Survey Analysis: Math 3210 Project 1

Noah G. Gallego

2024-11-11

Contents

1 Introduction

In this project, we will analyze survey data collected from classmates to gain insights into their financial habits and situations, including budgeting methods, spending patterns, saving behaviors, and concerns about debt. We'll conduct data cleaning and exploratory data analysis (EDA) to identify trends and patterns within financial practices. Using descriptive statistics, visualizations, and inferential analysis, we aim to highlight relationships between financial behaviors and various factors that may influence financial decision-making. This analysis will provide a deeper understanding of the financial habits and challenges among our peers.

1.1 Load Libraries:

```
library(dplyr)
library(ggplot2)
library(tm)
```

```
## Loading required package: NLP
```

```
##
## Attaching package: 'NLP'
```

```
## The following object is masked from 'package:ggplot2':
##
##     annotate
```

```
##
## Attaching package: 'tm'
```

```
## The following object is masked from 'package:mosaic':
##
##     inspect
```

```
library(wordcloud)
```

```
## Loading required package: RColorBrewer
```

```
library(RColorBrewer)
library(reshape2)
```

```
##
## Attaching package: 'reshape2'
```

```
## The following object is masked from 'package:tidyr':
##
##     smiths
```

```
library(cluster)
library(ggcorrplot)
library(tidyr)
library(tidyverse)
library(stats)
library(caret)
```

```
##
## Attaching package: 'caret'
```

```
## The following object is masked from 'package:mosaic':
##
##     dotPlot
```

```
## The following object is masked from 'package:purrr':
##
##     lift
```

1.2 Load Data & Basic Cleaning

```
# Load the DataFrame
df = read.csv("ProData.csv", header = TRUE)

# Remove Unnecessary Rows
df = df[-c(1, 2), ]

# Select Needed Cols
cols = c("RecordedDate", "LocationLatitude", "LocationLongitude", "Question.One.", "Question.Two._1", "
df = df[, cols]
df = df %>%
  rename(
    Date_Recorded = RecordedDate,
    Latitude = LocationLatitude,
    Longitude = LocationLongitude,
    Budget_Management_Method = Question.One.,
    Financial_Confidence = Question.Two._1,
    Student_Loans = Question.3.,
    Student_Debt_Concern = Question.4_1,
    Future_Financial_Decision_Impact = Question.5,
    Biggest_Expense = Question.6,
    Employment_Status = Question.7,
    Saving_For_Future_Expenses = Question.8,
    Saving_For_Retirement = Question.9,
    Financial_Independence_Confidence = Question.10,
    Job_Confidence_Post_Graduation = Question.11
```

```
  )

# Display First Few Rows
df
```

```
##          Date_Recorded Latitude Longitude     Budget_Management_Method
## 3  2024-10-03 00:31:22  35.3965 -119.1268        Using a budgeting app
## 4  2024-10-03 00:39:15  35.3407 -119.0596        Using a budgeting app
## 5  2024-10-18 23:03:12  33.7046 -117.8739      I don't manage a budget
## 6  2024-10-19 21:23:19  35.7768 -119.2414 Manually (notebook or Excel)
## 7  2024-10-19 21:32:32  35.3044 -119.1031          No budgeting at all
## 8  2024-10-20 20:39:40  35.3878  -118.936      I don't manage a budget
## 9  2024-10-20 22:28:16  35.3044 -119.1031      I don't manage a budget
## 10 2024-10-20 23:29:36  45.8491 -119.7143          No budgeting at all
## 11 2024-10-21 00:06:31  35.3879 -118.9861          No budgeting at all
## 12 2024-10-21 13:16:43  35.4145 -119.0403 Manually (notebook or Excel)
## 13 2024-10-22 23:02:49  35.3044 -119.1031 Manually (notebook or Excel)
##    Financial_Confidence Student_Loans Student_Debt_Concern
## 3                     5            No                    5
## 4                     3            No                    4
## 5                     4            No                    3
## 6                                  No                    1
## 7                     2            No                    0
## 8                     4            No
## 9                     4            No                    0
## 10                    4            No                    4
## 11                    3            No
## 12                    5           Yes                    5
## 13                    3           Yes                    4
##
## 3                         I think that student debt heavily impacts future financial decisions. Loa
## 4
## 5
## 6                                Too in my own head. conscious of my spending habits. inclined to sav
## 7  I think student debt will have me try to save more money and buy only what I need. This means bei
## 8
## 9                                                                                 I'm not taking on any debt whi
## 10
## 11
## 12                                                                                   They can be paid off immedi
## 13
##    Biggest_Expense Employment_Status Saving_For_Future_Expenses
## 3   Groceries/Food   Yes, part-time.                        Yes
## 4    Entertainment   Yes, part-time.                        Yes
## 5    Entertainment               No                         No
## 6   Groceries/Food   Yes, part-time.                         No
## 7            Other               No                         No
## 8    Rent/Mortgage   Yes, part-time.                        Yes
## 9   Transportation   Yes, part-time.                         No
## 10  Groceries/Food   Yes, full-time.                        Yes
## 11   Rent/Mortgage   Yes, part-time.                        Yes
## 12   Rent/Mortgage   Yes, part-time.                        Yes
## 13   Rent/Mortgage   Yes, part-time.                        Yes
##    Saving_For_Retirement Financial_Independence_Confidence
```

```
## 3                    Yes                         Confident
## 4                    Yes                         Confident
## 5                     No                     Not Confident
## 6                    Yes                         Confident
## 7                     No                         Confident
## 8                     No                         Confident
## 9                     No                         Confident
## 10                   Yes                         Confident
## 11                    No                     Not Confident
## 12                   Yes                     Not Confident
## 13                    No                         Confident
##    Job_Confidence_Post_Graduation
## 3                  Not Confident
## 4                  Not Confident
## 5                  Not Confident
## 6                  Not Confident
## 7                      Confident
## 8                      Confident
## 9                      Confident
## 10                     Confident
## 11                     Confident
## 12                     Confident
## 13                     Confident
```

