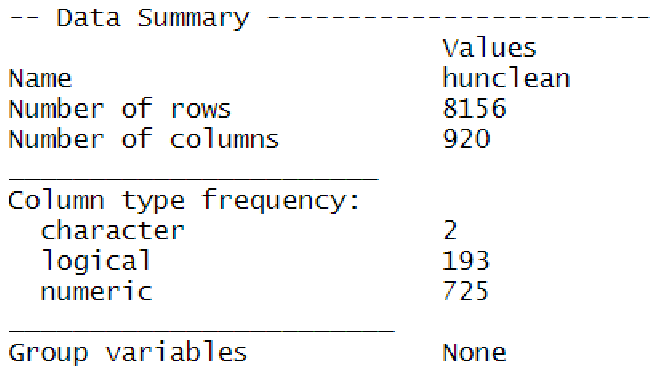
H1 Non Core Variables



```
-- Data Summary ------------------------
                           Values
Name                       hNonCore
Number of rows             8156
Number of columns          285
_______________________
Column type frequency:
  character                2
  logical                  120
  numeric                  163
________________________
Group variables            None
```

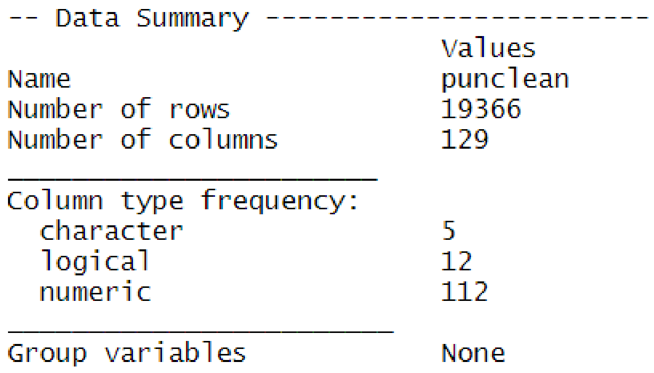
P1 Non core

H1



```
-- Data Summary ------------------------
                           Values
Name                       hunclean
Number of rows             8156
Number of columns          920
_______________________
Column type frequency:
  character                2
  logical                  193
  numeric                  725
________________________
Group variables            None
```

P1



```
-- Data Summary ------------------------
                           Values
Name                       punclean
Number of rows             19366
Number of columns          129
_______________________
Column type frequency:
  character                5
  logical                  12
  numeric                  112
________________________
Group variables            None
```

W

About the dataset

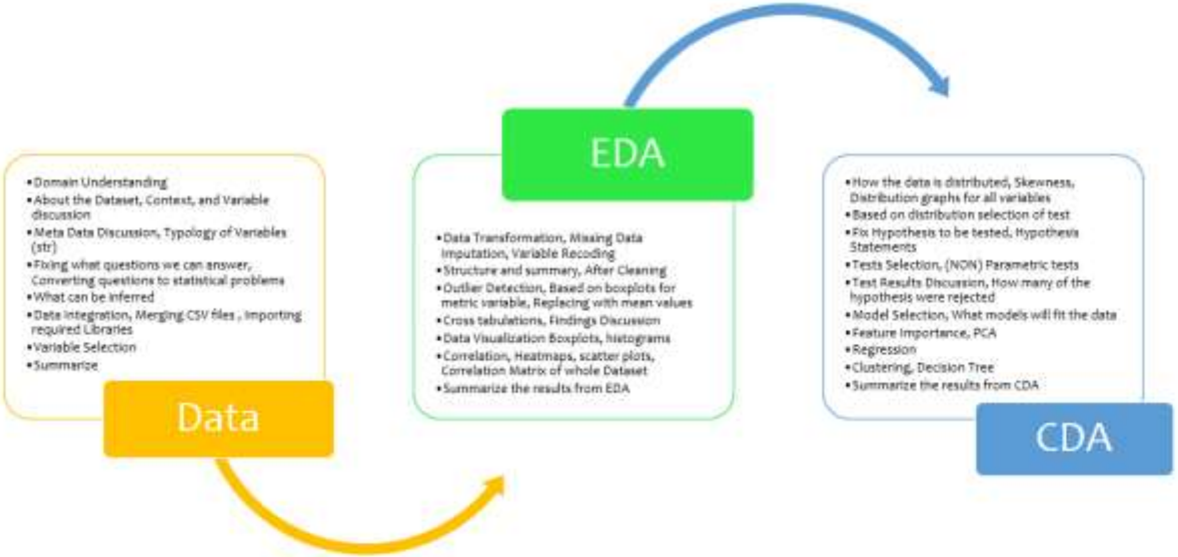
The dataset comprises several CSV files, some of which contain non-core and core variables. Among them, we focused on two CSV files, D1.csv (127 columns) and H1.csv (920 columns), which together had over 1000 columns in total. However, as many of the columns in D1 file were derived from H1 file, we narrowed our focus to approximately 20 columns in H1 file that had data about expenditure. To determine which variables to select and which statistical questions to ask, we brainstormed various possibilities based on the available columns.

Objectives of the study

- *To describe the distribution of financial and consumption-related variables among Italian households, including income, expenses, assets, debts, and savings.*
- *To identify patterns and trends in the financial behaviors and consumption habits of Italian households, such as how much they save, how much debt they have, and how they allocate their expenditures across different categories.*
- *To analyze the relationships between various financial and consumption-related variables, such as how income levels affect spending habits, how debt levels affect savings behavior, and how asset ownership relates to financial security.*
- *To identify factors that may influence household financial and consumption behaviors in Italy, such as age, education, employment status.*

Sprints

Sprint	*Session*	*Task*
Sprint 1	**Data**	Domain Understanding
		About the Dataset, Context, and Variable discussion
		Meta Data Discussion, Typology of Variables (str)
		Fixing what questions we can answer, Converting questions to statistical problems
		What can be inferred
		Data Integration, Merging CSV files , Importing required Libraries
Sprint 2		Variable Selection
		Summarize
	EDA	Data Transformation, Missing Data Imputation, Variable Recoding
		Structure and summary, After Cleaning
		Outlier Detection, Based on boxplots for metric variable, Replacing with mean values
Sprint 3		Cross tabulations, Findings Discussion
		Data Visualization Boxplots, histograms
		Correlation, Heatmaps, scatter plots, Correlation Matrix of whole Dataset
		Summarize the results from EDA
Sprint 4	**CDA**	How the data is distributed, Skewness, Distribution graphs for all variables
		Based on distribution selection of test
		Fix Hypothesis to be tested, Hypothesis Statements
		Tests Selection, (NON) Parametric tests
		Test Results Discussion, How many of the hypothesis were rejected
Sprint 5		Model Selection, What models will fit the data
		Regression
		PCA
		Clustering, Decision Tree
		Summarize the results from CDA



- Domain Understanding
- About the Dataset, Context, and Variable discussion
- Meta Data Discussion, Typology of Variables (str)
- Fixing what questions we can answer, Converting questions to statistical problems
- What can be inferred
- Data Integration, Merging CSV files , Importing required Libraries
- Variable Selection
- Summarize

Data

EDA

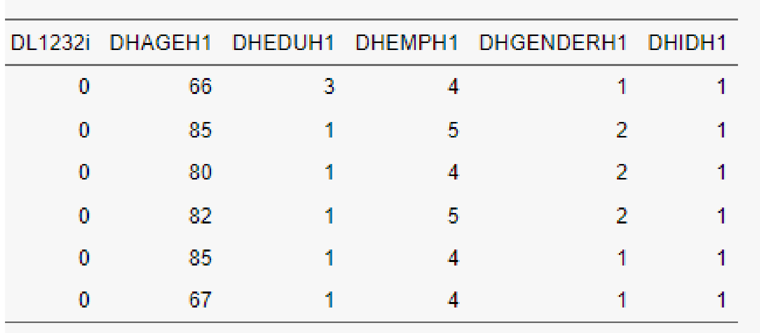
- Data Transformation, Missing Data Imputation, Variable Recoding
- Structure and summary, After Cleaning
- Outlier Detection, Based on boxplots for metric variable, Replacing with mean values
- Cross tabulations, Findings Discussion
- Data Visualization Boxplots, histograms
- Correlation, Heatmaps, scatter plots, Correlation Matrix of whole Dataset
- Summarize the results from EDA

- How the data is distributed, Skewness, Distribution graphs for all variables
- Based on distribution selection of test
- Fix Hypothesis to be tested, Hypothesis Statements
- Tests Selection, (NON) Parametric tests
- Test Results Discussion, How many of the hypothesis were rejected
- Model Selection, What models will fit the data
- Feature Importance, PCA
- Regression
- Clustering, Decision Tree
- Summarize the results from CDA

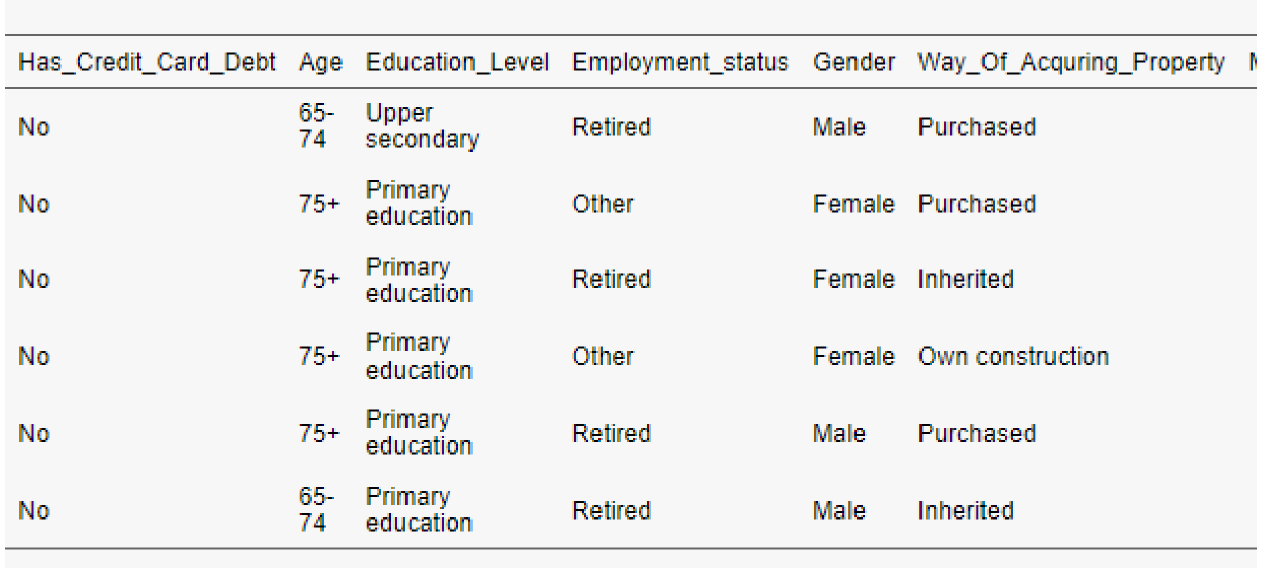
CDA

Workflow

Structure and Summary of the Dataset



DL1232i	DHAGEH1	DHEDUH1	DHEMPH1	DHGENDERH1	DHIDH1
0	66	3	4	1	1
0	85	1	5	2	1
0	80	1	4	2	1
0	82	1	5	2	1
0	85	1	4	1	1
0	67	1	4	1	1



Has_Credit_Card_Debt	Age	Education_Level	Employment_status	Gender	Way_Of_Acquring_Property	M
No	65-74	Upper secondary	Retired	Male	Purchased	
No	75+	Primary education	Other	Female	Purchased	
No	75+	Primary education	Retired	Female	Inherited	
No	75+	Primary education	Other	Female	Own construction	
No	75+	Primary education	Retired	Male	Purchased	
No	65-74	Primary education	Retired	Male	Inherited	

8156 Rows

62 Columns

31 Categorial Variables

30 Metric Variables

1 Factor Variable

Metric

- Total_Gross_Income
- AMount_Spent_on_Utilities
- Amount_Spent_on_Consumer_Goods_Services
- Employee_Income
- Self_Employment_income
- Financial_assets_Income
- Pension_Income
- Credit_Card_Debt
- Value_of_Saving_Accounts
- Value_of_Self_employment_Businesses
- Amount_spent_on_Food_at_Home
- Income_From_Other_Sources

Categorical

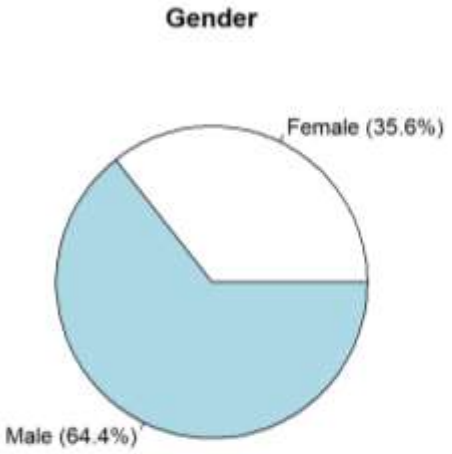
- Gender
- Education Level
- Investment Attitude
- Age
- Housing Status
- Employment Status
- Way of Acquiring Property

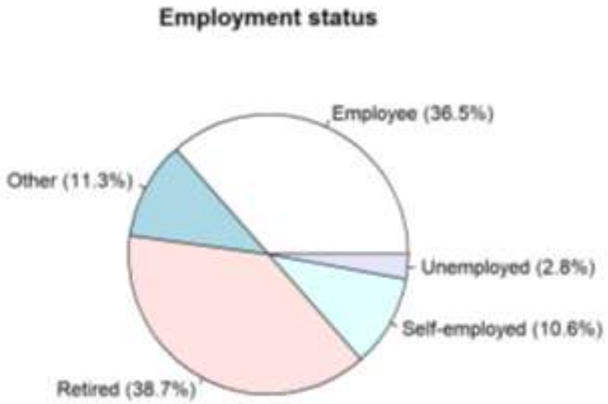
Binary Variables

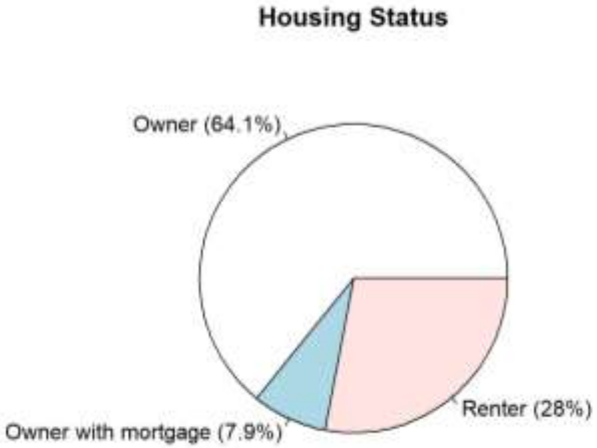
- Has_Real_Assets
- Has_Financial_Assets
- Has_Vehicles
- Has_Valuables
- Has_Real_Estate_Wealth
- Has_Deposits
- Has_Mutual_Funds
- Has_Bonds
- Has_Shares
- Has_Debt
- Has_Employee_Income
- Has_Self_Employee_Income
- Has_Financial_assets_Income
- Has_Income_From_Pensions
- Has_Income_From_Other_Sources
- Has_Credit_Card_Debt
- Has_Rental_Income
- Household_Has_a_Credit_Card
- Has_Private_Loans
- Has_Applied_for_Loan_Credit

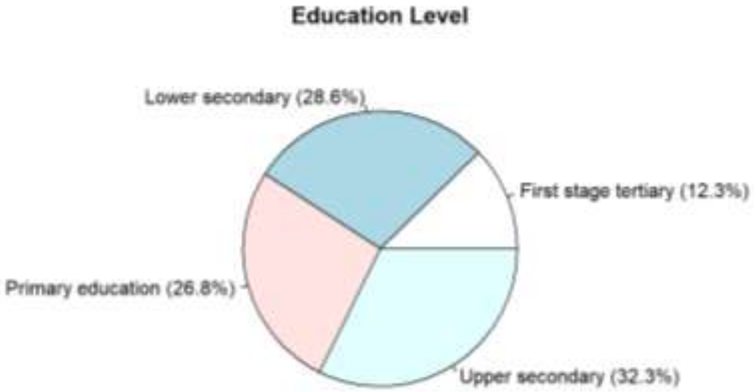
Data Visualization

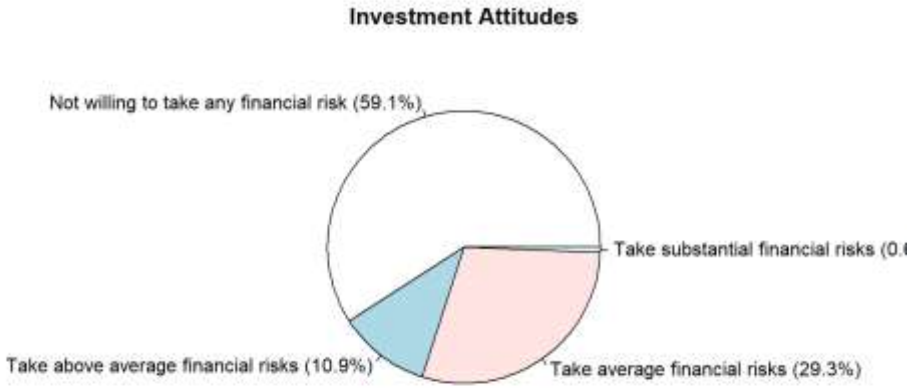
Pie Charts


Gender
Female (35.6%)
Male (64.4%)

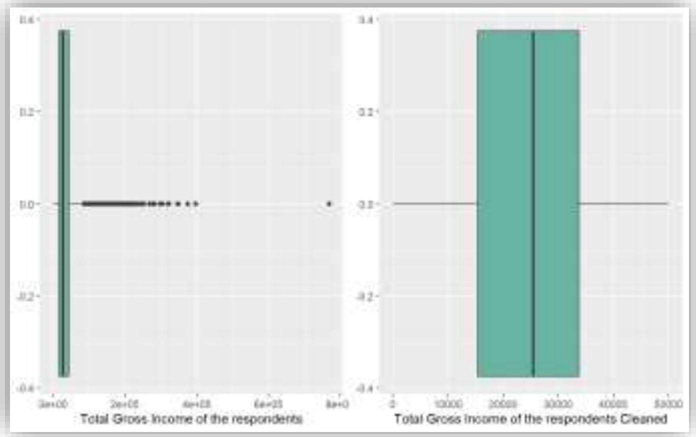

Employment status
Employee (36.5%)
Other (11.3%)
Unemployed (2.8%)
Self-employed (10.6%)
Retired (38.7%)

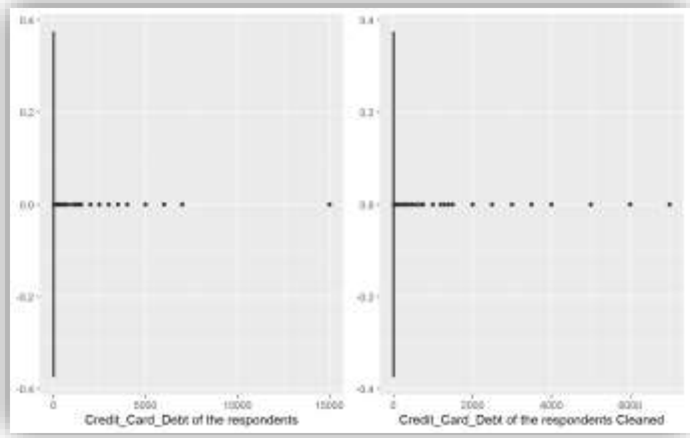

Housing Status
Owner (64.1%)
Renter (28%)
Owner with mortgage (7.9%)

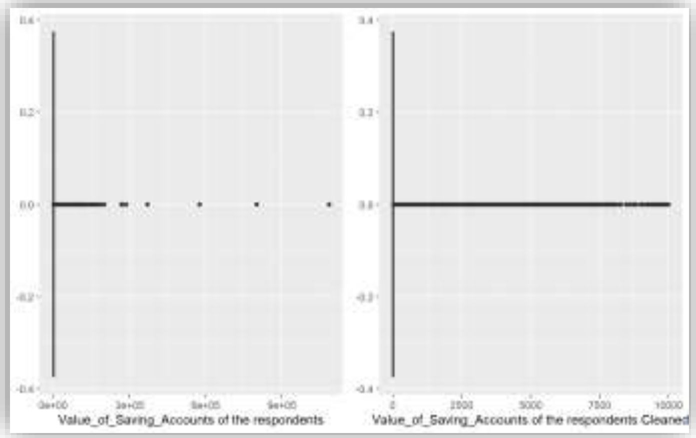

Education Level
Lower secondary (28.6%)
First stage tertiary (12.3%)
Primary education (26.8%)
Upper secondary (32.3%)


Investment Attitudes
Not willing to take any financial risk (59.1%)
Take substantial financial risks (0.
Take above average financial risks (10.9%)
Take average financial risks (29.3%)

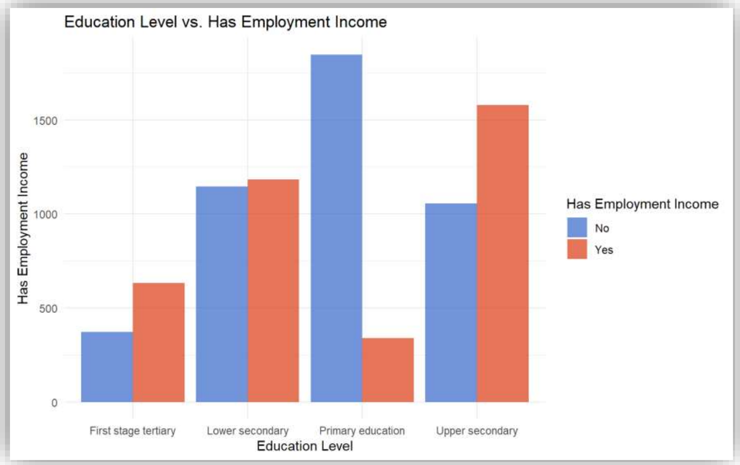
Outlier Detection using Box plots

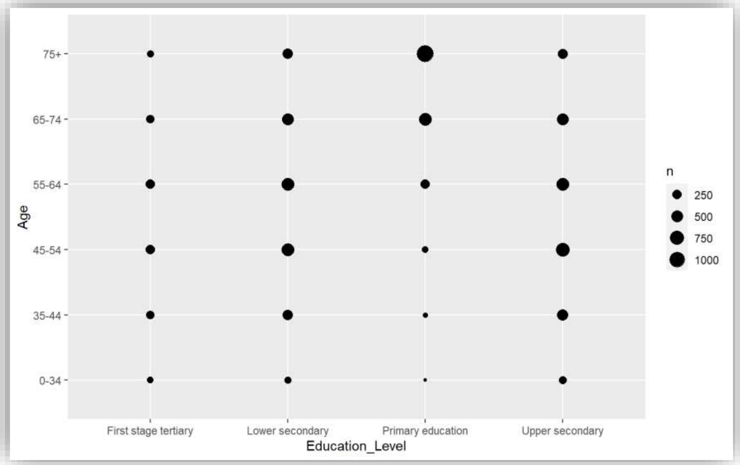


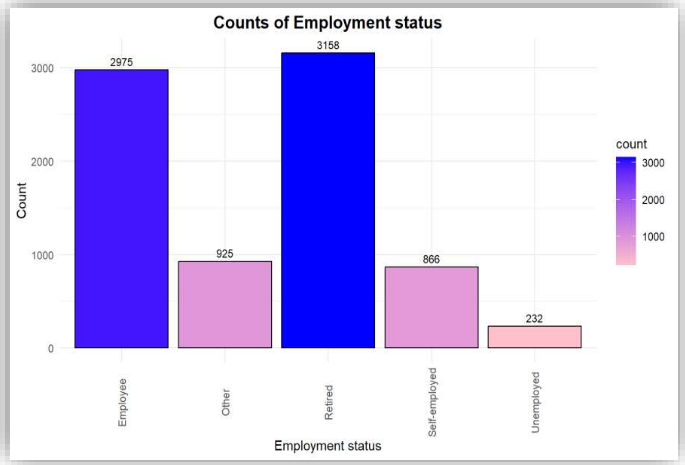


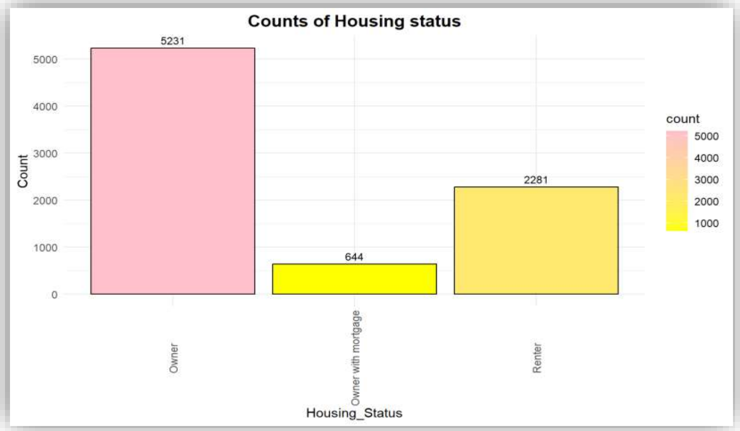


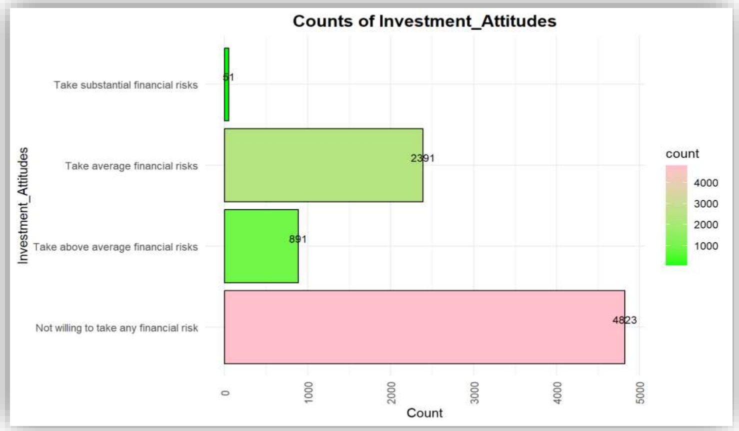
Graphs

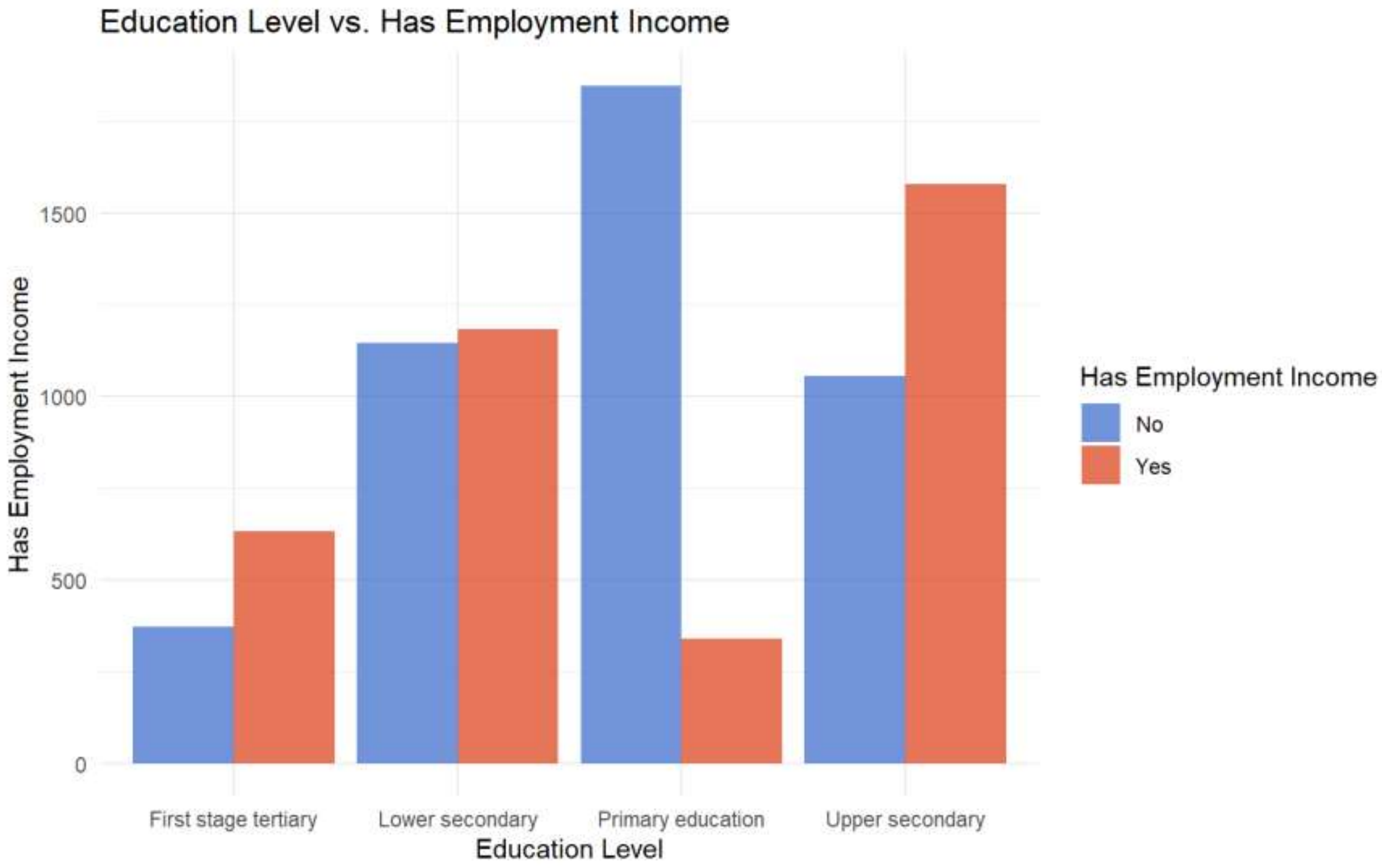

Education Level vs. Has Employment Income
Has Employment Income
1500
1000
500
0
First stage tertiary
Lower secondary
Primary education
Upper secondary
Education Level
Has Employment Income
No
Yes

