Technical Report

An exploratory study on Household Finance and Consumption of Italian households

A group Project for the course of Statistical Data Analysis

By
Mushtari Khan, Laila Arzuman Ara, Yuva Priyanka Manda,
Mahadevan KS, Hekmatullah Himmat

Introduction

The Household Finance and Consumption Survey data from European survey data of Banca D'Italia provides a comprehensive view of household balance sheets and related economic and demographic variables. It contains a vast array of variables related to household financial information, demographics, and assets. Each row of the dataset represents a value based on households' earnings, and the variables described in the report refer to different groups of households based on demographic or economic characteristics.

The dataset includes information on household weight, age of the reference person, number of household members, type of household, real estate properties' value, income from various sources such as self-employment and pensions, as well as information on net worth and debt. Additionally, it provides detailed information on the financial and non-financial assets, liabilities, income, and consumption of households in Italy.

The data was collected by sampling households from each region of the country, making it a representative dataset. With a broad range of variables related to household finances, including employment status, income, savings, debt, property, assets, and consumption patterns, this dataset can be a valuable resource for researchers and policymakers analyzing household finances in Italy.

Dataset Link : Household Finance and Consumption Survey (https://www.bancaditalia.it/statistiche/tematiche/indagini-famiglie-imprese/bilanci-famiglie/dati-indagine-europea/index.html)

Objective

The objective of this exploratory analysis on the dataset is to gain a better understanding of the financial and non-financial characteristics of households in Italy. This analysis will aim to identify trends in household income, wealth, debt, and consumption, as well as explore the relationships between these variables and demographic and economic characteristics of households. By analyzing the dataset, we hope to generate insights about household finance and consumption in Italy.

Workflow

We divided the pipeline of Statistical Data analysis into different sprints like in agile methodology to achieve desired results. The following are the tasks that were done as part of the each sprint.

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
Sprint 4		Summarize the results from EDA
	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
		Feature Importance, PCA
		Regression
		Clustering, Decision Tree, Confusion Matrix
		Summarize the results from CDA

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.

The raw sample data of the files D1 and H1 from the survey is as follows,

ID	survey	SA0010	SA0100	IM0100	HW0010	DWHOHO	DHAGEH1B	DH0001	DH0006	DH0004	DHHTYPE	DH0002	DA
IT100000173001	1	173	IT	1	898.7334	179.7467	65	1	1	0	52	1.0	15(
IT100000375001	1	375	IT	1	3,652.3074	730.4615	85	1	1	0	52	1.0	15(
IT100000633001	1	633	IT	1	958.0087	191.6017	80	1	1	0	52	1.0	13(
IT100000923001	1	923	IT	1	682.1561	136.4312	80	1	1	0	52	1.0	28(
IT100001367001	1	1,367	IT	1	890.2372	178.0474	85	2	2	0	7	1.5	6(
IT100001763001	1	1,763	IT	1	5,538.9744	1,107.7949	65	2	2	0	7	1.5	25

ID	survey	SA0010	SA0100	IM0100	HW0010	HB0100	fHB0100	hb0100_B	fhb0100_b	HB0200	fHB0200	HB0300	fHB030
IT100000173001	1	173	IT	1	898.7334	160	1	9	1	10	1	1	
IT100000375001	1	375	IT	1	3,652.3074	120	1	8	1	44	1	1	
IT100000633001	1	633	IT	1	958.0087	100	1	7	1	49	1	1	
IT100000923001	1	923	IT	1	682.1561	180	1	9	1	46	1	1	
IT100001367001	1	1,367	IT	1	890.2372	90	1	6	1	49	1	1	
IT100001763001	1	1,763	IT	1	5,538.9744	45	1	3	1	37	1	1	

The following table lists outs the variables that were considered for this study and a short description about them derived from the meta data of the dataset.

Variable	Description
Number_of_Household_Members	Number of household members, all household members included
Number_of_Household_Members_in_Employment	Number of persons for which PE0100a (Labour status, main) = 1 ('doing regular work for pay/ self-employed/ family business') or 2 ('sick, maternity/other leave, planning to return to work').
Household_Type	House hold Composition ; 51 - One adult, younger than 65 years 52 - One adult, 65 years and over 6 - Two adults younger than 65 years 7 - Two adults, at least one aged 65 years and over 8 - Three or more adults 9 - Single parent with dependent children 10 - Two adults with one dependent child 11 - Two adults with two dependent children 12 - Two adults with three or more dependent children 13 - Three or more adults with dependent children
Value_of_Household_VehicleS	Represents the value of household vehicles
Valuables	Value of other valuables
Deposits	Value of Deposits
Mutual_Funds	Value of Mutual Funds
Bonds	Value of Bonds
Employee_Income	Total employee income of the household
Self_Employment_income	Sum of gross self employment income
Rental_Income	Rental income from real estate property
Has_Rental_Income	Has rental income from real estate property
Financial_assets_Income	Value of Income through Financial assests
Pension_Income	Income received as Pensions
Total_Real_Assets	Total real assets 1 (incl. business wealth, vehicles and valuables)

Total_Financial_Assets	Total financial assets 1 (excl. public and occupational pension plans)
Has_Real_Assets	Do you have Real Assets?
Has_Financial_Assets	Do you have Financial assests?
Has_Vehicles	Do you have Vehicle?
Has_Valuables	Do you have any other valuables?
Value_of_Self_employment_Businesses	Income received from Self employment business
Has_Real_Estate_Wealth	Do you have real estate wealth?
Has_Deposits	Do you have deposits?
Has_Mutual_Funds	Do you have mutual funds?
Has_Bonds	Do you have bonds?
Has_Shares	Do you have shares?
Has_Debt	Do you have debt?
Housing_Status	Households housing status 1 - Owner - outright 2 - Owner - with mortgage 3 - Renter/Other
Has_Employee_Income	Do you have income through employment?
Has_Self_Employee_Income	Do you have income through self employment?
Has_Financial_assets_Income	Do you have income through financial assets?
Has_Income_From_Pensions	Do you have income through pensions?
Income_From_Other_Sources	Value of income through other sources
Has_Income_From_Other_Sources	Do you income from other sources?
Credit_Card_Debt	Value of Credit card debt
Has_Credit_Card_Debt	Do you have credit card employment?
Way_Of_Acquring_Property	How (did you/your household) acquire the (part of the) residence (you own/your household owns): did you purchase it, did you construct it yourself, did you inherit it or did you receive it as a gift?
Monthly_Amount_Paid_As_Rent	What is the monthly amount paid as rent (please exclude utilities, heating, etc. if possible)?
Ownership_of_Cars	(Do you/Does anyone in your household) own any cars?
Total_Value_of_Cars	For the cars that you/your household own, if you sold them now, about how much do you think you could get?
Has_Other_Vehicles	(Do you/does anyone in your household) own any other type of vehicle, such as motorbikes, trucks, vans, planes, boats or yachts or any other vehicle such as trailers, caravans, etc.?

Value_Of_Other_Vehicles	If (you/your household) decided to sell (this vehicle/these vehicles) now, how much do you think you would get?
Ownership_Of_Other_Valuables	(Do you/Does you household) own any valuables such as jewellery, works of art, antiques, etc.?
Value_Of_Other_Valuables	In total, approximately how much do you think all these valuables would bring if you sold them?
Household_Has_a_Credit_Card	Do you or any other member of the household have credit cards other than ones paid by employers? (Do not consider here debit cards, i.e. cards where the money spent is immediately deducted from your bank account).
Has_Private_Loans	Do you have loans from relatives or friends that you are expected to repay?
No_of_PrivateLoans	Number of private loans a household has taken
Has_Applied_for_Loan_Credit	In the last three years, have you (or any member of your household) applied for a loan or other credit?
Household_Owns_Saving_accounts	Do you/does anyone in your household) have any saving accounts, time deposits, certificates of deposit or other such deposits?
Value_of_Saving_Accounts	Positive account balances are summed up as part of assets in HD1210
Investment_Attitudes	the amount of financial risk that you (and your husband/wife/partner) are willing to take when you save or make investments? 1- Take substantial financial risks expecting to earn substantial returns 2 - Take above average financial risks expecting to earn above average returns 3 - Take average financial risks expecting to earn average returns 4 - Not willing to take any financial risk
Amount_spent_on_Food_at_Home	How much does (you/your household) spend in a typical month on food and beverages at home?
Amount_Spent_on_Food_Outside_Home	How much does (you/your household) spend in a typical month on food and beverages outside the home? I mean expenses at restaurants, lunches, canteens, coffee shops and the like. Please, include only the amounts (you/your household) pay out i.e. net of any employer subsidy/discount/promotion etc.
AMount_Spent_on_Utilities	How much does your household spend on utilities (electricity, water, gas, telephone, internet and television) in a typical month?
Amount_Spent_on_Consumer_Goods_Services	How much does a household spend in a typical month on all consumer goods and services? Includes all household expenses including food, utilities, etc. but excluding consumer durables (e.g. cars, household appliances, etc.), rent, loan repayments, insurance policies, renovation, etc
Total_Gross_Income	Total gross annual household income aggregate.

Data Cleaning and Transformation

Variable Recoding

Filtering the required columns from the dataset files provided us a dataset with necessary columns that define the income and expenditure of the households. Since the data was raw, recoding and renaming was performed for most of the columns. Each row in the dataset represents a household and as there were no rows with many missing values data cleaning did not take much effort apart from skipping columns that do not have much relevant information to be considered.

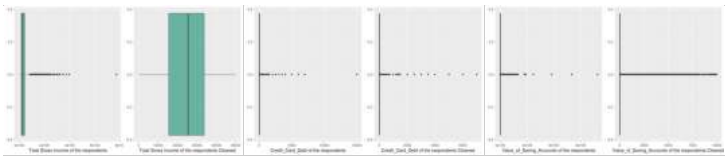
```
# Identify numeric columns
numeric_cols <- sapply(hcfs, is.numeric)

# Replace NAs with 0 in numeric columns
hcfs[numeric_cols][is.na(hcfs[numeric_cols])] <- 0

hcfs <- hcfs %>%
```

```
  rename(Gender = DHGENDERH1, Age = DHAGEH1, Education_Level = DHEDUH1) %>%
  filter(Gender %in% c(1,2)) %>%
  mutate(Gender = recode(Gender, `1` = "Male", `2` ="Female")) %>%
  mutate(Age = age_groups(Age, split_at = c(35, 45, 55, 65, 75), na.rm = FALSE)) %>%
  mutate(Education_Level = recode(Education_Level,
                                  `0` = "No formal education",
                                  `1` = "Primary education",
                                  `2` = "Lower secondary",
                                  `3` = "Upper secondary",
                                  `4` = "Post-secondary",
                                  `5` = "First stage tertiary",
                                  `6` = "Second stage tertiary"))

hcfs <- hcfs %>%
  rename(Employment_status = DHEMPH1) %>%
  mutate(Employment_status = recode(Employment_status,
                                  `1` = "Employee",
                                  `2` = "Self-employed",
                                  `3` = "Unemployed",
                                  `4` = "Retired",
                                  `5` = "Other"))

hcfs <- hcfs %>%
  rename(
     Number_of_Household_Members = DH0001,
     Number_of_Household_Members_in_Employment = DH0004,
     Household_Type =  DHHTYPE,
     Value_of_Household_Vehicles = DA1130,
     Valuables = DA1131,
     Deposits = DA2101,
     Mutual_Funds = DA2102,
     Bonds = DA2103,
     Employee_Income = DI1100,
     Self_Employment_income = DI1200,
     Rental_Income = DI1300,
     Has_Rental_Income = DI1300i,
     Financial_assets_Income = DI1400,
     Pension_Income =  DI1500,
     Total_Real_Assets =  DA1000,
     Total_Financial_Assets = DA2100,
     Has_Real_Assets = DA1000i,
     Has_Financial_Assets = DA2100i,
     Has_Vehicles = DA1130i,
     Has_Valuables = DA1131i,
     Value_of_Self_employment_Businesses = DA1140i,
     Has_Real_Estate_Wealth = DA1400i,
     Has_Deposits = DA2101i,
     Has_Mutual_Funds = DA2102i,
     Has_Bonds = DA2103i,
     Has_Shares = DA2105i,
     Has_Debt = DL1000i,
     Housing_Status = DHHST,
     Has_Employee_Income = DI1100i,
     Has_Self_Employee_Income = DI1200i,
     Has_Financial_assets_Income =  DI1400i,
     Has_Income_From_Pensions = DI1500i,
     Income_From_Other_Sources =  DI1800,
     Has_Income_From_Other_Sources =  DI1800i,
     Credit_Card_Debt = DL1220,
     Has_Credit_Card_Debt =  DL1220i,
     Way_Of_Acquring_Property = HB0600,
     Monthly_Amount_Paid_As_Rent = HB2300,
     Ownership_of_Cars = HB4300,
     Total_Value_of_Cars = HB4400,
     Has_Other_Vehicles = HB4500,
     Value_Of_Other_Vehicles =  HB4600,
     Ownership_Of_Other_Valuables = HB4700,
     Value_Of_Other_Valuables = HB4710,
     Household_Has_a_Credit_Card = HC0300,
     Has_Private_Loans = HC0330,
     No_of_PrivateLoans = HC0340,
     Has_Applied_for_Loan_Credit = HC1300,
     Household_Owns_Saving_accounts = HD1200,
     Value_of_Saving_Accounts = HD1210,
     Investment_Attitudes = HD1800,
     Amount_spent_on_Food_at_Home = HI0100,
     Amount_Spent_on_Food_Outside_Home = HI0200,
     AMount_Spent_on_Utilities = HI0210,
     Amount_Spent_on_Consumer_Goods_Services = HI0220,
     Total_Gross_Income=DI2000
  )

hcfs <- hcfs %>%
    mutate(Household_Owns_Saving_accounts = recode(Household_Owns_Saving_accounts, `1` = "Yes", `2` ="No", .default = "No")) %>%
    mutate(Ownership_of_Cars = recode(Ownership_of_Cars, `1` = "Yes", `2` ="No", .default = "No")) %>%
    mutate(Ownership_Of_Other_Valuables = recode(Ownership_Of_Other_Valuables, `1` = "Yes", `2` ="No", .default = "No")) %>%
    mutate(Has_Real_Assets = recode(Has_Real_Assets, `1` = "Yes", `2` ="No", .default = "No")) %>%
    mutate(Has_Financial_Assets = recode(Has_Financial_Assets, `1` = "Yes", `2` ="No", .default = "No")) %>%
    mutate(Has_Vehicles = recode(Has_Vehicles, `1` = "Yes", `2` ="No", .default = "No")) %>%
    mutate(Has_Valuables = recode(Has_Valuables, `1` = "Yes", `2` ="No", .default = "No")) %>%
    mutate(Has_Real_Estate_Wealth = recode(Has_Real_Estate_Wealth, `1` = "Yes", `2` ="No", .default = "No")) %>%
    mutate(Has_Deposits = recode(Has_Deposits, `1` = "Yes", `2` ="No", .default = "No")) %>%
    mutate(Has_Mutual_Funds = recode(Has_Mutual_Funds, `1` = "Yes", `2` ="No", .default = "No")) %>%
    mutate(Has_Bonds = recode(Has_Bonds, `1` = "Yes", `2` ="No", .default = "No")) %>%
    mutate(Has_Shares = recode(Has_Shares, `1` = "Yes", `2` ="No", .default = "No")) %>%
    mutate(Has_Debt = recode(Has_Debt, `1` = "Yes", `2` ="No", .default = "No")) %>%
    mutate(Has_Employee_Income = recode(Has_Employee_Income, `1` = "Yes", `2` ="No", .default = "No")) %>%
    mutate(Has_Self_Employee_Income = recode(Has_Self_Employee_Income, `1` = "Yes", `2` ="No", .default = "No")) %>%
    mutate(Has_Financial_assets_Income = recode(Has_Financial_assets_Income, `1` = "Yes", `2` ="No", .default = "No")) %>%
    mutate(Has_Income_From_Pensions = recode(Has_Income_From_Pensions, `1` = "Yes", `2` ="No", .default = "No")) %>%
    mutate(Has_Income_From_Other_Sources = recode(Has_Income_From_Other_Sources, `1` = "Yes", `2` ="No", .default = "No")) %>%
    mutate(Has_Credit_Card_Debt = recode(Has_Credit_Card_Debt, `1` = "Yes", `2` ="No", .default = "No")) %>%
    mutate(Has_Other_Vehicles = recode(Has_Other_Vehicles, `1` = "Yes", `2` ="No", .default = "No")) %>%
```

```
  mutate(Household_Has_a_Credit_Card = recode(Household_Has_a_Credit_Card, `1` = "Yes", `2` ="No", .default = "No")) %>%
  mutate(Has_Private_Loans = recode(Has_Private_Loans, `1` = "Yes", `2` ="No", .default = "No")) %>%
  mutate(Has_Applied_for_Loan_Credit = recode(Has_Applied_for_Loan_Credit, `1` = "Yes", `2` ="No", .default = "No")) %>%
  mutate(Has_Rental_Income = recode(Has_Rental_Income, `1` = "Yes", `2` ="No", .default = "No"))

hcfs <- hcfs %>%
 mutate(Way_Of_Acquring_Property = recode(Way_Of_Acquring_Property,
                        `1` = "Purchased",
                        `2` = "Own construction",
                        `3` = "Inherited",
                        `4` = "Gift",
                        .default = "Inherited"))


hcfs <- hcfs %>%
 mutate(Housing_Status = recode(Housing_Status,
                        `1` = "Owner",
                        `2` = "Owner with mortgage",
                        `3` = "Renter"))


hcfs <- hcfs %>%
 mutate(Investment_Attitudes = recode(Investment_Attitudes,
                        `1` = "Take substantial financial risks",
                        `2` = "Take above average financial risks",
                        `3` = "Take average financial risks",
                        `4` = "Not willing to take any financial risk"))


hcfs <- hcfs %>%
 mutate(Household_Type = recode(Household_Type,
                        `51` = "One adult, younger than 65 years",
                        `52` = "One adult, 65 years and over",
                        `6` = "Two adults younger than 65 years",
                        `7` = "Two adults, at least one aged 65 years and over",
                        `8` = "Three or more adults",
                        `9` = "Single parent with dependent children",
                        `10` = "Two adults with one dependent child",
                        `11` = "Two adults with two dependent children",
                        `12` = "Two adults with three or more dependent children",
                        `13` = "Three or more adults with dependent children"))
```