1.3 Further Cleaning

```r
# Convert Cols to Ints where Needed
df = df %>%
  mutate(across(c(Latitude, Longitude, Financial_Confidence, Student_Debt_Concern), as.numeric))

# Convert Response to DT
df$Date_Recorded = as.Date(df$Date_Recorded)

# Convert Binary Responses to 1/0 For Analysis
df = df %>%
  mutate(
    Student_Loans = ifelse(trimws(tolower(Student_Loans)) == "yes", 1, 0),
    Financial_Independence_Confidence = ifelse(trimws(tolower(Financial_Independence_Confidence)) == "c
    Job_Confidence_Post_Graduation = ifelse(trimws(tolower(Job_Confidence_Post_Graduation)) == "confide
    Saving_For_Future_Expenses = ifelse(trimws(tolower(Saving_For_Future_Expenses)) == "yes", 1, 0),
    Saving_For_Retirement = ifelse(trimws(tolower(Saving_For_Retirement)) == "yes", 1, 0),
    Employment_Status = ifelse(trimws(tolower(Employment_Status)) == "yes, part-time" |
                               trimws(tolower(Employment_Status)) == "yes, full-time", 1, 0)
  )

# Function to calculate the mode
get_mode = function(x) {
  unique_x = na.omit(x)  # Remove NA values
  unique_x[which.max(tabulate(match(x, unique_x)))]
}

# Replace NAs with mode for each column
df = df %>%
  mutate(across(everything(), ~ ifelse(is.na(.), get_mode(.), .)))
```

```r
# Re-Display Dataframe
head(df)
```

```
##   Date_Recorded Latitude Longitude     Budget_Management_Method
## 3         19999  35.3965 -119.1268        Using a budgeting app
## 4         19999  35.3407 -119.0596        Using a budgeting app
## 5         20014  33.7046 -117.8739      I don't manage a budget
## 6         20015  35.7768 -119.2414 Manually (notebook or Excel)
## 7         20015  35.3044 -119.1031          No budgeting at all
## 8         20016  35.3878 -118.9360      I don't manage a budget
##   Financial_Confidence Student_Loans Student_Debt_Concern
## 3                    5             0                    5
## 4                    3             0                    4
## 5                    4             0                    3
## 6                    4             0                    1
## 7                    2             0                    0
## 8                    4             0                    4
##
## 3                        I think that student debt heavily impacts future financial decisions. Loans
## 4
## 5
## 6                              Too in my own head. conscious of my spending habits. inclined to save
## 7 I think student debt will have me try to save more money and buy only what I need. This means being
## 8
##   Biggest_Expense Employment_Status Saving_For_Future_Expenses
## 3  Groceries/Food                 0                          1
## 4   Entertainment                 0                          1
## 5   Entertainment                 0                          0
## 6  Groceries/Food                 0                          0
## 7           Other                 0                          0
## 8   Rent/Mortgage                 0                          1
##   Saving_For_Retirement Financial_Independence_Confidence
## 3                     1                                1
## 4                     1                                1
## 5                     0                                0
## 6                     1                                1
## 7                     0                                1
## 8                     0                                1
##   Job_Confidence_Post_Graduation
## 3                              0
## 4                              0
## 5                              0
## 6                              0
## 7                              1
## 8                              1
```