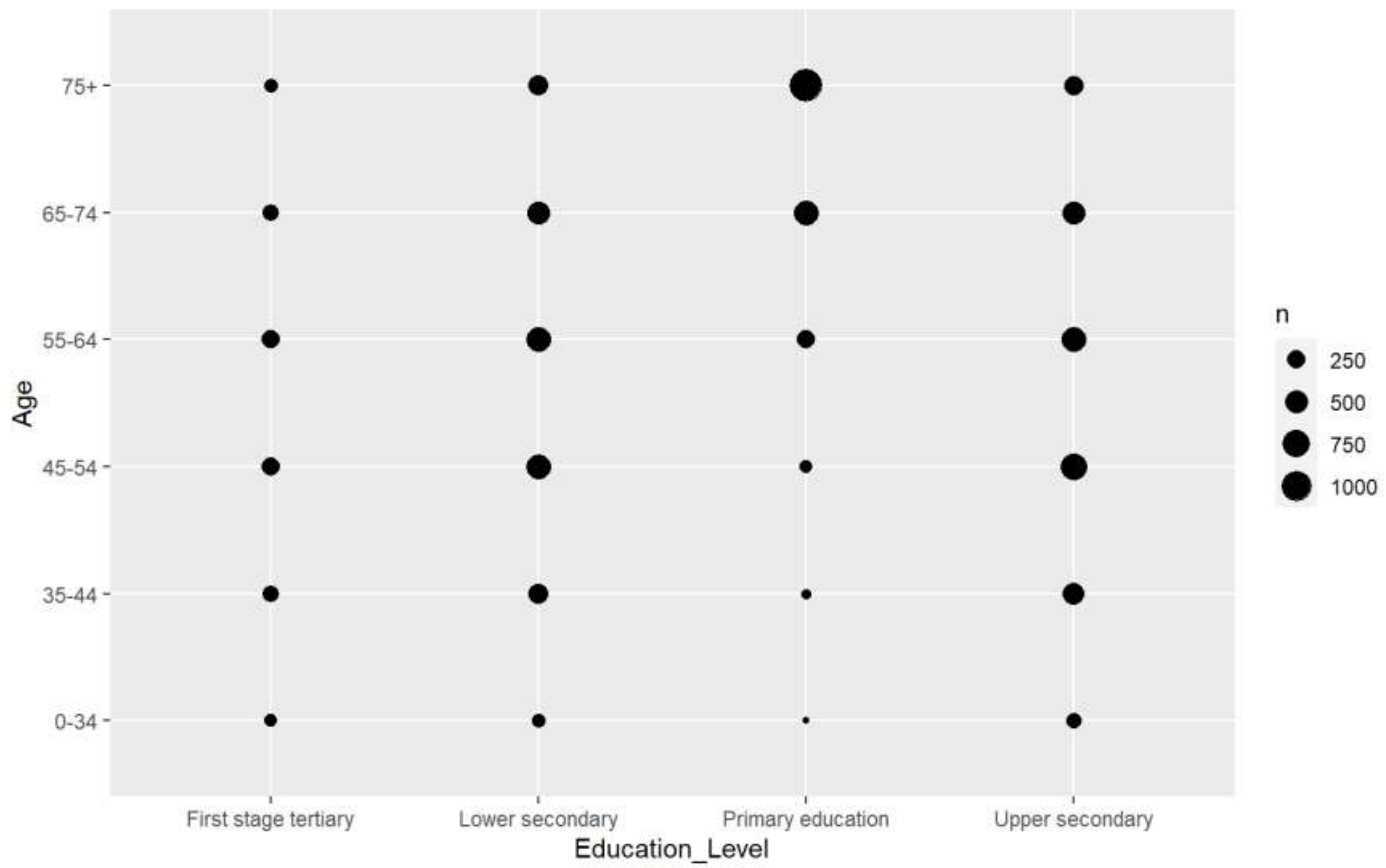

Age
75+
65-74
55-64
45-54
35-44
0-34
First stage tertiary
Lower secondary
Primary education
Upper secondary
Education_Level
n
250
500
750
1000


Counts of Employment status
Count
3000
2000
1000
0
2975
925
3158
866
232
Employee
Other
Retired
Self-employed
Unemployed
Employment status
count
3000
2000
1000

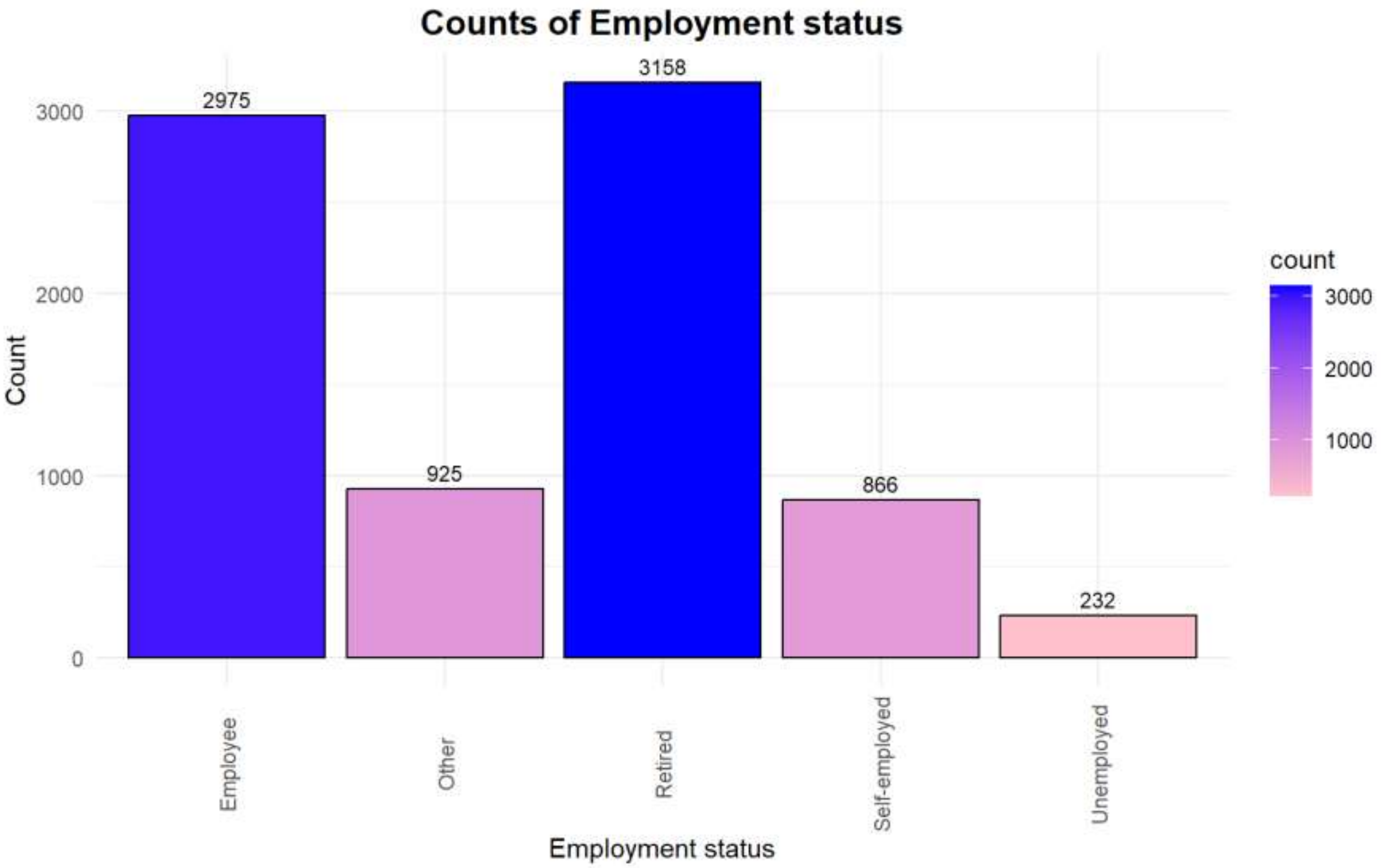

Counts of Housing status
Count
5000
4000
3000
2000
1000
0
5231
644
2281
Owner
Owner with mortgage
Renter
Housing_Status
count
5000
4000
3000
2000
1000


Counts of Investment_Attitudes
Investment_Attitudes
Take substantial financial risks
Take average financial risks
Take above average financial risks
Not willing to take any financial risk
51
2391
891
4823
0
1000
2000
3000
4000
5000
Count
count
4000
3000
2000
1000

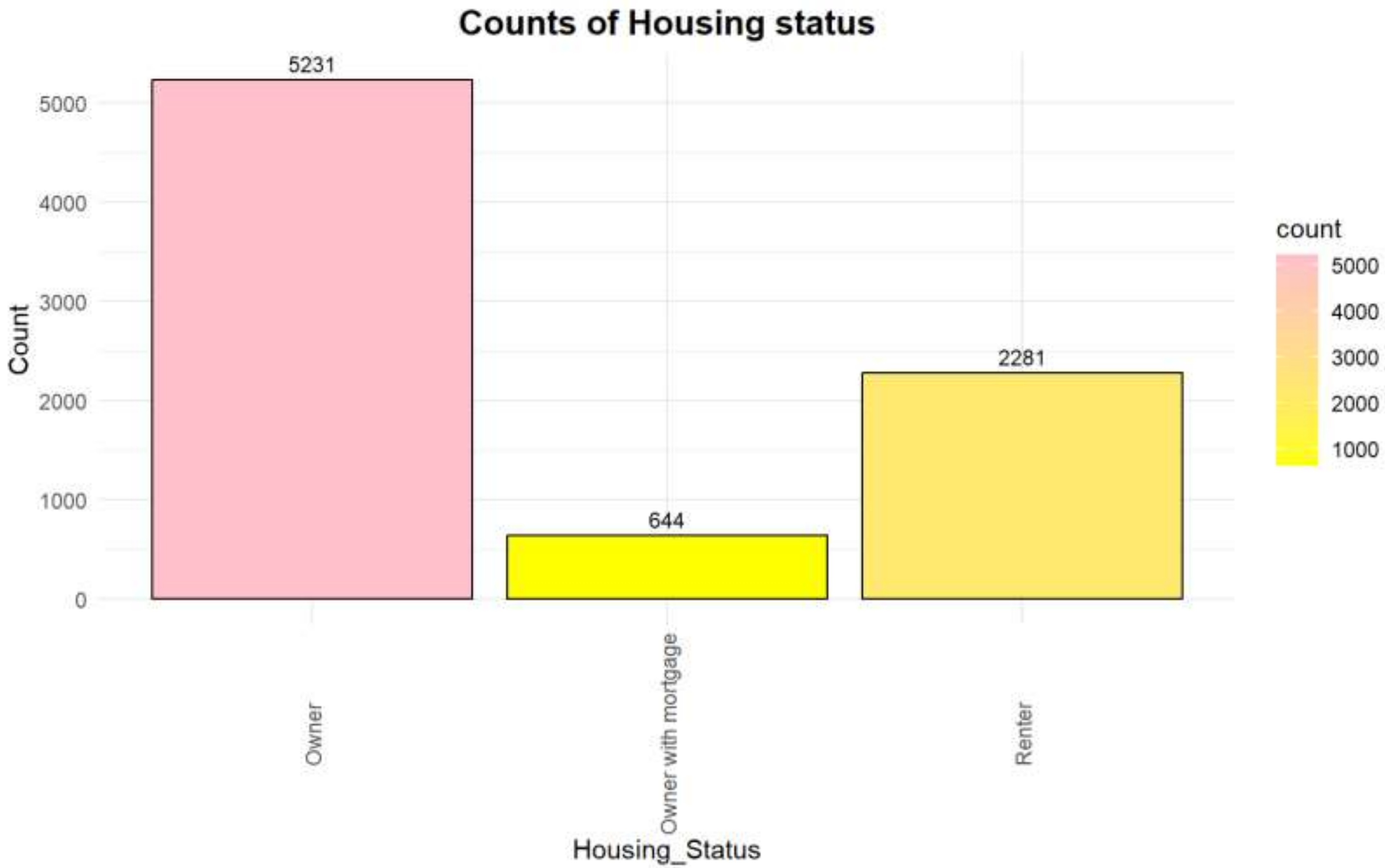

Education Level vs. Has Employment Income
Has Employment Income
1500
1000
500
0
First stage tertiary
Lower secondary
Primary education
Upper secondary
Education Level
Has Employment Income
No
Yes

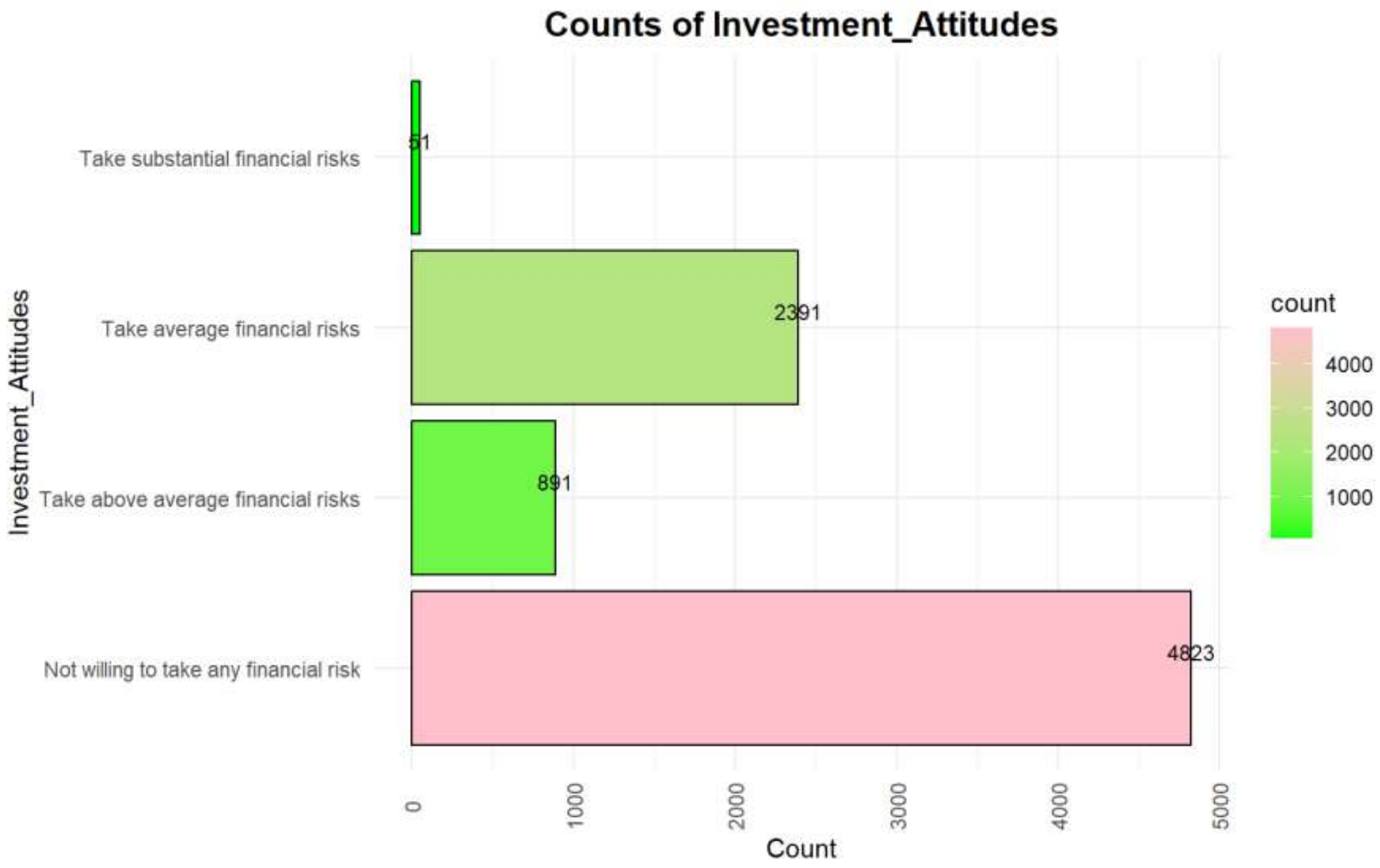


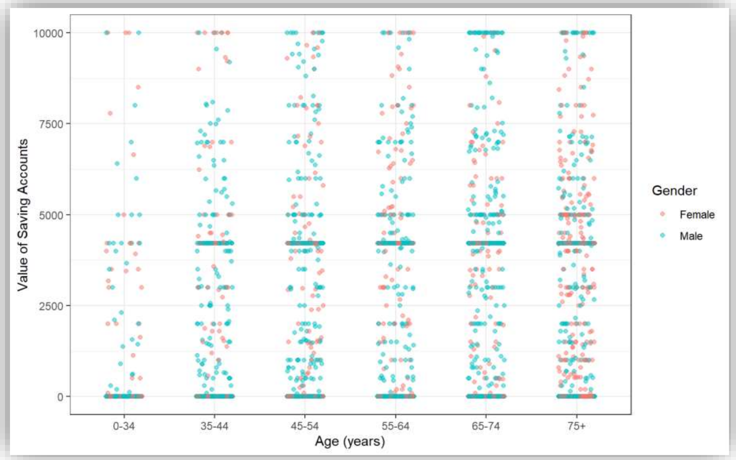
75+
65-74
55-64
45-54
35-44
0-34
Age
First stage tertiary
Lower secondary
Primary education
Upper secondary
Education_Level
n
250
500
750
1000

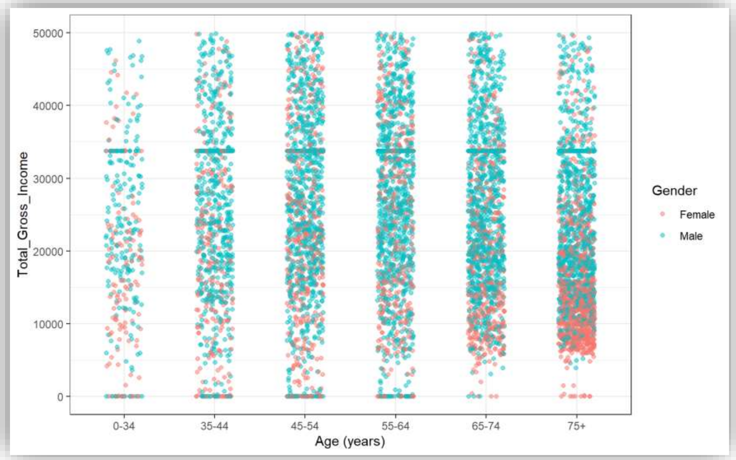


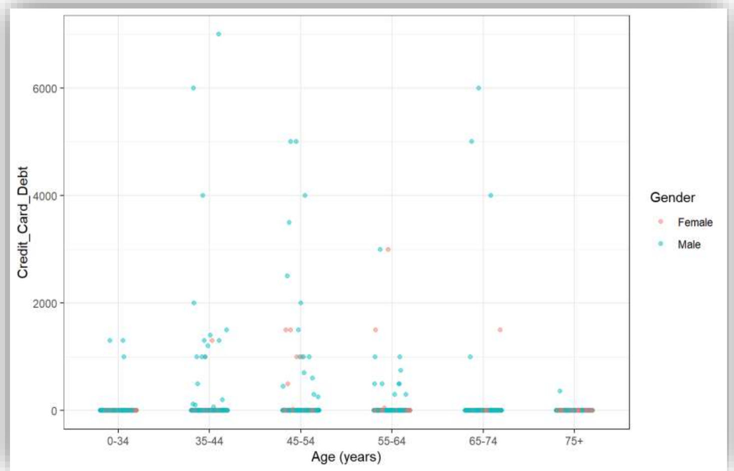
Counts of Employment status

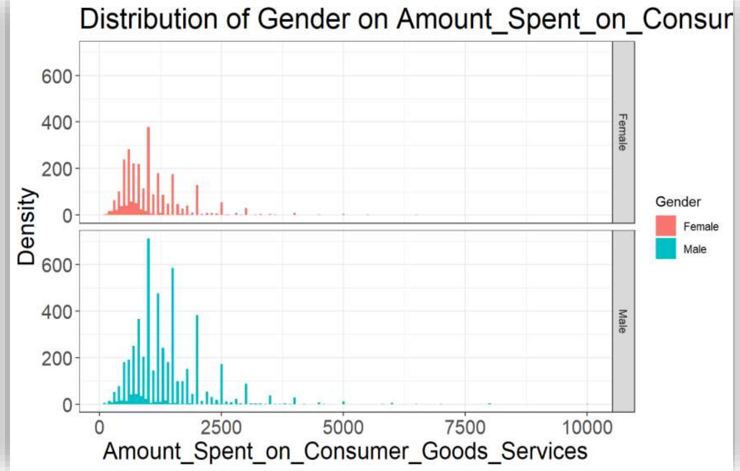

Counts of Housing status
5231
644
2281
5000
4000
3000
2000
1000
0
Count
Owner
Owner with mortgage
Renter
Housing_Status
count
5000
4000
3000
2000
1000

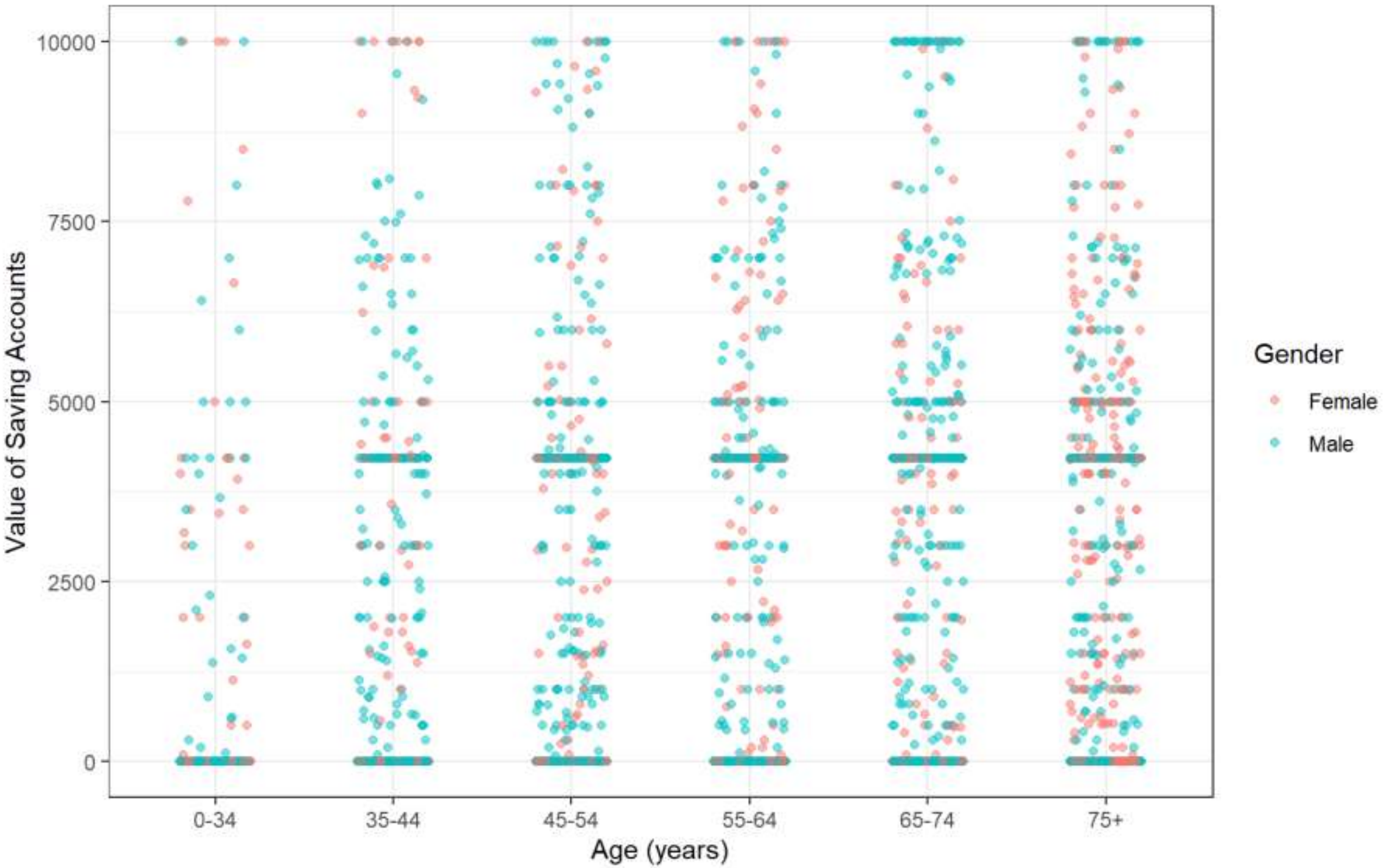

Counts of Investment_Attitudes
Investment_Attitudes
Take substantial financial risks
Take average financial risks
Take above average financial risks
Not willing to take any financial risk
51
2391
891
4823
0
1000
2000
3000
4000
5000
Count
count
4000
3000
2000
1000

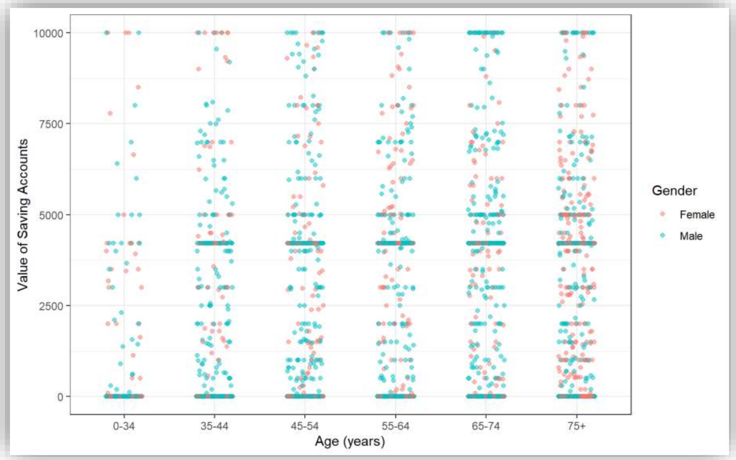

Value of Saving Accounts
10000
7500
5000
2500
0
0-34
35-44
45-54
55-64
65-74
75+
Age (years)
Gender
Female
Male

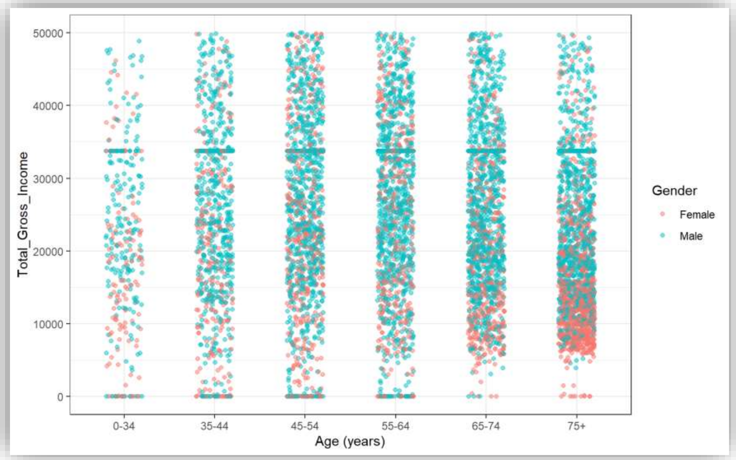

Total_Gross_Income
50000
40000
30000
20000
10000
0
0-34
35-44
45-54
55-64
65-74
75+
Age (years)
Gender
Female
Male

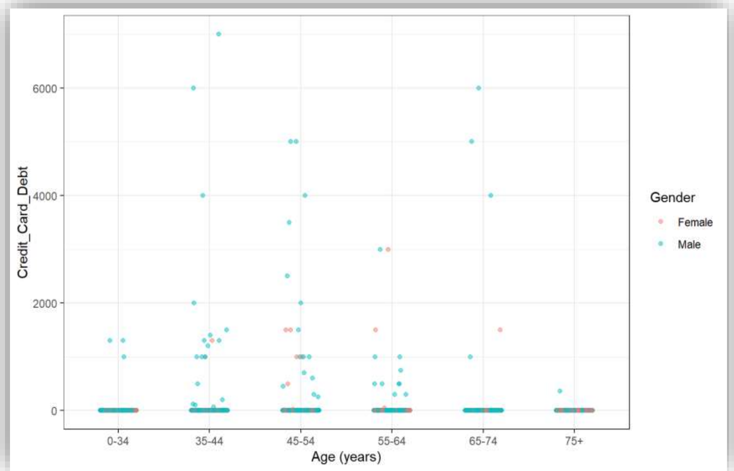

Credit_Card_Debt
6000
4000
2000
0
0-34
35-44
45-54
55-64
65-74
75+
Age (years)
Gender
Female
Male