Outlier detection

To detect outliers in the dataset the boxplot of the major metric variables were considered and the outlier values were replaced by mean values.



Structure and Summaries

The dataset has 8156 rows with 62 columns after cleaning and recoding. The structure and the distribution of values from the cleaned data is as follows,

ID	Number_of_Household_Members	Number_of_Household_Members_in_Employment	Household_Type	Value_of_Household_Vehicles	Valuables	Dep
IT100000173001	1	0	One adult, 65 years and over	10,000	1,000	20,00
IT100000375001	1	0	One adult, 65 years and over	0	500	
IT100000633001	1	0	One adult, 65 years and over	1,000	2,000	60
IT100000923001	1	0	One adult, 65 years and over	0	200	50
IT100001367001	2	0	Two adults, at least one aged 65 years and over	0	1,000	13,00
IT100001763001	2	0	Two adults, at least one aged 65 years and over	1,000	0	8,62

```
## 'data.frame':    8156 obs. of  61 variables:
##  $ ID                                  : chr  "IT100000173001" "IT100000375001" "IT100000633001" "IT100000923001" ...
##  $ Number_of_Household_Members         : int  1 1 1 1 2 2 1 1 3 1 ...
##  $ Number_of_Household_Members_in_Employment: int  0 0 0 0 0 0 0 0 1 1 ...
##  $ Household_Type                      : chr  "One adult, 65 years and over" "One adult, 65 years and over" "One adult, 65 years and over" "One ad
ult, 65 years and over" ...
##  $ Value_of_Household_Vehicles         : num  10000 0 1000 0 0 1000 0 200 8500 0 ...
##  $ Valuables                           : num  1000 500 2000 200 1000 0 3000 5000 500 200 ...
##  $ Deposits                            : num  20000 0 600 500 13000 ...
##  $ Mutual_Funds                        : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Bonds                               : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Employee_Income                     : num  0 0 0 0 0 ...
##  $ Self_Employment_income              : num  0 0 0 0 0 ...
##  $ Rental_Income                       : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Financial_assets_Income             : num  172.05 0 5.16 4.3 111.83 ...
##  $ Pension_Income                      : num  31932 9345 15653 7150 12053 ...
##  $ Total_Real_Assets                   : num  211000 150500 133000 870200 61000 ...
##  $ Total_Financial_Assets              : num  20000 0 600 500 13000 ...
##  $ Total_Gross_Income                  : num  32104 9345 15658 10154 12165 ...
##  $ Has_Real_Assets                     : chr  "Yes" "Yes" "Yes" "Yes" ...
##  $ Has_Financial_Assets                : chr  "Yes" "Yes" "Yes" "Yes" ...
##  $ Has_Vehicles                        : chr  "Yes" "No" "Yes" "No" ...
##  $ Has_Valuables                       : chr  "Yes" "Yes" "Yes" "Yes" ...
##  $ Value_of_Self_employment_Businesses : int  0 0 0 0 0 0 0 0 0 1 ...
##  $ Has_Real_Estate_Wealth              : chr  "Yes" "Yes" "Yes" "Yes" ...
##  $ Has_Deposits                        : chr  "Yes" "Yes" "Yes" "Yes" ...
##  $ Has_Mutual_Funds                    : chr  "No" "No" "No" "No" ...
##  $ Has_Bonds                           : chr  "No" "No" "No" "No" ...
##  $ Has_Shares                          : chr  "No" "No" "No" "No" ...
##  $ Has_Debt                            : chr  "No" "No" "No" "No" ...
##  $ Has_Rental_Income                   : chr  "No" "No" "No" "No" ...
##  $ Housing_Status                      : chr  "Owner" "Owner" "Owner" "Owner" ...
##  $ Has_Employee_Income                 : chr  "No" "No" "No" "No" ...
##  $ Has_Self_Employee_Income            : chr  "No" "No" "No" "No" ...
##  $ Has_Financial_assets_Income         : chr  "Yes" "No" "Yes" "Yes" ...
##  $ Has_Income_From_Pensions            : chr  "Yes" "Yes" "Yes" "Yes" ...
##  $ Income_From_Other_Sources           : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Has_Income_From_Other_Sources       : chr  "No" "No" "No" "No" ...
##  $ Credit_Card_Debt                    : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Has_Credit_Card_Debt                : chr  "No" "No" "No" "No" ...
##  $ Age                                 : Ord.factor w/ 6 levels "0-34"<"35-44"<..: 5 6 6 6 6 5 6 4 4 4 ...
##  $ Education_Level                     : chr  "Upper secondary" "Primary education" "Primary education" "Primary education" ...
##  $ Employment_status                   : chr  "Retired" "Other" "Retired" "Other" ...
##  $ Gender                              : chr  "Male" "Female" "Female" "Female" ...
##  $ Way_Of_Acquring_Property            : chr  "Purchased" "Purchased" "Inherited" "Own construction" ...
##  $ Monthly_Amount_Paid_As_Rent         : num  0 0 0 0 0 0 318 118 0 400 ...
##  $ Ownership_of_Cars                   : chr  "Yes" "No" "Yes" "No" ...
##  $ Total_Value_of_Cars                 : num  10000 0 1000 0 0 1000 0 0 8500 0 ...
##  $ Has_Other_Vehicles                  : chr  "No" "No" "No" "No" ...
##  $ Value_Of_Other_Vehicles             : num  0 0 0 0 0 0 0 200 0 0 ...
##  $ Ownership_Of_Other_Valuables        : chr  "Yes" "Yes" "Yes" "Yes" ...
##  $ Value_Of_Other_Valuables            : num  1000 500 2000 200 1000 0 3000 5000 500 200 ...
##  $ Household_Has_a_Credit_Card         : chr  "No" "No" "No" "No" ...
##  $ Has_Private_Loans                   : chr  "No" "No" "No" "No" ...
##  $ No_of_PrivateLoans                  : num  0 0 0 0 0 0 0 0 0 1 ...
##  $ Has_Applied_for_Loan_Credit         : chr  "No" "No" "No" "No" ...
##  $ Household_Owns_Saving_accounts      : chr  "No" "No" "Yes" "No" ...
##  $ Value_of_Saving_Accounts            : num  0 0 600 0 4214 ...
##  $ Investment_Attitudes                : chr  "Take above average financial risks" "Not willing to take any financial risk" "Take average financial risk
s" "Not willing to take any financial risk" ...
##  $ Amount_spent_on_Food_at_Home        : int  300 300 400 350 250 300 200 250 600 250 ...
##  $ Amount_Spent_on_Food_Outside_Home   : int  200 0 30 0 50 0 0 50 100 0 ...
##  $ AMount_Spent_on_Utilities           : num  250 16.7 150 91.7 83.3 ...
##  $ Amount_Spent_on_Consumer_Goods_Services  : int  1000 790 1000 800 750 800 400 600 2500 400 ...
```

Data summary

Name	hcfs
Number of rows	8156
Number of columns	61
_____	
Column type frequency:	
character	32
factor	1
numeric	28
_____	
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
ID	0	1	14	14	0	8156	0
Household_Type	0	1	20	48	0	10	0
Has_Real_Assets	0	1	2	3	0	2	0
Has_Financial_Assets	0	1	2	3	0	2	0

Has_Vehicles	0	1	2	3	0	2	0
Has_Valuables	0	1	2	3	0	2	0
Has_Real_Estate_Wealth	0	1	2	3	0	2	0
Has_Deposits	0	1	2	3	0	2	0
Has_Mutual_Funds	0	1	2	3	0	2	0
Has_Bonds	0	1	2	3	0	2	0
Has_Shares	0	1	2	3	0	2	0
Has_Debt	0	1	2	3	0	2	0
Has_Rental_Income	0	1	2	3	0	2	0
Housing_Status	0	1	5	19	0	3	0
Has_Employee_Income	0	1	2	3	0	2	0
Has_Self_Employee_Income	0	1	2	3	0	2	0
Has_Financial_assets_Income	0	1	2	3	0	2	0
Has_Income_From_Pensions	0	1	2	3	0	2	0
Has_Income_From_Other_Sources	0	1	2	3	0	2	0
Has_Credit_Card_Debt	0	1	2	3	0	2	0
Education_Level	0	1	15	20	0	4	0
Employment_status	0	1	5	13	0	5	0
Gender	0	1	4	6	0	2	0
Way_Of_Acquring_Property	0	1	4	16	0	4	0
Ownership_of_Cars	0	1	2	3	0	2	0
Has_Other_Vehicles	0	1	2	3	0	2	0
Ownership_Of_Other_Valuables	0	1	2	3	0	2	0
Household_Has_a_Credit_Card	0	1	2	3	0	2	0
Has_Private_Loans	0	1	2	3	0	2	0
Has_Applied_for_Loan_Credit	0	1	2	3	0	2	0
Household_Owns_Saving_accounts	0	1	2	3	0	2	0
Investment_Attitudes	0	1	28	38	0	4	0

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
Age	0	1	TRUE	6	75+: 1917, 65-: 1687, 55-: 1679, 45-: 1602

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
Number_of_Household_Members	0	1	2.37	1.25	1	1.00	2.00	3.00	10.00	
Number_of_Household_Members_in_Employment	0	1	0.77	0.86	0	0.00	1.00	1.00	5.00	
Value_of_Household_Vehicles	0	1	6361.55	9529.77	0	500.00	4000.00	9000.00	275000.00	
Valuables	0	1	3741.27	16739.98	0	500.00	1500.00	3500.00	1000000.00	
Deposits	0	1	14581.40	48904.24	0	1000.00	5000.00	13500.00	2000000.00	
Mutual_Funds	0	1	3472.32	27768.04	0	0.00	0.00	0.00	1000001.00	
Bonds	0	1	6896.40	41288.50	0	0.00	0.00	0.00	2145209.43	
Employee_Income	0	1	14168.89	22396.91	0	0.00	0.00	22949.66	379084.70	
Self_Employment_income	0	1	5445.80	22072.92	0	0.00	0.00	0.00	761181.41	
Rental_Income	0	1	451.93	3214.03	0	0.00	0.00	0.00	120000.00	
Financial_assets_Income	0	1	399.98	1775.43	0	9.10	53.32	251.05	79133.90	
Pension_Income	0	1	12935.45	16120.91	0	0.00	8923.77	20374.16	128532.26	
Total_Real_Assets	0	1	221416.25	345197.20	0	29225.00	155500.00	272212.50	13600000.00	
Total_Financial_Assets	0	1	30715.41	113895.15	0	1352.10	6819.18	26590.27	5143000.00	
Total_Gross_Income	0	1	24905.01	11758.10	0	15379.55	25396.86	33734.56	49987.81	
Value_of_Self_employment_Businesses	0	1	0.15	0.35	0	0.00	0.00	0.00	1.00	

Income_From_Other_Sources	0	1	127.14	1349.77	0	0.00	0.00	0.00	53295.00	
Credit_Card_Debt	0	1	12.35	200.93	0	0.00	0.00	0.00	7000.00	
Monthly_Amount_Paid_As_Rent	0	1	61.35	152.31	0	0.00	0.00	0.00	1300.00	
Total_Value_of_Cars	0	1	5868.75	7749.18	0	100.00	4000.00	8000.00	150000.00	
Value_Of_Other_Vehicles	0	1	492.80	4478.09	0	0.00	0.00	0.00	250000.00	
Value_Of_Other_Valuables	0	1	3741.27	16739.98	0	500.00	1500.00	3500.00	1000000.00	
No_of_PrivateLoans	0	1	0.03	0.17	0	0.00	0.00	0.00	1.00	
Value_of_Saving_Accounts	0	1	1058.32	2178.47	0	0.00	0.00	0.00	10000.00	
Amount_spent_on_Food_at_Home	0	1	444.45	222.34	50	300.00	400.00	600.00	2000.00	
Amount_Spent_on_Food_Outside_Home	0	1	82.59	123.93	0	0.00	50.00	100.00	3000.00	
AMount_Spent_on_Utilities	0	1	171.88	134.30	0	83.33	166.67	250.00	2083.33	
Amount_Spent_on_Consumer_Goods_Services	0	1	1232.54	714.39	100	750.00	1000.00	1500.00	10000.00	

Cross tabulations

To get a quick overview of the distribution of the data and to identify any patterns or relationships that may exist, the following cross tabulations were drawn and the results have been discussed accordingly.