1.4 Exploratory Data Analysis (EDA)

1.4.1 Descriptive Statistics

```r
# Summary Statistics
summary(df)
```

```
##  Date_Recorded      Latitude       Longitude      Budget_Management_Method
```

```
##  Min.   :19999   Min.   :33.70   Min.   :-119.7   Length:11
##  1st Qu.:20014   1st Qu.:35.30   1st Qu.:-119.1   Class :character
##  Median :20016   Median :35.39   Median :-119.1   Mode  :character
##  Mean   :20013   Mean   :36.20   Mean   :-119.0
##  3rd Qu.:20016   3rd Qu.:35.41   3rd Qu.:-119.0
##  Max.   :20018   Max.   :45.85   Max.   :-117.9
##  Financial_Confidence Student_Loans    Student_Debt_Concern
##  Min.   :2.000        Min.   :0.0000   Min.   :0.000
##  1st Qu.:3.000        1st Qu.:0.0000   1st Qu.:2.000
##  Median :4.000        Median :0.0000   Median :4.000
##  Mean   :3.727        Mean   :0.1818   Mean   :3.091
##  3rd Qu.:4.000        3rd Qu.:0.0000   3rd Qu.:4.000
##  Max.   :5.000        Max.   :1.0000   Max.   :5.000
##  Future_Financial_Decision_Impact Biggest_Expense    Employment_Status
##  Length:11                        Length:11          Min.   :0
##  Class :character                 Class :character   1st Qu.:0
##  Mode  :character                 Mode  :character   Median :0
##                                                      Mean   :0
##                                                      3rd Qu.:0
##                                                      Max.   :0
##  Saving_For_Future_Expenses Saving_For_Retirement
##  Min.   :0.0000             Min.   :0.0000
##  1st Qu.:0.0000             1st Qu.:0.0000
##  Median :1.0000             Median :0.0000
##  Mean   :0.6364             Mean   :0.4545
##  3rd Qu.:1.0000             3rd Qu.:1.0000
##  Max.   :1.0000             Max.   :1.0000
##  Financial_Independence_Confidence Job_Confidence_Post_Graduation
##  Min.   :0.0000                    Min.   :0.0000
##  1st Qu.:0.5000                    1st Qu.:0.0000
##  Median :1.0000                    Median :1.0000
##  Mean   :0.7273                    Mean   :0.6364
##  3rd Qu.:1.0000                    3rd Qu.:1.0000
##  Max.   :1.0000                    Max.   :1.0000
```

```r
# Mean, Median, Mode, Standard Deviation for Numerical Columns
mean_values = sapply(df %>% select_if(is.numeric), mean, na.rm = TRUE)
median_values = sapply(df %>% select_if(is.numeric), median, na.rm = TRUE)
mode_values = sapply(df %>% select_if(is.numeric), get_mode)
std_dev_values = sapply(df %>% select_if(is.numeric), sd, na.rm = TRUE)

list(mean = mean_values, median = median_values, mode = mode_values, sd = std_dev_values)
```

```
## $mean
##                      Date_Recorded                           Latitude
##                      20012.9090909                       36.1973727
##                          Longitude                Financial_Confidence
##                       -119.0261545                          3.7272727
##                      Student_Loans                Student_Debt_Concern
##                          0.1818182                          3.0909091
##                  Employment_Status          Saving_For_Future_Expenses
##                          0.0000000                          0.6363636
##              Saving_For_Retirement Financial_Independence_Confidence
##                          0.4545455                          0.7272727
##     Job_Confidence_Post_Graduation
```

```
##                         0.6363636
##
## $median
##                     Date_Recorded                          Latitude
##                        20016.0000                           35.3878
##                         Longitude              Financial_Confidence
##                         -119.1031                            4.0000
##                     Student_Loans              Student_Debt_Concern
##                            0.0000                            4.0000
##                 Employment_Status        Saving_For_Future_Expenses
##                            0.0000                            1.0000
##             Saving_For_Retirement Financial_Independence_Confidence
##                            0.0000                            1.0000
##    Job_Confidence_Post_Graduation
##                            1.0000
##
## $mode
##                     Date_Recorded                          Latitude
##                        20016.0000                           35.3044
##                         Longitude              Financial_Confidence
##                         -119.1031                            4.0000
##                     Student_Loans              Student_Debt_Concern
##                            0.0000                            4.0000
##                 Employment_Status        Saving_For_Future_Expenses
##                            0.0000                            1.0000
##             Saving_For_Retirement Financial_Independence_Confidence
##                            0.0000                            1.0000
##    Job_Confidence_Post_Graduation
##                            1.0000
##
## $sd
##                     Date_Recorded                          Latitude
##                         6.9635414                         3.2440356
##                         Longitude              Financial_Confidence
##                         0.4343394                         0.9045340
##                     Student_Loans              Student_Debt_Concern
##                         0.4045199                         1.8683975
##                 Employment_Status        Saving_For_Future_Expenses
##                         0.0000000                         0.5045250
##             Saving_For_Retirement Financial_Independence_Confidence
##                         0.5222330                         0.4670994
##    Job_Confidence_Post_Graduation
##                         0.5045250
```