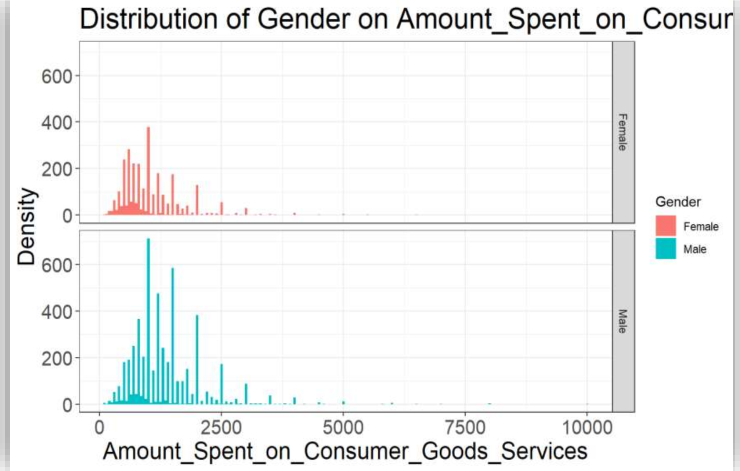

Distribution of Gender on Amount_Spent_on_Consur
Density
600
400
200
0
600
400
200
0
Female
Male
0
2500
5000
7500
10000
Amount_Spent_on_Consumer_Goods_Services
Gender
Female
Male

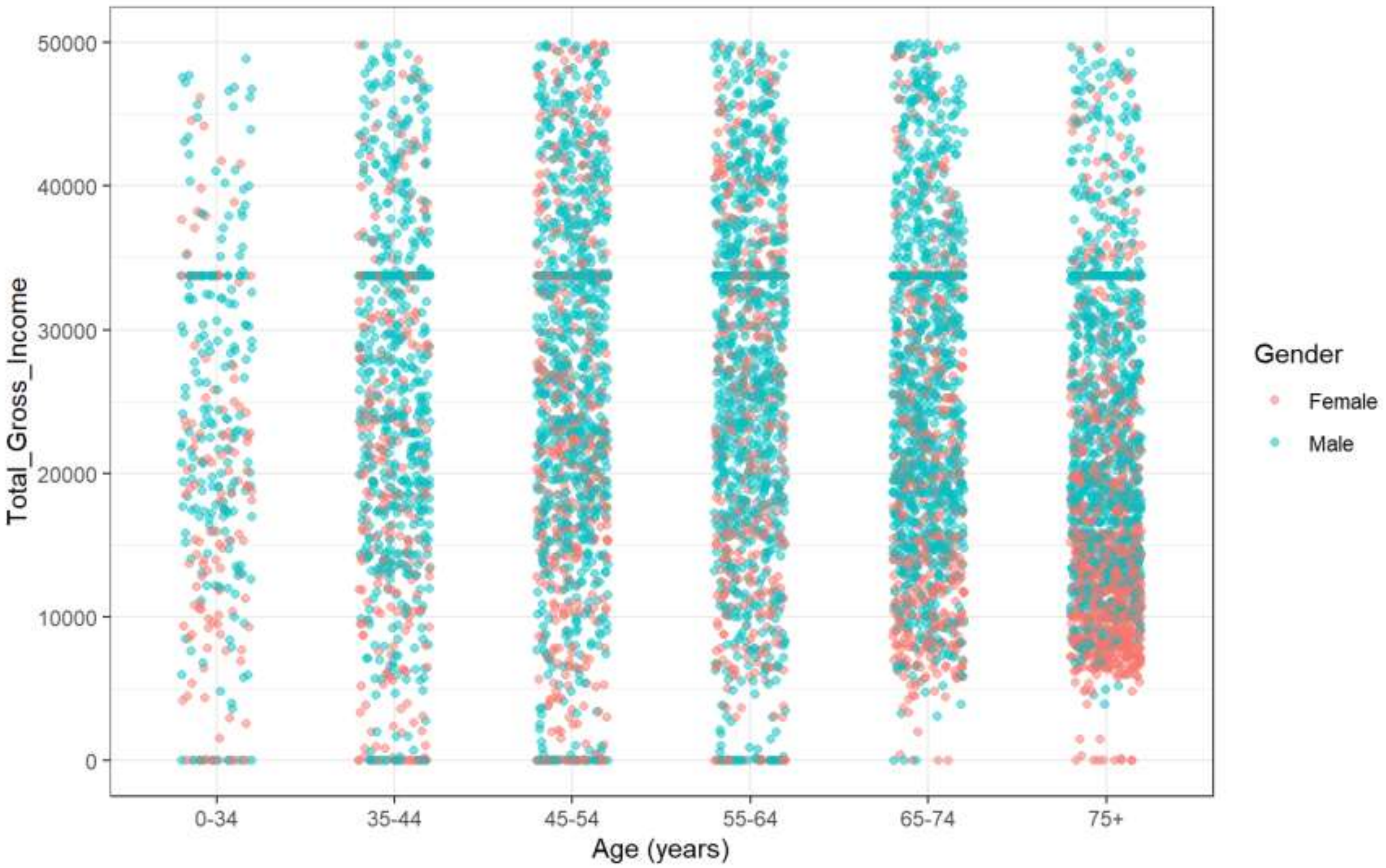

10000
7500
5000
2500
0
Value of Saving Accounts
0-34
35-44
45-54
55-64
65-74
75+
Age (years)
Gender
Female
Male


Value of Saving Accounts
10000
7500
5000
2500
0
0-34
35-44
45-54
55-64
65-74
75+
Age (years)
Gender
Female
Male

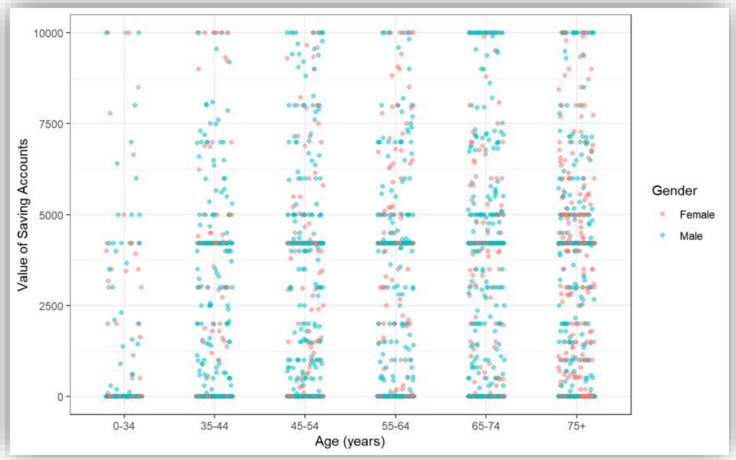

Total_Gross_Income
50000
40000
30000
20000
10000
0
0-34
35-44
45-54
55-64
65-74
75+
Age (years)
Gender
Female
Male

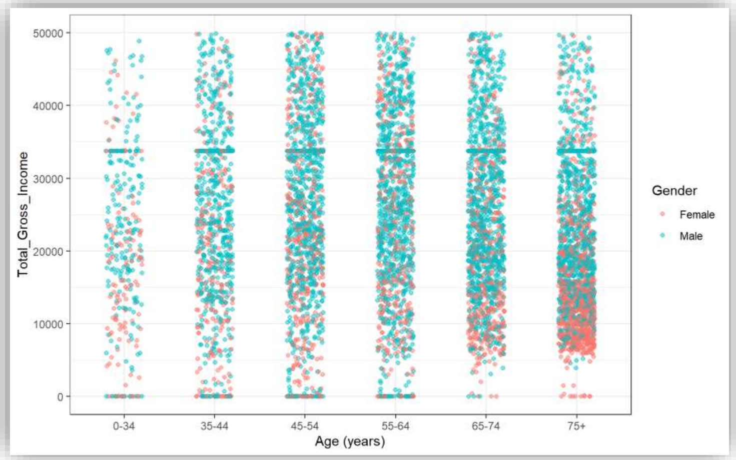

Credit_Card_Debt
6000
4000
2000
0
0-34
35-44
45-54
55-64
65-74
75+
Age (years)
Gender
Female
Male

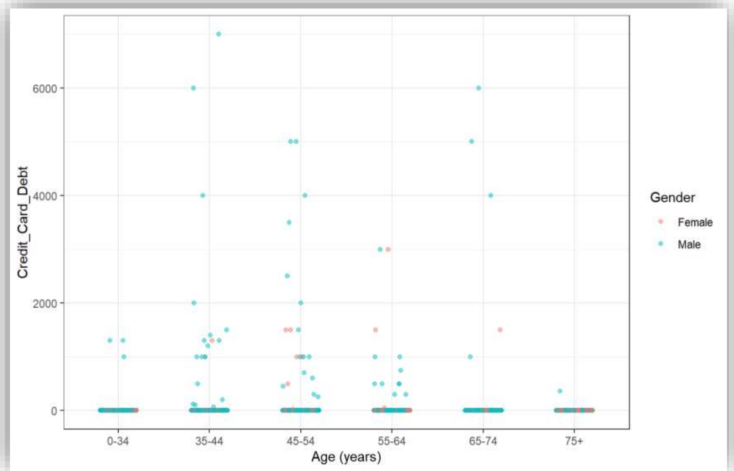

Distribution of Gender on Amount_Spent_on_Consur
Density
600
400
200
0
600
400
200
0
Female
Male
0
2500
5000
7500
10000
Amount_Spent_on_Consumer_Goods_Services
Gender
Female
Male

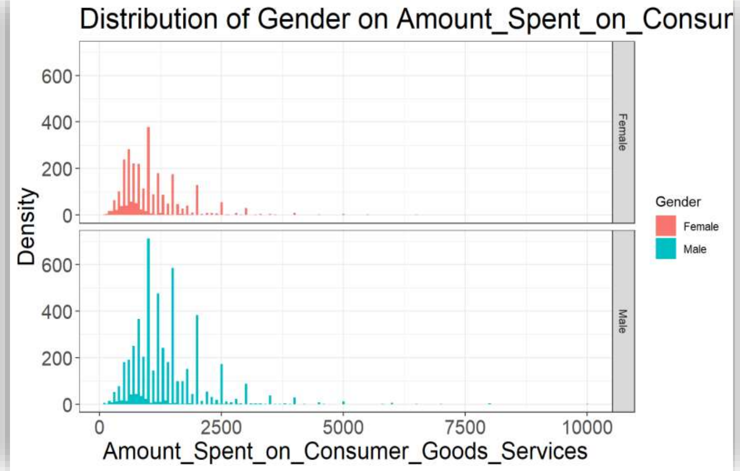


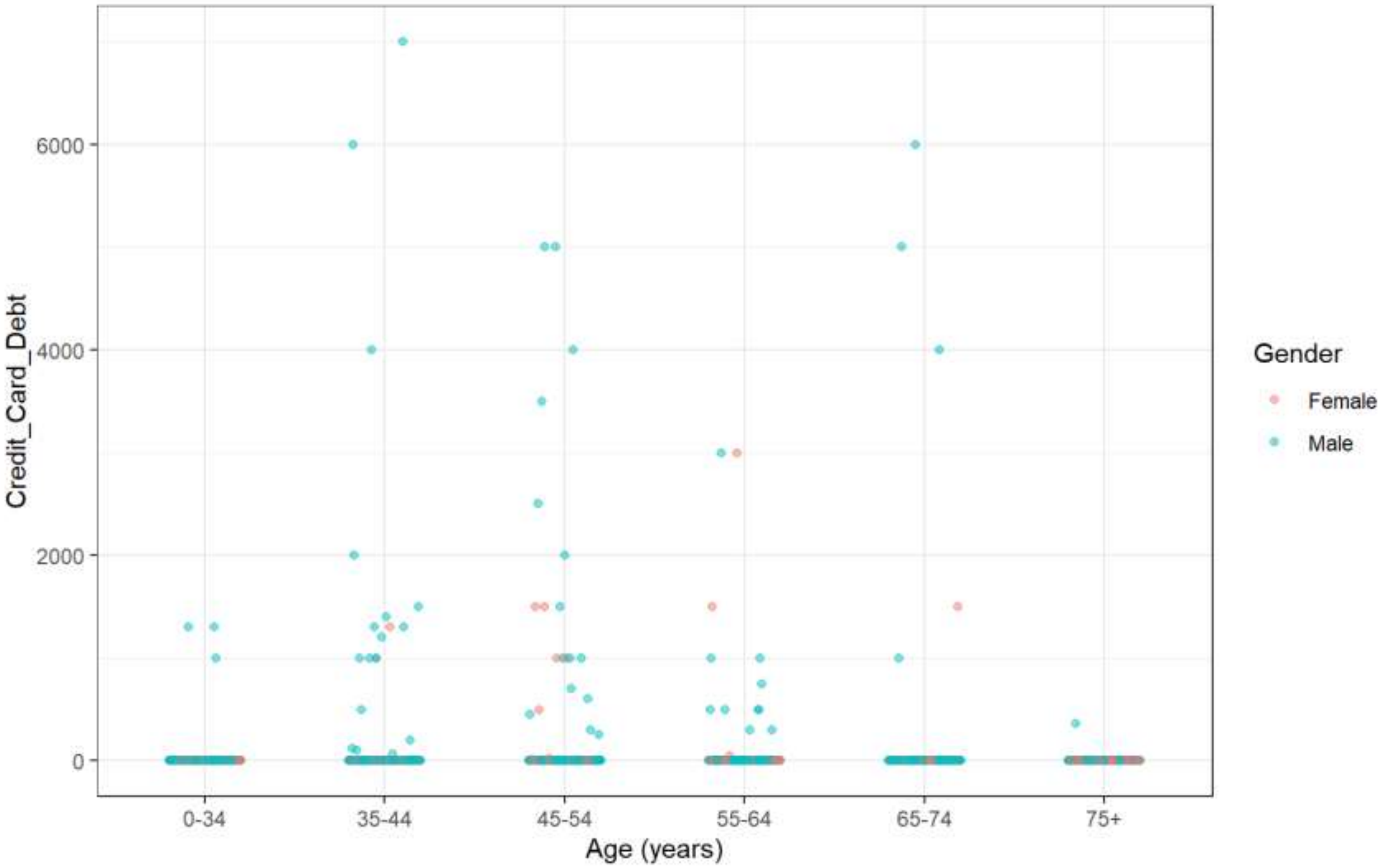
50000
40000
30000
20000
10000
0
Total_Gross_Income
0-34
35-44
45-54
55-64
65-74
75+
Age (years)
Gender
Female
Male

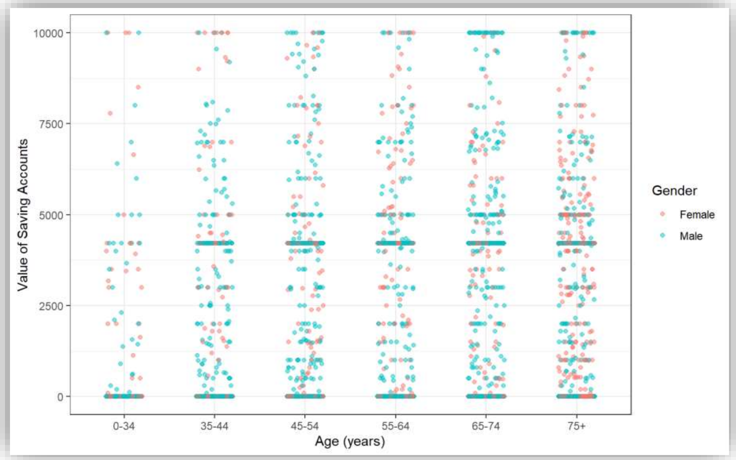

Value of Saving Accounts
10000
7500
5000
2500
0
0-34
35-44
45-54
55-64
65-74
75+
Age (years)
Gender
Female
Male

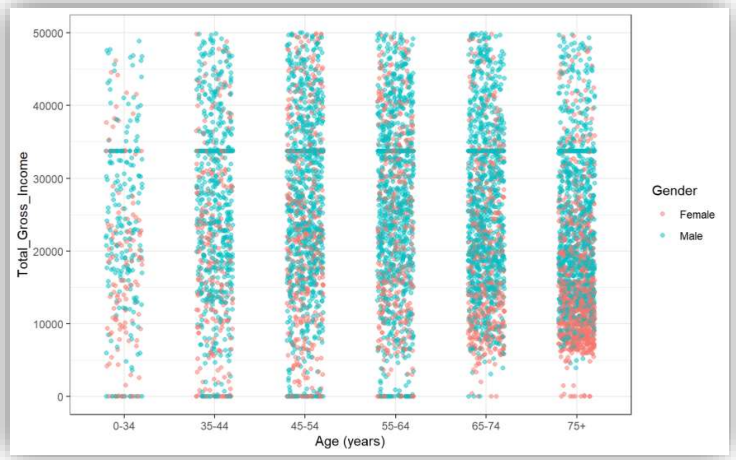

Total_Gross_Income
50000
40000
30000
20000
10000
0
0-34
35-44
45-54
55-64
65-74
75+
Age (years)
Gender
Female
Male

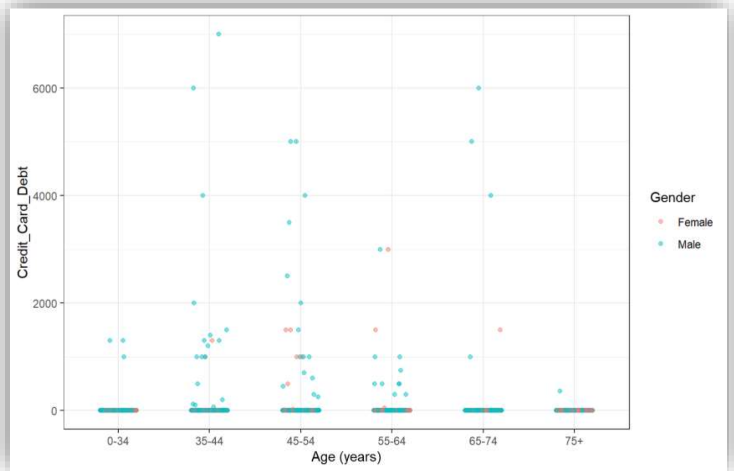

Credit_Card_Debt
6000
4000
2000
0
0-34
35-44
45-54
55-64
65-74
75+
Age (years)
Gender
Female
Male

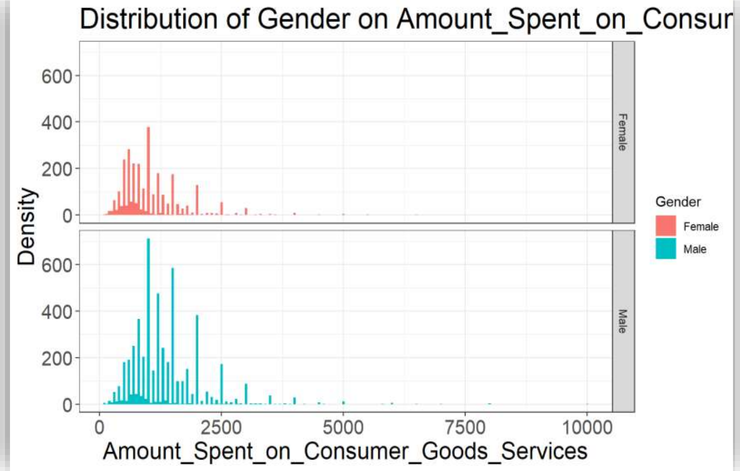

Distribution of Gender on Amount_Spent_on_Consur
Density
600
400
200
0
Female
Male
0
2500
5000
7500
10000
Amount_Spent_on_Consumer_Goods_Services
Gender
Female
Male

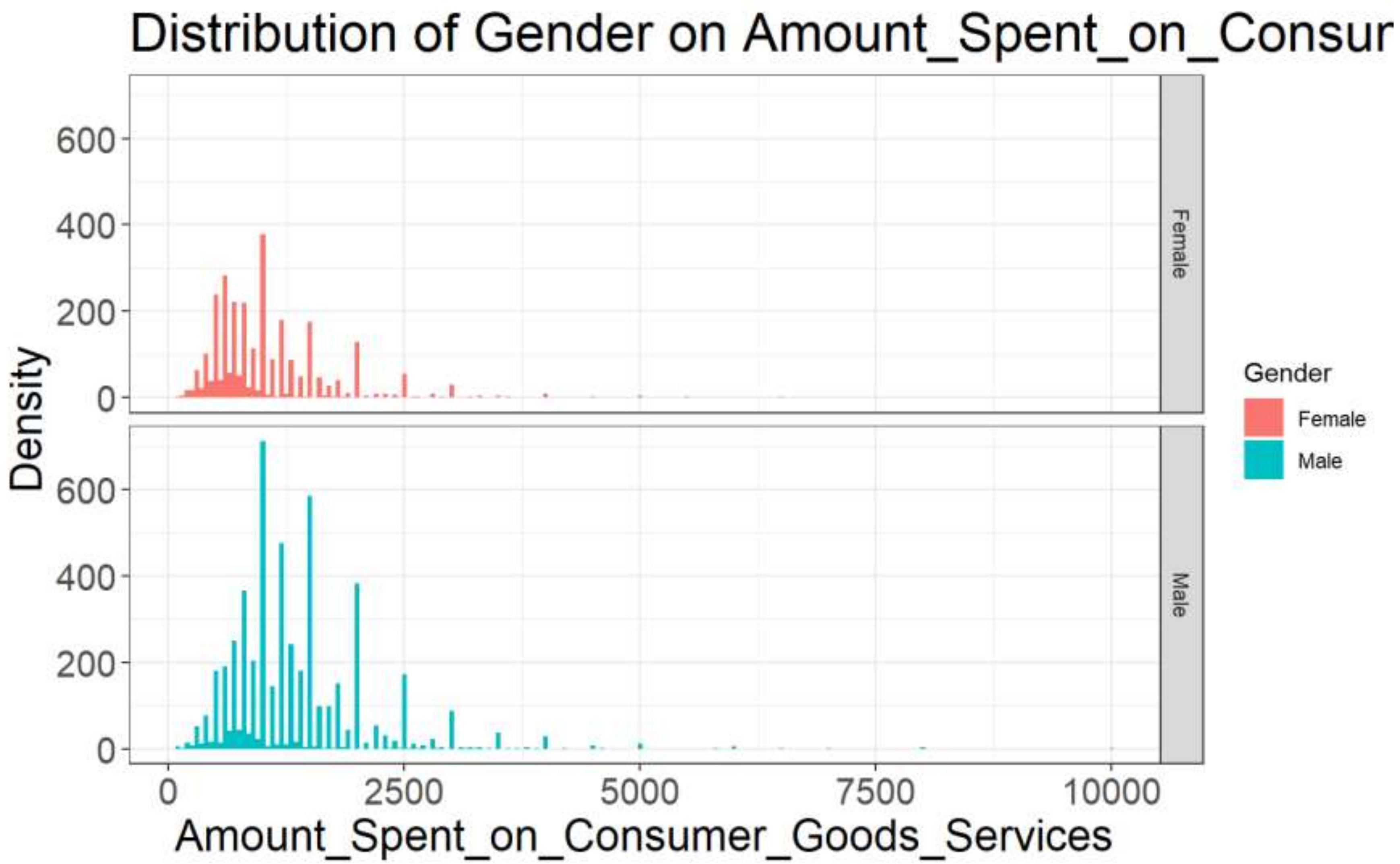

Credit_Card_Debt
6000
4000
2000
0
0-34
35-44
45-54
55-64
65-74
75+
Age (years)
Gender
Female
Male

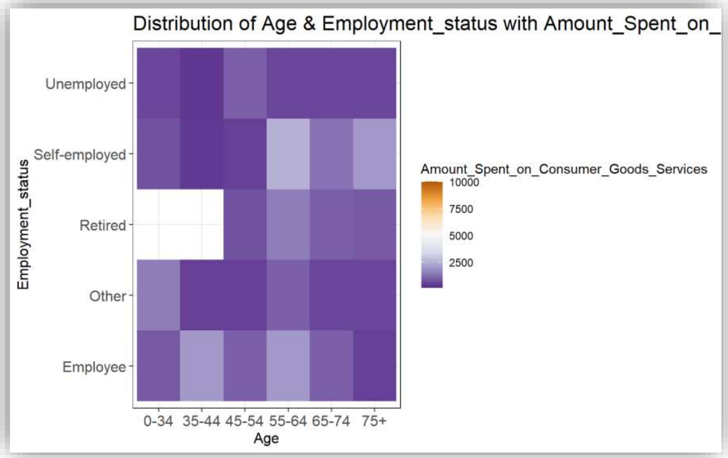

Value of Saving Accounts
10000
7500
5000
2500
0
0-34
35-44
45-54
55-64
65-74
75+
Age (years)
Gender
Female
Male

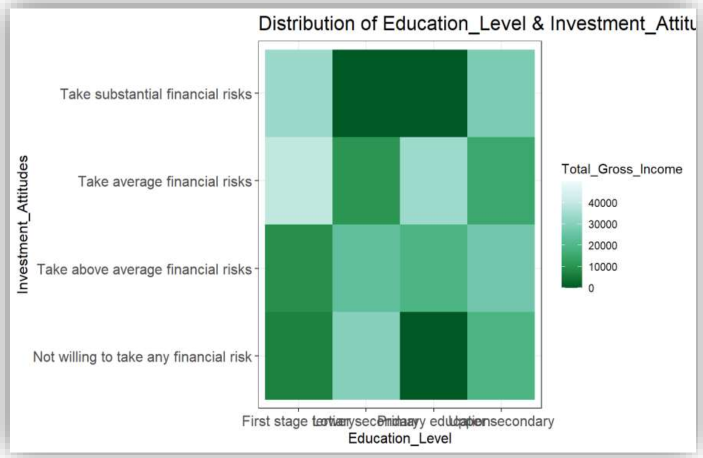

Total_Gross_Income
50000
40000
30000
20000
10000
0
0-34
35-44
45-54
55-64
65-74
75+
Age (years)
Gender
Female
Male

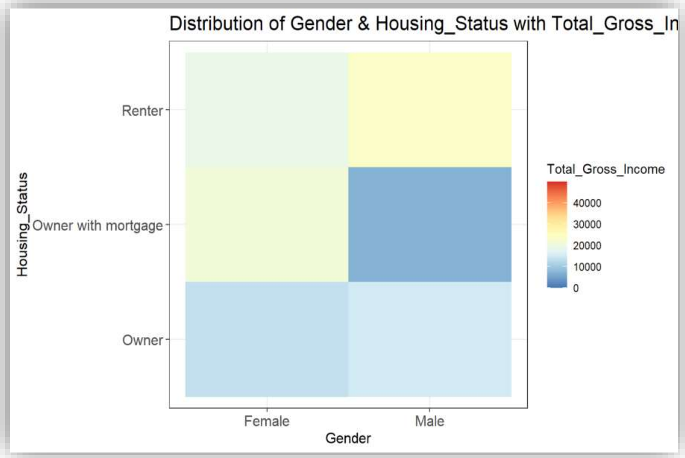

Credit_Card_Debt
6000
4000
2000
0
0-34
35-44
45-54
55-64
65-74
75+
Age (years)
Gender
Female
Male

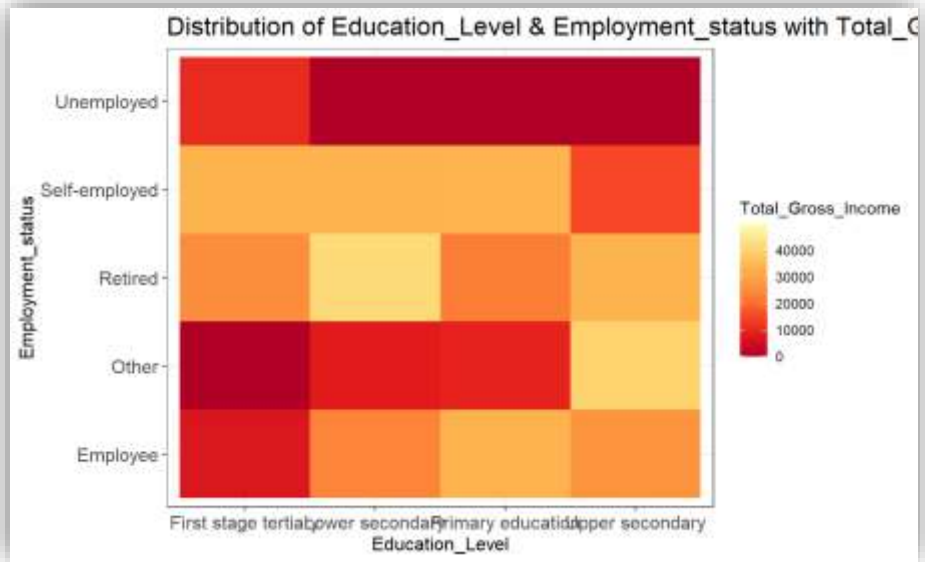

Distribution of Gender on Amount_Spent_on_Consur
Density
600
400
200
0
600
400
200
0
Female
Male
0
2500
5000
7500
10000
Amount_Spent_on_Consumer_Goods_Services
Gender
Female
Male