Age Category/Gender	Female	Male	Total
0-34	129 (4.4%)	206 (3.9%)	335 (4.1%)
35-44	299 (10.3%)	637 (12.1%)	936 (11.5%)
45-54	521 (17.9%)	1,081 (20.6%)	1,602 (19.6%)
55-64	521 (17.9%)	1,158 (22.0%)	1,679 (20.6%)
65-74	531 (18.3%)	1,156 (22.0%)	1,687 (20.7%)
75+	902 (31.1%)	1,015 (19.3%)	1,917 (23.5%)

Table 1 : Age vs Gender

The table displays the number and percentage of individuals by age and gender category ranging from 0-34 to 75+ years. It can be seen that the respondents are majorly Males and of age group 55-64.

Level of Education/Gender	Female	Male	Total
First stage tertiary	384 (13.2%)	620 (11.8%)	1,004 (12.3%)
Lower secondary	644 (22.2%)	1,685 (32.1%)	2,329 (28.6%)
Primary education	1,010 (34.8%)	1,178 (22.4%)	2,188 (26.8%)
Upper secondary	865 (29.8%)	1,770 (33.7%)	2,635 (32.3%)

Table 2 : Education level vs Gender

The table displays the distribution of education levels among females and males. We observe that the majority of females have completed primary education (34.8%), followed by lower secondary (22.2%), while the majority of males have completed upper secondary (33.7%), followed by lower secondary (32.1%).

Employment_status/Gender	Female	Male	Total
Employee	968 (33.3%)	2,007 (38.2%)	2,975 (36.5%)

Other	798 (27.5%)	127 (2.4%)	925 (11.3%)
Retired	898 (30.9%)	2,260 (43.0%)	3,158 (38.7%)
Self-employed	163 (5.6%)	703 (13.4%)	866 (10.6%)
Unemployed	76 (2.6%)	156 (3.0%)	232 (2.8%)

Table 3 : Employment status vs Gender

The above table presents the distribution of employment status among females and males in the population being studied. The results indicate that a higher proportion of males are employed compared to females (38.2% vs. 33.3%). On the other hand, a higher proportion of females are retired compared to males (30.9% vs. 43.0%). Additionally, a small proportion of both males and females are self-employed or unemployed.

Education_Level/Employment_status	Employee	Other	Retired	Self-employed	Unemployed	Total
First stage tertiary	546	12	252	181	13	1,004
Lower secondary	939	198	792	283	117	2,329
Primary education	156	614	1,319	58	41	2,188
Upper secondary	1,334	101	795	344	61	2,635

Table 4 : Education Level vs Employment Status

The table presents the cross-tabulation between Education Level and Employment Status. 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).

Investment Attitude/Gender	Female	Male	Total
Not willing to take any financial risk	1,942 (66.9%)	2,881 (54.8%)	4,823 (59.1%)
Take above average financial risks	247 (8.5%)	644 (12.3%)	891 (10.9%)
Take average financial risks	708 (24.4%)	1,683 (32.0%)	2,391 (29.3%)
Take substantial financial risks	6 (0.2%)	45 (0.9%)	51 (0.6%)

Table 5 : Investment Attitude by Gender

The Cross tabulation presents the distribution of Investment Attitude by Gender. 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. The percentage of females who were willing to take average financial risks (24.4%) was slightly higher than males (32.0%). Finally, a negligible percentage of respondents, both female (0.2%) and male (0.9%), were willing to take substantial financial risks. Overall, the results suggest that both female and male respondents were generally risk-averse regarding their investment attitude.

Number of Household Members/Number of Household Members in Employment	0	1	2	3	4	5
1	1,732	662	0	0	0	0
2	1,637	667	284	0	0	0
3	296	646	505	53	0	0
4	124	450	537	94	17	0
5	43	141	111	34	9	2
6	10	30	30	10	5	1

6	10	39	20	10	5	1
7	6	10	2	3	0	0
8	2	1	1	0	0	0
9	0	1	0	0	0	0
10	0	0	1	0	0	0

Table 6 : Frequency of households by the number of household members vs number of household members in employment

The table represents the frequency of households by the number of household members and the number of household members in employment. The majority of households have no members in employment, and this is more common among females. As the number of household members in employment increases, the frequency of households decreases. The highest frequency of households is observed in the category of one household member with one household member in employment (662), followed by two household members with two household members in employment (537).

Gender	N	Employement	Self	Rental	Financial	Pension	Total_Gross_Income
Female	2,903	10,287.83	3,407.447	334.9150	283.8476	11,585.23	20,823.63
Male	5,253	16,313.70	6,572.261	516.6009	464.1626	13,681.64	27,160.53

Table 7 : Gender and their Mean Income

The table represents the income from different sources of the households based on gender. There are 2,903 households with a female head and 5,253 households with a male head. 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. The difference in total gross income between male and female-headed households may be due to various factors such as differences in education, work experience, and job opportunities. However, without further analysis, it is difficult to draw any definite conclusions.

Gender	N	Food	Consumer_Goods	Utilities
Female	2,903	374.9983	1,035.087	151.5144
Male	5,253	482.8249	1,341.658	183.1321

Table 8 : Gender and their expenditure

The table presents the average expenditures on food, consumer goods, and utilities for females and males. From the table, we can see that 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.

Household_Type	N	Food	Consumer_Goods	Utilities
One adult, 65 years and over	1,494	294.4311	820.3313	138.4351
One adult, younger than 65 years	900	264.2100	833.2056	123.4713
Single parent with dependent children	222	359.8649	971.9144	149.6734
Three or more adults	1,052	582.5095	1,572.0608	204.5954
Three or more adults with dependent children	452	597.4115	1,571.5044	192.7240
Two adults with one dependent child	653	492.7259	1,431.0628	192.0776
Two adults with three or more dependent children	198	604.0404	1,475.7576	181.7508
Two adults with two dependent children	727	549.4498	1,551.8982	202.6058

Two adults younger than 65 years	714	437.1429	1,273.7759	175.5048
Two adults, at least one aged 65 years and over	1,744	476.8291	1,280.2993	180.2190

Table 9 : Household Type and their expenditure

The cross-tabulation result presented in the table shows the distribution of Household_Type in the hcfs dataset with respect to three expenditure categories: Food, Consumer_Goods, and Utilities. The table also provides the number of observations (N) in each category. From the results, we can observe that the category with the largest number of observations is "Two adults, at least one aged 65 years and over" with 1,744 observations, while the category with the smallest number of observations is "Two adults with three or more dependent children" with only 198 observations.

```
## `summarise()` has grouped output by 'Gender'. You can override using the
## `.groups` argument.
```

Gender	Investment_Attitudes	N	Employement	Self	Rental	Financial	Pension	Total_Gross_Income
Female	Not willing to take any financial risk	1,942	8,501.438	1,875.688	146.1869	185.6105	11,414.420	19,047.92
Female	Take above average financial risks	247	13,064.831	6,729.370	559.6761	415.9566	9,573.682	21,380.09
Female	Take average financial risks	708	14,084.207	6,416.304	774.4678	500.9052	12,774.641	25,397.86
Female	Take substantial financial risks	6	26,189.951	7,389.354	300.0000	1,028.6376	9,329.190	32,893.97
Male	Not willing to take any financial risk	2,881	13,668.282	4,721.677	284.1855	239.1875	13,319.018	25,885.58
Male	Take above average financial risks	644	18,437.799	9,590.465	756.4887	783.5533	13,569.556	27,321.31
Male	Take average financial risks	1,683	19,980.768	8,394.089	787.3876	697.6551	14,440.865	29,313.07
Male	Take substantial financial risks	45	18,132.896	13,720.563	1,835.8667	1,564.1136	10,106.111	25,978.98

Table 10 : Investment Attitude with Gender and their mean Income

The table presents investment attitudes and total gross income of males and females in four categories of investment attitudes (Not willing to take any financial risk, Take above average financial risks, Take average financial risks, and Take substantial financial risks). From the table, it can be seen that:

Males tend to have a higher total gross income than females across all categories of investment attitudes. Both males and females who are willing to take above average or substantial financial risks tend to have higher total gross income than those who are not willing to take any financial risks. Males tend to have higher income in the categories of Take above average financial risks and Take substantial financial risks, while females tend to have higher income in the category of Take average financial risks.

Housing_Status/Gender	Female	Male	Total
Owner	1,795 (61.8%)	3,436 (65.4%)	5,231 (64.1%)
Owner with mortgage	166 (5.7%)	478 (9.1%)	644 (7.9%)
Renter	942 (32.4%)	1,339 (25.5%)	2,281 (28.0%)

Table 11 : Gender and Housing status

From the cross tabulation of gender and housing status, it can be inferred that the majority of respondents were owners (64.1%) followed by renters (28%) and those with a mortgage (7.9%). 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%). When looking at investment attitudes, a larger percentage of females were not willing to take any financial risks (56.9%) compared to males (52.1%). However, a larger percentage of males were willing to take above-average financial risks (21.9%) compared to females (10.8%).

Has_Private_Loans/Education_Level	First stage tertiary	Lower secondary	Primary education	Upper secondary	Total
No	983 (97.9%)	2,228 (95.7%)	2,136 (97.6%)	2,572 (97.6%)	7,919 (97.1%)

No	983 (97.9%)	2,228 (95.7%)	2,136 (97.6%)	2,372 (97.6%)	7,919 (97.1%)
Yes	21 (2.1%)	101 (4.3%)	52 (2.4%)	63 (2.4%)	237 (2.9%)

Table 12 : Private loans by Education Level

The table represents the distribution of individuals based on their education level and whether they have private loans. It can be observed that a vast majority of individuals with different education levels do not have private loans. 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%).

```
## `summarise()` has grouped output by 'Gender'. You can override using the
## `.groups` argument.
```

Gender	Has_Credit_Card_Debt	N	Debt
Female	No	2,891	0.000
Female	Yes	12	1,155.833
Male	No	5,201	0.000
Male	Yes	52	1,669.542

Table 13 : Who has credit card debit?

This table shows the relationship between gender and having credit card debt, as well as the amount of debt for those who have it. 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.

```
## `summarise()` has grouped output by 'Gender', 'Education_Level'. You can
## override using the `.groups` argument.
```

Gender	Education_Level	Age	N	Savings	Vehicles	Business
Female	First stage tertiary	0-34	38	1,504.18504	5,643.4211	0.15789474
Female	First stage tertiary	35-44	68	1,616.13056	9,181.4706	0.26470588
Female	First stage tertiary	45-54	104	1,014.89171	9,559.3846	0.29807692
Female	First stage tertiary	55-64	101	1,356.14387	10,817.3960	0.17821782
Female	First stage tertiary	65-74	48	1,547.08783	5,378.1250	0.06250000
Female	First stage tertiary	75+	25	1,579.19583	4,164.0000	0.04000000
Female	Lower secondary	0-34	30	881.47529	2,153.3333	0.16666667
Female	Lower secondary	35-44	83	292.73803	6,420.4819	0.12048193
Female	Lower secondary	45-54	143	581.13098	5,044.3357	0.18881119
Female	Lower secondary	55-64	145	1,355.99080	5,226.6207	0.08965517
Female	Lower secondary	65-74	142	1,272.99020	2,409.8592	0.02816901
Female	Lower secondary	75+	101	1,283.24751	1,251.0891	0.02970297
Female	Primary education	0-34	2	0.00000	1,790.0000	0.00000000

Female	Primary education	35-44	9	473.13545	2,088.8889	0.00000000
Female	Primary education	45-54	30	225.63807	1,916.6667	0.10000000
Female	Primary education	55-64	80	850.58586	3,589.7500	0.03750000
Female	Primary education	65-74	210	892.45802	1,299.5333	0.02380952
Female	Primary education	75+	679	1,274.00131	428.7069	0.00736377
Female	Upper secondary	0-34	59	484.85266	3,692.5254	0.11864407
Female	Upper secondary	35-44	139	1,043.91575	6,704.9640	0.15107914
Female	Upper secondary	45-54	244	815.20489	6,512.8689	0.15983607
Female	Upper secondary	55-64	195	1,169.10265	7,463.3846	0.16410256
Female	Upper secondary	65-74	131	1,117.80542	3,968.8550	0.05343511
Female	Upper secondary	75+	97	1,241.76545	1,326.7010	0.02061856
Male	First stage tertiary	0-34	38	657.70545	8,897.3684	0.18421053
Male	First stage tertiary	35-44	106	1,190.48946	10,458.3019	0.36792453
Male	First stage tertiary	45-54	136	1,305.27686	13,926.1029	0.42647059
Male	First stage tertiary	55-64	148	1,040.88955	13,406.3514	0.34459459
Male	First stage tertiary	65-74	129	1,452.11997	11,596.1938	0.25581395
Male	First stage tertiary	75+	63	1,020.31785	7,983.3333	0.15873016
Male	Lower secondary	0-34	68	130.15139	4,734.1176	0.08823529
Male	Lower secondary	35-44	227	747.03433	6,462.4141	0.17621145
Male	Lower secondary	45-54	435	806.93091	7,964.8276	0.22988506
Male	Lower secondary	55-64	434	836.39428	7,612.5783	0.24193548
Male	Lower secondary	65-74	314	1,307.01788	5,931.9427	0.06687898
Male	Lower secondary	75+	207	1,332.23324	3,662.2754	0.03381643
Male	Primary education	0-34	6	0.00000	2,841.6667	0.16666667
Male	Primary education	35-44	17	65.19703	2,070.5882	0.11764706
Male	Primary education	45-54	49	136.57966	3,440.0000	0.12244898
Male	Primary education	55-64	158	1,064.51704	6,178.7975	0.17088608

Male	Primary education	65-74	378	1,156.10350	5,374.2857	0.10582011
Male	Primary education	75+	570	1,357.05538	2,944.0982	0.01929825
Male	Upper secondary	0-34	94	822.45969	7,127.9787	0.17021277
Male	Upper secondary	35-44	287	1,023.07805	8,768.8850	0.16376307
Male	Upper secondary	45-54	461	1,027.79348	12,224.7722	0.29501085
Male	Upper secondary	55-64	418	840.13717	9,359.5024	0.23923445
Male	Upper secondary	65-74	335	1,188.00964	9,153.1970	0.16119403
Male	Upper secondary	75+	175	1,000.65858	6,104.9143	0.08571429

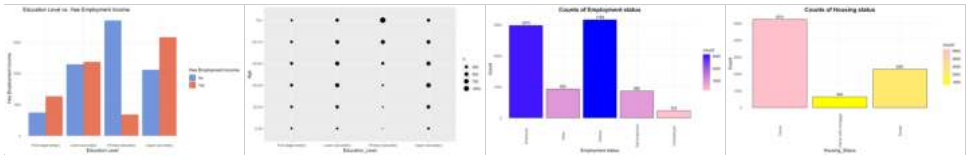
Table 14 : Who has savings, vehicles, and business ownership?