1.4.2 Missing Values Analysis

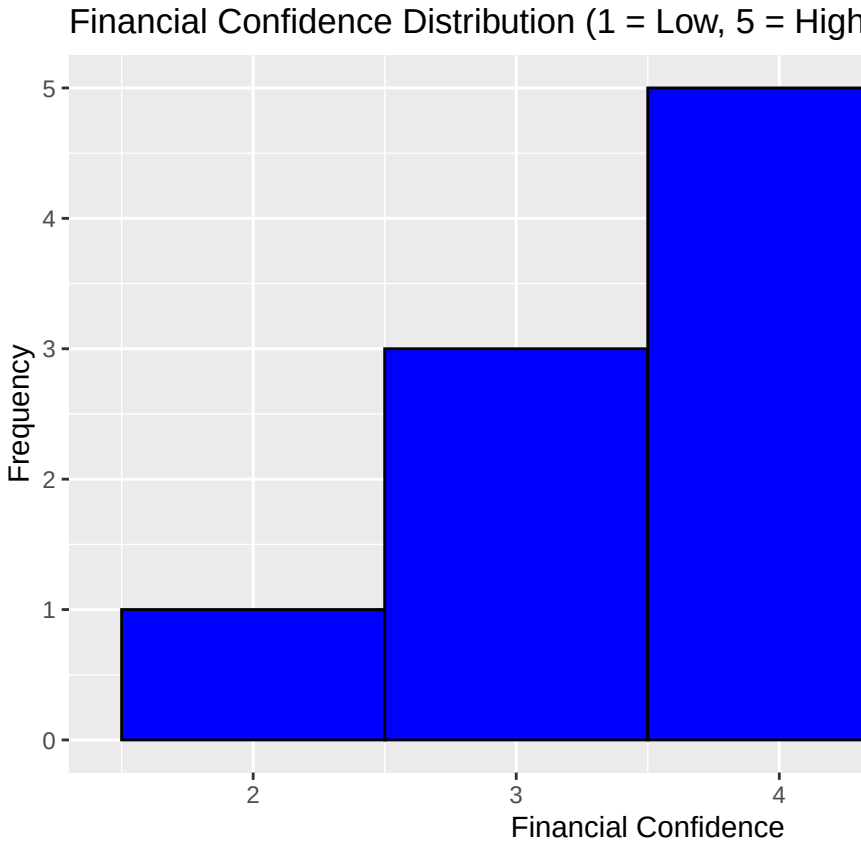
```
# Check for Missing Values
missing_values = colSums(is.na(df))
print(missing_values)
```

```
##                     Date_Recorded                          Latitude
##                                 0                                 0
##                         Longitude           Budget_Management_Method
##                                 0                                 0
```

```
##              Financial_Confidence                          Student_Loans
##                                 0                                      0
##              Student_Debt_Concern       Future_Financial_Decision_Impact
##                                 0                                      0
##                   Biggest_Expense                     Employment_Status
##                                 0                                      0
##        Saving_For_Future_Expenses                 Saving_For_Retirement
##                                 0                                      0
## Financial_Independence_Confidence        Job_Confidence_Post_Graduation
##                                 0                                      0
```

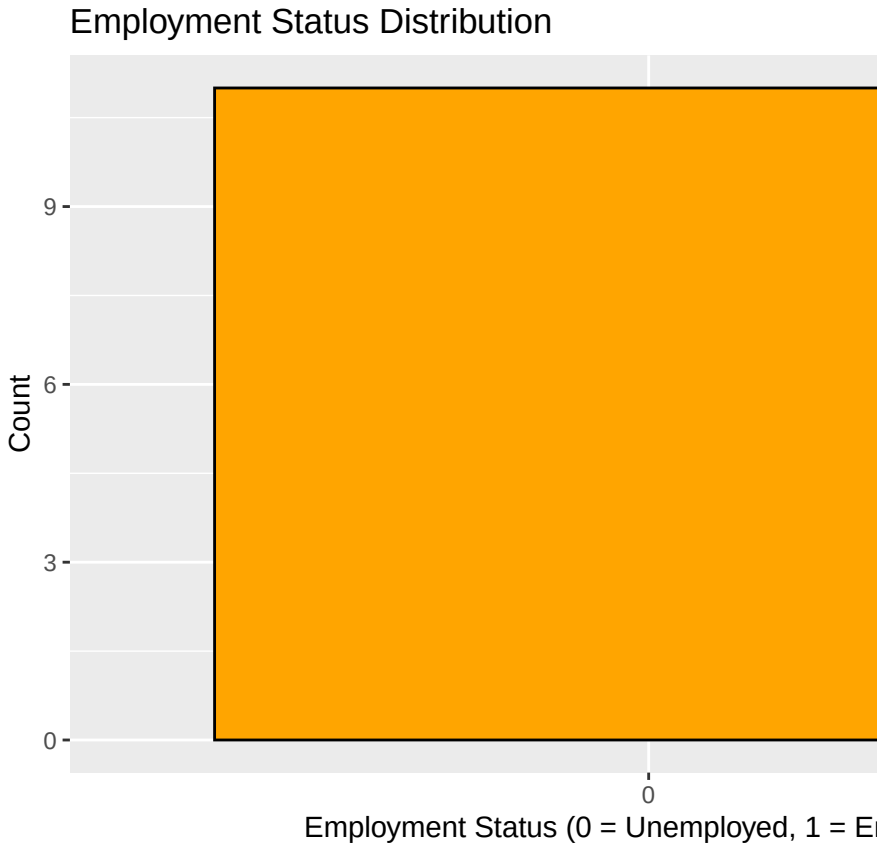
1.4.3 Visualizations

```
ggplot(df, aes(x = Financial_Confidence)) +
  geom_histogram(binwidth = 1, fill = "blue", color = "black") +
  labs(title = "Financial Confidence Distribution (1 = Low, 5 = High)", x = "Financial Confidence", y =
```