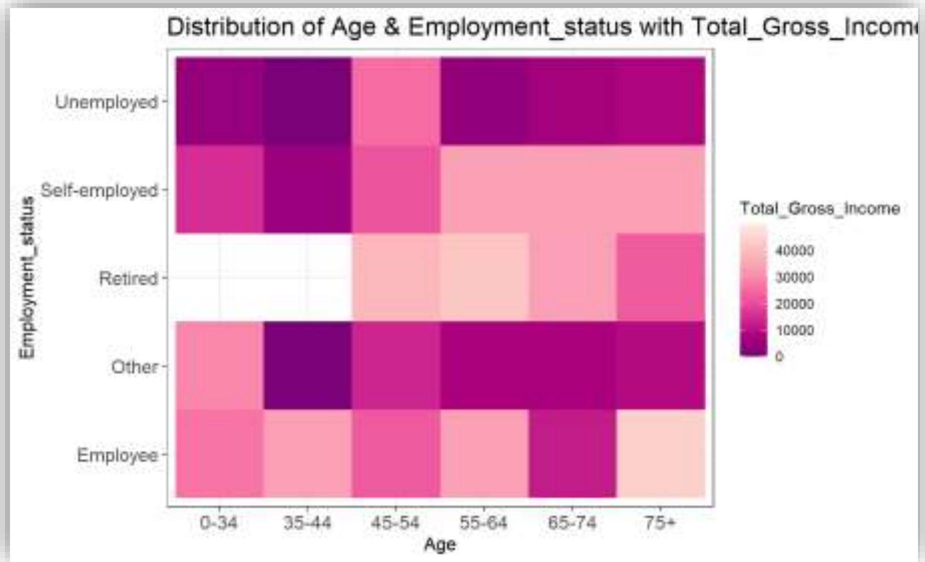

Distribution of Gender on Amount_Spent_on_Consur
Density
600
400
200
0
600
400
200
0
Female
Male
0
2500
5000
7500
10000
Amount_Spent_on_Consumer_Goods_Services
Gender
Female
Male

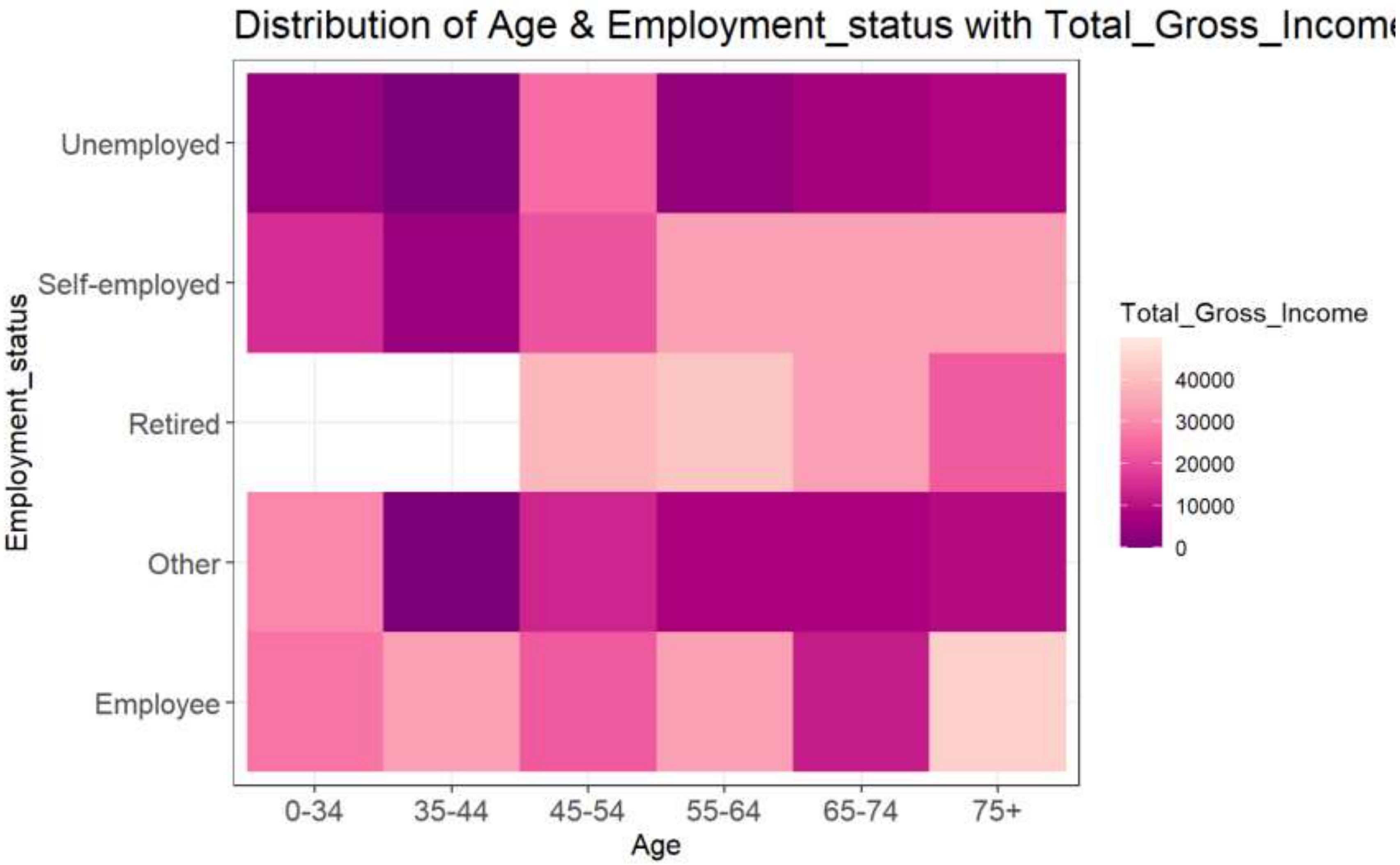

Distribution of Age & Employment_status with Amount_Spent_on_
Employment_status
Unemployed
Self-employed
Retired
Other
Employee
0-34
35-44
45-54
55-64
65-74
75+
Age
Amount_Spent_on_Consumer_Goods_Services
10000
7500
5000
2500

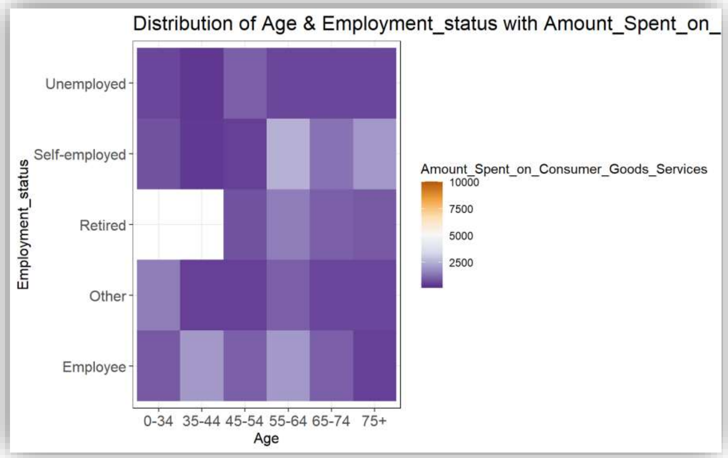

Distribution of Education_Level & Investment_Attitu
Investment_Attitudes
Take substantial financial risks
Take average financial risks
Take above average financial risks
Not willing to take any financial risk
First stage tertiary
Lower secondary
Primary education
Upper secondary
Education_Level
Total_Gross_Income
40000
30000
20000
10000
0

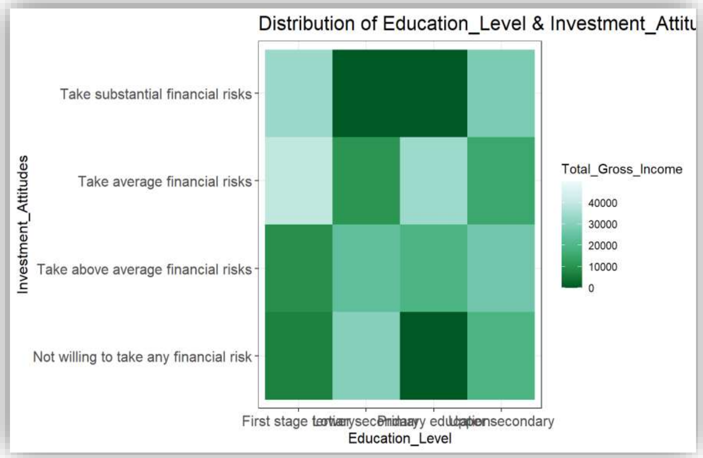

Distribution of Gender & Housing_Status with Total_Gross_In
Housing_Status
Renter
Owner with mortgage
Owner
Female
Male
Gender
Total_Gross_Income
40000
30000
20000
10000
0

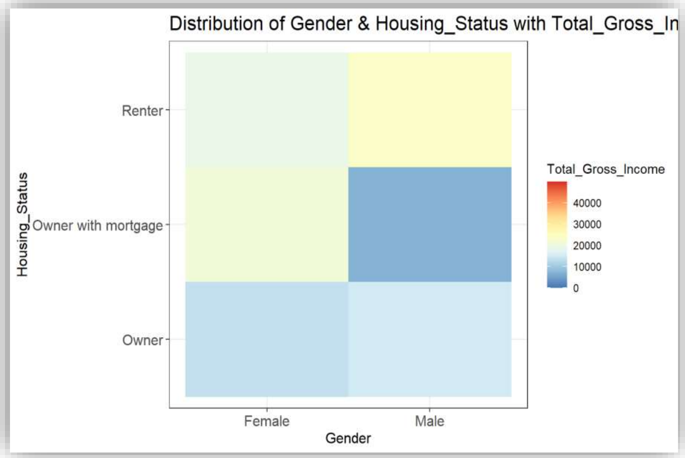

Distribution of Education_Level & Employment_status with Total_G
Employment_status
Unemployed
Self-employed
Retired
Other
Employee
First stage tertiary
Lower secondary
Primary education
Upper secondary
Education_Level
Total_Gross_Income
40000
30000
20000
10000
0

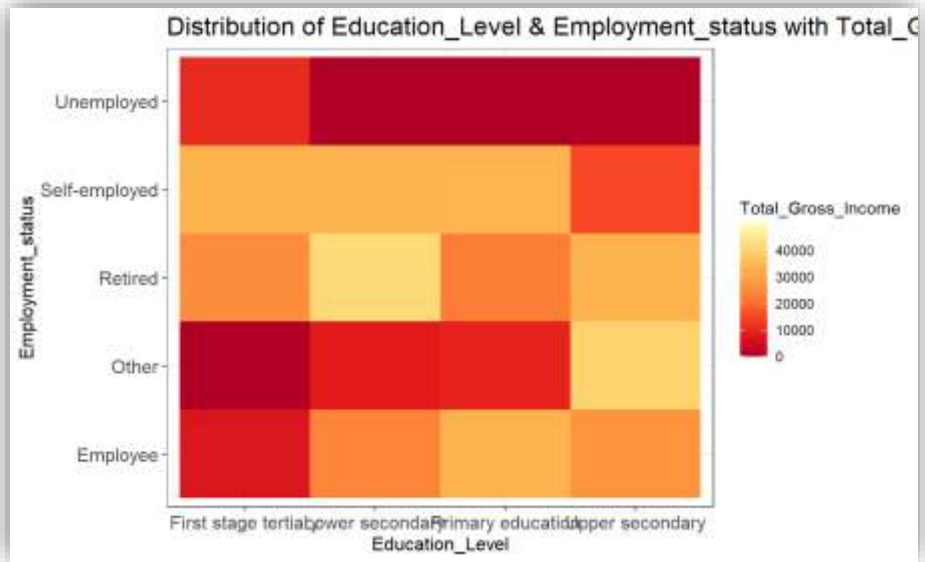

Distribution of Age & Employment_status with Total_Gross_Income
Employment_status
Unemployed
Self-employed
Retired
Other
Employee
0-34
35-44
45-54
55-64
65-74
75+
Age
Total_Gross_Income
40000
30000
20000
10000
0

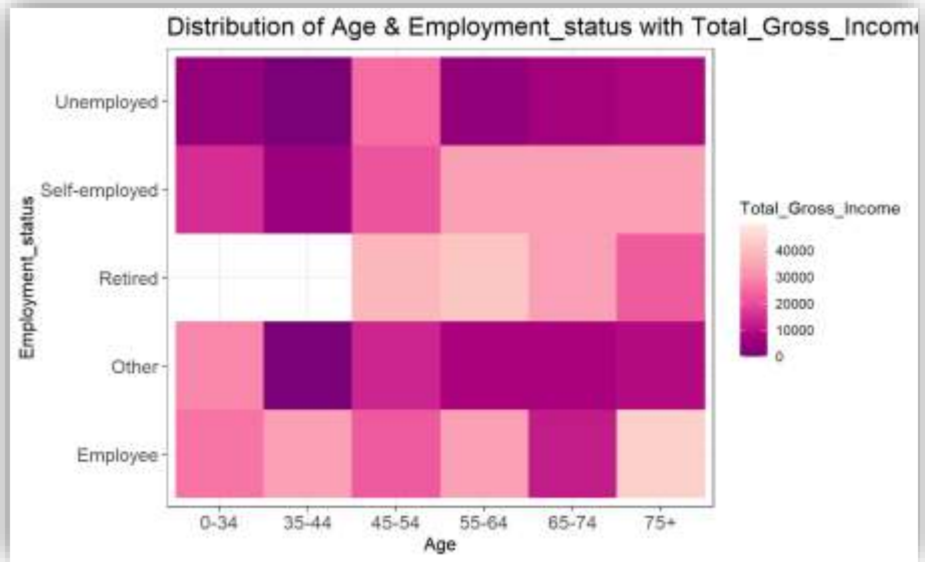

Distribution of Age & Employment_status with Total_Gross_Income
Employment_status
Unemployed
Self-employed
Retired
Other
Employee
0-34
35-44
45-54
55-64
65-74
75+
Age
Total_Gross_Income
40000
30000
20000
10000
0

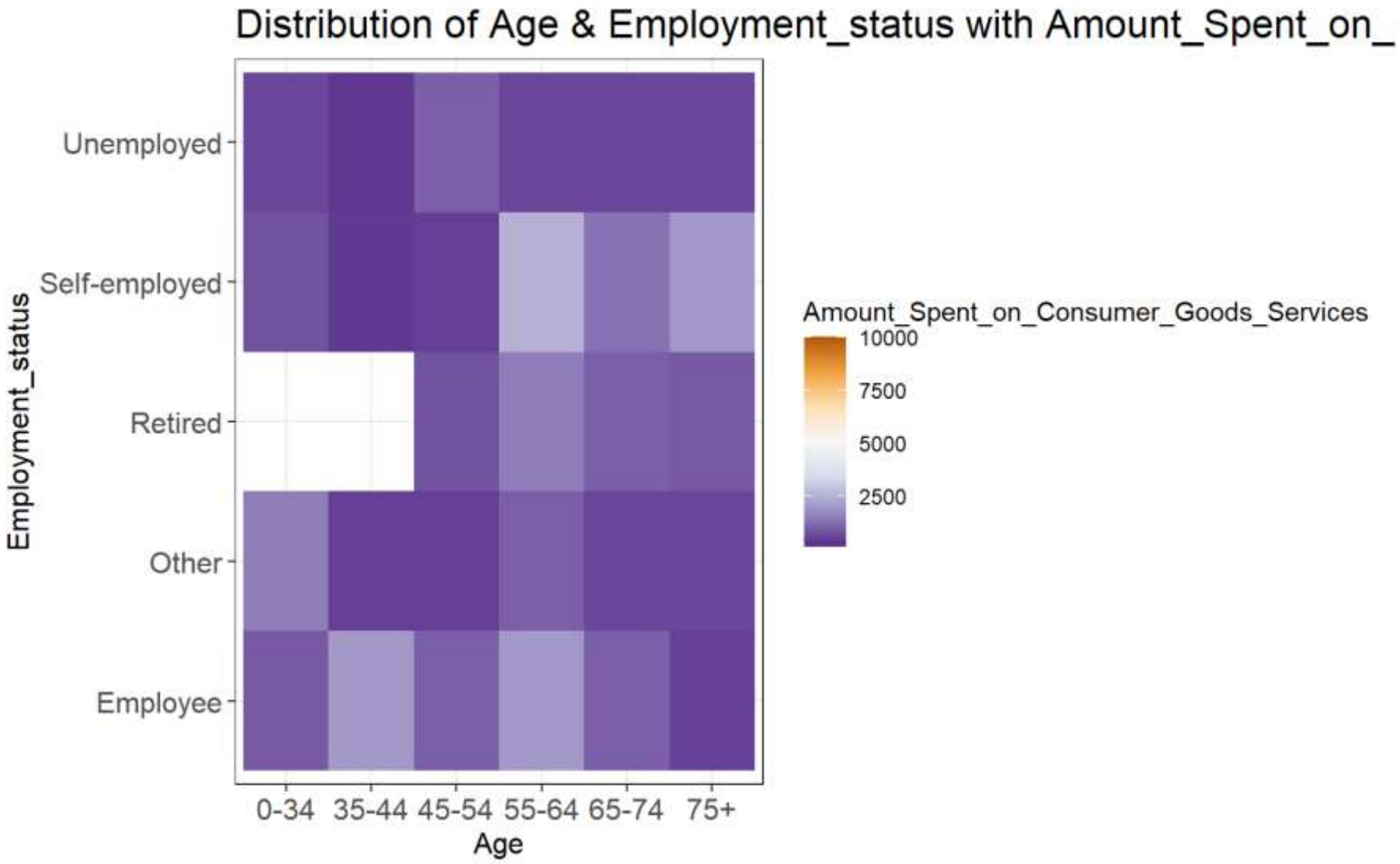

Distribution of Age & Employment_status with Amount_Spent_on_
Employment_status
Unemployed
Self-employed
Retired
Other
Employee
0-34
35-44
45-54
55-64
65-74
75+
Age
Amount_Spent_on_Consumer_Goods_Services
10000
7500
5000
2500

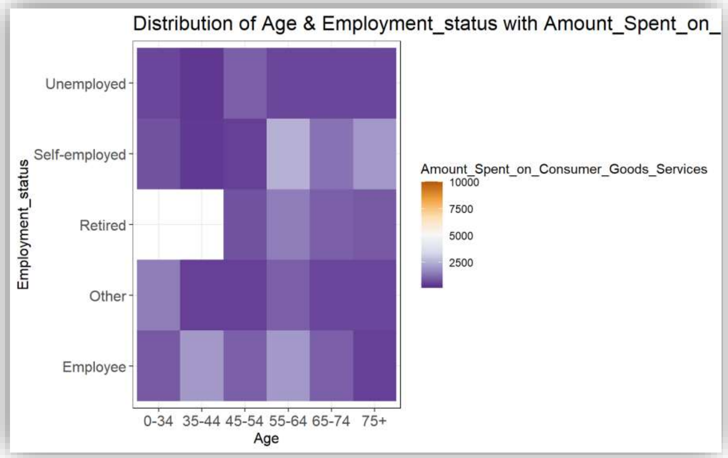

Distribution of Education_Level & Investment_Attitu
Investment_Attitudes
Take substantial financial risks
Take average financial risks
Take above average financial risks
Not willing to take any financial risk
First stage tertiary
Lower secondary
Primary education
Upper secondary
Education_Level
Total_Gross_Income
40000
30000
20000
10000
0

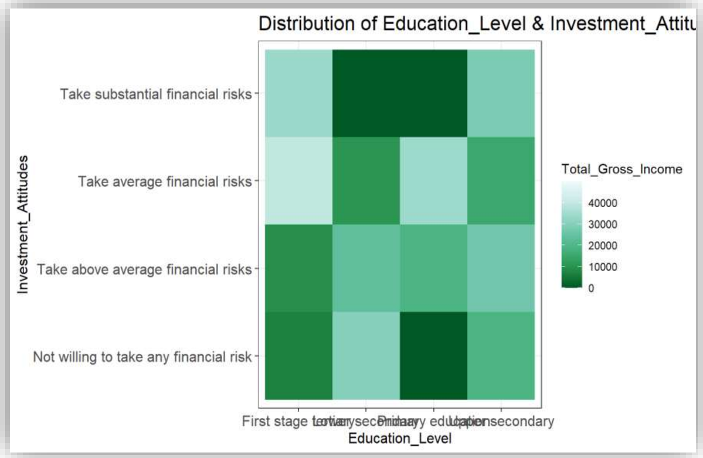

Distribution of Gender & Housing_Status with Total_Gross_In
Housing_Status
Renter
Owner with mortgage
Owner
Female
Male
Gender
Total_Gross_Income
40000
30000
20000
10000
0

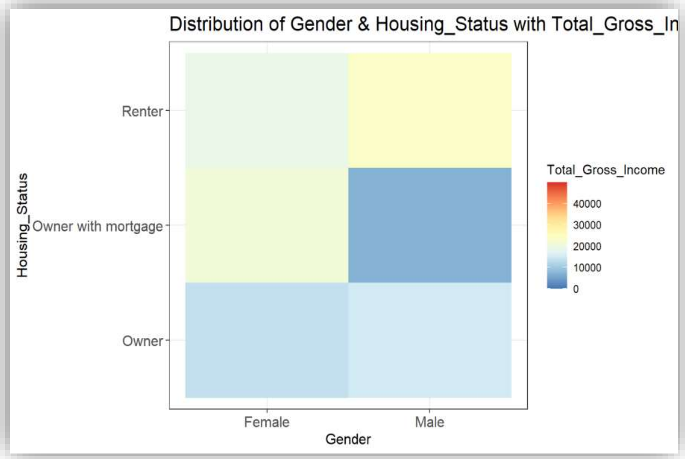

Distribution of Education_Level & Employment_status with Total_G
Employment_status
Unemployed
Self-employed
Retired
Other
Employee
First stage tertiary
Lower secondary
Primary education
Upper secondary
Education_Level
Total_Gross_Income
40000
30000
20000
10000
0

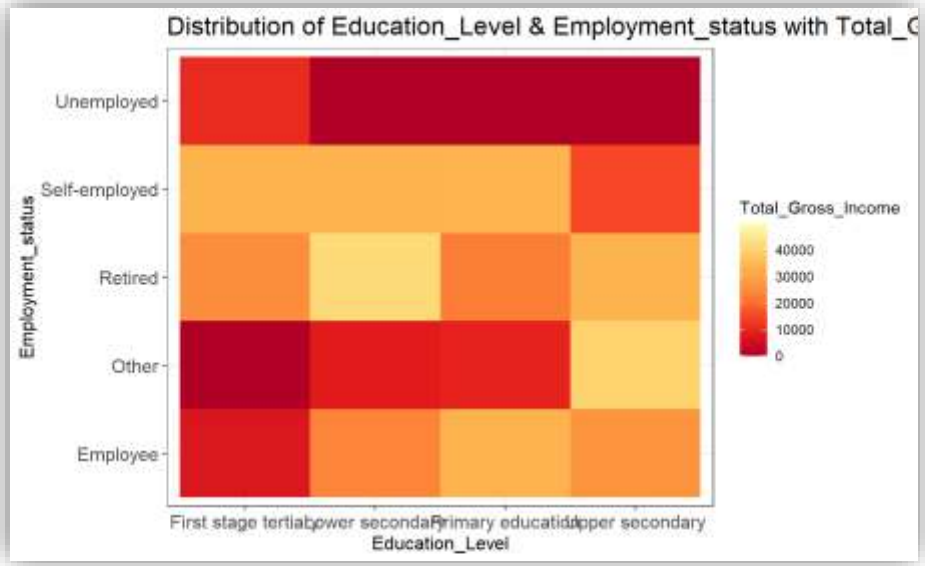

Distribution of Age & Employment_status with Total_Gross_Income
Employment_status
Unemployed
Self-employed
Retired
Other
Employee
0-34
35-44
45-54
55-64
65-74
75+
Age
Total_Gross_Income
40000
30000
20000
10000
0

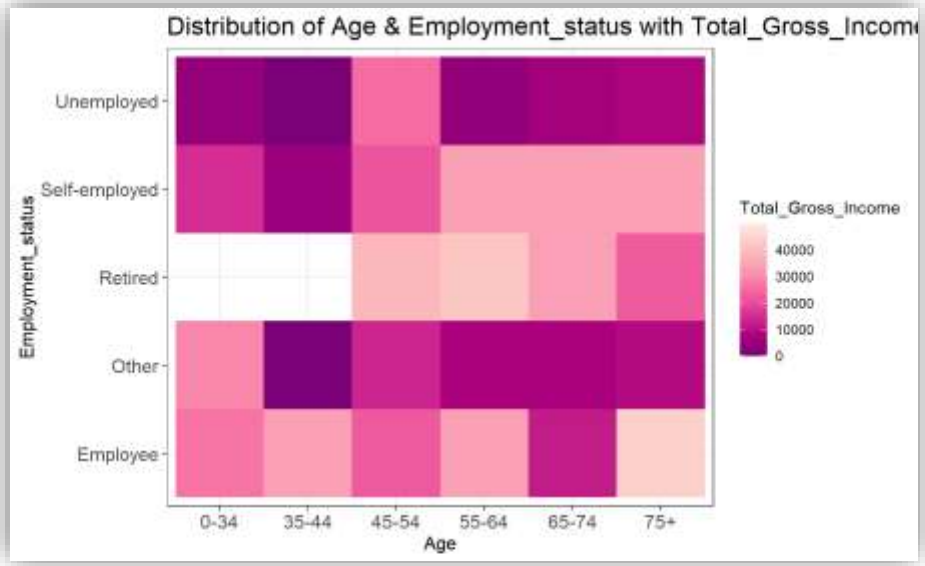

Distribution of Age & Employment_status with Amount_Spent_on_
Employment_status
Unemployed
Self-employed
Retired
Other
Employee
0-34
35-44
45-54
55-64
65-74
75+
Age
Amount_Spent_on_Consumer_Goods_Services
10000
7500
5000
2500

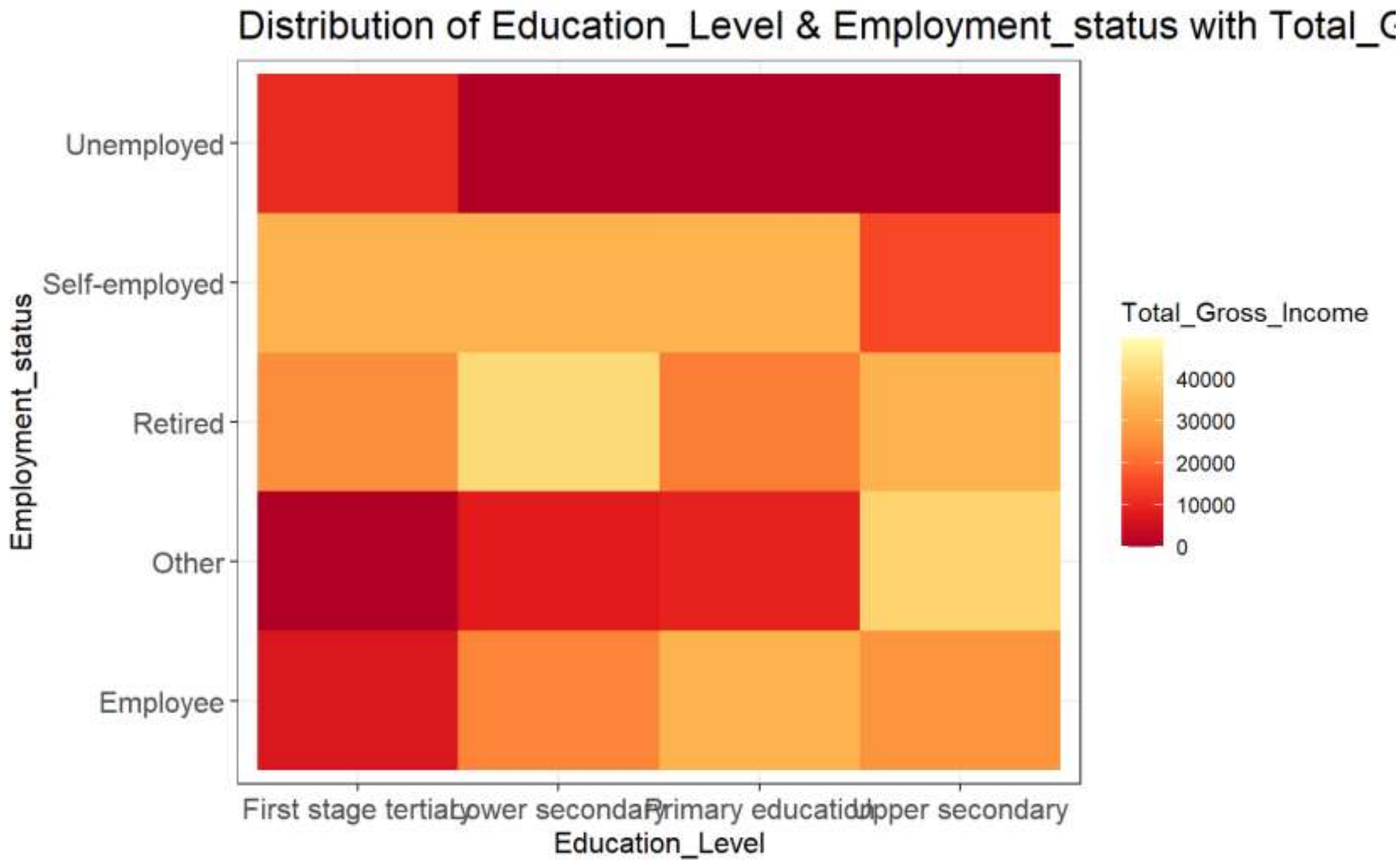

Distribution of Age & Employment_status with Amount_Spent_on_
Unemployed
Self-employed
Retired
Other
Employee
Employment_status
0-34
35-44
45-54
55-64
65-74
75+
Age
Amount_Spent_on_Consumer_Goods_Services
10000
7500
5000
2500