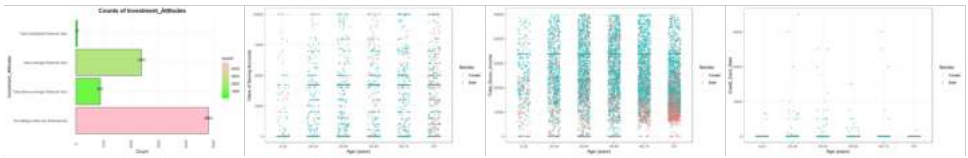
The table displays the summary statistics for savings, vehicles, and business ownership for different demographic groups, including gender, education level, and age. In general, females tend to save less and own fewer vehicles and businesses compared to males. Additionally, individuals with higher education levels tend to have more savings and own more vehicles and businesses compared to those with lower education levels. Finally, older individuals tend to have more savings and own more vehicles and businesses compared to younger individuals.

Data Visualization

Graphs

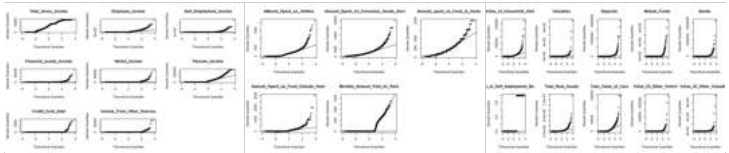
The following graphs for the data were drawn to visualize the dataset and obtain insights.





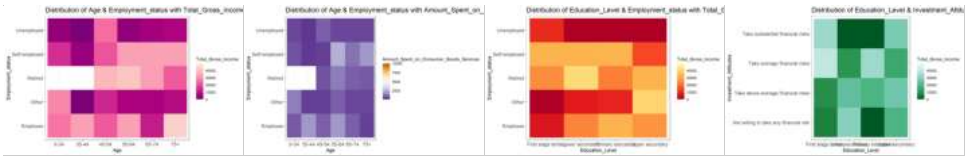
Q-Q Plots

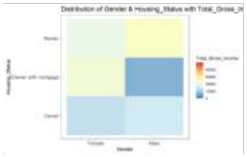
To obtain a series of QQ (Quantile-Quantile) plots for the data we considered different subsets of data. Subset 1 includes income-related variables, subset 2 includes variables related to household expenses, and subset 3 includes variables related to household assets. For each subset, a loop is used to create multiple QQ plots, one for each variable in the subset. The resulting QQ plots were used to check whether the data follow a normal distribution.



Tiles

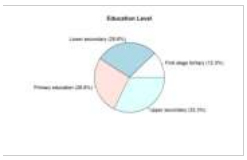
The ggplot graphs, display the distribution of a different set of variables. The graphs display the relationship between two variables with a color-coded tile. The color of the tile represents the value of a third variable Total_Gross_Income, Amount_Spent_on_Consumer_Goods_Services, or Investment_Attitudes



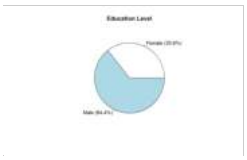


Pie Charts

```
# Create a table of education levels
edu_table <- table(hcfs$Education_Level)

# Plot a pie chart with the frequency count and percentage labels
pie(edu_table, main = "Education Level", labels = paste(names(edu_table), " (", round(100*edu_table/sum(edu_table),1), "%)", sep = ""))
```

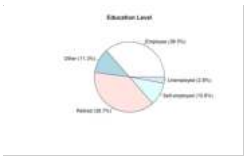


```
# Create a table of education levels
edu_table <- table(hcfs$Gender)

# Plot a pie chart with the frequency count and percentage labels
pie(edu_table, main = "Education Level", labels = paste(names(edu_table), " (", round(100*edu_table/sum(edu_table),1), "%)", sep = ""))
```



```
# Create a table of education levels
edu_table <- table(hcfs$Investment_Attitudes)

# Plot a pie chart with the frequency count and percentage labels
pie(edu_table, main = "Education Level", labels = paste(names(edu_table), " (", round(100*edu_table/sum(edu_table),1), "%)", sep = ""))
```

```
# Create a table of education levels
edu_table <- table(hcfs$Employment_status)

# Plot a pie chart with the frequency count and percentage labels
pie(edu_table, main = "Education Level", labels = paste(names(edu_table), " (", round(100*edu_table/sum(edu_table),1), "%)", sep = ""))
```

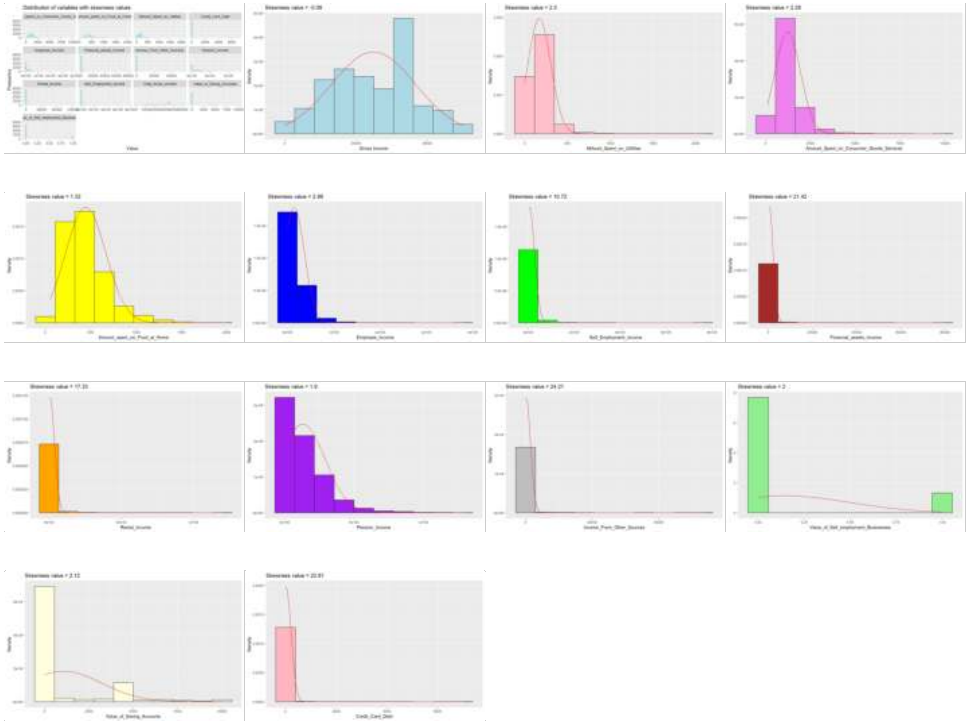


Skewness

To understand how the data is distributed, the shape and center we have computed the skew values of the major metric columns and plotted the following histograms to visualize them.