1.4.3.1 Financial Confidence Distribution

```
# Bar Plot for Employment Status
ggplot(df, aes(x = as.factor(Employment_Status))) +
  geom_bar(fill = "orange", color = "black") +
  labs(title = "Employment Status Distribution", x = "Employment Status (0 = Unemployed, 1 = Employed)"
```



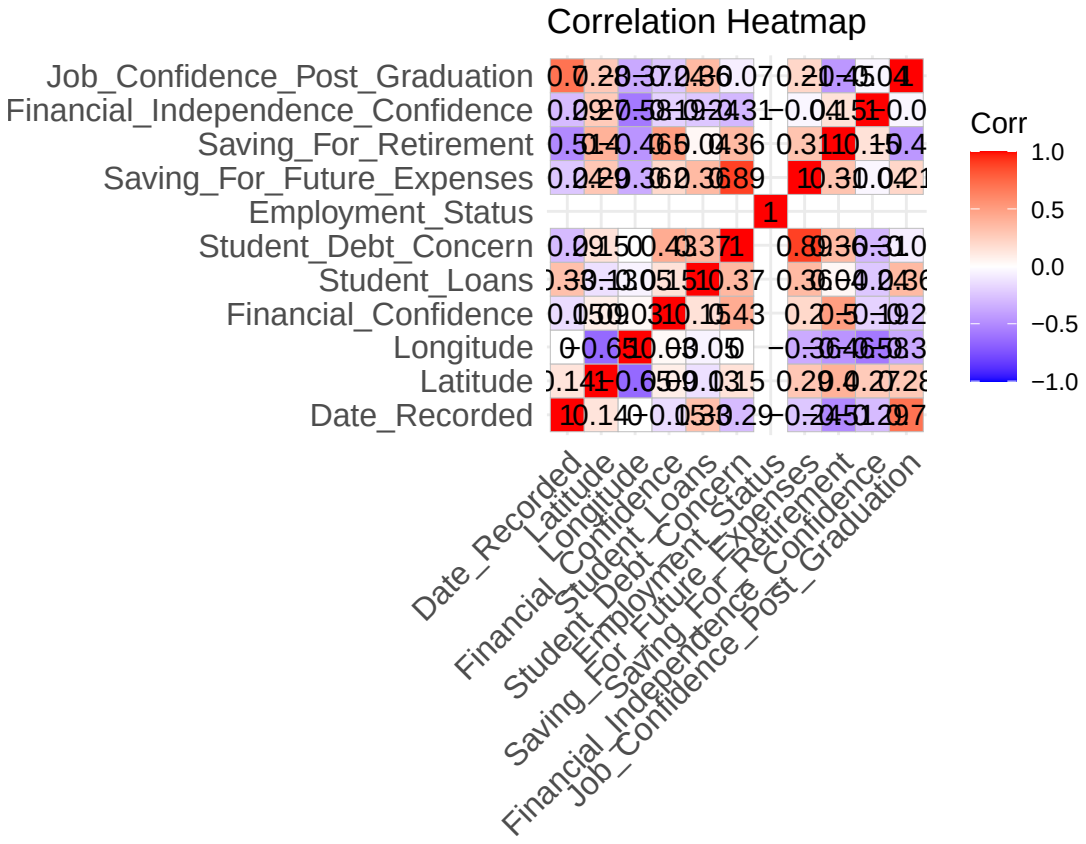
1.4.3.2 Employment Status Distribution

```
# Correlation Plot (for numerical variables)
numeric_cols = df %>% select_if(is.numeric)
correlation_matrix = cor(numeric_cols, use = "complete.obs")
```