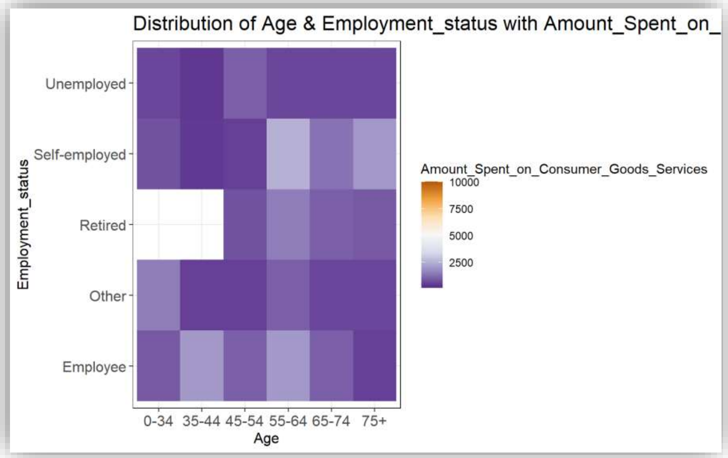

Distribution of Education_Level & Investment_Attitu
Take substantial financial risks
Take average financial risks
Take above average financial risks
Not willing to take any financial risk
Investment_Attitudes
First stage tertiary
Lower secondary
Primary education
Upper secondary
Education_Level
Total_Gross_Income
40000
30000
20000
10000
0


Distribution of Gender & Housing_Status with Total_Gross_In
Renter
Owner with mortgage
Owner
Housing_Status
Female
Male
Gender
Total_Gross_Income
40000
30000
20000
10000
0

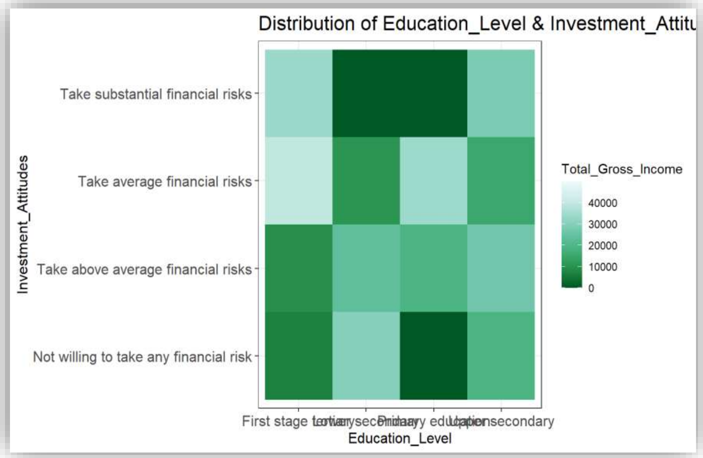

Distribution of Education_Level & Employment_status with Total_G
Unemployed
Self-employed
Retired
Other
Employee
Employment_status
First stage tertia
Lower seconda
Primary educatio
Upper secondary
Education_Level
Total_Gross_Income
40000
30000
20000
10000
0


Distribution of Age & Employment_status with Total_Gross_Income
Unemployed
Self-employed
Retired
Other
Employee
Employment_status
0-34
35-44
45-54
55-64
65-74
75+
Age
Total_Gross_Income
40000
30000
20000
10000
0

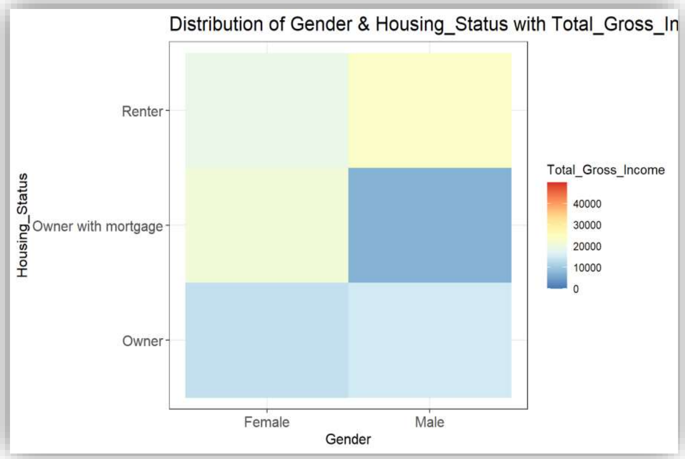

Distribution of Education_Level & Employment_status with Total_G



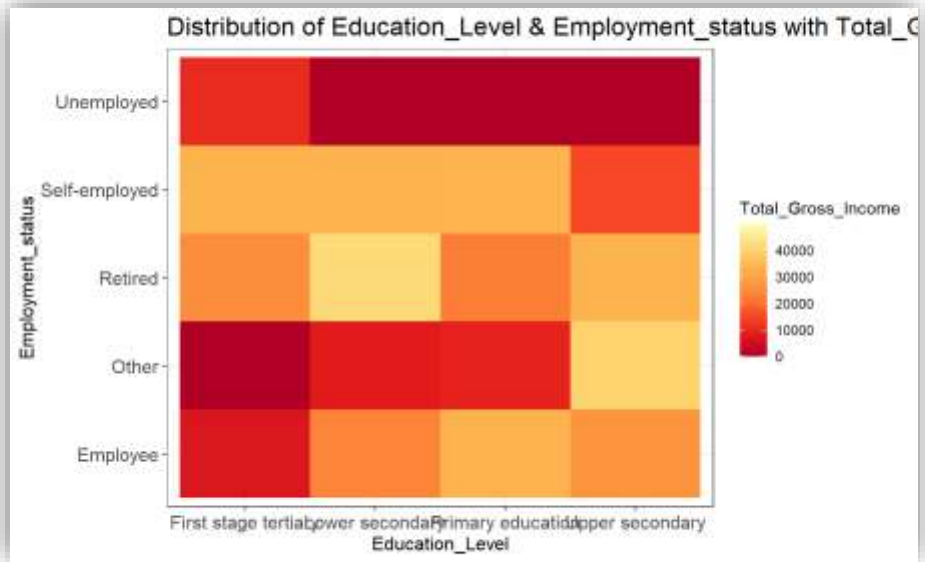
Distribution of Age & Employment_status with Amount_Spent_on_



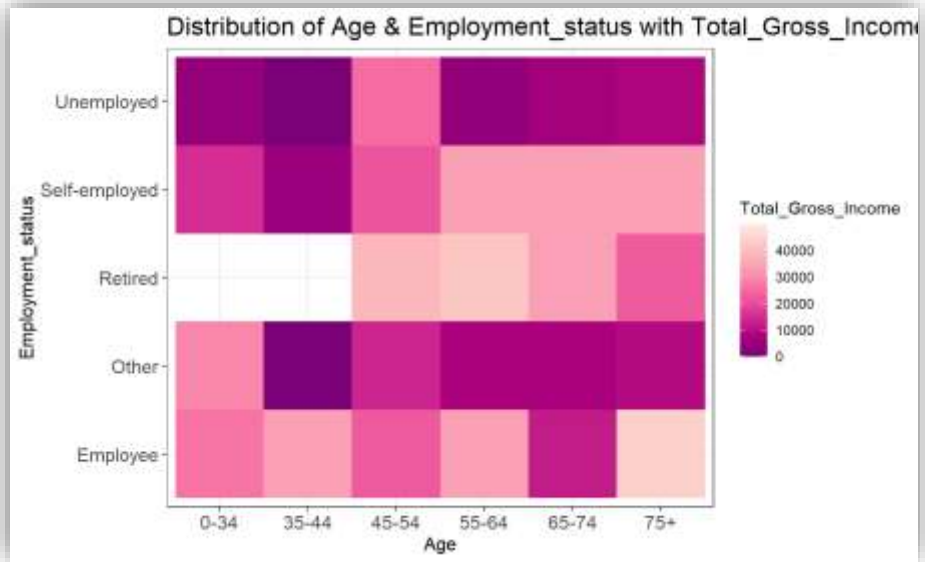
Distribution of Education_Level & Investment_Attitu



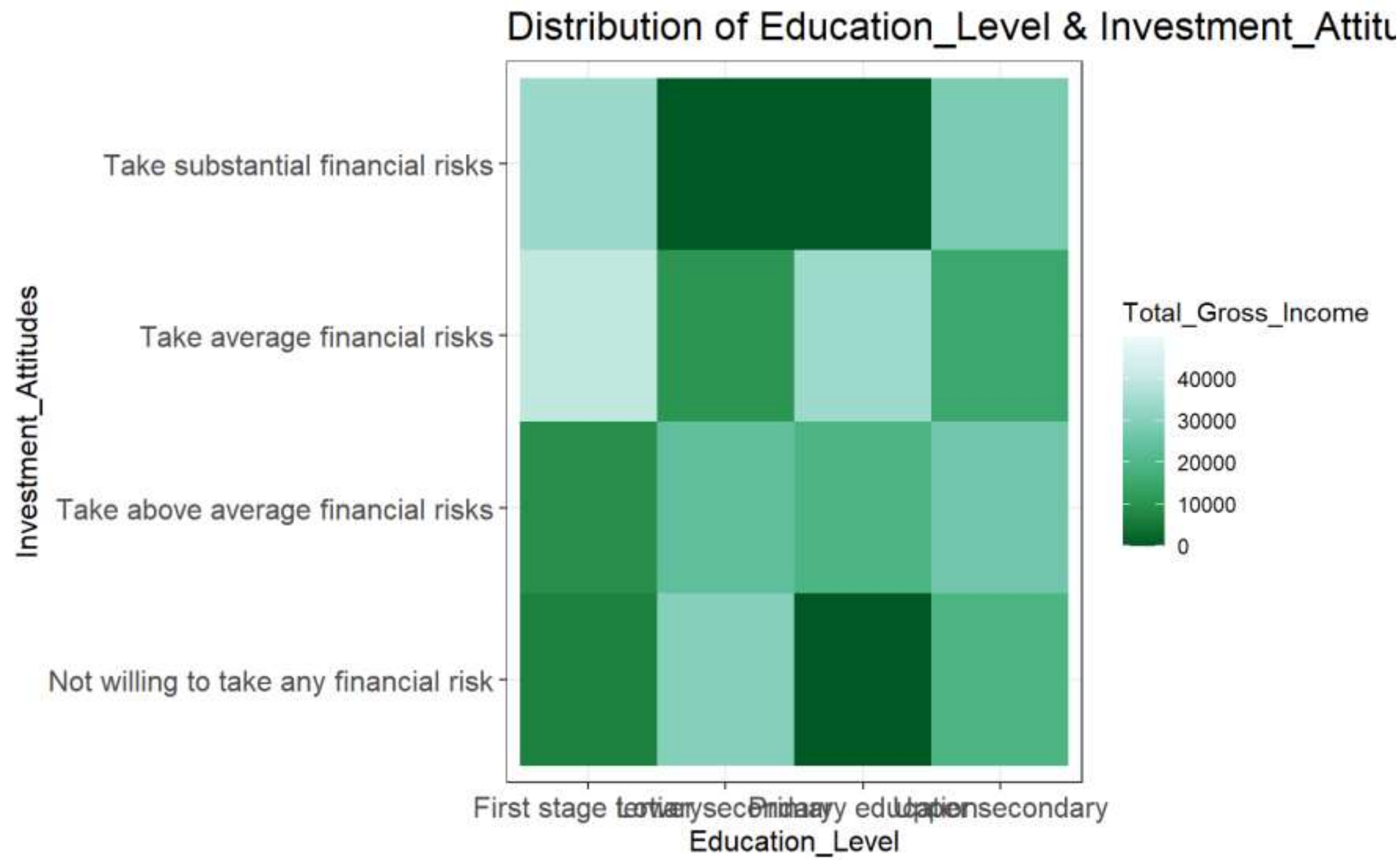
Distribution of Gender & Housing_Status with Total_Gross_In

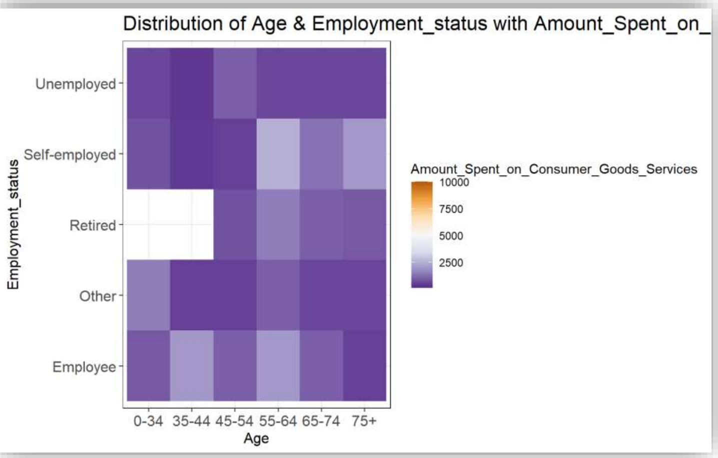


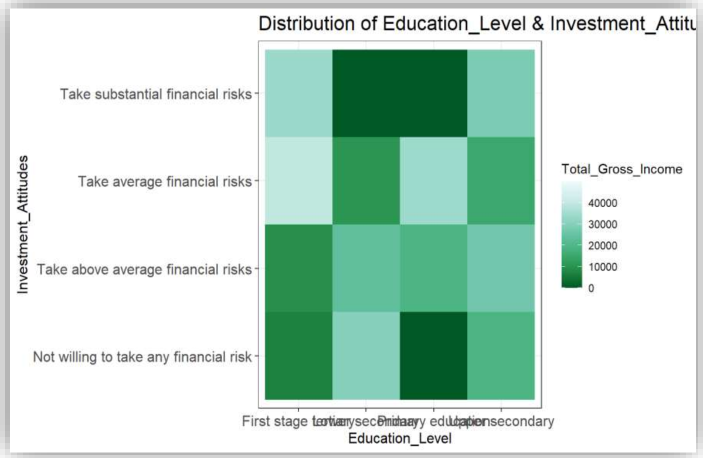
Distribution of Education_Level & Employment_status with Total_G

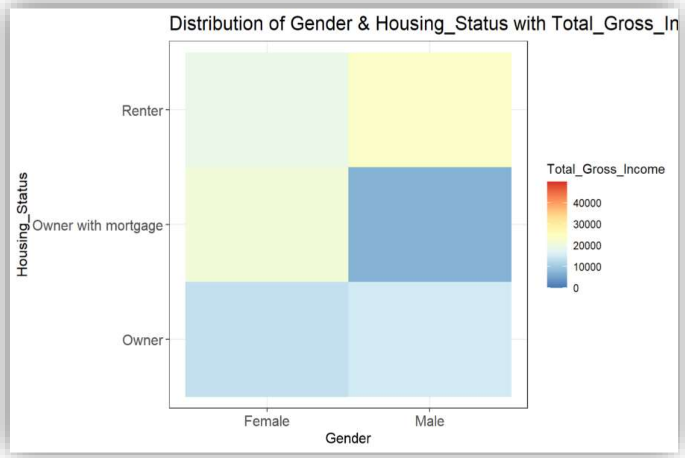


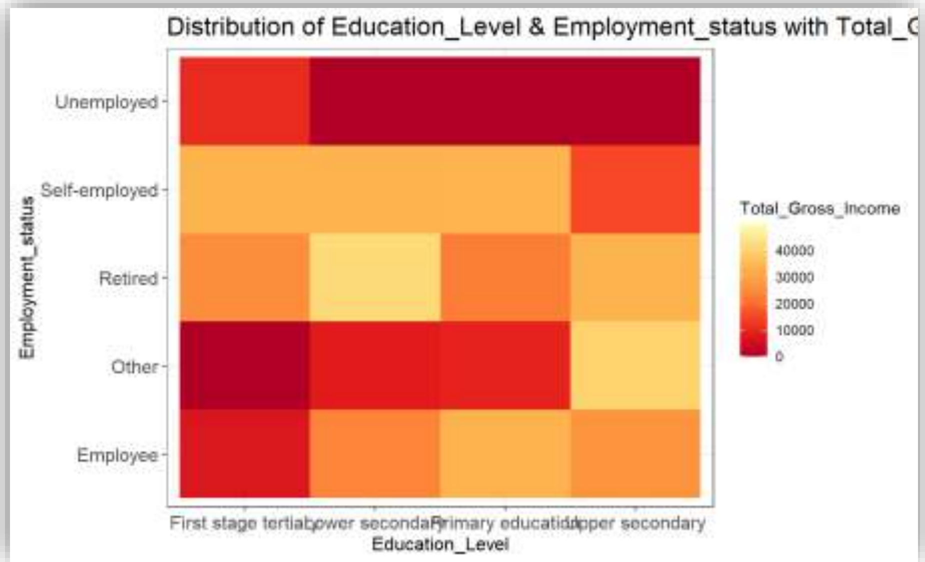
Distribution of Age & Employment_status with Total_Gross_Incom

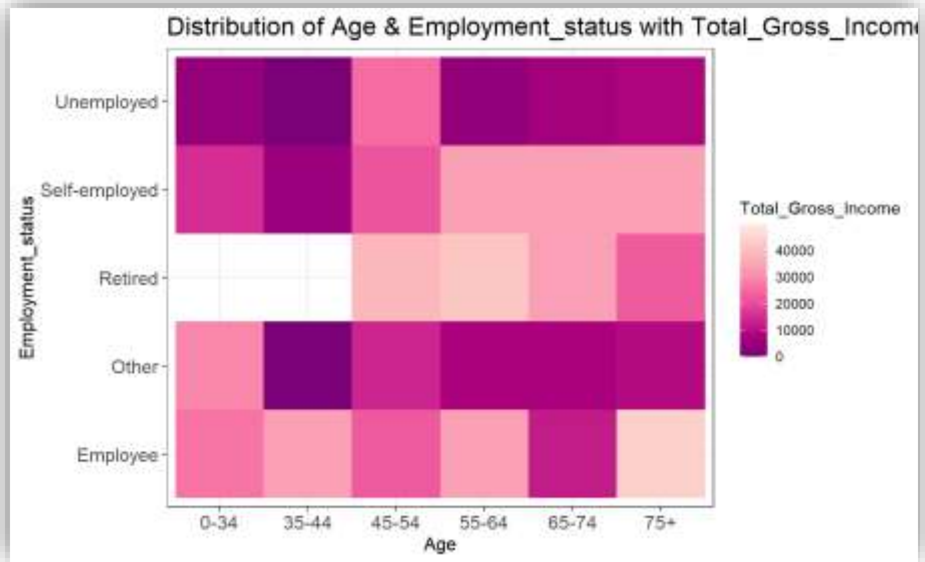

Distribution of Education_Level & Investment_Attitu
Investment_Attitudes
Take substantial financial risks
Take average financial risks
Take above average financial risks
Not willing to take any financial risk
First stage tertiary
Lower secondary
Primary education
Upper secondary
Education_Level
Total_Gross_Income
40000
30000
20000
10000
0

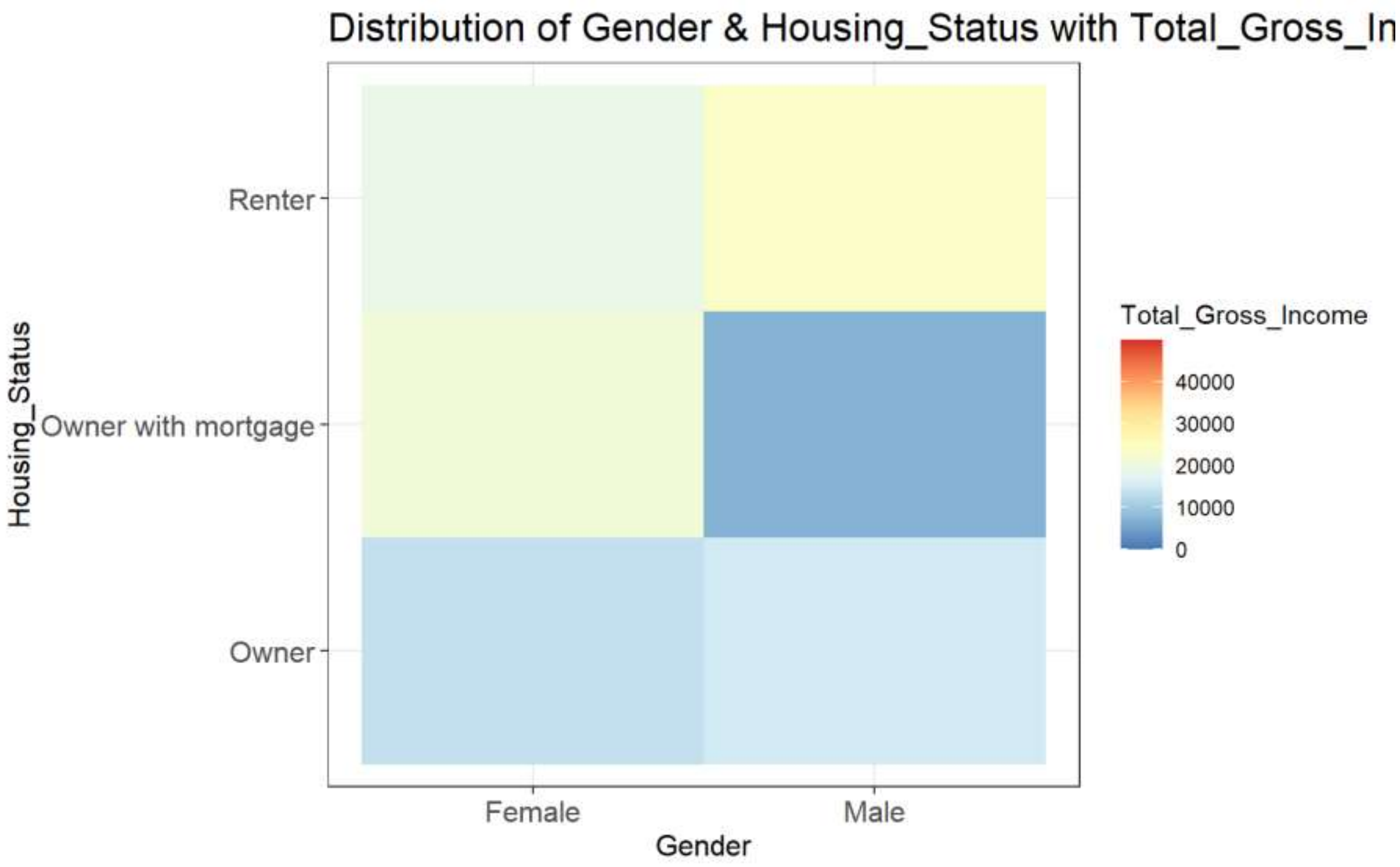

Distribution of Age & Employment_status with Amount_Spent_on_
Employment_status
Unemployed
Self-employed
Retired
Other
Employee
0-34
35-44
45-54
55-64
65-74
75+
Age
Amount_Spent_on_Consumer_Goods_Services
10000
7500
5000
2500


Distribution of Education_Level & Investment_Attitu
Investment_Attitudes
Take substantial financial risks
Take average financial risks
Take above average financial risks
Not willing to take any financial risk
First stage tertiary
Lower secondary
Primary education
Upper secondary
Education_Level
Total_Gross_Income
40000
30000
20000
10000
0

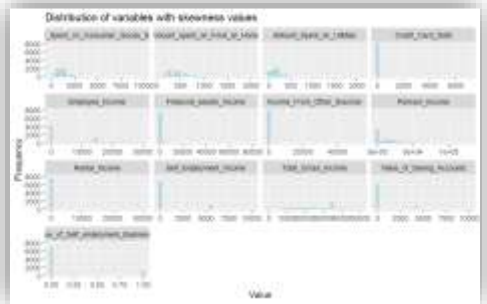

Distribution of Gender & Housing_Status with Total_Gross_In
Housing_Status
Renter
Owner with mortgage
Owner
Female
Male
Gender
Total_Gross_Income
40000
30000
20000
10000
0

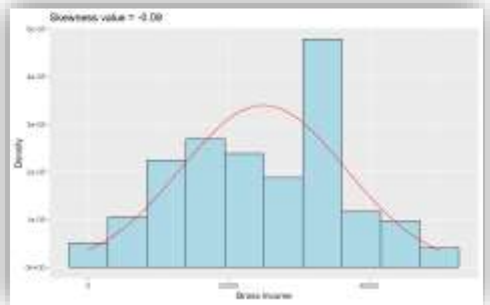

Distribution of Education_Level & Employment_status with Total_G
Employment_status
Unemployed
Self-employed
Retired
Other
Employee
First stage tertiary
Lower secondary
Primary education
Upper secondary
Education_Level
Total_Gross_Income
40000
30000
20000
10000
0

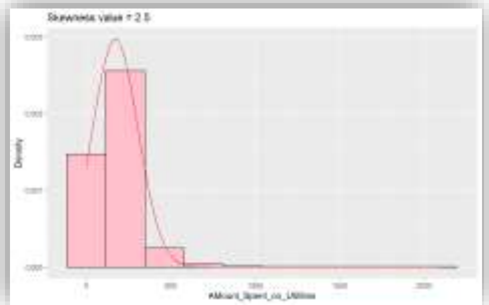

Distribution of Age & Employment_status with Total_Gross_Income
Employment_status
Unemployed
Self-employed
Retired
Other
Employee
0-34
35-44
45-54
55-64
65-74
75+
Age
Total_Gross_Income
40000
30000
20000
10000
0

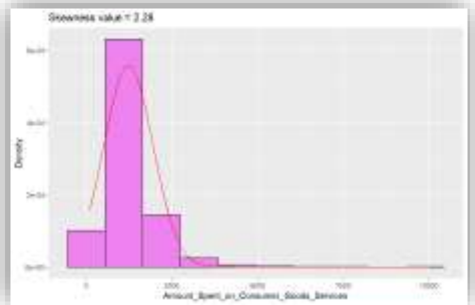

Distribution of Gender & Housing_Status with Total_Gross_In
Renter
Owner with mortgage
Owner
Housing_Status
Female
Male
Gender
Total_Gross_Income
40000
30000
20000
10000
0

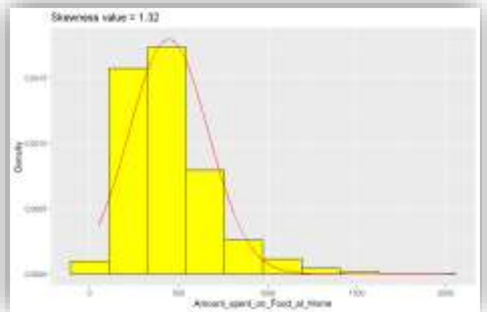
Data Distribution

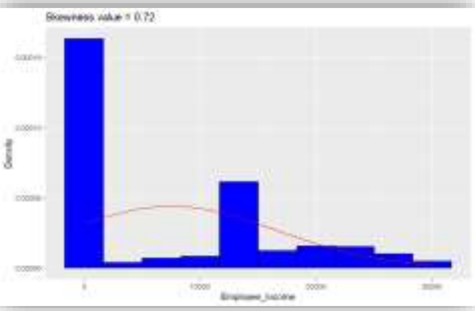


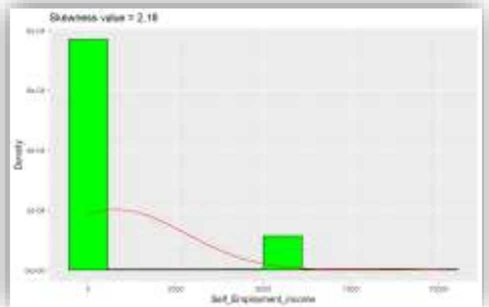


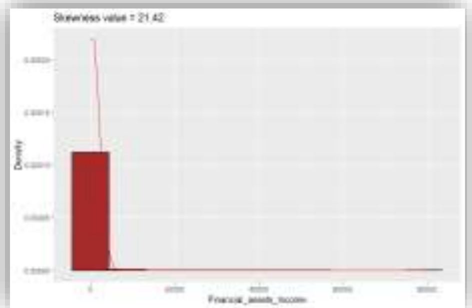


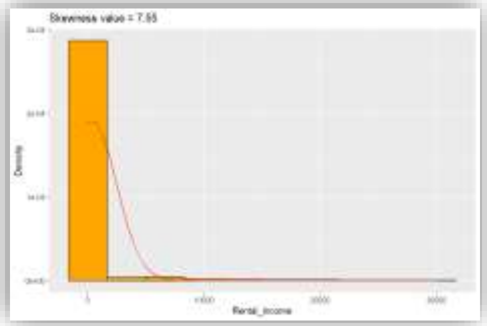


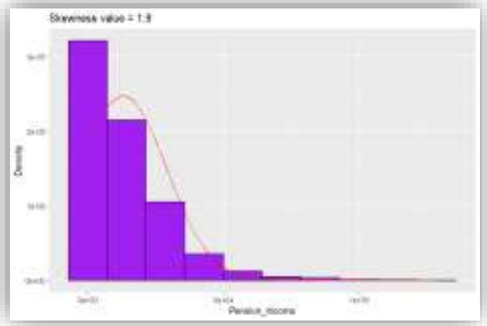


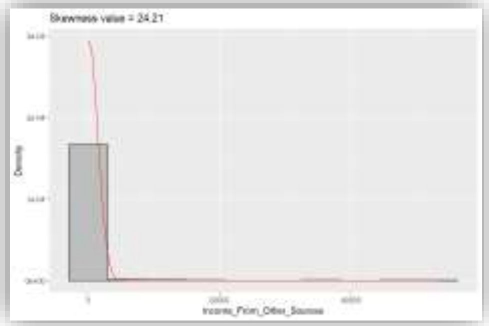


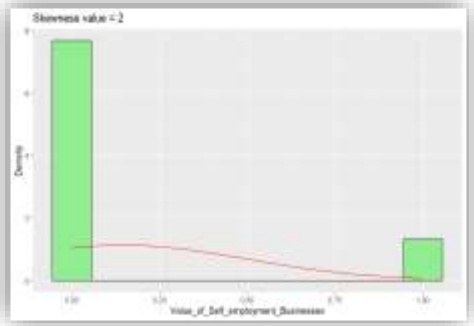












Outcome

The skewness value of Total Gross Income was -0.089 indicating that the data is approximately symmetric, as the value is close to zero.