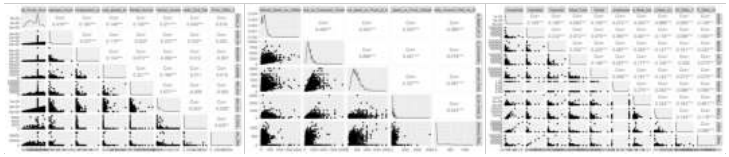
```
##                Total_Gross_Income            AMount_Spent_on_Utilities
##                       -0.08913195                           2.50346726
## Amount_Spent_on_Consumer_Goods_Services                   Employee_Income
##                           2.28148430                           2.99082831
##              Self_Employment_income              Financial_assets_Income
##                          10.72115257                          21.42052488
##    Value_of_Self_employment_Businesses                    Pension_Income
##                           1.99882600                           1.90081580
##          Amount_spent_on_Food_at_Home                   Rental_Income
##                           1.32405321                          17.32616287
##                 Credit_Card_Debt             Value_of_Saving_Accounts
##                          22.80546422                           2.12286317
##            Income_From_Other_Sources
##                          24.21245486
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

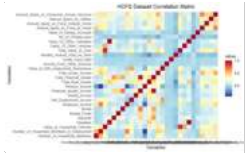


Correlation

The following set of scatter plots show the pairwise relationships between the numeric variables in the dataset. Considering three subsets of the dataset that include different combinations of variables related to income, expenses, and assets the plots visually identify patterns and correlations.

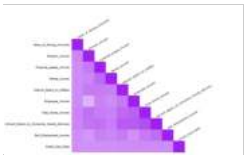


To further visualize the relationship between the variables correlation matrix, and correlation plot using the corrplot package are drawn.



	Total_Gross_Income	AMount_Spent_on_Utilities	Amount_Spent_on_Consumer_Goods_Services	Employee_Income	Self_Employment_i
Total_Gross_Income	1.0000000	0.3377350	0.5648751	0.4186945	0.1826003
AMount_Spent_on_Utilities	0.3377350	1.0000000	0.4557343	0.1940775	0.1776377
Amount_Spent_on_Consumer_Goods_Services	0.5648751	0.4557343	1.0000000	0.4438192	0.2858661
Employee_Income	0.4186945	0.1940775	0.4438192	1.0000000	-0.0374893
Self_Employment_income	0.1826003	0.1776377	0.2858661	-0.0374893	1.0000000
Financial_assets_Income	0.1475132	0.1753224	0.2763094	0.1185655	0.1339790
Rental_Income	0.1049226	0.1274115	0.1417755	0.0219828	0.0725096

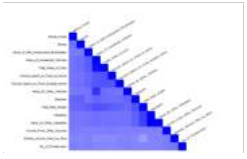
Pension_Income	0.2712045	0.1862575	0.2389673	-0.3372886	-0.0818505
Credit_Card_Debt	0.0395201	0.0007820	0.0147197	0.0296488	0.0115530
Value_of_Saving_Accounts	0.0858689	0.0373934	0.0794036	0.0241524	0.0046256



The correlation matrix shows the pairwise correlations between the variables in the dataset. The correlation coefficient ranges from -1 to 1, where a value of 1 indicates a perfect positive correlation, a value of 0 indicates no correlation, and a value of -1 indicates a perfect negative correlation.

- 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.
- The variables Employee Income and Pension Income have a weak positive correlation with Total Gross Income 0.36 and 0.27, respectively, while Self Employment Income and Financial Assets Income have a weak positive correlation with Total Gross Income 0.23 and 0.15, respectively.
- The variables Credit Card Debt and Value of Saving Accounts have weak positive correlations with Total Gross Income 0.04 and 0.09, respectively.

	Value_of_Household_Vehicles	Valuables	Deposits	Mutual_Funds	Bonds	Value_of_Self_employment_Businesses	Total_Rea
Value_of_Household_Vehicles	1.0000000	0.1386356	0.1951227	0.0925590	0.1498322	0.2718373	0.340057
Valuables	0.1386356	1.0000000	0.2320160	0.0725267	0.0754184	0.0895533	0.490791
Deposits	0.1951227	0.2320160	1.0000000	0.2019328	0.2248474	0.0871600	0.352278
Mutual_Funds	0.0925590	0.0725267	0.2019328	1.0000000	0.1058151	0.0569125	0.176647
Bonds	0.1498322	0.0754184	0.2248474	0.1058151	1.0000000	0.0459544	0.191033
Value_of_Self_employment_Businesses	0.2718373	0.0895533	0.0871600	0.0569125	0.0459544	1.0000000	0.274775
Total_Real_Assets	0.3400578	0.4907918	0.3522789	0.1766471	0.1910334	0.2747753	1.000000
Income_From_Other_Sources	0.0011845	0.0065577	-0.0035471	0.0213245	0.0168866	-0.0004089	-0.00337
Monthly_Amount_Paid_As_Rent	-0.0923150	-0.0507739	-0.0706288	-0.0374539	-0.0466830	-0.0128197	-0.22306
Total_Value_of_Cars	0.8856921	0.1194925	0.1468712	0.1021025	0.1419004	0.2831667	0.324039
Value_Of_Other_Vehicles	0.5954281	0.0882509	0.1610826	0.0202889	0.0733025	0.0884834	0.162933
Value_Of_Other_Valuables	0.1386356	1.0000000	0.2320160	0.0725267	0.0754184	0.0895533	0.490791
No_of_PrivateLoans	-0.0468559	-0.0256304	-0.0430908	-0.0195802	-0.0260860	0.0005682	-0.05832
Amount_spent_on_Food_at_Home	0.3098402	0.1098660	0.1272698	0.0837952	0.0916031	0.1714202	0.275064
Amount_Spent_on_Food_Outside_Home	0.3050785	0.1048881	0.1355410	0.0979383	0.1339986	0.2137338	0.230040



The correlation matrix suggests that the 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).

The other variables have low or moderate correlations with each other. For example, Income from Other Sources has a very low correlation with most other variables, while No of Private Loans has a moderate positive correlation with Valuables (0.23).

Hypothesis Testing

Hypothesis Statements

To perform Hypothesis testing on the dataset the following set of questions were considered to test whether there is a significant difference between two groups or whether there is a significant relationship between two variables. Based on previous computations of data distribution and skew values the parametric and non parametric tests were selected.

Hypothesis Statements of Independent Variable Gender

S.No	Dependent Variable	Statistical Question	Null Hypothesis	Alternative Hypothesis	Test

1	Total_Gross_Income	Is there a significant difference in Total Gross income between males and females?	There is no difference between the Male and Female groups with respect to the dependent variable Total Gross Income.	There is a difference between the Male and Female groups with respect to the dependent variable Total Gross Income.	T test
2	AMount_Spent_on_Utilities	Does the Amount Spent on Utilities significantly differ between males and females?	There is no difference between the Male and Female groups with respect to the dependent variable Amount spent on utilities.	There is a difference between the Male and Female groups with respect to the dependent variable Amount spent on utilities	Wilcoxon-Mann-Whitney test
3	Amount_Spent_on_Consumer_Goods_Services	Is there a significant difference in the amount spent on Consumer goods and services among males and females?	There is no difference between the male and female groups with respect to the dependent variable Amount spent on Consumer goods and services	There is difference between the male and female groups with respect to the dependent variable Amount spent on Consumer goods and services	Wilcoxon-Mann-Whitney test
4	Employee_Income	Does the dependent variable Employee income show a significant difference between males and females?	There is no difference between the Male and Female groups with respect to the dependent variable Employee_Income.	There is a difference between the Male and Female groups with respect to the dependent variable Employee_Income.	Wilcoxon-Mann-Whitney test
5	Self_Employment_income	Is there a significant difference in self-employee income between males and females?	There is no difference between the Male and Female groups with respect to the dependent variable Self_Employee_Income	There is a difference between the Male and Female groups with respect to the dependent variable Self_Employee_Income	Wilcoxon-Mann-Whitney test
6	Financial_assets_Income	Does the Financial asset income significantly differ between males and females?	There is no difference between the Male and Female groups with respect to the dependent variable Financial_assets_Income	There is a difference between the Male and Female groups with respect to the dependent variable Financial_assets_Income	Wilcoxon-Mann-Whitney test
7	Pension_Income	Is there a significant difference in the pension income earned between males and females?	There is no difference between the Male and Female groups with respect to the dependent variable Pension_Income.	There is no difference between the Male and Female groups with respect to the dependent variable Pension_Income.	Wilcoxon-Mann-Whitney test
8	Rental_Income	Does the Rental income earned significantly differ between males and females?	There is no difference between the Male and Female groups with respect to the dependent variable Rental_Income.	There is a difference between the Male and Female groups with respect to the dependent variable Rental_Income.	Wilcoxon-Mann-Whitney test
9	Credit_Card_Debt	Does the pattern of Credit Card debt significantly differ between males and females?	There is no difference between the Male and Female groups with respect to the dependent variable Credit_Card_Debt.	There is a difference between the Male and Female groups with respect to the dependent variable Credit_Card_Debt.	Wilcoxon-Mann-Whitney test
10	Value_of_Saving_Accounts	Is there a significant difference in the value of saving account held between males and females?	There is no difference between the Male and Female groups with respect to the dependent variable Value_of_Saving_Accounts.	There is a difference between the Male and Female groups with respect to the dependent variable Value_of_Saving_Accounts.	Wilcoxon-Mann-Whitney test
		Is there a significant difference in the value of self-employment			

11	Value_of_Self_employment_Businesses	business held between males and females?	There is no difference between the Male and Female groups with respect to the dependent variable Value_of_Self_employment_Businesses.	There is a difference between the Male and Female groups with respect to the dependent variable Value_of_Self_employment_Businesses.	Wilcoxon-Mann-Whitney test
12	Amount_spent_on_Food_at_Home	Does the Amount Spent on Food at home significantly differ between males and females?	There is no difference between the Male and Female groups with respect to the dependent variable Amount_spent_on_Food_at_Home.	There is a difference between the Male and Female groups with respect to the dependent variable Amount_spent_on_Food_at_Home.	Wilcoxon-Mann-Whitney test
13	Income_From_Other_Sources	Is there a significant difference in the income earned from other sources between males and females?	There is no difference between the Male and Female groups with respect to the dependent variable Income_From_Other_Sources.	There is a difference between the Male and Female groups with respect to the dependent variable Income_From_Other_Sources.	Wilcoxon-Mann-Whitney test

Hypothesis Statements of Independent Variable Age

S.No	Dependent Variable	Statistical Question	Null Hypothesis	Alternative Hypothesis	Test
1	Total_Gross_Income	Is there a statistically significant difference in Total Gross Income across different age groups?	There is no difference between the 6 categories of the independent variable Age with respect to the dependent variable Total_Gross_Income.	There is a difference between the 6 categories of the independent variable Age with respect to the dependent variable Total_Gross_Income.	ANOVA
2	AMount_Spent_on_Utilities	Is there a statistically significant change in the amount spent on utilities across different age groups?	There is no difference between the 6 categories of the independent variable Age with respect to the dependent variable Amount_Spent_on_Utilities.	There is a difference between the 6 categories of the independent variable Age with respect to the dependent variable Amount_Spent_on_Utilities.	Kruskal-Wallis rank sum test
3	Amount_Spent_on_Consumer_Goods_Services	Does the amount spent on consumer goods & services show statistically significant variation with age?	There is no difference between the 6 categories of the independent variable Age with respect to the dependent variable Amount_Spent_on_Consumer_Goods_Services.	There is no difference between the 6 categories of the independent variable Age with respect to the dependent variable Amount_Spent_on_Consumer_Goods_Services.	Kruskal-Wallis rank sum test
4	Employee_Income	Is there a statistically significant difference in employee income across different age groups?	There is no difference between the 6 categories of the independent variable Age with respect to the dependent variable Employee_Income.	There is a difference between the 6 categories of the independent variable Age with respect to the dependent variable Employee_Income.	Kruskal-Wallis rank sum test
5	Self_Employment_income	Is there a statistically significant difference in self-employment income between different age groups?	There is no difference between the 6 categories of the independent variable Age with respect to the dependent variable Self_Employment_income.	There is a difference between the 6 categories of the independent variable Age with respect to the dependent variable Self_Employment_income.	Kruskal-Wallis rank sum test
6	Financial_assets_Income	Does the Financial assets income show statistically significant difference across different age groups?	There is no difference between the 6 categories of the independent variable Age with respect to the dependent variable Financial_assets_Income.	There is a difference between the 6 categories of the independent variable Age with respect to the dependent variable Financial_assets_Income.	Kruskal-Wallis rank sum test
		Is there a statistically significant			

7	Rental_Income	change in rental income across different age groups?	There is no difference between the 6 categories of the independent variable Age with respect to the dependent variable Rental_Income.	There is a difference between the 6 categories of the independent variable Age with respect to the dependent variable Rental_Income.	Kruskal-Wallis rank sum test
8	Credit_Card_Debt	Does Credit Card debt show statistically significant difference across different age groups?	There is no difference between the 6 categories of the independent variable Age with respect to the dependent variable Credit_Card_Debt.	There is a difference between the 6 categories of the independent variable Age with respect to the dependent variable Credit_Card_Debt.	Kruskal-Wallis rank sum test
9	Value_of_Saving_Accounts	Is there a statistically significant difference in the value of savings account among different age groups?	There is no difference between the 6 categories of the independent variable Age with respect to the dependent variable Value_of_Saving_Accounts.	There is a difference between the 6 categories of the independent variable Age with respect to the dependent variable Value_of_Saving_Accounts.	Kruskal-Wallis rank sum test
10	Value_of_Self_employment_Businesses	Is there a statistically significant difference in the value of self emplomentbusiness among different age groups?	There is no difference between the 6 categories of the independent variable Age with respect to the dependent variable Value_of_Self_employment_Businesses.	There is a difference between the 6 categories of the independent variable Age with respect to the dependent variable Value_of_Self_employment_Businesses.	Kruskal-Wallis rank sum test
11	Amount_spent_on_Food_at_Home	Is there a statistically significant change in the amount spent on food across different age groups?	There is no difference between the 6 categories of the independent variable Age with respect to the dependent variable Amount_spent_on_Food_at_Home.	There is a difference between the 6 categories of the independent variable Age with respect to the dependent variable Amount_spent_on_Food_at_Home.	Kruskal-Wallis rank sum test
12	Income_From_Other_Sources	Is there a statistically significant difference in Income from other sources across different age groups?	There is no difference between the 6 categories of the independent variable Age with respect to the dependent variable Income_From_Other_Sources.	There is a difference between the 6 categories of the independent variable Age with respect to the dependent variable Income_From_Other_Sources.	Kruskal-Wallis rank sum test

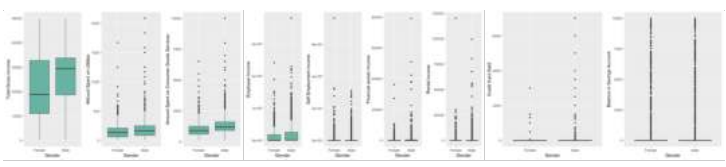
Hypothesis Statements of Independent Variable Education Level

S.No	Dependent Variable	Statistical Question	Null Hypothesis	Alternative Hypothesis	Test
1	Total_Gross_Income	Is there a significant relationship between education level and total gross income?	There is no difference between the 4 categories of the independent variable Education_Level with respect to the dependent variable Total_Gross_Income.	There is a difference between the 4 categories of the independent variable Education_Level with respect to the dependent variable Total_Gross_Income.	ANOVA
2	AMount_Spent_on_Utilities	Is there a significant difference in utility expenses among different education levels?	There is no difference between the 4 categories of the independent variable Education_Level with respect to the dependent variable Amount_Spent_on_Utilities.	There is a difference between the 4 categories of the independent variable Education_Level with respect to the dependent variable Amount_Spent_on_Utilities.	Kruskal-Wallis rank sum test
3	Amount_Spent_on_Consumer_Goods_Services	Is there a significant difference in consumer goods and services expenses among different education levels?	There is no difference between the 4 categories of the independent variable Education_Level with respect to the dependent variable Amount_spent_on_consumer_goods.	There is a difference between the 4 categories of the independent variable Education_Level with respect to the dependent variable Amount_spent_on_consumer_goods.	Kruskal-Wallis rank sum test
4	Employee_Income	Does education level have a significant	There is no difference between the 4 categories of the independent variable Education_Level with respect to the dependent variable Employee_Income.	There is a difference between the 4 categories of the independent variable Education_Level with respect to the dependent variable Employee_Income.	Kruskal-Wallis rank sum test

		impact on employee income?			
5	Self_Employment_income	Is there a significant difference in self-employment income among different education levels?	There is no difference between the 4 categories of the independent variable Education_Level with respect to the dependent variable Self_Employment_income.	There is a difference between the 4 categories of the independent variable Education_Level with respect to the dependent variable Self_Employment_income.	Kruskal-Wallis rank sum test
6	Financial_assets_Income	Is there a significant relationship between education level and financial asset income?	There is no difference between the 4 categories of the independent variable Education_Level with respect to the dependent variable Financial_assets_Income.	There is a difference between the 4 categories of the independent variable Education_Level with respect to the dependent variable Financial_assets_Income.	Kruskal-Wallis rank sum test
7	Pension_Income	Is there a significant difference in pension income among different education levels?	There is no difference between the 4 categories of the independent variable Education_Level with respect to the dependent variable Pension_Income.	There is a difference between the 4 categories of the independent variable Education_Level with respect to the dependent variable Pension_Income.	Kruskal-Wallis rank sum test
8	Credit_Card_Debt	Is there a significant difference in Credit Card Debt among different education levels?	There is no difference between the 4 categories of the independent variable Education_Level with respect to the dependent variable Credit_Card_Debt.	There is a difference between the 4 categories of the independent variable Education_Level with respect to the dependent variable Credit_Card_Debt.	Kruskal-Wallis rank sum test
9	Value_of_Saving_Accounts	Is there a significant relationship between education level and value of savings account?	There is no difference between the 4 categories of the independent variable Education_Level with respect to the dependent variable Value_of_Saving_Accounts.	There is a difference between the 4 categories of the independent variable Education_Level with respect to the dependent variable Value_of_Saving_Accounts.	Kruskal-Wallis rank sum test
10	Value_of_Self_employment_Businesses	Is there a significant relationship between education level and value of self employment business?	There is no difference between the 4 categories of the independent variable Education_Level with respect to the dependent variable Value_of_Self_employment_Businesses.	There is a difference between the 4 categories of the independent variable Education_Level with respect to the dependent variable Value_of_Self_employment_Businesses.	Kruskal-Wallis rank sum test
11	Amount_spent_on_Food_at_Home	Is there a significant relationship between education level and amount spent on food at home?	There is no difference between the 4 categories of the independent variable Education_Level with respect to the dependent variable Amount_spent_on_Food_at_Home.	There is a difference between the 4 categories of the independent variable Education_Level with respect to the dependent variable Amount_spent_on_Food_at_Home.	Kruskal-Wallis rank sum test
12	Income_From_Other_Sources	Is there a significant relationship between education level and income earned from other sources?	There is no difference between the 4 categories of the independent variable Education_Level with respect to the dependent variable Income_From_Other_Sources.	There is a difference between the 4 categories of the independent variable Education_Level with respect to the dependent variable Income_From_Other_Sources.	Kruskal-Wallis rank sum test

T-tests

The following box plots show how the means of the metric variables for male and female.



```
t.test(Total_Gross_Income ~ Gender, data = hcfs)
```

```
##
##  Welch Two Sample t-test
##
## data:  Total_Gross_Income by Gender
## t = -23.467, df = 5521.1, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group Female and group Male is not equal to 0
## 95 percent confidence interval:
##  -6866.277 -5807.520
## sample estimates:
## mean in group Female   mean in group Male
##             20823.63             27160.53
```

We conducted a Welch Two Sample t-test to determine if there is a difference in the Total Gross Income between males and females. The test was performed with a significance level of 0.05. The test showed that the t-value was -23.467 with a degrees of freedom (df) of 5521.1 and a p-value of less than 2.2e-16, which is much smaller than the significance level. This indicates strong evidence against the null hypothesis and suggests that there is a statistically significant difference in the means of Total Gross Income between males and females. The 95 percent confidence interval for the difference in means ranged from -6866.277 to -5807.520. The sample mean for females was 20823.63 and for males it was 27160.53.

Non-Parametric Tests with Gender

```
hcfs_subset <- hcfs[, c("AMount_Spent_on_Utilities", "Amount_Spent_on_Consumer_Goods_Services", "Employee_Income", "Self_Employment_I
ncome", "Financial_assets_Income", "Value_of_Self_employment_Businesses", "Pension_Income", "Amount_spent_on_Food_at_Home", "Rental_I
ncome", "Credit_Card_Debt", "Value_of_Saving_Accounts", "Income_From_Other_Sources", "Gender")]

cols_of_interest <- c("AMount_Spent_on_Utilities", "Amount_Spent_on_Consumer_Goods_Services", "Employee_Income",
                      "Self_Employment_income", "Financial_assets_Income", "Value_of_Self_employment_Businesses",
                      "Pension_Income", "Amount_spent_on_Food_at_Home", "Rental_Income", "Credit_Card_Debt",
                      "Value_of_Saving_Accounts", "Income_From_Other_Sources")

# Perform Wilcoxon-Mann-Whitney test for each column
for(col in cols_of_interest) {
  test_res <- hcfs_subset %>%
    wilcox.test(formula = as.formula(paste(col, "~ Gender")), data = .)