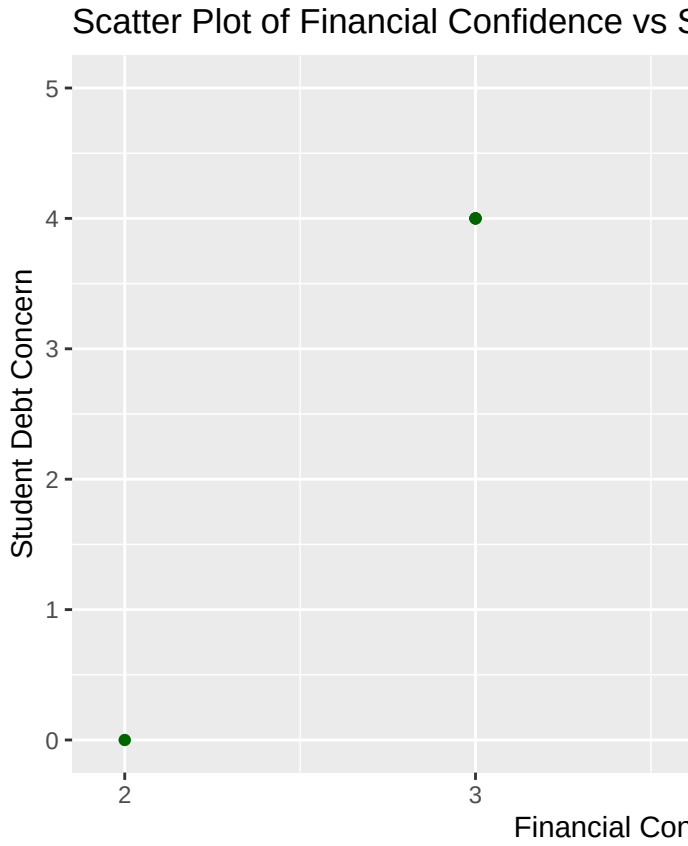
1.4.3.3 Correlation Analysis

```
## Warning in cor(numeric_cols, use = "complete.obs"): the standard deviation is
## zero
```

```
ggcorrplot(correlation_matrix, lab = TRUE, title = "Correlation Heatmap")
```



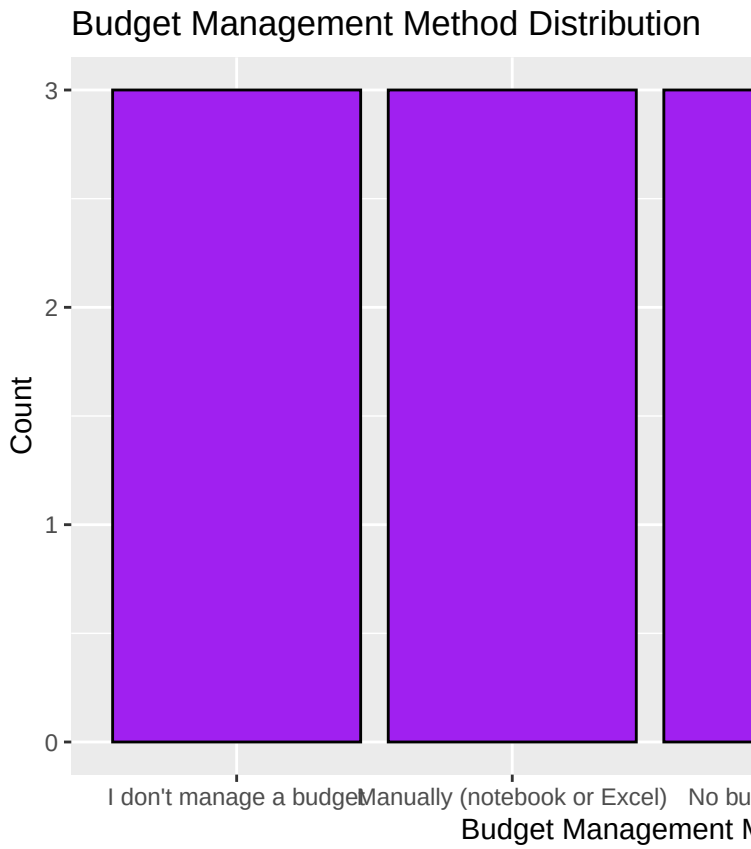

```r
# Scatter Plot of Financial Confidence vs. Student Debt Concern
ggplot(df, aes(x = Financial_Confidence, y = Student_Debt_Concern)) +
  geom_point(color = "darkgreen") +
  labs(title = "Scatter Plot of Financial Confidence vs Student Debt Concern",
       x = "Financial Confidence", y = "Student Debt Concern")
```