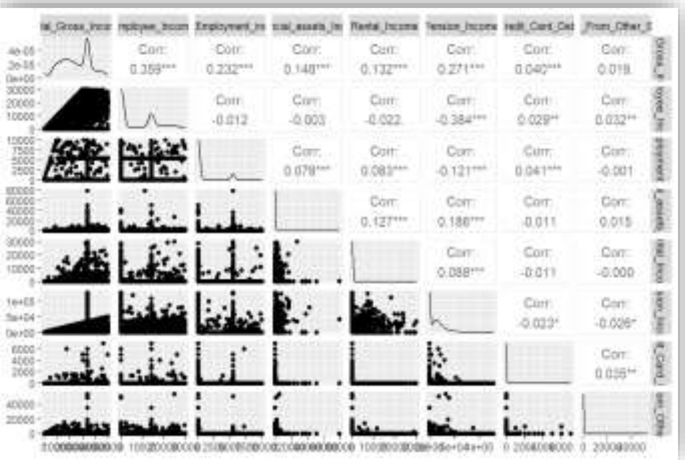
Highly skewed to the right (positive skewness)

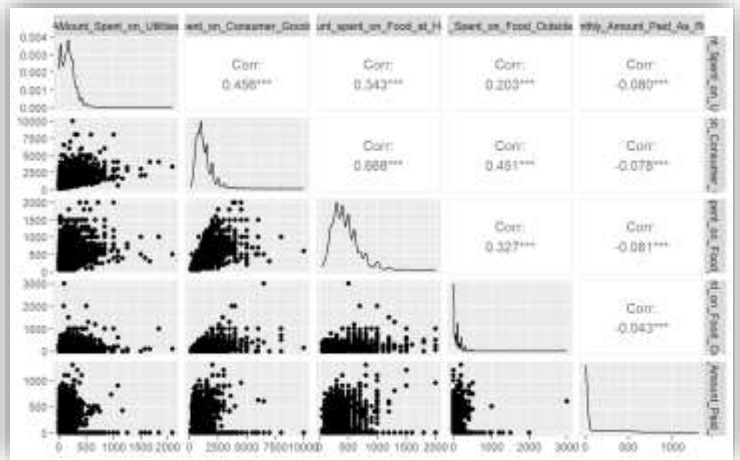
- AMount_Spent_on_Utilities (2.503)
- Amount_Spent_on_Consumer_Goods_Services (2.281)
- Self_Employment_income (2.184)
- Financial_assets_Income (21.421)
- Credit_Card_Debt (22.805)
- Income_From_Other_Sources (24.212)

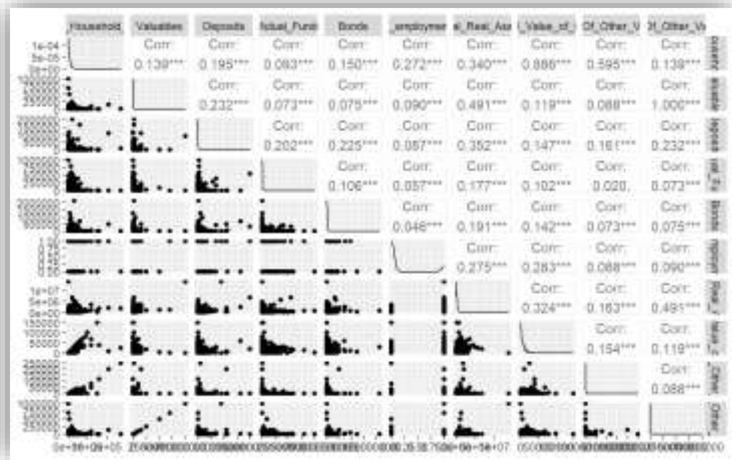
Moderately skewed to the right (positive skewness)

- Employee_Income (0.716)
- Pension_Income (1.901)
- Value_of_Self_employment_Businesses (1.999)
- Rental_Income (7.553)

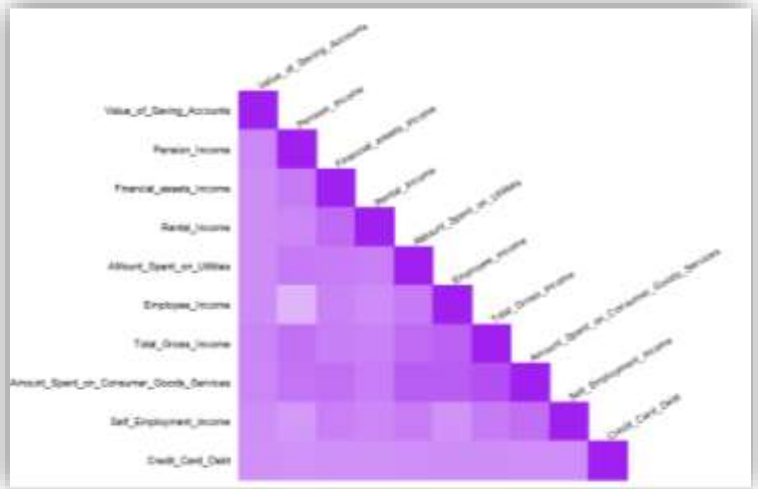
Correlation

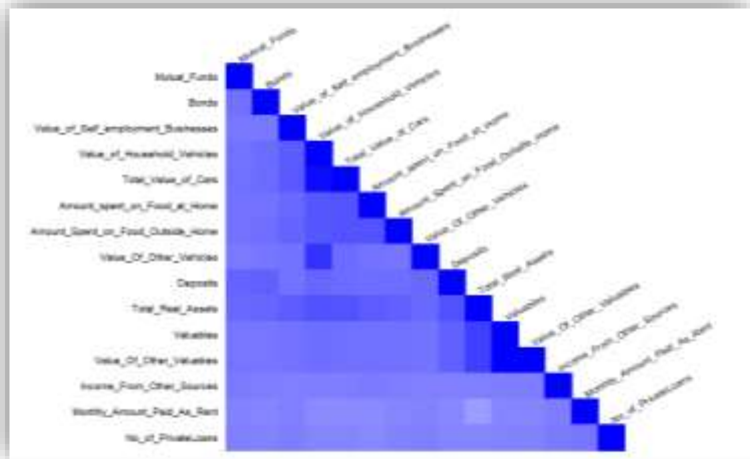


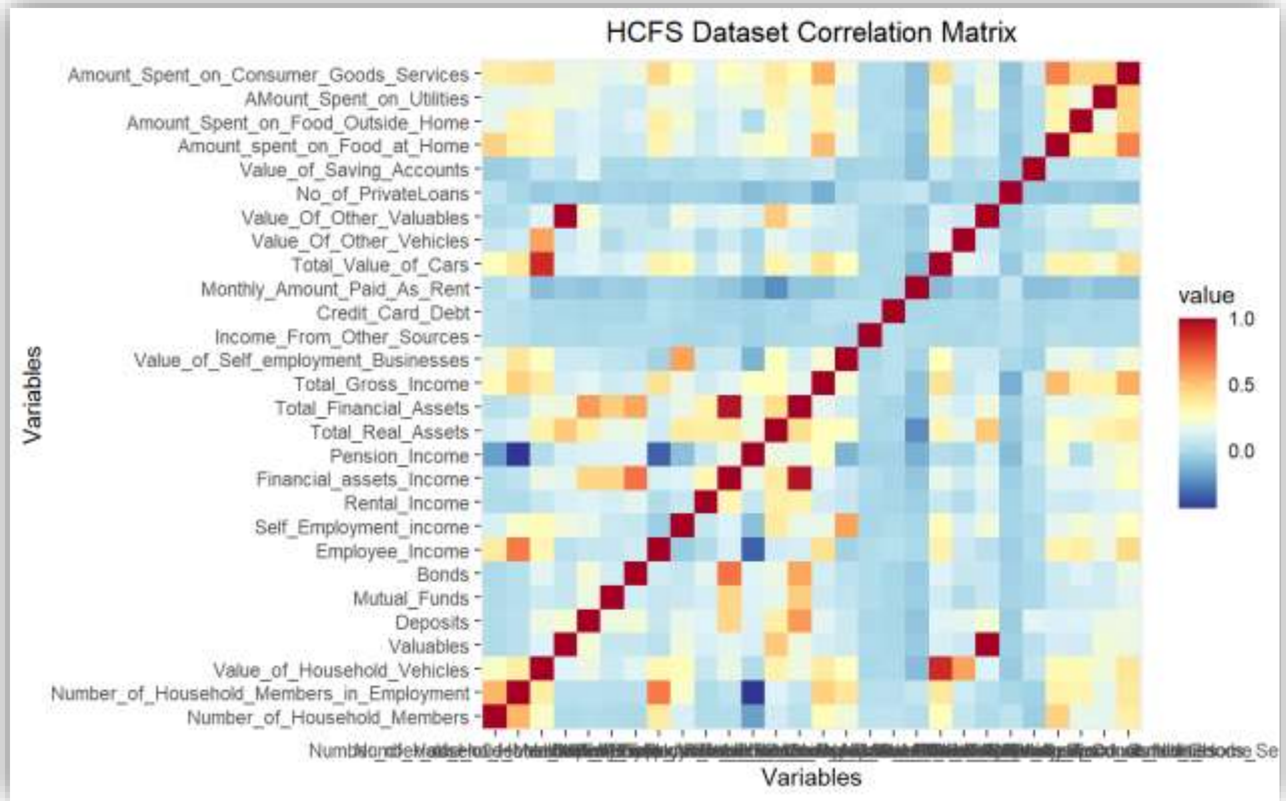




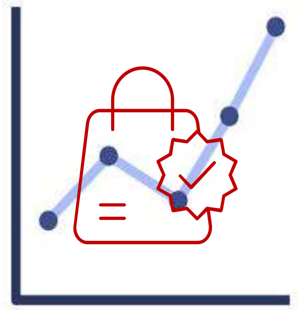
Correlation







Outcome

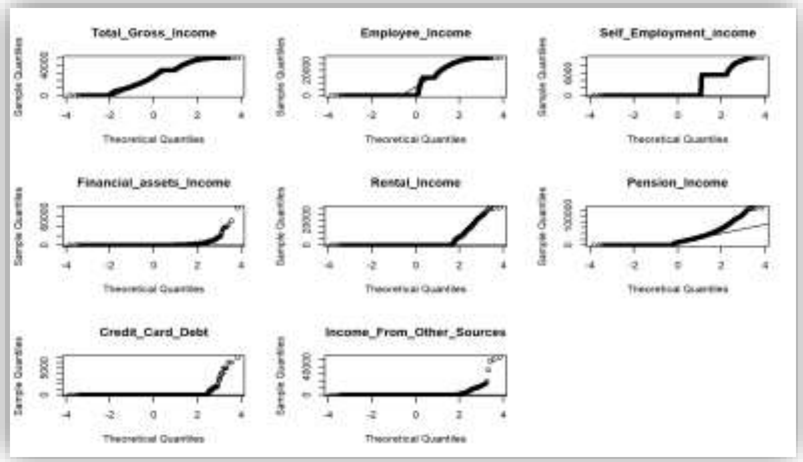


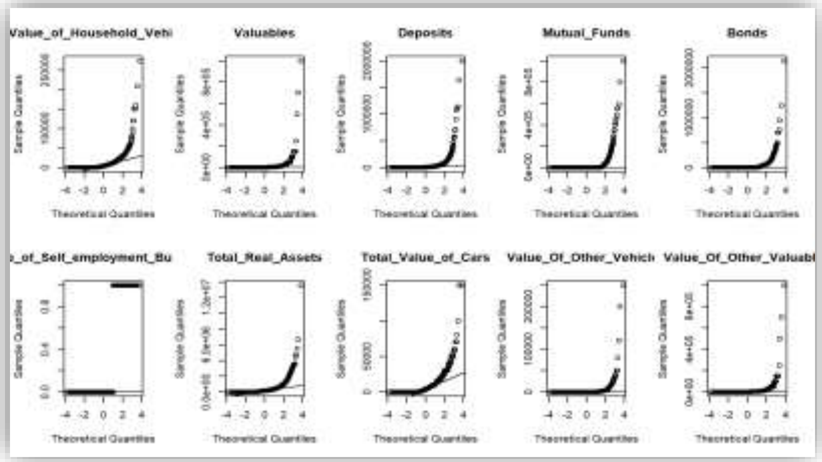
The variables Total Gross Income and Amount Spent on Consumer Goods and Services have a moderate positive correlation 0.56, while the variables Amount Spent on Utilities and Amount Spent on Consumer Goods and Services have a moderate positive correlation 0.46.

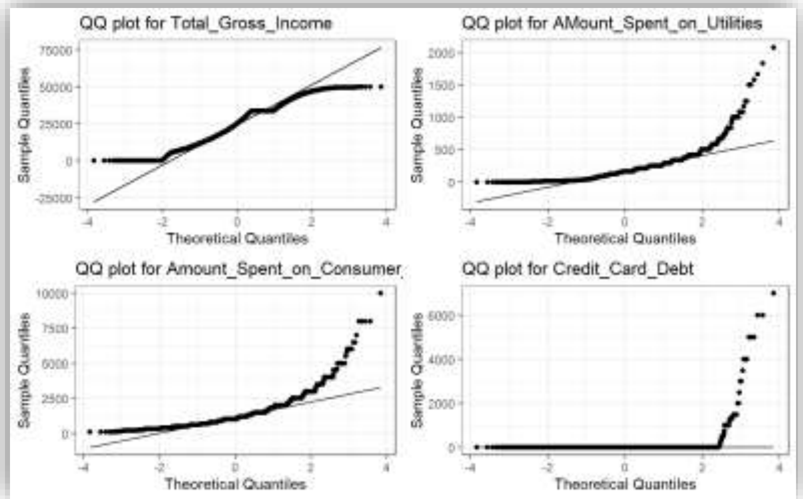
Total Real Assets have a strong positive correlation with the Value of Self-employment Businesses (0.27), Valuables (0.49), and Total Value of Cars (0.32).

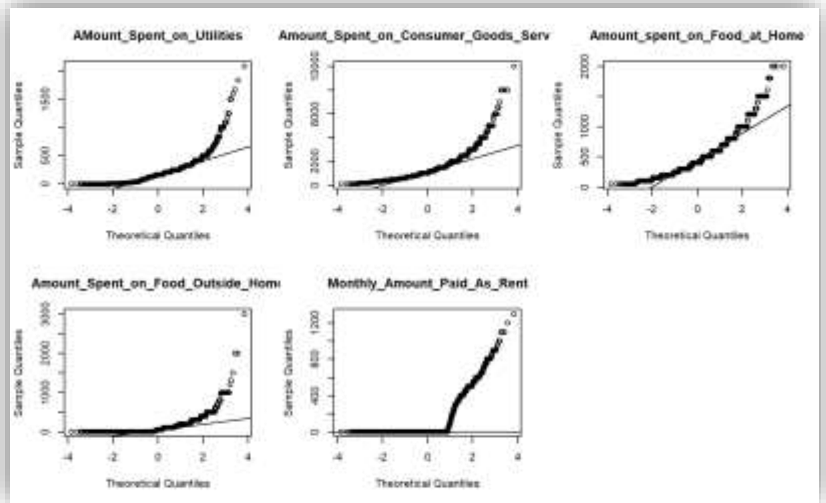
The Monthly Amount Paid as Rent has a negative correlation with Total Real Assets (-0.22) and Value of Household Vehicles (-0.09).

QQ Plots


Total_Gross_Income
Employee_Income
Self_Employment_Income
Financial_assets_Income
Rental_Income
Pension_Income
Credit_Card_Debt
Income_From_Other_Sources
Sample Quantiles
Theoretical Quantiles


Value_of_Household_Vehi
Valuables
Deposits
Mutual_Funds
Bonds
_of_Self_employment_Bu
Total_Real_Assets
Total_Value_of_Cars
Value_Of_Other_Vehicl
Value_Of_Other_Valuabl
Sample Quantiles
Theoretical Quantiles


QQ plot for Total_Gross_Income
QQ plot for AMount_Spent_on_Utilities
QQ plot for Amount_Spent_on_Consumer
QQ plot for Credit_Card_Debt
Sample Quantiles
Theoretical Quantiles
75000
50000
25000
0
-25000
2000
1500
1000
500
10000
7500
5000
2500
6000
4000
2000
-4
-2
0
2
4


AMount_Spent_on_Utilities
Amount_Spent_on_Consumer_Goods_Serv
Amount_spent_on_Food_at_Home
Amount_Spent_on_Food_Outside_Hom
Monthly_Amount_Paid_As_Rent
Sample Quantiles
Theoretical Quantiles

Hypothesis Testing

Hypothesis Testing

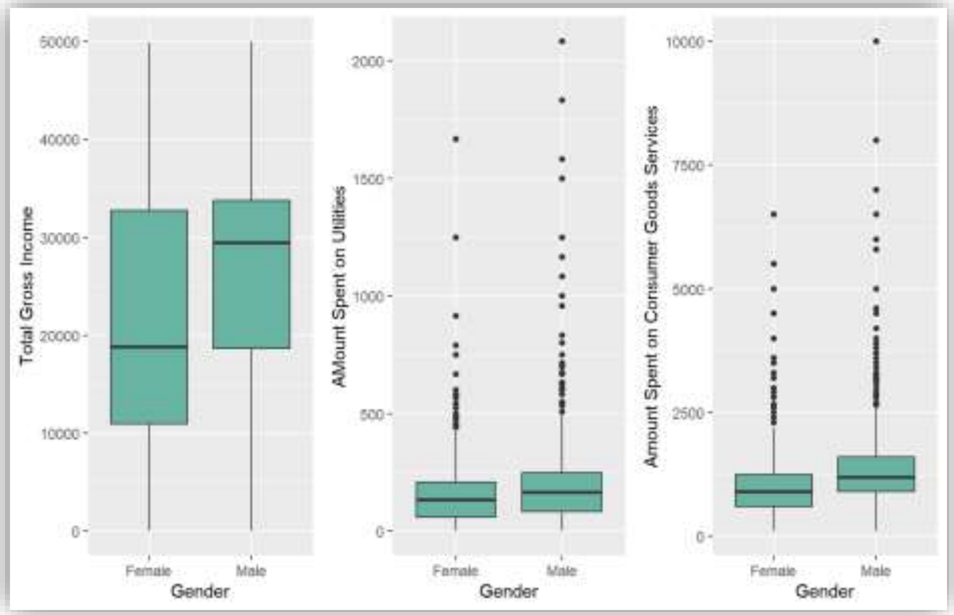
Hypothesis testing was performed on the dataset to determine if there is a significant association between demographic variables like Gender, Age and Education Level and various metric and categorical variables related to income, assets, financial status and expenditure. This helped to identify if there were any disparities or significant differences between variables with regard to income, financial assets or expenditure behavior.

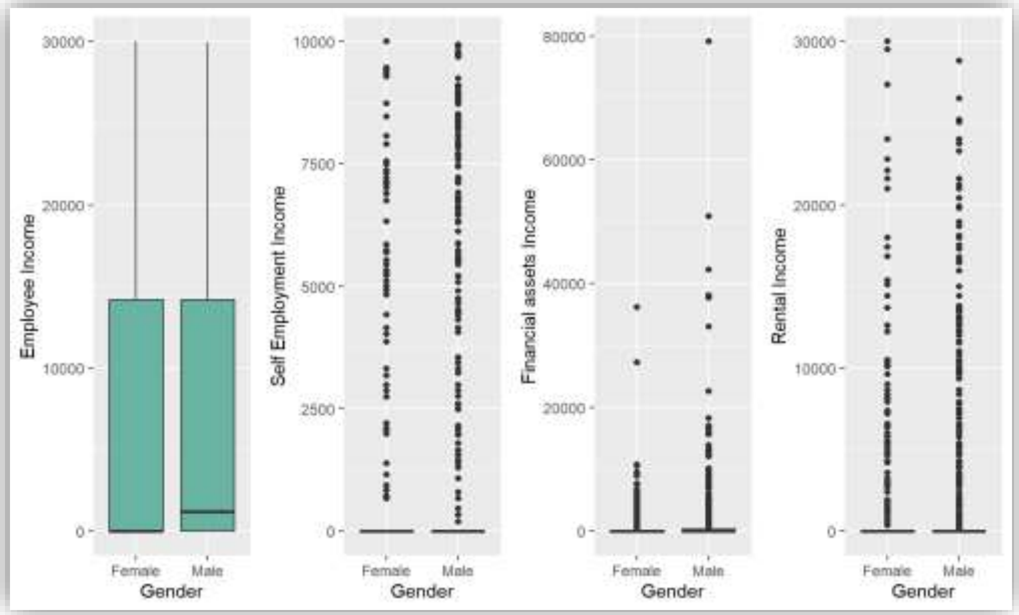
We performed the following parametric and non parametric tests on the dataset,

- T test
- Mann Whitney U Test
- ANOVA
- Kruskal Wallis Test

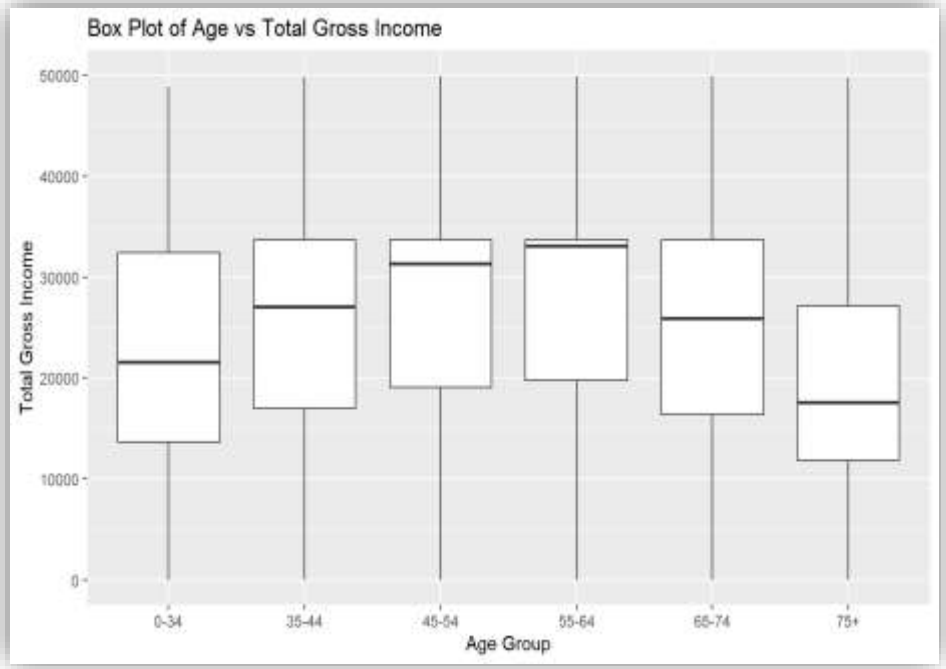
- Null Hypothesis:
  - There is no difference between the Male and Female groups with respect to the dependent variable.
  - There is no difference between the 6 categories of the independent variable Age with respect to the dependent variable.
  - There is no difference between the 4 categories of the independent variable Education Level with respect to the dependent. Variable.
- Alternative Hypothesis:
  - There is a difference between the Male and Female groups with respect to the dependent variable
  - There is a difference between the 6 categories of the independent variable Age with respect to the dependent variable
  - There is a difference between the 4 categories of the independent variable Education Level with respect to the dependent variable

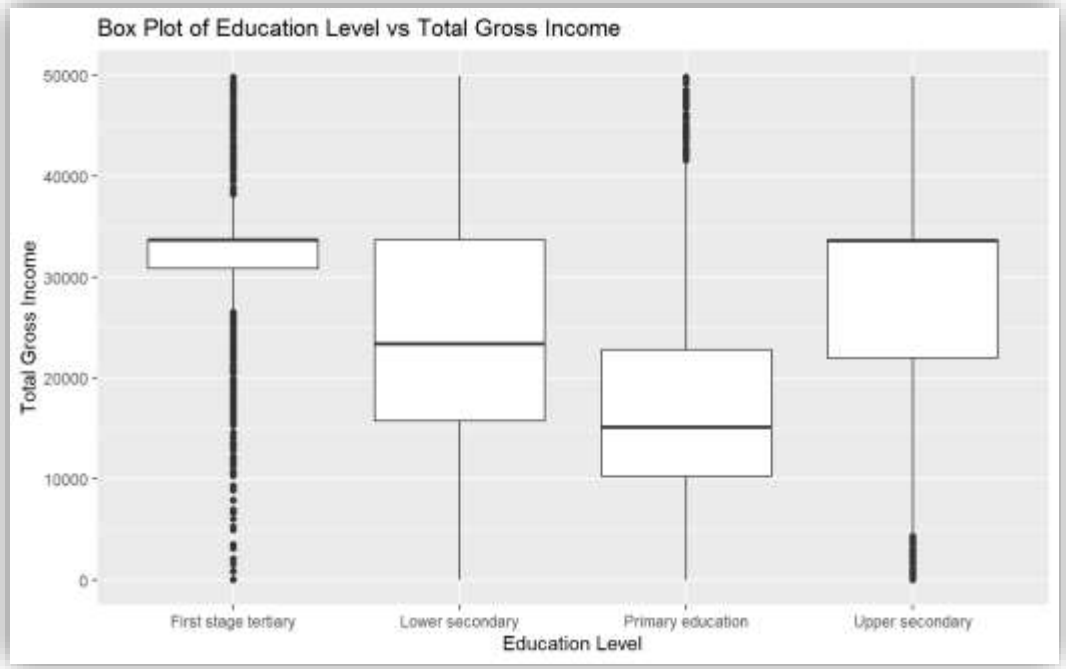
Boxplots for T test





Boxplots for ANOVA





Hypothesis Tests with Gender

Moderator	Dependent Variable	Test	P value	Result
Gender	Total_Gross_Income	T test	0.002	Rejected
	AMount_Spent_on_Utilities	Wilcoxon-Mann-Whitney test	0.002	Rejected
	Amount_Spent_on_Consumer_Goods_Services		0.002	Rejected
	Employee_Income		0.002	Rejected
	Self_Employment_income		0.002	Rejected
	Financial_assets_Income		0.002	Rejected
	Pension_Income		0.9046	Accepted
	Rental_Income		0.002982	Rejected
	Credit_Card_Debt		0.004729	Rejected
	Value_of_Saving_Accounts		0.3272	Accepted
	Value_of_Self_employment_Businesses		0.002	Rejected
	Amount_spent_on_Food_at_Home		0.002	Rejected
	Income_From_Other_Sources		0.201	Accepted

Hypothesis Tests with Gender

Moderator	Dependent Variable	Test	P value	Result
Gender	Has_Real_Assets	Chi 2 Test	0.000653	Rejected
	Has_Financial_Assets		0.000442	Rejected
	Has_Vehicles		0.000516	Rejected
	Has_Valuables		0.1143	Accepted
	Has_Real_Estate_Wealth		0.00028	Rejected
	Has_Deposits		0.00064	Rejected
	Has_Mutual_Funds		0.00014	Rejected
	Has_Bonds		0.00058	Rejected
	Has_Shares		0.000185	Rejected
	Has_Debt		0.000944	Rejected
	Has_Credit_Card_Debt		0.007	Rejected
	Has_Private_Loans		0.59	Accepted
	Has_Applied_for_Loan_Credit		0.000615	Rejected

Hypothesis Tests with Age

Moderator	Dependent Variable	Test	P/h value	Result
Age	Total_Gross_Income	ANOVA	0.002	Rejected
	AMount_Spent_on_Utilities	Kruskal-Wallis rank sum test	0.002	Rejected
	Amount_Spent_on_Consumer_Goods_Services		0.002	Rejected
	Employee_Income		0.002	Rejected
	Self_Employment_income		0.002	Rejected
	Financial_assets_Income		0.002	Rejected
	Credit_Card_Debt		0.0004	Rejected
	Value_of_Saving_Accounts		0.0002	Rejected
	Value_of_Self_employment_Businesses		0.002	Rejected
	Amount_spent_on_Food_at_Home		0.002	Rejected
	Income_From_Other_Sources		0.002	Rejected

Hypothesis Tests with Education Level

Moderator	Dependent Variable	Test	P/h value	Result
Education Level	Total_Gross_Income	ANOVA	0.002	Rejected
	AMount_Spent_on_Utilities	Kruskal-Wallis rank sum test	0.002	Rejected
	Amount_Spent_on_Consumer_Goods_Services		0.002	Rejected
	Employee_Income		0.002	Rejected
	Self_Employment_income		0.002	Rejected
	Financial_assets_Income		0.002	Rejected
	Pension_Income		0.002	Rejected
	Credit_Card_Debt		0.002	Rejected
	Value_of_Saving_Accounts		0.00046	Rejected
	Value_of_Self_employment_Businesses		0.002	Rejected
	Amount_spent_on_Food_at_Home		0.002	Rejected
	Income_From_Other_Sources		0.00015	Rejected

Regression

Linear Regression

What is the impact of total gross income on the expenditure of consumer goods and services?