  print(paste0("Column: ", col))
  print(test_res)
}
```

```
## [1] "Column: AMount_Spent_on_Utilities"
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  AMount_Spent_on_Utilities by Gender
## W = 6497246, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
##
## [1] "Column: Amount_Spent_on_Consumer_Goods_Services"
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Amount_Spent_on_Consumer_Goods_Services by Gender
## W = 5289522, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
##
## [1] "Column: Employee_Income"
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Employee_Income by Gender
## W = 6492337, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
##
## [1] "Column: Self_Employment_income"
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Self_Employment_income by Gender
## W = 6947133, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
##
## [1] "Column: Financial_assets_Income"
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Financial_assets_Income by Gender
## W = 6668344, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
##
## [1] "Column: Value_of_Self_employment_Businesses"
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Value_of_Self_employment_Businesses by Gender
## W = 6962701, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
##
## [1] "Column: Pension_Income"
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Pension_Income by Gender
## W = 7612983, p-value = 0.9046
## alternative hypothesis: true location shift is not equal to 0
##
## [1] "Column: Amount_spent_on_Food_at_Home"
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Amount_spent_on_Food_at_Home by Gender
## W = 5153925, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
##
## [1] "Column: Rental_Income"
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Rental_Income by Gender
## W = 7509804, p-value = 0.002774
## alternative hypothesis: true location shift is not equal to 0
##
## [1] "Column: Credit_Card_Debt"
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Credit_Card_Debt by Gender
## W = 7580774, p-value = 0.004729
## alternative hypothesis: true location shift is not equal to 0
##
## [1] "Column: Value_of_Saving_Accounts"
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Value_of_Saving_Accounts by Gender
## W = 7699595, p-value = 0.3272
## alternative hypothesis: true location shift is not equal to 0
##
## [1] "Column: Income_From_Other_Sources"
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Income_From_Other_Sources by Gender
## W = 7585036, p-value = 0.201
## alternative hypothesis: true location shift is not equal to 0
```

Each test is a Wilcoxon rank sum test with continuity correction that is used to compare two groups: male and female, for each variable. The null hypothesis is that there is no difference between the two groups and the alternative hypothesis is that there is a difference between the two groups.For all variables except "Pension Income", the p-value is less than 0.05, which means that we reject the null hypothesis and conclude that

there is a significant difference between the two groups for these variables. The p-values for "Pension Income", "Value of Saving Accounts" and "Income From Other Sources" are greater than 0.05, which means that we fail to reject the null hypothesis and conclude that there is no significant difference between the two groups for these variables. Numeric values for the test statistic (W) and p-value are given for each variable.

The test statistic measures the difference between the median values of the two groups, and the p-value represents the probability of obtaining the observed test statistic, or one more extreme, assuming the null hypothesis is true. In conclusion, the results of the Wilcoxon rank sum tests suggest that there is a significant difference between male and female for most of the variables except "Pension Income", "Value of Saving Accounts" and "Income From Other Sources".

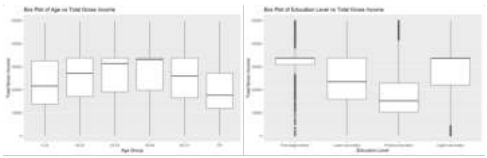
These results suggest that gender is a significant factor in determining the differences in the variables such as "Amount Spent on Utilities", "Amount Spent on Consumer Goods and Services", "Employee Income", "Self Employment Income", "Financial Assets Income", "Value of Self Employment Businesses", "Amount Spent on Food at Home", "Rental Income" and "Credit Card Debt".

Hypothesis Test Results

Null Hypothesis	Test	P value	Result
There is no difference between the Male and Female groups with respect to the dependent variable Total Gross Income.	T test	0.002000	Rejected
There is no difference between the Male and Female groups with respect to the dependent variable Amount spent on utilities.	Wilcoxon-Mann-Whitney test	0.002000	Rejected
There is no difference between the male and female groups with respect to the dependent variable Amount spent on Consumer goods and services	Wilcoxon-Mann-Whitney test	0.002000	Rejected
There is no difference between the Male and Female groups with respect to the dependent variable Employee_Income.	Wilcoxon-Mann-Whitney test	0.002000	Rejected
There is no difference between the Male and Female groups with respect to the dependent variable Self_Employee_Income	Wilcoxon-Mann-Whitney test	0.002000	Rejected
There is no difference between the Male and Female groups with respect to the dependent variable Financial_assets_Income	Wilcoxon-Mann-Whitney test	0.002000	Rejected
There is no difference between the Male and Female groups with respect to the dependent variable Pension_Income.	Wilcoxon-Mann-Whitney test	0.904600	Accepted
There is no difference between the Male and Female groups with respect to the dependent variable Rental_Income.	Wilcoxon-Mann-Whitney test	0.002982	Rejected
There is no difference between the Male and Female groups with respect to the dependent variable Credit_Card_Debt.	Wilcoxon-Mann-Whitney test	0.004729	Rejected
There is no difference between the Male and Female groups with respect to the dependent variable Value_of_Saving_Accounts.	Wilcoxon-Mann-Whitney test	0.327200	Accepted
There is no difference between the Male and Female groups with respect to the dependent variable Value_of_Self_employment_Businesses.	Wilcoxon-Mann-Whitney test	0.002000	Rejected
There is no difference between the Male and Female groups with respect to the dependent variable Amount_spent_on_Food_at_Home.	Wilcoxon-Mann-Whitney test	0.002000	Rejected
There is no difference between the Male and Female groups with respect to the dependent variable Income_From_Other_Sources.	Wilcoxon-Mann-Whitney test	0.201000	Accepted

ANOVA

Before performing the ANOVA test the following plots were drawn,



```
fit <- lm(Total_Gross_Income ~ Age, data = hcfs)

anova(fit)
```

```
## Analysis of Variance Table
##
## Response: Total_Gross_Income
##             Df     Sum Sq    Mean Sq F value    Pr(>F)
## Age          5 7.0898e+10 1.4180e+10  109.38 < 2.2e-16 ***
## Residuals 8150 1.0566e+12 1.2964e+08
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This analysis presents an ANOVA table for the response variable Total_Gross_Income, which is being analyzed in terms of the Age group. The null hypothesis in this test is that there is no significant difference in the mean Total_Gross_Income across the different age groups, and the alternative hypothesis is that there is a significant difference in the mean Total_Gross_Income across at least one age group. The table shows that there are 5 degrees of freedom for Age and 8150 degrees of freedom for Residuals. The sum of squares for Age is 7.0898e+10, while the sum of squares for Residuals is 1.0566e+12. The mean sum of squares for Age is 1.4180e+10, while the mean sum of squares for Residuals is 1.2964e+08. The F-statistic for this test is 109.38, which has a p-value less than 2.2e-16, indicating that there is significant evidence to reject the null hypothesis. Therefore, we can conclude that there is a significant difference in the mean Total_Gross_Income across at least one age group. The age group variable is a significant predictor of the Total_Gross_Income, and we can reject the null hypothesis.

```
fit <- lm(Total_Gross_Income ~ Education_Level, data = hcfs)
anova(fit)
```

```
## Analysis of Variance Table
##
## Response: Total_Gross_Income
##                  Df     Sum Sq    Mean Sq F value    Pr(>F)
## Education_Level   3 2.1248e+11 7.0827e+10  631.04 < 2.2e-16 ***
## Residuals      8152 9.1497e+11 1.1224e+08
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Above we have analysis of variance table for the response variable Total_Gross_Income, where the data has been divided by Education_Level. The table shows the results of the one-way ANOVA, which tests whether there are any statistically significant differences in the mean Total_Gross_Income between the different Education_Level groups.

The table reports two degrees of freedom (df) values: Df for Education_Level and Df for Residuals. The Sum Sq column displays the sum of squares for each source of variation, while the Mean Sq column displays the mean sum of squares for each source of variation. The F-value and its associated p-value (Pr(>F)) test whether there is a significant difference between groups, with a lower p-value indicating a greater likelihood that the differences are statistically significant. In this case, the p-value is less than 0.001 (< 2.2e-16), which indicates strong evidence that there are statistically significant differences between the mean Total_Gross_Income for the different Education_Level groups. Therefore, we reject the null hypothesis that there is no difference in mean Total_Gross_Income between the groups.

Non-Parametric Tests with Age

```
hcfs_selected <- hcfs %>% select(Age, AMount_Spent_on_Utilities, Amount_Spent_on_Consumer_Goods_Services,
                                 Employee_Income, Self_Employment_income, Financial_assets_Income,
                                 Value_of_Self_employment_Businesses, Amount_spent_on_Food_at_Home, Credit_Card_Debt,
                                 Value_of_Saving_Accounts, Income_From_Other_Sources)

for (col in 2:ncol(hcfs_selected)) {
  kw_result <- kruskal.test(as.formula(paste(colnames(hcfs_selected)[col], "~", "Age")), data = hcfs_selected)
  print(paste("Column:", colnames(hcfs_selected)[col]))
  print(kw_result)
}
```

```
## [1] "Column: AMount_Spent_on_Utilities"
##
##  Kruskal-Wallis rank sum test
##
## data:  AMount_Spent_on_Utilities by Age
## Kruskal-Wallis chi-squared = 131.27, df = 5, p-value < 2.2e-16
##
## [1] "Column: Amount_Spent_on_Consumer_Goods_Services"
##
##  Kruskal-Wallis rank sum test
##
## data:  Amount_Spent_on_Consumer_Goods_Services by Age
## Kruskal-Wallis chi-squared = 456.43, df = 5, p-value < 2.2e-16
##
## [1] "Column: Employee_Income"
##
##  Kruskal-Wallis rank sum test
##
## data:  Employee_Income by Age
## Kruskal-Wallis chi-squared = 3296.3, df = 5, p-value < 2.2e-16
##
## [1] "Column: Self_Employment_income"
##
##  Kruskal-Wallis rank sum test
##
## data:  Self_Employment_income by Age
## Kruskal-Wallis chi-squared = 455.15, df = 5, p-value < 2.2e-16
##
## [1] "Column: Financial_assets_Income"
##
##  Kruskal-Wallis rank sum test
##
## data:  Financial_assets_Income by Age
## Kruskal-Wallis chi-squared = 239.95, df = 5, p-value < 2.2e-16
##
## [1] "Column: Value_of_Self_employment_Businesses"
##
##  Kruskal-Wallis rank sum test
##
## data:  Value_of_Self_employment_Businesses by Age
## Kruskal-Wallis chi-squared = 445.62, df = 5, p-value < 2.2e-16
##
## [1] "Column: Amount_spent_on_Food_at_Home"
##
##  Kruskal-Wallis rank sum test
##
## data:  Amount_spent_on_Food_at_Home by Age
## Kruskal-Wallis chi-squared = 443.74, df = 5, p-value < 2.2e-16
##
## [1] "Column: Credit_Card_Debt"
##
##  Kruskal-Wallis rank sum test
##
## data:  Credit_Card_Debt by Age
## Kruskal-Wallis chi-squared = 47.515, df = 5, p-value = 4.461e-09
##
## [1] "Column: Value_of_Saving_Accounts"
##
##  Kruskal-Wallis rank sum test
##
## data:  Value_of_Saving_Accounts by Age
## Kruskal-Wallis chi-squared = 58.31, df = 5, p-value = 2.714e-11
##
## [1] "Column: Income_From_Other_Sources"
##
##  Kruskal-Wallis rank sum test
##
## data:  Income_From_Other_Sources by Age
## Kruskal-Wallis chi-squared = 152.8, df = 5, p-value < 2.2e-16
```

The Kruskal-Wallis rank sum test was performed on 10 different columns of data categorized by age groups. The test was used to determine whether there were statistically significant differences between the medians of each age group for each variable. For the column "Amount_Spent_on_Utilities", the Kruskal-Wallis chi-squared value was 131.27 with 5 degrees of freedom and a p-value of less than 2.2e-16, indicating strong evidence of a significant difference in median amount spent on utilities across age groups. Similarly, for the remaining nine columns, the Kruskal-Wallis test yielded chi-squared values and p-values that strongly suggested significant differences in median values across age groups. Based on these results, we reject the null hypothesis that there are no differences in median values across age groups for each variable, and conclude that age is a significant factor in determining the median value for each variable.

Non-Parametric Tests with Education level

```
hcfs_selected <- hcfs %>% select(Education_Level, AMount_Spent_on_Utilities, Amount_Spent_on_Consumer_Goods_Services,
                                 Employee_Income, Self_Employment_income, Financial_assets_Income,
                                 Value_of_Self_employment_Businesses, Pension_Income,
                                 Amount_spent_on_Food_at_Home, Credit_Card_Debt,
                                 Value_of_Saving_Accounts, Income_From_Other_Sources)


for (col in 2:ncol(hcfs_selected)) {
  kw_result <- kruskal.test(as.formula(paste(colnames(hcfs_selected)[col], "~", "Education_Level")), data = hcfs_selected)
  print(paste("Column:", colnames(hcfs_selected)[col]))
  print(kw_result)
}
```

```
## [1] "Column: AMount_Spent_on_Utilities"
##
##  Kruskal-Wallis rank sum test
##
## data:  AMount_Spent_on_Utilities by Education_Level
## Kruskal-Wallis chi-squared = 453.25, df = 3, p-value < 2.2e-16
##
## [1] "Column: Amount_Spent_on_Consumer_Goods_Services"
##
##  Kruskal-Wallis rank sum test
##
## data:  Amount_Spent_on_Consumer_Goods_Services by Education_Level
## Kruskal-Wallis chi-squared = 1184.9, df = 3, p-value < 2.2e-16
##
## [1] "Column: Employee_Income"
##
##  Kruskal-Wallis rank sum test
##
## data:  Employee_Income by Education_Level
## Kruskal-Wallis chi-squared = 1371.7, df = 3, p-value < 2.2e-16
##
## [1] "Column: Self_Employment_income"
##
##  Kruskal-Wallis rank sum test
##
## data:  Self_Employment_income by Education_Level
## Kruskal-Wallis chi-squared = 336.57, df = 3, p-value < 2.2e-16
##
## [1] "Column: Financial_assets_Income"
##
##  Kruskal-Wallis rank sum test
##
## data:  Financial_assets_Income by Education_Level
## Kruskal-Wallis chi-squared = 681.79, df = 3, p-value < 2.2e-16
##
## [1] "Column: Value_of_Self_employment_Businesses"
##
##  Kruskal-Wallis rank sum test
##
## data:  Value_of_Self_employment_Businesses by Education_Level
## Kruskal-Wallis chi-squared = 327.81, df = 3, p-value < 2.2e-16
##
## [1] "Column: Pension_Income"
##
##  Kruskal-Wallis rank sum test
##
## data:  Pension_Income by Education_Level
## Kruskal-Wallis chi-squared = 342.78, df = 3, p-value < 2.2e-16
##
## [1] "Column: Amount_spent_on_Food_at_Home"
##
##  Kruskal-Wallis rank sum test
##
## data:  Amount_spent_on_Food_at_Home by Education_Level
## Kruskal-Wallis chi-squared = 622.09, df = 3, p-value < 2.2e-16
##
## [1] "Column: Credit_Card_Debt"
##
##  Kruskal-Wallis rank sum test
##
## data:  Credit_Card_Debt by Education_Level
## Kruskal-Wallis chi-squared = 26.302, df = 3, p-value = 8.246e-06
##
## [1] "Column: Value_of_Saving_Accounts"
##
##  Kruskal-Wallis rank sum test
##
## data:  Value_of_Saving_Accounts by Education_Level
## Kruskal-Wallis chi-squared = 27.467, df = 3, p-value = 4.699e-06
##
## [1] "Column: Income_From_Other_Sources"
##
##  Kruskal-Wallis rank sum test
##
## data:  Income_From_Other_Sources by Education_Level
## Kruskal-Wallis chi-squared = 29.733, df = 3, p-value = 1.571e-06
```

The test results can be used to evaluate whether there is a statistically significant difference between the education levels in terms of the amount spent on utilities, amount spent on consumer goods and services, employee income, self-employment income, financial assets income, value of self-employment businesses, pension income, amount spent on food at home, credit card debt, value of savings accounts, and income from other sources. For all columns, the p-value is less than 0.05, indicating that there is a statistically significant difference between the education levels with respect to the amount spent on utilities, amount spent on consumer goods and services, employee income, self-employment income, financial assets income, value of self-employment businesses, pension income, amount spent on food at home, credit card debt, value of savings accounts, and income from other sources.

Therefore, we reject the null hypothesis that there is no significant difference between the education levels with respect to the amount spent on utilities, amount spent on consumer goods and services, employee income, self-employment income, financial assets income, value of self-employment businesses, pension income, amount spent on food at home, credit card debt, value of savings accounts, and income from other sources. The alternative hypothesis is that there is a significant difference between the education levels for these variables.

For example, for the column "Amount_Spent_on_Utilities," the Kruskal-Wallis chi-squared value is 453.25, with 3 degrees of freedom and a p-value of less than 2.2e-16, which is less than the significance level of 0.05. Thus, we reject the null hypothesis.

Hypothesis Test Results - Age

Null Hypothesis	Test	P value	Result

There is no difference between the 6 categories of the independent variable Age with respect to the dependent variable Total_Gross_Income.	ANOVA	0.0020	Rejected
There is no difference between the 6 categories of the independent variable Age with respect to the dependent variable Amount_Spent_on_Utilities.	Kruskal-Wallis rank sum test	0.0020	Rejected
There is no difference between the 6 categories of the independent variable Age with respect to the dependent variable Amount_Spent_on_Consumer_Goods_Services.	Kruskal-Wallis rank sum test	0.0020	Rejected
There is no difference between the 6 categories of the independent variable Age with respect to the dependent variable Employee_Income.	Kruskal-Wallis rank sum test	0.0020	Rejected
There is no difference between the 6 categories of the independent variable Age with respect to the dependent variable Self_Employment_income.	Kruskal-Wallis rank sum test	0.0020	Rejected
There is no difference between the 6 categories of the independent variable Age with respect to the dependent variable Financial_assets_Income.	Kruskal-Wallis rank sum test	0.0020	Rejected
There is no difference between the 6 categories of the independent variable Age with respect to the dependent variable Credit_Card_Debt.	Kruskal-Wallis rank sum test	0.0004	Rejected
There is no difference between the 6 categories of the independent variable Age with respect to the dependent variable Value_of_Saving_Accounts.	Kruskal-Wallis rank sum test	0.0002	Rejected
There is no difference between the 6 categories of the independent variable Age with respect to the dependent variable Value_of_Self_employment_Businesses.	Kruskal-Wallis rank sum test	0.0020	Rejected
There is no difference between the 6 categories of the independent variable Age with respect to the dependent variable Amount_spent_on_Food_at_Home.	Kruskal-Wallis rank sum test	0.0020	Rejected
There is no difference between the 6 categories of the independent variable Age with respect to the dependent variable Income_From_Other_Sources.	Kruskal-Wallis rank sum test	0.0020	Rejected

Hypothesis Test Results - Education Level

Null Hypothesis	Test	P value	Result
There is no difference between the 4 categories of the independent variable Education_Level with respect to the dependent variable Total_Gross_Income.	ANOVA	0.00200	Rejected
There is no difference between the 4 categories of the independent variable Education_Level with respect to the dependent variable Amount_Spent_on_Utilities.	Kruskal-Wallis rank sum test	0.00200	Rejected
There is no difference between the 4 categories of the independent variable Education_Level with respect to the dependent variable Amount_spent_on_consumer_goods.	Kruskal-Wallis rank sum test	0.00200	Rejected
There is no difference between the 4 categories of the independent variable Education_Level with respect to the dependent variable Employee_Income.	Kruskal-Wallis rank sum test	0.00200	Rejected
There is no difference between the 4 categories of the independent variable Education_Level with respect to the dependent variable Self_Employment_income.	Kruskal-Wallis rank sum test	0.00200	Rejected
There is no difference between the 4 categories of the independent variable Education_Level with respect to the dependent variable Financial_assets_Income.	Kruskal-Wallis rank sum test	0.00200	Rejected
There is no difference between the 4 categories of the independent variable Education_Level with respect to the dependent variable Pension_Income.	Kruskal-Wallis rank sum test	0.00200	Rejected
There is no difference between the 4 categories of the independent variable Education_Level with respect to the dependent variable Credit_Card_Debt.	Kruskal-Wallis rank sum test	0.00200	Rejected
There is no difference between the 4 categories of the independent variable Education_Level with respect to the dependent variable Value_of_Saving_Accounts.	Kruskal-Wallis rank sum test	0.00046	Rejected
There is no difference between the 4 categories of the independent variable Education_Level with respect to the dependent variable Value_of_Self_employment_Businesses.	Kruskal-Wallis rank sum test	0.00200	Rejected

There is no difference between the 4 categories of the independent variable Education_Level with respect to the dependent variable Amount_spent_on_Food_at_Home.	Kruskal-Wallis rank sum test	0.00200	Rejected
There is no difference between the 4 categories of the independent variable Education_Level with respect to the dependent variable Income_From_Other_Sources.	Kruskal-Wallis rank sum test	0.00015	Rejected

Chi-Square

To determine if there is significant association between gender and the categorical variables in our dataset we performed chi square test on selected subset of dataset with 12 different HAS variables.

The following Hypothesis statements are considered,

- Null hypothesis (H0): There is no association between gender and any of the listed categorical variables related to assets and financial status.
- Alternative hypothesis (HA): There is an association between gender and at least one of the listed categorical variables related to assets and financial status..

The variables are listed in the first column and are of character data type as yes or no responses. The test statistic value and p-value are provided for each variable.

```
##                          variable           test   statistic
## X-squared             Has_Real_Assets Chi-square test  33.6683070
## X-squared1         Has_Financial_Assets Chi-square test  70.5789591
## X-squared2               Has_Vehicles Chi-square test 997.8911625
## X-squared3              Has_Valuables Chi-square test   2.4927087
## X-squared4       Has_Real_Estate_Wealth Chi-square test  62.3978583
## X-squared5               Has_Deposits Chi-square test  69.9052432
## X-squared6            Has_Mutual_Funds Chi-square test  23.2308071
## X-squared7                 Has_Bonds Chi-square test  11.8020341
## X-squared8                Has_Shares Chi-square test  40.6088141
## X-squared9                 Has_Debt Chi-square test  50.9562701
## X-squared10       Has_Credit_Card_Debt Chi-square test   7.2595582
## X-squared11          Has_Private_Loans Chi-square test   0.2819215
## X-squared12 Has_Applied_for_Loan_Credit Chi-square test  29.3152621
##                  p.value
## X-squared    6.535688e-09
## X-squared1   4.422084e-17
## X-squared2  5.160023e-219
## X-squared3   1.143747e-01
## X-squared4   2.806282e-15
## X-squared5   6.222284e-17
## X-squared6   1.436772e-06
## X-squared7   5.916604e-04
## X-squared8   1.859659e-10
## X-squared9   9.444683e-13
## X-squared10  7.052464e-03
## X-squared11  5.954445e-01
## X-squared12  6.150933e-08
```

From the results, we can see that for most of the variables, the p-value is less than 0.05, which is the commonly used significance level. This means that we can reject the null hypothesis and conclude that there is a significant association between gender and the listed variables. However, for the variable Has_Valuables and Has_Private_Loans the p-value is greater than 0.05, which means that we cannot reject the null hypothesis and conclude that there is no significant association between gender and Has_Valuables or Has_Private_Loans.

Chi Square Test Results

Gender vs	P value	Result
Has_Real_Assets	0.000653	Rejected
Has_Financial_Assets	0.000442	Rejected
Has_Vehicles	0.000516	Rejected
Has_Valuables	0.114300	Accepted
Has_Real_Estate_Wealth	0.000280	Rejected
Has_Deposits	0.000640	Rejected
Has_Mutual_Funds	0.000140	Rejected
Has_Bonds	0.000580	Rejected
Has_Shares	0.000185	Rejected

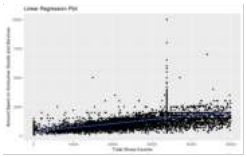
Has_Debt	0.000944	Rejected
Has_Credit_Card_Debt	0.007000	Rejected
Has_Private_Loans	0.590000	Accepted
Has_Applied_for_Loan_Credit	0.000615	Rejected

Regression

Linear Regression

To check how does total gross income affect the amount spent on consumer goods and services we considered a linear regression model assuming that there is linear relationship between the variables we performed the test and obtained the following result.

```
##
## Call:
## lm(formula = Amount_Spent_on_Consumer_Goods_Services ~ Total_Gross_Income,
##     data = hcfs)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1600.0  -309.3   -50.3   169.7  8464.4
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       3.778e+02  1.529e+01   24.71   <2e-16 ***
## Total_Gross_Income 3.432e-02  5.552e-04   61.81   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 589.5 on 8154 degrees of freedom
## Multiple R-squared:  0.3191, Adjusted R-squared:  0.319
## F-statistic:  3821 on 1 and 8154 DF,  p-value: < 2.2e-16


## `geom_smooth()` using formula = 'y ~ x'
```



A linear regression analysis was performed to examine the influence of the variable Total_Gross_Income on the variable Amount_Spent_on_Consumer_Goods_Services.

The regression model showed that the variable Total_Gross_Income explained 31.91% of the variance from the variable Amount_Spent_on_Consumer_Goods_Services. 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=3821.04, p = <.001, R2 = 0.32.

The following regression model is obtained,

Amount_Spent_on_Consumer_Goods_Services = 377.79 +0.03 · Total_Gross_Income

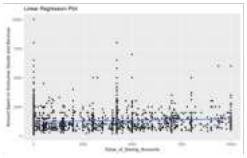
When all independent variables are zero, the value of the variable Amount_Spent_on_Consumer_Goods_Services is 377.79. If the value of the variable Total_Gross_Income changes by one unit, the value of the variable Amount_Spent_on_Consumer_Goods_Services changes by 0.03.

The standardized coefficients beta are independent of the measured variable and are always between -1 and 1. The larger the amount of beta, the greater the contribution of the respective independent variable to explain the dependent variable Amount_Spent_on_Consumer_Goods_Services . In this model, the variable Total_Gross_Income has the greatest influence on the variable Amount_Spent_on_Consumer_Goods_Services.

A linear regression analysis was performed to examine the influence of the variable Value_of_Saving_Accounts on the variable Amount_Spent_on_Consumer_Goods_Services.

```
##
## Call:
## lm(formula = Amount_Spent_on_Consumer_Goods_Services ~ Value_of_Saving_Accounts,
##     data = hcfs)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -1115   -505   -205    295   8795
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)              1.205e+03  8.767e+00 137.440  < 2e-16 ***
## Value_of_Saving_Accounts 2.604e-02  3.620e-03   7.193 6.91e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 712.2 on 8154 degrees of freedom
## Multiple R-squared:  0.006305,   Adjusted R-squared:  0.006183
## F-statistic: 51.74 on 1 and 8154 DF,  p-value: 6.909e-13
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



The regression model showed that the variable Value_of_Saving_Accounts explained 0.63% of the variance from the variable Amount_Spent_on_Consumer_Goods_Services. 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=51.74, p = <.001, R2 = 0.01.

The following regression model is obtained,

Amount_Spent_on_Consumer_Goods_Services = 1204.98 +0.03 · Value_of_Saving_Accounts

When all independent variables are zero, the value of the variable Amount_Spent_on_Consumer_Goods_Services is 1204.98. If the value of the variable Value_of_Saving_Accounts changes by one unit, the value of the variable Amount_Spent_on_Consumer_Goods_Services changes by 0.03. In this model, the variable Value_of_Saving_Accounts has the greatest influence on the variable Amount_Spent_on_Consumer_Goods_Services.

Multiple Linear Regression

A multiple linear regression analysis was performed to examine the influence of the variables Total_Gross_Income and Value_of_Saving_Accounts on the variable Total_Real_Assets.

```
## [1] 0
```

```
##
## Call:
## lm(formula = Total_Real_Assets ~ Total_Gross_Income + Value_of_Saving_Accounts,
##     data = hcfs_subset)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -423647  -137611   -53872    47574 13297724
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)             9570.1977  8647.2241   1.107    0.268
## Total_Gross_Income         8.4182     0.3125  26.938   <2e-16 ***
## Value_of_Saving_Accounts   2.0699     1.6867   1.227    0.220
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 330600 on 8153 degrees of freedom
## Multiple R-squared:  0.08303,    Adjusted R-squared:  0.08281
## F-statistic: 369.1 on 2 and 8153 DF,  p-value: < 2.2e-16
```

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.

The following regression model is obtained,

Total_Real_Assets = 9570.2 +8.42 · Total_Gross_Income +2.07 · Value_of_Saving_Accounts

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.

A multiple linear regression analysis was performed to examine the influence of the variables Employee_Income, Self_Employment_income, Rental_Income, Financial_assets_Income and Pension_Income on the variable Value_of_Household_Vehicles.

```
## [1] 0
```

```
##
## Call:
## lm(formula = Value_of_Household_Vehicles ~ Employee_Income +
##     Self_Employment_income + Rental_Income + Financial_assets_Income +
##     Pension_Income, data = hcfs_subset)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -81561  -3474  -1920   1950 253160
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)             1.976e+03  1.536e+02  12.866  < 2e-16 ***
## Employee_Income         1.610e-01  4.506e-03  35.721  < 2e-16 ***
## Self_Employment_income  1.355e-01  4.287e-03  31.609  < 2e-16 ***
## Rental_Income           3.051e-02  3.047e-02   1.001    0.317
## Financial_assets_Income 3.358e-01  5.765e-02   5.824 5.95e-09 ***
## Pension_Income          9.425e-02  6.357e-03  14.825  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8369 on 8150 degrees of freedom
## Multiple R-squared:  0.2292, Adjusted R-squared:  0.2287
## F-statistic: 484.7 on 5 and 8150 DF,  p-value: < 2.2e-16
```

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.

The following regression model is obtained,

Value_of_Household_Vehicles = 1975.53 +0.16 · Employee_Income +0.14 · Self_Employment_income +0.03 · Rental_Income +0.34 · Financial_assets_Income +0.09 · Pension_Income

- When all independent variables are zero, the value of the variable Value_of_Household_Vehicles is 1975.53.
- If the value of the variable Employee_Income changes by one unit, the value of the variable Value_of_Household_Vehicles changes by 0.16.
- If the value of the variable Self_Employment_income changes by one unit, the value of the variable Value_of_Household_Vehicles changes by 0.14.
- 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 perform Logistic Regression we considered the following questions,

1. What is the relationship between age, education level, employment status and the likelihood of having credit card debt?
2. Can likelihood of having mutual funds be predicted based on Gender, Education level and other variables?

```
##
## Call:
## glm(formula = Has_Credit_Card_Debt ~ Education_Level + Employment_status,
##     family = binomial, data = hcfs_subset)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.2218  -0.1580  -0.0994  -0.0602   3.6331
##
## Coefficients:
##                                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)                         -3.8523     0.2578 -14.943  < 2e-16 ***
## Education_LevelLower secondary      -0.7243     0.3433  -2.110  0.03487 *
## Education_LevelPrimary education    -1.8149     0.6636  -2.735  0.00624 **
## Education_LevelUpper secondary      -0.5250     0.3099  -1.694  0.09022 .
## Employment_statusOther              -1.7361     1.0411  -1.668  0.09541 .
## Employment_statusRetired            -0.9312     0.3572  -2.607  0.00914 **
## Employment_statusSelf-employed       0.1594     0.3243   0.492  0.62302
## Employment_statusUnemployed         -0.8654     1.0193  -0.849  0.39585
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 747.99  on 8155  degrees of freedom
## Residual deviance: 706.28  on 8148  degrees of freedom
## AIC: 722.28
##
## Number of Fisher Scoring iterations: 9
```

The logistic regression model includes Education_Level and Employment_status as predictors of whether a person has credit card debt. The coefficients show the direction and magnitude of the effect of each predictor on the outcome variable.

The intercept coefficient is -3.8523, which is the log-odds of having credit card debt when all predictors are zero.