1. The estimated intercept in the model is 3.778e+02, which says the expected amount spent on consumer goods and services is 377.8 units when total gross income is zero.
2. However, this value may not have practical significance, given that total gross income is unlikely to be zero in practice.
3. The estimated intercept is the value of the dependent variable when all independent variables in the model are equal to zero, holding all other variables constant.
4. The estimated coefficient for Total_Gross_Income is 3.432e-02, which means that for each unit increase in total gross income, the estimated amount spent on consumer goods and services increases by 0.03432 units, holding all other variables constant.
5. The estimated coefficient represents the amount of change in the dependent variable associated with a one-unit increase in the independent variable, holding all other independent variables constant.

Linear Regression

What is the relationship between the amount spent on consumer goods and services and the value of saving accounts?

- These Coefficient of intercept is 0.02604 which suggests that, for every one-unit increase in Value of Saving Accounts, Amount Spent on Consumer Goods Services increases by 0.02604 units.
- Residuals: These are the differences between the actual values of Amount Spent on Consumer Goods Services and the predicted values from the regression model. The minimum residual is -1115, the first quartile is -505, the median is -205, the third quartile is 295, and the maximum residual is 8795.

Multi-Linear Regression

What is the relationship between the Total Gross Income and Value of Saving Accounts with the Total Real Assets?

- The regression model showed that the variables Total_Gross_Income and Value_of_Saving_Accounts explained 8.3% of the variance from the variable Total_Real_Assets.
- An ANOVA was used to test whether this value was significantly different from zero. Using the present sample, it was found that the effect was significantly different from zero, F=369.14, p = <.001, R2 = 0.08.
- When all independent variables are zero, the value of the variable Total_Real_Assets is 9570.2. If the value of the variable Total_Gross_Income changes by one unit, the value of the variable Total_Real_Assets changes by 8.42. If the value of the variable Value_of_Saving_Accounts changes by one unit, the value of the variable Total_Real_Assets changes by 2.07. In this model, the variable Total_Gross_Income has the greatest influence on the variable Total_Real_Assets.

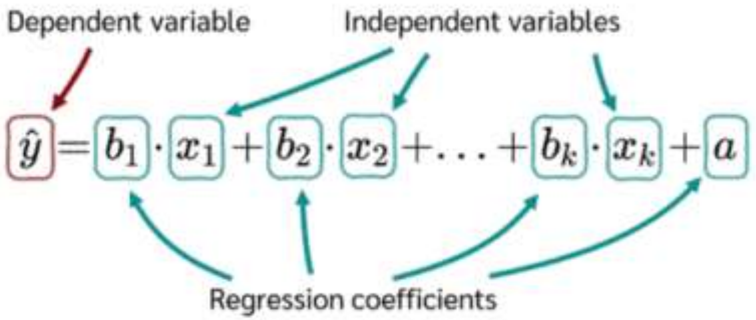
Multi-Linear Regression

What is the relationship between the value of household vehicles and various sources of income, such as employee income, self-employment income, rental income, financial assets income, and pension income, among respondents?

- The regression model showed that the variables Employee_Income, Self_Employment_income, Rental_Income, Financial_assets_Income and Pension_Income explained 22.92% of the variance from the variable Value_of_Household_Vehicles.
- An ANOVA was used to test whether this value was significantly different from zero. Using the present sample, it was found that the effect was significantly different from zero, F=484.68, p = <.001, R2 = 0.23.
- When all independent variables are zero, the value of the variable Value_of_Household_Vehicles is 1975.53.
- If the value of the variable Rental_Income changes by one unit, the value of the variable Value_of_Household_Vehicles changes by 0.03.
- If the value of the variable Financial_assets_Income changes by one unit, the value of the variable Value_of_Household_Vehicles changes by 0.34.
- If the value of the variable Pension_Income changes by one unit, the value of the variable Value_of_Household_Vehicles changes by 0.09.
- In this model, the variable Employee_Income has the greatest influence on the variable Value_of_Household_Vehicles.

Logistic Regression

To predict the likelihood of a household having certain financial assets or liabilities based on their demographic and employment characteristics. The response variable of interest, such as whether a household has mutual funds or credit card debt, is binary in nature (yes/no), and logistic regression is well-suited for modeling binary outcomes. By fitting a logistic regression model, we can identify the significant predictors and quantify their impact on the likelihood of having the financial asset or liability of interest, controlling for other covariates



Dependent variable

Independent variables

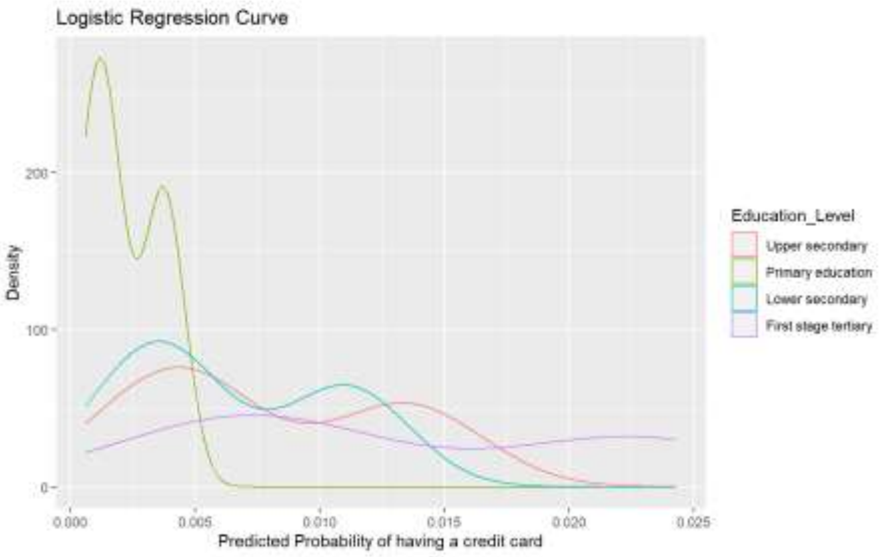
$$\boxed{\hat{y}} = \boxed{b_1} \cdot \boxed{x_1} + \boxed{b_2} \cdot \boxed{x_2} + \ldots + \boxed{b_k} \cdot \boxed{x_k} + \boxed{a}$$

Regression coefficients

Logistic Regression

What is the relationship between age, education level, employment status and the likelihood of having credit card debt?

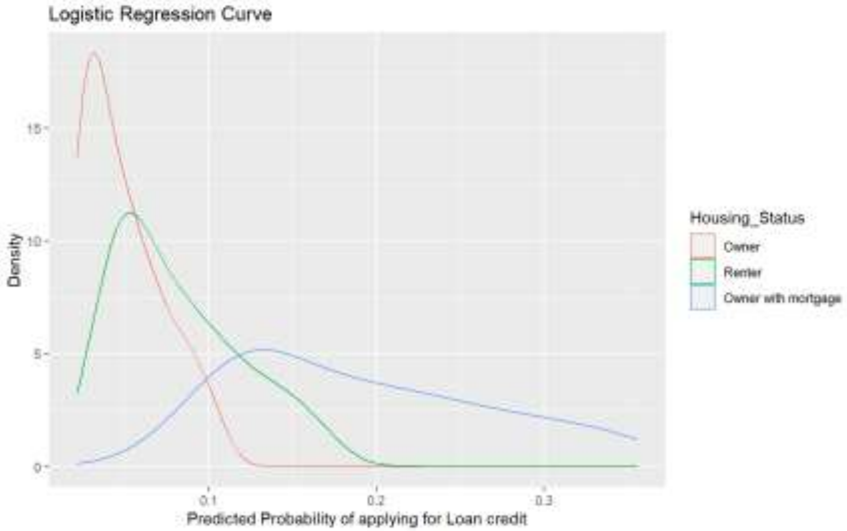
- The Education Level coefficients show that compared to having a tertiary education level, having a lower secondary or primary education level is associated with a lower log-odds of having credit card debt. The coefficient for upper secondary education level is not statistically significant.
- The Employment status coefficients show that compared to being employed full-time, being retired is associated with a lower log-odds of having credit card debt, while being self-employed or unemployed is not significantly associated with credit card debt. The coefficient for "other" employment status is not statistically significant.
- The deviance residuals indicate that the model fits the data reasonably well, and the AIC is 722.28, which suggests that the model is a good fit.



Logistic Regression

What is the relationship between Education Level, Employment Status, Housing Status, and the probability of having applied for a loan/credit?

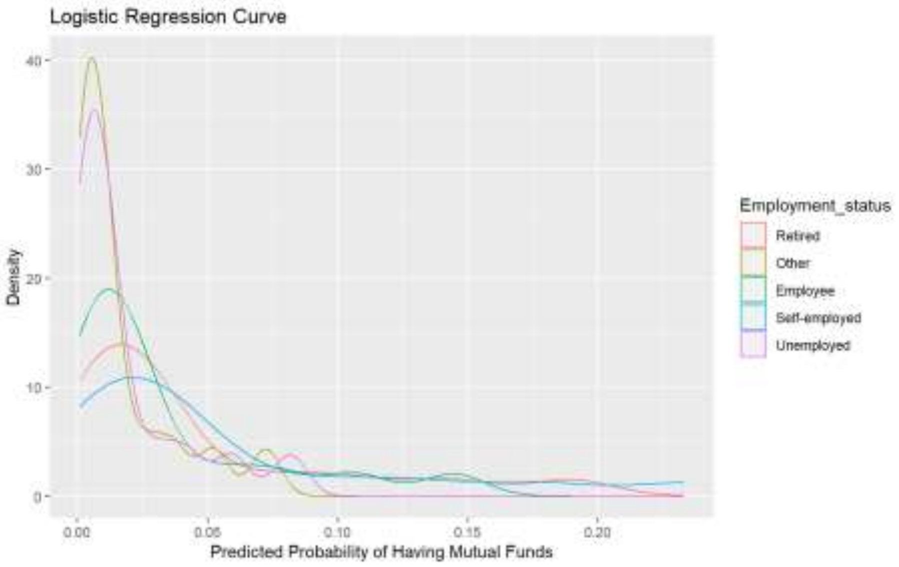
- The p-values associated with the coefficients of the independent variables show that education level, employment status, and housing status are all significant predictors of having applied for a loan credit. Among the education levels, those with lower secondary education are more likely to apply for loan credit compared to those with primary education. Similarly, among employment status, those who are self-employed and those who have other employment status are more likely to apply for loan credit compared to those who are unemployed. Among housing status, those who own a house with a mortgage and those who rent are more likely to apply for loan credit compared to those who live in other housing arrangements.
- The null and residual deviance and AIC show that the model provides a good fit to the data.



Logistic Regression

Can likelihood of having mutual funds be predicted based on Gender, Education level and Employment status?

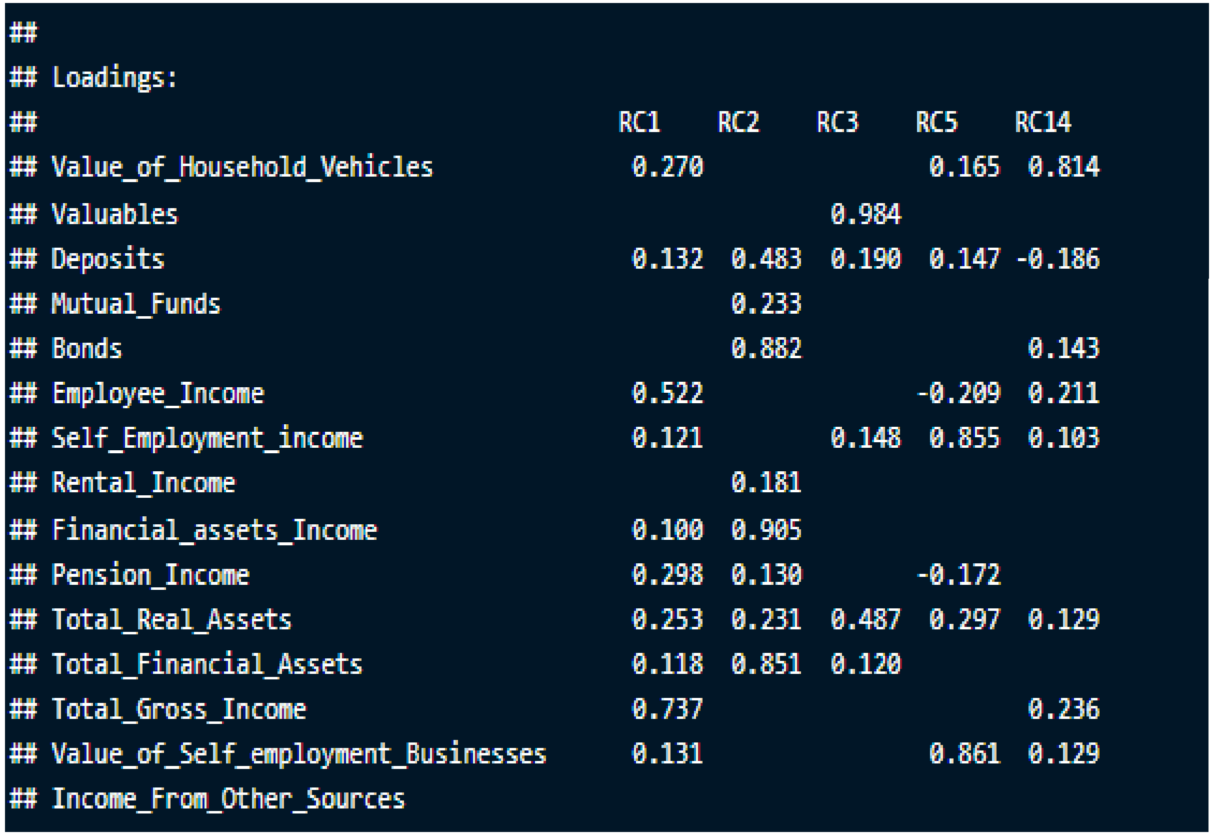
- The coefficients of the model reveal that individuals with lower education levels are less likely to have mutual funds, with estimates of -1.61 for “Lower secondary,” -2.05 for “Primary education,” and -0.70 for “Upper secondary” education levels. Retired individuals and those who are self-employed are more likely to have mutual funds, with estimates of 0.33 and 0.59, respectively, while individuals in other employment status categories are less likely to have mutual funds.
- Moreover, individuals with real assets are more likely to have mutual funds, with an estimate of 2.04, and male individuals are more likely to have mutual funds than female individuals, with an estimate of 0.34. The significance codes reveal that all coefficients are statistically significant, except for “Employment_statusUnemployed” and “Has_Real_AssetsYes” at the 0.05 significance level.



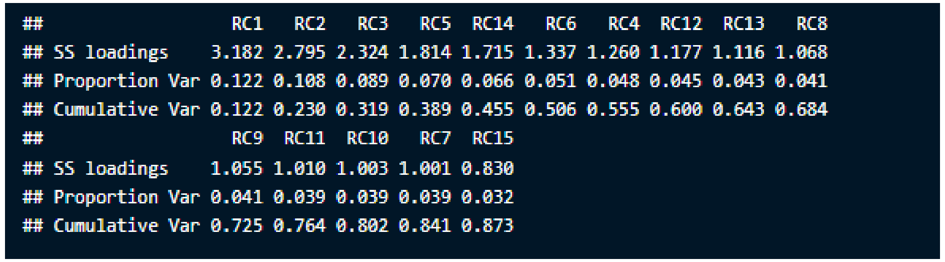
Principal Component Analysis

- PCA (Principal Component Analysis) can be considered for the HCFS dataset because it is a multivariate technique that can be used to identify patterns and relationships among a large number of variables.
- PCA can be used to reduce the dimensionality of a dataset.

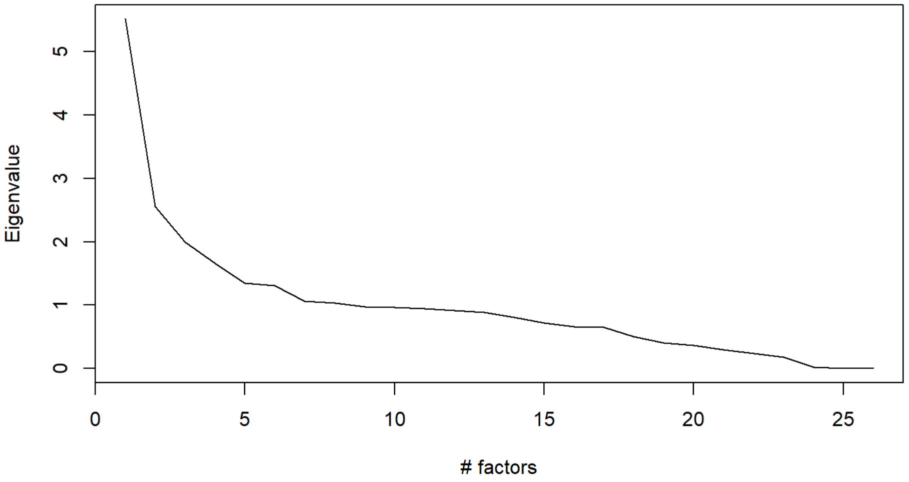
Principal Component Analysis



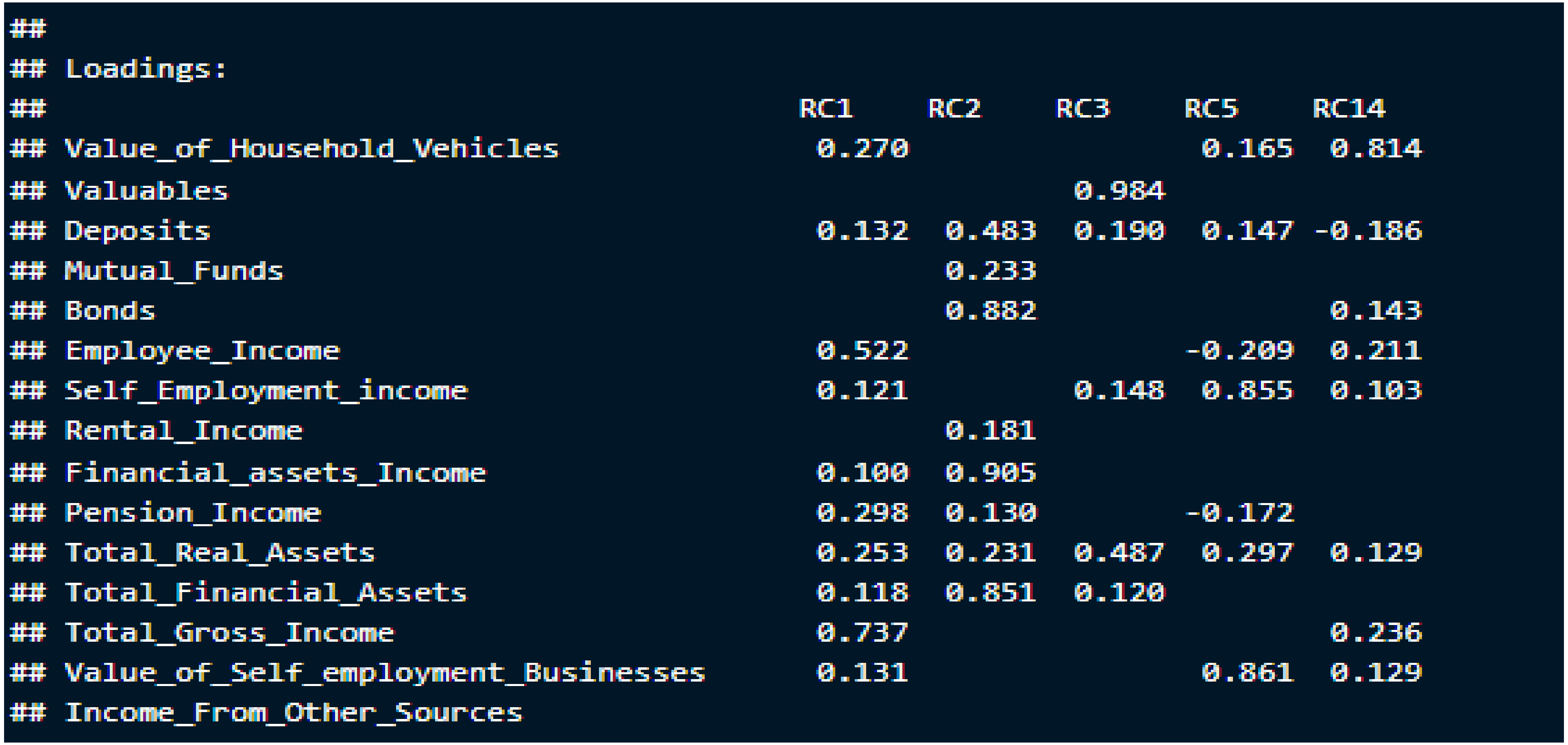
```
##
## Loadings:
##                                         RC1    RC2    RC3    RC5    RC14
## Value_of_Household_Vehicles              0.270                0.165  0.814
## Valuables                                              0.984
## Deposits                                 0.132  0.483  0.190  0.147 -0.186
## Mutual_Funds                                    0.233
## Bonds                                           0.882                0.143
## Employee_Income                          0.522               -0.209  0.211
## Self_Employment_income                   0.121         0.148  0.855  0.103
## Rental_Income                                   0.181
## Financial_assets_Income                  0.100  0.905
## Pension_Income                           0.298  0.130        -0.172
## Total_Real_Assets                        0.253  0.231  0.487  0.297  0.129
## Total_Financial_Assets                   0.118  0.851  0.120
## Total_Gross_Income                       0.737                       0.236
## Value_of_Self_employment_Businesses      0.131                0.861  0.129
## Income_From_Other_Sources
```



```
##                   RC1   RC2   RC3   RC5  RC14   RC6   RC4  RC12  RC13   RC8
## SS loadings     3.182 2.795 2.324 1.814 1.715 1.337 1.260 1.177 1.116 1.068
## Proportion Var 0.122 0.108 0.089 0.070 0.066 0.051 0.048 0.045 0.043 0.041
## Cumulative Var 0.122 0.230 0.319 0.389 0.455 0.506 0.555 0.600 0.643 0.684
##                   RC9  RC11  RC10   RC7  RC15
## SS loadings     1.055 1.010 1.003 1.001 0.830
## Proportion Var 0.041 0.039 0.039 0.039 0.032
## Cumulative Var 0.725 0.764 0.802 0.841 0.873
```

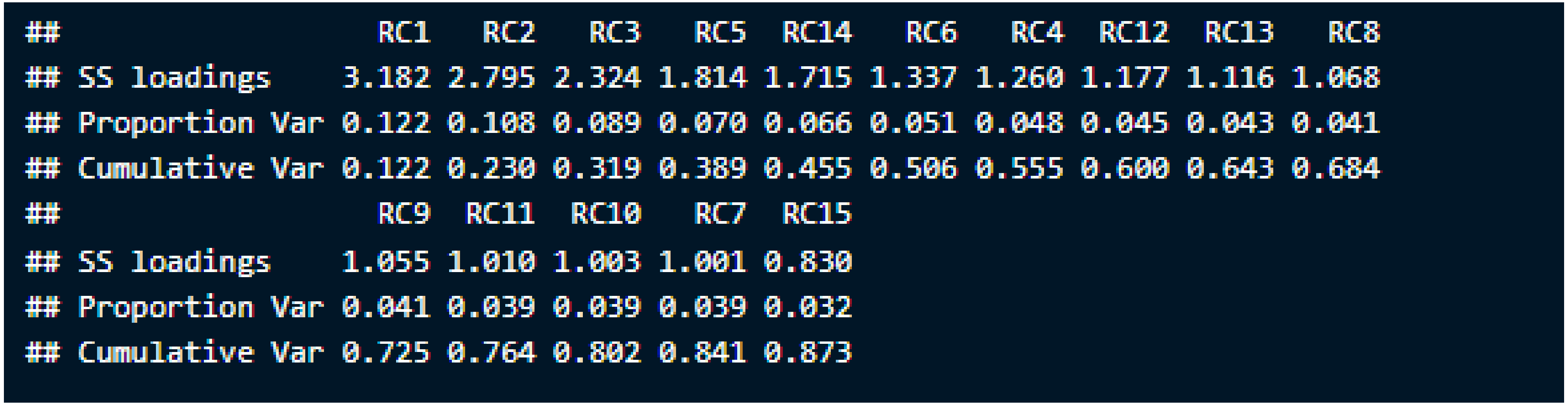


Principal Component Analysis



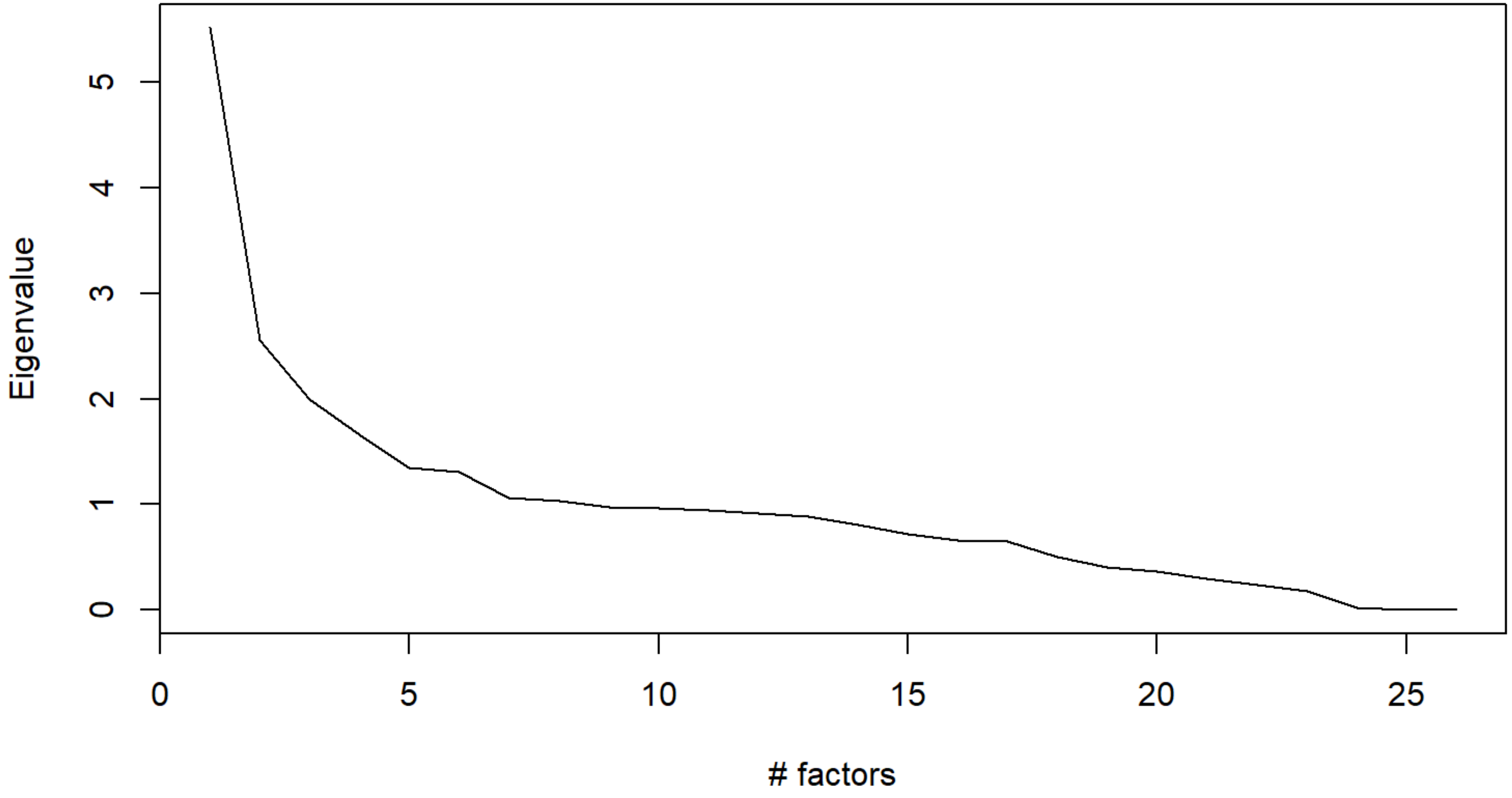
```
##
## Loadings:
##                                         RC1    RC2    RC3    RC5    RC14
## Value_of_Household_Vehicles              0.270                0.165  0.814
## Valuables                                              0.984
## Deposits                                 0.132  0.483  0.190  0.147 -0.186
## Mutual_Funds                                    0.233
## Bonds                                           0.882                0.143
## Employee_Income                          0.522               -0.209  0.211
## Self_Employment_income                   0.121         0.148  0.855  0.103
## Rental_Income                                   0.181
## Financial_assets_Income                  0.100  0.905
## Pension_Income                           0.298  0.130        -0.172
## Total_Real_Assets                        0.253  0.231  0.487  0.297  0.129
## Total_Financial_Assets                   0.118  0.851  0.120
## Total_Gross_Income                       0.737                       0.236
## Value_of_Self_employment_Businesses      0.131                0.861  0.129
## Income_From_Other_Sources
```