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.

```
# Subset the data for relevant columns
hcfs_subset <- hcfs[, c("Education_Level", "Employment_status", "Has_Applied_for_Loan_Credit", "Housing_Status")]

hcfs_subset <- hcfs_subset %>%
  mutate(Has_Applied_for_Loan_Credit = recode(Has_Applied_for_Loan_Credit, "Yes" = 1, "No" = 0, .default = 0))

# Fit a logistic regression model
logit_model <- glm(Has_Applied_for_Loan_Credit  ~ Education_Level + Employment_status + Housing_Status, data=hcfs_subset, family=binomi
al)

# Summarize the model
summary(logit_model)
```

```
##
## Call:
## glm(formula = Has_Applied_for_Loan_Credit ~ Education_Level +
##     Employment_status + Housing_Status, family = binomial, data = hcfs_subset)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9373  -0.3976  -0.3104  -0.2351   2.7752
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                       -2.9943     0.1444 -20.731  < 2e-16 ***
## Education_LevelLower secondary     0.4076     0.1496   2.725 0.006431 **
## Education_LevelPrimary education   0.1868     0.1880   0.993 0.320533
## Education_LevelUpper secondary     0.2550     0.1453   1.755 0.079330 .
## Employment_statusOther            -0.7568     0.2052  -3.688 0.000226 ***
## Employment_statusRetired          -0.8352     0.1315  -6.350 2.16e-10 ***
## Employment_statusSelf-employed     0.4114     0.1206   3.413 0.000644 ***
## Employment_statusUnemployed       -0.3140     0.2642  -1.189 0.234591
## Housing_StatusOwner with mortgage  1.5803     0.1208  13.081  < 2e-16 ***
## Housing_StatusRenter               0.5554     0.1049   5.293 1.20e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4137.8  on 8155  degrees of freedom
## Residual deviance: 3791.5  on 8146  degrees of freedom
## AIC: 3811.5
##
## Number of Fisher Scoring iterations: 6
```

The logistic regression model assesses the association between the probability of having applied for a loan credit and the independent variables education level, employment status, and housing status. The coefficients and standard errors of the model indicate the direction and strength of the relationship between the dependent variable and independent variables.

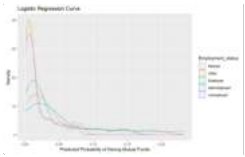
The p-values associated with the coefficients of the independent variables show that education level, employment status, and housing status are all significant predictors of having applied for a loan credit. Among the education levels, those with lower secondary education are more likely to apply for 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. The number of Fisher Scoring iterations indicates the number of iterations required to fit the model to the data.

```
# Subset the data for relevant columns
hcfs_subset <- hcfs[, c("Education_Level", "Employment_status", "Has_Mutual_Funds", "Gender", "Has_Real_Assets")]

hcfs_subset <- hcfs_subset %>%
  mutate(Has_Mutual_Funds = recode(Has_Mutual_Funds, "Yes" = 1, "No" = 0, .default = 0))

# Fit a logistic regression model
logit_model <- glm(Has_Mutual_Funds  ~ Education_Level + Employment_status + Has_Real_Assets + Gender, data=hcfs_subset, family=binomi
al)

# Summarize the model
summary(logit_model)
```

```
##
## Call:
## glm(formula = Has_Mutual_Funds ~ Education_Level + Employment_status +
##     Has_Real_Assets + Gender, family = binomial, data = hcfs_subset)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7282  -0.4003  -0.2578  -0.2078   3.1954
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                       -4.1608     1.0087  -4.125 3.71e-05 ***
## Education_LevelLower secondary    -1.6085     0.1462 -11.000  < 2e-16 ***
## Education_LevelPrimary education  -2.0458     0.1882 -10.872  < 2e-16 ***
## Education_LevelUpper secondary    -0.7048     0.1143  -6.164 7.09e-10 ***
## Employment_statusOther            -0.7731     0.3402  -2.273  0.02303 *
## Employment_statusRetired           0.3260     0.1131   2.883  0.00393 **
## Employment_statusSelf-employed     0.5870     0.1333   4.405 1.06e-05 ***
## Employment_statusUnemployed       -0.6422     0.4625  -1.389  0.16495
## Has_Real_AssetsYes                 2.0374     1.0039   2.030  0.04240 *
## GenderMale                         0.3442     0.1112   3.094  0.00197 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3716.8  on 8155  degrees of freedom
## Residual deviance: 3401.3  on 8146  degrees of freedom
## AIC: 3421.3
##
## Number of Fisher Scoring iterations: 7
```

```
library(ggplot2)

# Create a data frame with predictor values
newdata <- expand.grid(
  Education_Level = unique(hcfs_subset$Education_Level),
  Employment_status = unique(hcfs_subset$Employment_status),
  Gender = unique(hcfs_subset$Gender),
  Has_Real_Assets = unique(hcfs_subset$Has_Real_Assets)
)

# Add predicted probabilities to data frame
newdata$prob <- predict(logit_model, newdata, type="response")

# Plot the logistic regression curve
ggplot(newdata, aes(x=prob, color=Employment_status)) +
  geom_density() +
  xlab("Predicted Probability of Having Mutual Funds") +
  ylab("Density") +
  ggtitle("Logistic Regression Curve")
```



The logistic regression model tests the association between the binary response variable "Has_Mutual_Funds" and the predictor variables "Education_Level," "Employment_status," "Has_Real_Assets," and "Gender." The model's deviance residuals indicate that the model fits the data well.

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.

The null and residual deviances of the model suggest that the model explains a substantial amount of the variation in the data. The Akaike information criterion (AIC) value of 3421.3 indicates that this model is better than other candidate models with higher AIC values.

Principal Component Analysis

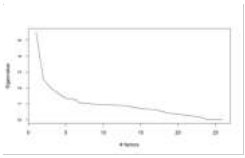
To reduce the dimensionality of the data by identifying the most important variables, which can be used to represent the data, Principal Component Analysis of the numeric columns in the dataset was performed.

```
# Select only the numeric variables from hcfs
hcfs_numeric <- hcfs %>%
  select_if(is.numeric)

#str(hcfs_numeric)

keep1<-subset(hcfs_numeric, select = 3:ncol(hcfs_numeric))

# # scale the data
# hcfs_scaled <- scale(keep1)
#
# # perform PCA
# hcfs_pca <- prcomp(hcfs_scaled, center = TRUE, scale. = TRUE)
#
# # view summary of the PCA results
# summary(hcfs_pca)
#
# # plot the PCA results
# biplot(hcfs_pca)

# Principal Components Analysis creating 15 principal components (i.e. artificial variables)

cat("Doing PCA\n")


## Doing PCA


# change the number "15" in the code below this line if you want to adjust the number of principal components to be created from your data
pc <- principal(keep1, nfactors=min(ncol(keep1),15), rotate="varimax") #rotated
print(summary(pc)) # print the variance accounted for by each principal component


##
## Factor analysis with Call: principal(r = keep1, nfactors = min(ncol(keep1), 15), rotate = "varimax")
##
## Test of the hypothesis that 15 factors are sufficient.
## The degrees of freedom for the model is 40  and the objective function was  41.33
## The number of observations was  8156  with Chi Square =  336272.5  with prob <  0
##
## The root mean square of the residuals (RMSA) is  0.04
## NULL


print(loadings(pc)) # pc loadings for each observed variable
```

```
##
## 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
## Credit_Card_Debt
## Monthly_Amount_Paid_As_Rent
## Total_Value_of_Cars              0.315                0.193  0.870
## Value_Of_Other_Vehicles                                0.227
## Value_Of_Other_Valuables                               0.984
## No_of_PrivateLoans
## Value_of_Saving_Accounts
## Amount_spent_on_Food_at_Home     0.802
## Amount_Spent_on_Food_Outside_Home  0.611               0.192
## AMount_Spent_on_Utilities        0.490         0.168
## Amount_Spent_on_Consumer_Goods_Services  0.826  0.156         0.134  0.139
##                                  RC6    RC4    RC12   RC13   RC8
## Value_of_Household_Vehicles      0.455
## Valuables
## Deposits                         0.390  0.218         0.343
## Mutual_Funds                                0.923
## Bonds                                      -0.233
## Employee_Income                  -0.713
## Self_Employment_income
## Rental_Income                               0.957
## Financial_assets_Income                     0.263  0.158
## Pension_Income                   0.870
## Total_Real_Assets                0.135         0.363
## Total_Financial_Assets                      0.118  0.350  0.171
## Total_Gross_Income
## Value_of_Self_employment_Businesses
## Income_From_Other_Sources
## Credit_Card_Debt
## Monthly_Amount_Paid_As_Rent
## Total_Value_of_Cars
## Value_Of_Other_Vehicles                     0.920
## Value_Of_Other_Valuables
## No_of_PrivateLoans
## Value_of_Saving_Accounts                                  0.955
## Amount_spent_on_Food_at_Home
## Amount_Spent_on_Food_Outside_Home  -0.172
## AMount_Spent_on_Utilities
## Amount_Spent_on_Consumer_Goods_Services
##                                  RC9    RC11   RC10   RC7    RC15
## Value_of_Household_Vehicles
## Valuables
## Deposits                                    0.137
## Mutual_Funds
## Bonds
## Employee_Income
## Self_Employment_income
## Rental_Income
## Financial_assets_Income
## Pension_Income
## Total_Real_Assets                -0.259                0.129
## Total_Financial_Assets
## Total_Gross_Income                          -0.141
## Value_of_Self_employment_Businesses
## Income_From_Other_Sources                                 0.998
## Credit_Card_Debt                            0.996
## Monthly_Amount_Paid_As_Rent      0.982
## Total_Value_of_Cars
## Value_Of_Other_Vehicles
## Value_Of_Other_Valuables
## No_of_PrivateLoans                          0.990
## Value_of_Saving_Accounts
## Amount_spent_on_Food_at_Home
## Amount_Spent_on_Food_Outside_Home                        -0.533
## AMount_Spent_on_Utilities                                0.693
## Amount_Spent_on_Consumer_Goods_Services
##
##                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


print(pc$values)
```

```
##  [1]  5.519263e+00  2.555186e+00  1.992212e+00  1.659175e+00  1.345562e+00
##  [6]  1.311511e+00  1.063026e+00  1.033489e+00  9.752188e-01  9.606216e-01
## [11]  9.469383e-01  9.119303e-01  8.846765e-01  8.061963e-01  7.230774e-01
## [16]  6.555034e-01  6.462307e-01  5.033627e-01  4.068576e-01  3.666032e-01
## [21]  2.955986e-01  2.363339e-01  1.833365e-01  1.809050e-02  2.010047e-16
## [26] -2.410907e-16
```

```
# create scree plot (to help decide how many PCs to keep)
plot(pc$values,type="l",main="", xlab="# factors", ylab="Eigenvalue") # scree plot
```



The test of the hypothesis that 15 factors are sufficient shows that the degrees of freedom for the model is 40 and the objective function was 41.33. The number of observations was 8156, and the Chi-Square value was 336272.5 with prob <0, which indicates that the model is a good fit for the dataset.

The loadings table provides information about the correlation between each variable in the dataset and each of the principal components. The values in the table represent the factor loadings, which indicate the strength and direction of the relationship between the variables and the components. The higher the absolute value of the factor loading, the stronger the relationship between the variable and the component.The retained components are listed in the columns, and the variables are listed in the rows. The values in the table represent the correlation between the variable and the component, with values close to 1 or -1 indicating a strong correlation.

For example, the variable "Value_of_Household_Vehicles" has a loading of 0.27 on the first principal component (RC1), indicating a moderate positive correlation. The variable "Valuables" has a loading of 0.98 on RC2, indicating a strong positive correlation. The variable "Credit_Card_Debt" has a loading of 1 on RC4, indicating a perfect correlation. The variable "Amount_Spent_on_Food_Outside_Home" has a loading of -0.53 on RC15, indicating a moderate negative correlation.

The table also shows the eigenvalues and the proportion and cumulative proportion of variance explained by each component. The eigenvalues indicate the amount of variance in the data that is explained by each component. The proportion and cumulative proportion of variance explained indicate the proportion of total variance in the data that is explained by each component and the cumulative proportion of variance explained by each successive component. For example, the first principal component (RC1) explains 12.2% of the total variance in the data, while the first two components (RC1 and RC2) together explain 23% of the total variance.

The root mean square of the residuals (RMSA) is 0.04, which is low, indicating that the model has a good fit to the data.

Overall, the output suggests that 15 components are sufficient to explain the variability in the dataset. The loadings table shows that several variables are strongly correlated with the first few components, indicating that these components capture important information in the data.

Decision Trees

To identify the important factors that determine whether a person has private loans, deposits, real estate wealth, vehicles, financial assets, debt, or real assets we used decision trees.

We converted the categorical variables to factors and considered a subset of data to include variables, such as gender, age, employment status, education level, total gross income, and whether the household owns saving accounts or has credit card debt.

We then fit a decision tree model using the rpart() function to predict whether an individual has private loans, deposits, real estate wealth, vehicles, financial assets, debt, or real assets based on all available predictors, and then plotted the decision tree using the rpart.plot() function to visualize the important predictors that influence the target variable.

```
library(rpart)
library(rpart.plot)

#Subset the data
hcfs_subset <- hcfs[, c("Gender", "Age", "Employment_status", "Education_Level",
                        "Total_Gross_Income", "Has_Debt", "Household_Owns_Saving_accounts",
                        "Has_Credit_Card_Debt",
                        "Has_Mutual_Funds", "Has_Shares", "Has_Bonds","Has_Real_Assets", "Has_Financial_Assets", "Has_Vehicles", "Has_Valuables",
                "Has_Real_Estate_Wealth", "Has_Deposits",
                "Has_Private_Loans",
                "Has_Applied_for_Loan_Credit")]

# Convert categorical variables to factors
hcfs_subset$Gender <- as.factor(hcfs_subset$Gender)
hcfs_subset$Employment_status <- as.factor(hcfs_subset$Employment_status)
hcfs_subset$Education_Level <- as.factor(hcfs_subset$Education_Level)
hcfs_subset$Has_Debt <- as.factor(hcfs_subset$Has_Debt)
hcfs_subset$Household_Owns_Saving_accounts <- as.factor(hcfs_subset$Household_Owns_Saving_accounts)
hcfs_subset$Has_Credit_Card_Debt <- as.factor(hcfs_subset$Has_Credit_Card_Debt)
hcfs_subset$Has_Mutual_Funds <- as.factor(hcfs_subset$Has_Mutual_Funds)
hcfs_subset$Has_Shares <- as.factor(hcfs_subset$Has_Shares)
hcfs_subset$Has_Bonds <- as.factor(hcfs_subset$Has_Bonds)
hcfs_subset$Has_Real_Assets <- as.factor(hcfs_subset$Has_Real_Assets)
hcfs_subset$Has_Financial_Assets <- as.factor(hcfs_subset$Has_Financial_Assets)
hcfs_subset$Has_Vehicles <- as.factor(hcfs_subset$Has_Vehicles)
hcfs_subset$Has_Real_Estate_Wealth <- as.factor(hcfs_subset$Has_Real_Estate_Wealth)
hcfs_subset$Has_Deposits <- as.factor(hcfs_subset$Has_Deposits)
hcfs_subset$Has_Private_Loans <- as.factor(hcfs_subset$Has_Private_Loans)
hcfs_subset$Has_Applied_for_Loan_Credit <- as.factor(hcfs_subset$Has_Applied_for_Loan_Credit)
```