Principal Component Analysis



```
##                   RC1   RC2   RC3   RC5  RC14   RC6   RC4  RC12  RC13   RC8
## SS loadings     3.182 2.795 2.324 1.814 1.715 1.337 1.260 1.177 1.116 1.068
## Proportion Var 0.122 0.108 0.089 0.070 0.066 0.051 0.048 0.045 0.043 0.041
## Cumulative Var 0.122 0.230 0.319 0.389 0.455 0.506 0.555 0.600 0.643 0.684
##                   RC9  RC11  RC10   RC7  RC15
## SS loadings     1.055 1.010 1.003 1.001 0.830
## Proportion Var 0.041 0.039 0.039 0.039 0.032
## Cumulative Var 0.725 0.764 0.802 0.841 0.873
```

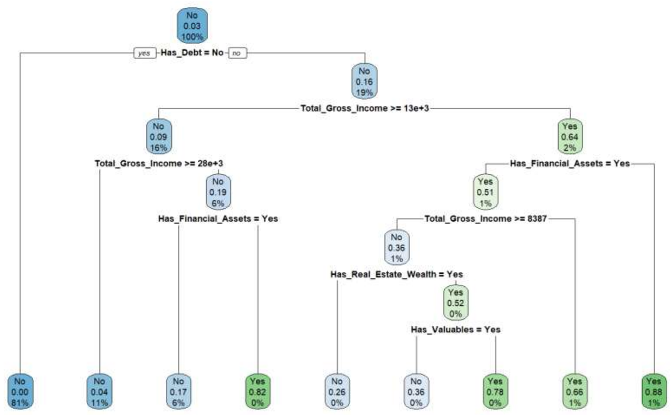
Principal Component Analysis

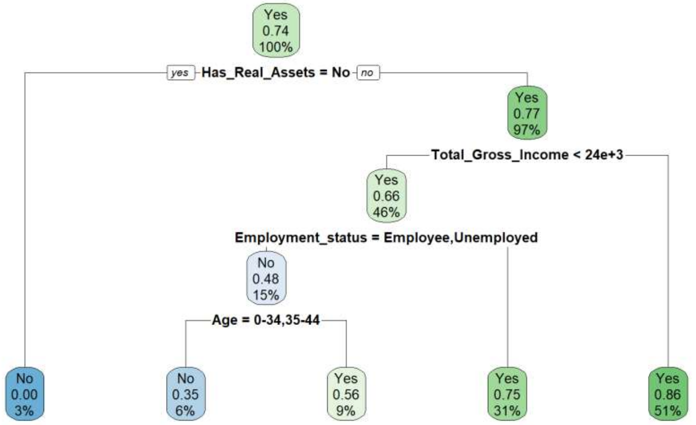


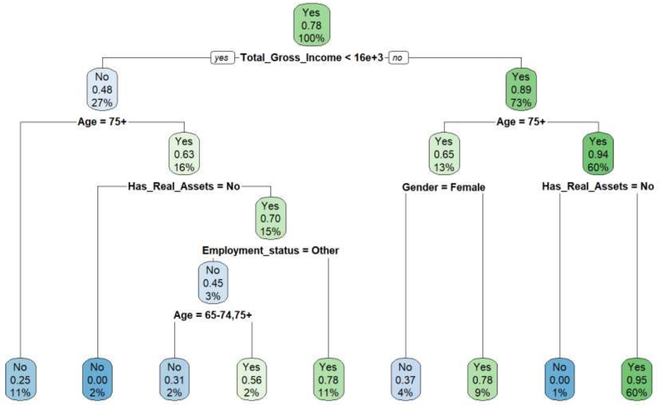
Decision Tree

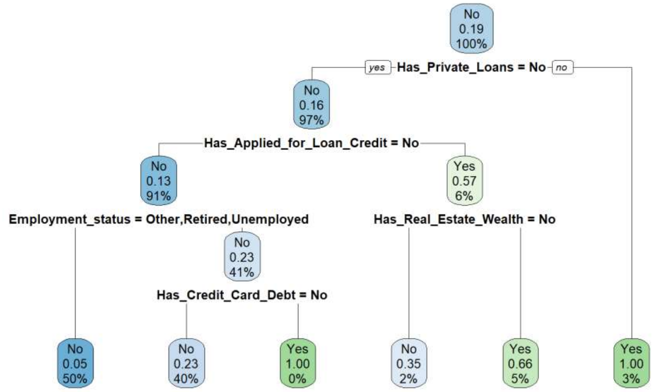
- Decision trees can be considered for the HCFS dataset because they are a useful technique for predicting outcomes based on a set of input variables. In the HCFS dataset, there are multiple variables related to household finance and consumption that may be associated with each other, such as income, assets, expenses, and debts. Decision trees can be used to identify the most important variables that predict financial outcomes, such as default on loans or bankruptcy, and can help policymakers design targeted interventions to prevent financial distress.
- For example, decision trees can be used to identify the most important predictors of loan default, such as income, credit score, or debt-to-income ratio, which can help lenders make more informed lending decisions and design loan products that are more accessible to low-income households. Decision trees can also be used to identify the most effective interventions to prevent financial distress, such as financial education, counseling, or targeted subsidies, based on the financial characteristics of households.
- Additionally, decision trees can be used to identify the most important factors that drive household financial behavior, such as attitudes towards saving, investment, or debt management, which can inform the design of financial education and counseling programs. Decision trees can also be used to identify the most effective communication channels for financial information, such as social media, mobile apps, or community events, based on the preferences of different households.
- Overall, decision trees can be a powerful tool for predicting financial outcomes and designing targeted interventions to support financial inclusion and economic growth. By identifying the most important predictors of financial behavior and outcomes, decision trees can help policymakers develop more effective policies and programs to support household financial well-being.

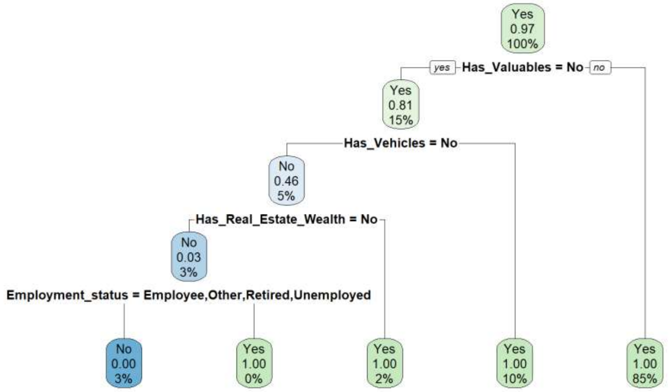
Decision Tree



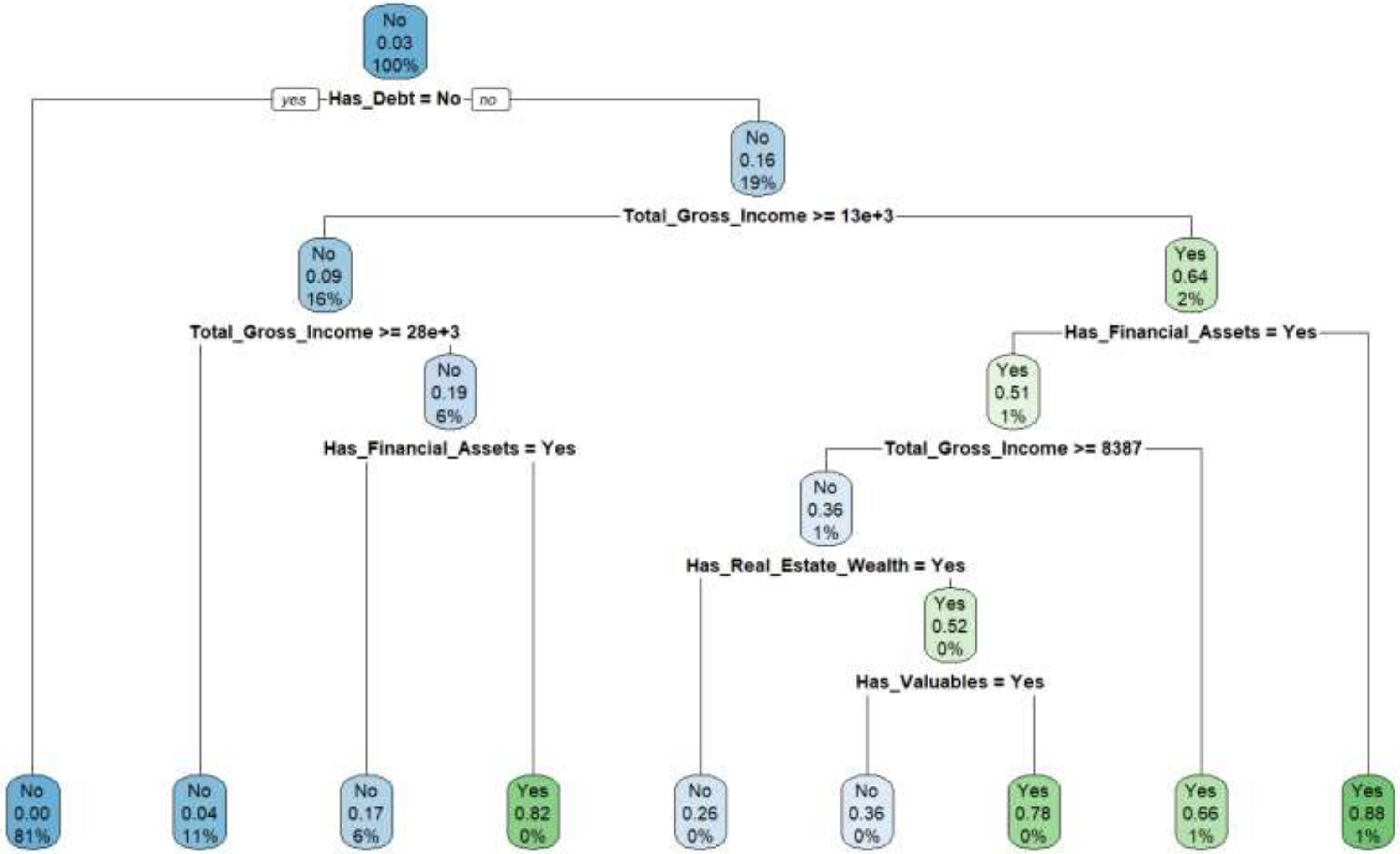




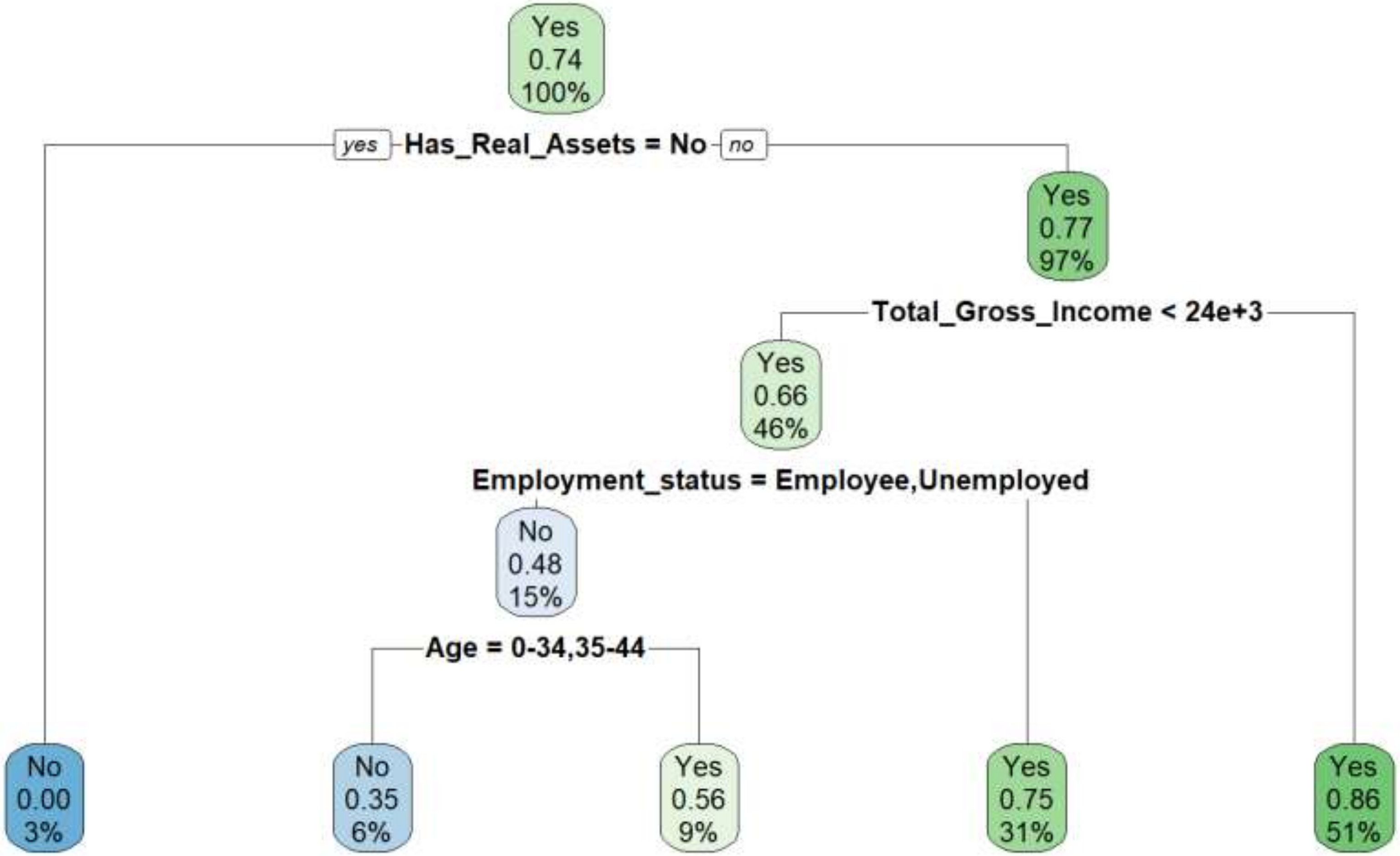




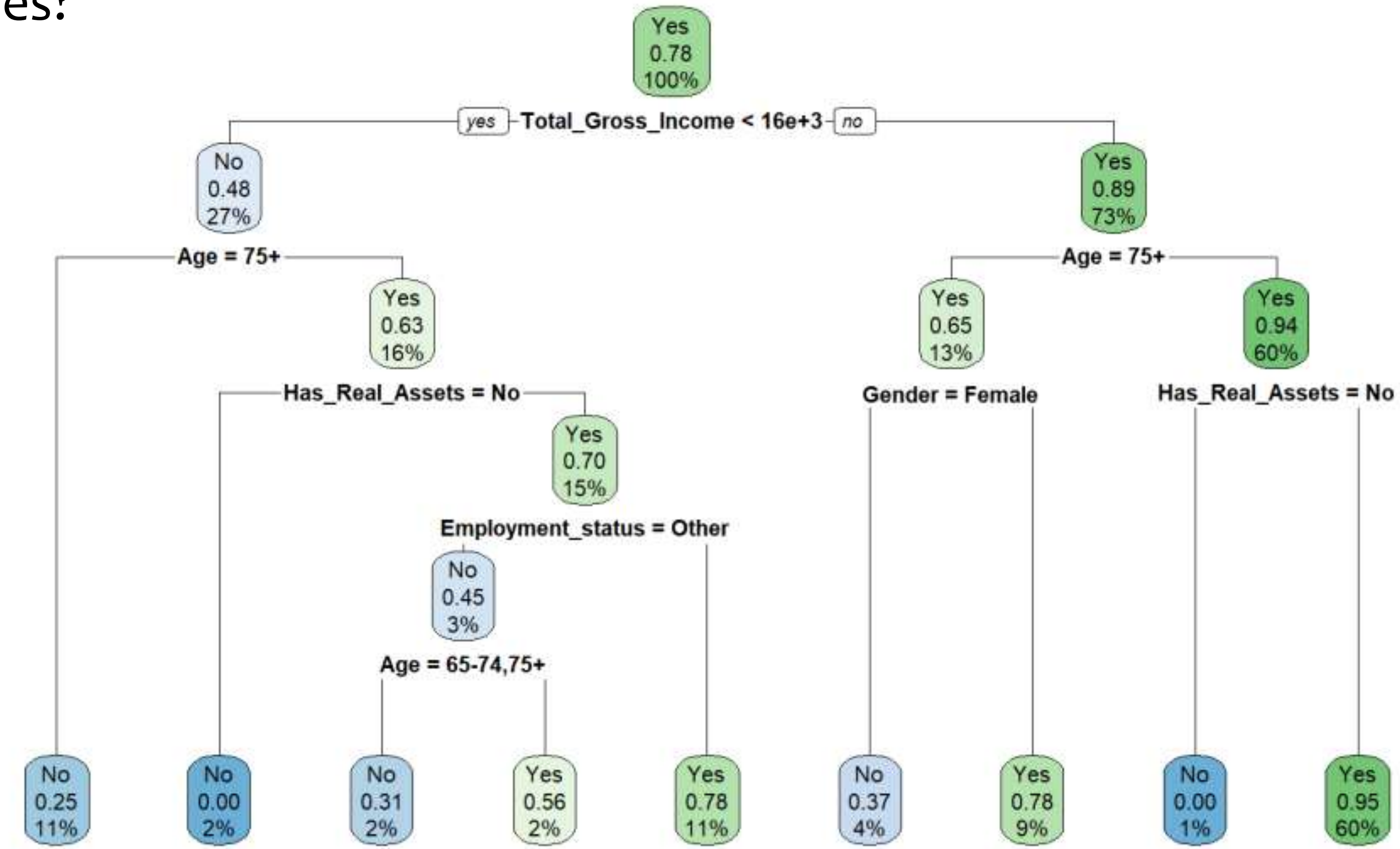
Who has Private Loans?



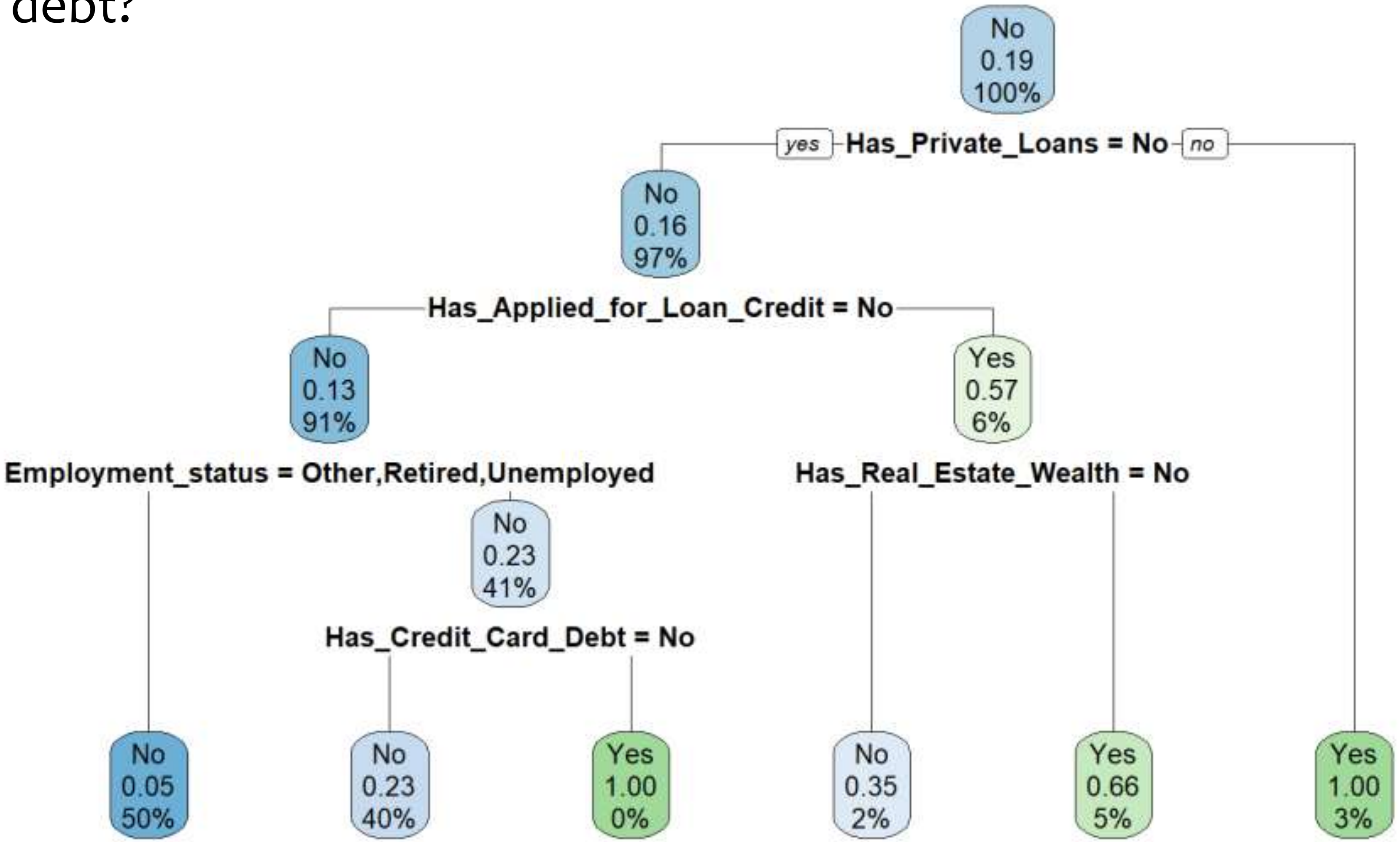
Who has Real Estate Wealth ?



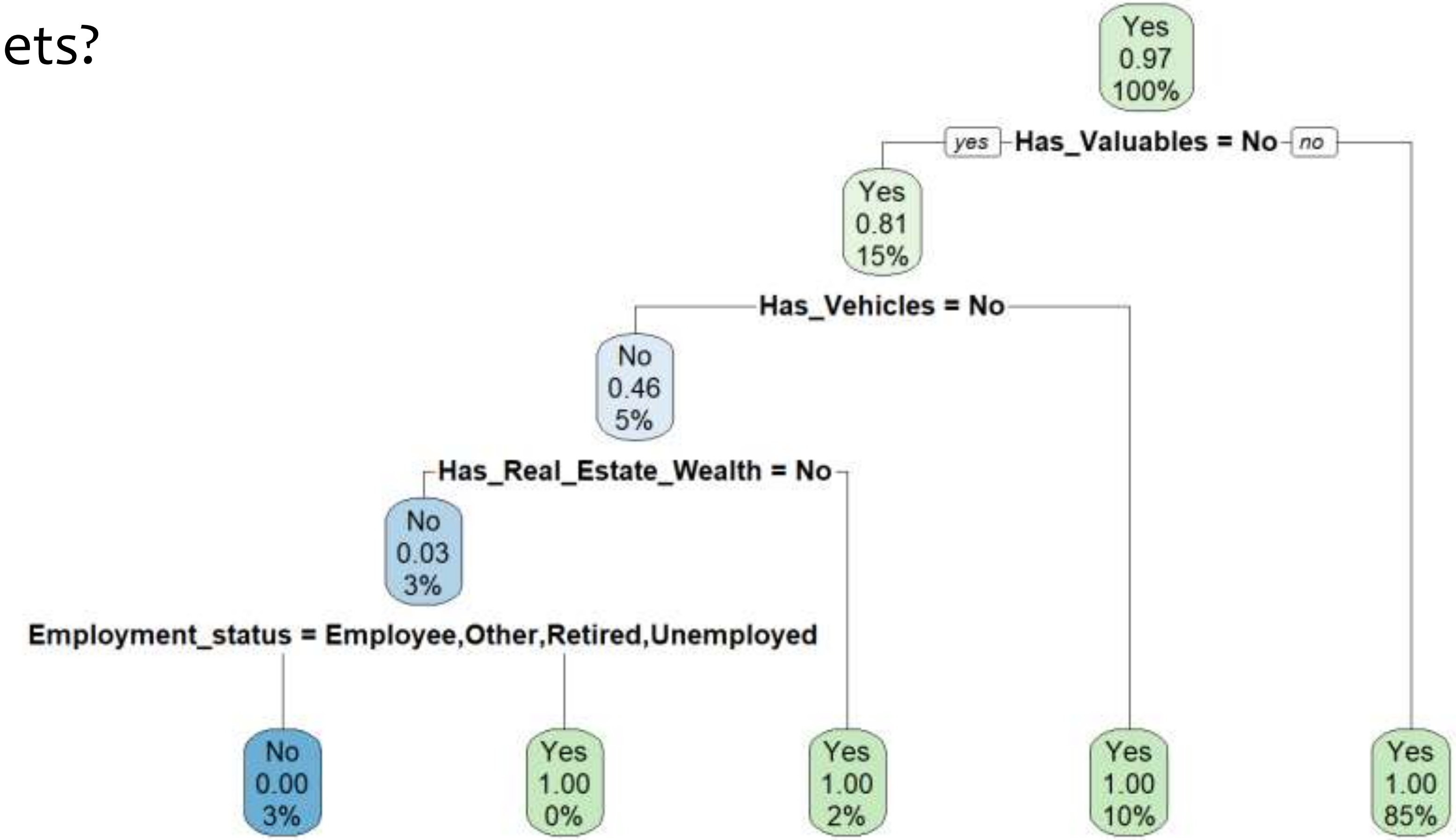
Who has Vehicles?



Who can have debt?


No
0.19
100%
yes
Has_Private_Loans = No
no
No
0.16
97%
Has_Applied_for_Loan_Credit = No
No
0.13
91%
Yes
0.57
6%
Employment_status = Other,Retired,Unemployed
No
0.23
41%
Has_Real_Estate_Wealth = No
Has_Credit_Card_Debt = No
No
0.05
50%
No
0.23
40%
Yes
1.00
0%
No
0.35
2%
Yes
0.66
5%
Yes
1.00
3%

Who has real assets?



Findings

chi cerca trova!

Findings

Gender

Female 2,903

Male 5,253

- The majority of the respondents, both female (66.9%) and male (54.8%), were not willing to take any financial risk.
- A small percentage of respondents, both female (8.5%) and male (12.3%), were willing to take above-average financial risks.
- Both female and male respondents were generally risk-averse regarding their investment attitude.

Females tend to save less and own fewer vehicles and businesses compared to males.

- 31.1% of the female respondents were aged 75+ and 4.4% fell in the age group of 0-34.
- 44% of the male respondents were of the age group 55-74 and 3.9% in the category of 0-34 years.

- 34.8% of the female respondents were found to have attained primary level of education with the lowest being 13.2% who attained first stage tertiary level of education.
- 33.7% of the male respondents had attained upper secondary education and 11.8% attained first stage tertiary level of education.

- Employed female respondents constituted of 33.3% and unemployed were found to be 2.6%.
- Majority of the male respondents were retired(43%) and only 3% were found to be unemployed.

Findings

Gender

✓

Among females, 2,891 have no credit card debt, while 12 have an average debt of 1,155.833. Among males, 5,201 have no credit card debt, while 52 have an average debt of 1,669.542.

✓

Females were more likely to be renters (32.4%) compared to males (25.5%) and males were more likely to be owners (65.4%) compared to females (61.8%).

✓

Males have higher average expenditures in all categories than females. Specifically, males spend on average 482.8249 more on food, 1,341.658 more on consumer goods, and 183.1321 more on utilities than females.

✓

On average, male-headed households have higher total gross income than female-headed households. Male-headed households also have higher average income from self-employment, rental, financial, and pension sources.

Findings

Education Level

✓

The highest count of individuals falls in the Education Level category of Lower Secondary (2,329) and Employment Status category of Employee (2,635).

✓

The lowest count of individuals falls in the Education Level category of First Stage Tertiary and Employment Status category of Unemployed (13).

✓

The highest count of individuals in Education Level category of First Stage Tertiary is employed in the Self-Employed category (181), while the highest count of individuals in Education Level category of Lower Secondary and Primary Education are employed in the Employee category (939 and 1,319, respectively).

✓

The highest percentage of individuals without private loans was observed among those with a first stage tertiary education (97.9%). On the other hand, the highest percentage of individuals with private loans was observed among those with lower secondary education (4.3%).

Findings

Hypothesis

✓ The null hypothesis that there is no difference between male and female groups was rejected for most of the dependent variables, indicating that there is a significant difference between genders in terms of total gross income, amount spent on utilities, consumer goods and services, employee income, self-employment income, financial assets income, rental income, credit card debt, value of self-employment businesses, and amount spent on food at home. However, the null hypothesis was accepted for pension income and income from other sources.

✓ There is a significant difference between the 6 categories of the independent variable Age and all of the dependent variables, indicating that age is a significant predictor of these variables.

✓ The null hypothesis that there is no difference between the categories of the independent variable Education_Level and the dependent variables is rejected for all variables except for Pension_Income.

Suggestions and Future research

The data available was for just 1 year. If we had the privilege of having the data for 3-5 years we could have done a comparative analysis between various years and analyzed if patterns and behavior of households changed over different time lines.

Suggestions and Future research

- For example, we could have analyzed, as the years progressed if there were an increase in the number of households getting into higher levels of education. And by getting higher education, if there is an increase in the Total Gross Income of the households and in turn if their savings were increasing and if they were investing more and purchasing more assets or if an increase in income is leading to an increase in the money spent on expenses.
- And if the above mentioned questions showed a positive impact, we could have inferred that the Italian households have been giving more importance to education over the years and their savings, investing attitude is securing them financially in a good way and their increased spending is also contributing positively to the Italian economy as the government would get more taxes, once the Total Income increase.

Thank You