1.4.3.4 Financial Confidence vs. Student Debt Concern

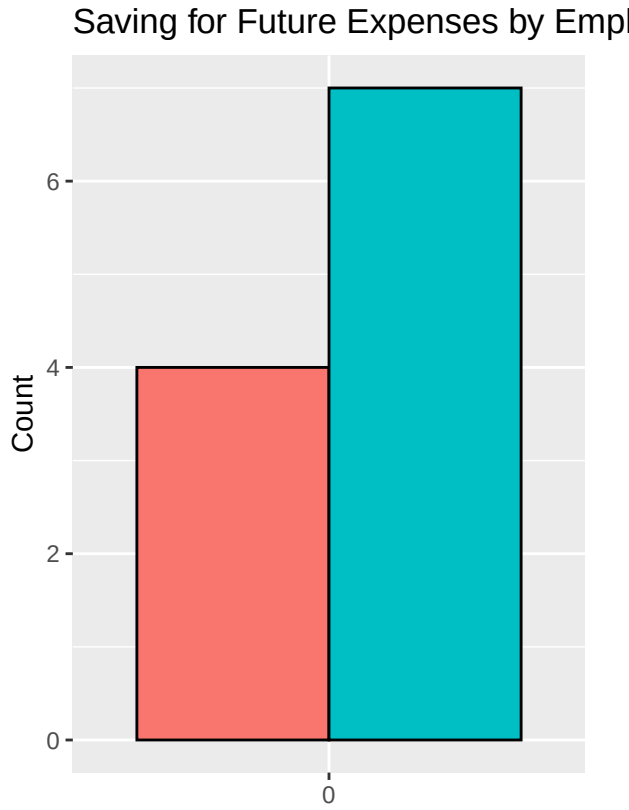
1.4.4 Additional Visualizations

```
# Bar Plot for Budget Management Methods
ggplot(df, aes(x = Budget_Management_Method)) +
  geom_bar(fill = "purple", color = "black") +
  labs(title = "Budget Management Method Distribution", x = "Budget Management Method", y = "Count")
```



1.4.4.1 Budget Management Method Distribution

```
# Bar Plot for Saving for Future Expenses vs. Employment Status
ggplot(df, aes(x = as.factor(Employment_Status), fill = as.factor(Saving_For_Future_Expenses))) +
  geom_bar(position = "dodge", color = "black") +
  labs(title = "Saving for Future Expenses by Employment Status", x = "Employment Status (0 = Unemployed
```



1.4.4.2 Saving for Future Expenses vs. Employment Status

```
# Bar Plot for Financial Independence Confidence by Employment Status
ggplot(df, aes(x = as.factor(Employment_Status), fill = as.factor(Financial_Independence_Confidence))) +
  geom_bar(position = "dodge", color = "black") +
  labs(title = "Financial Independence Confidence by Employment Status", x = "Employment Status (0 = Un
```

1.4.4.3 Financial Independence Confidence by Employment Status

1.4.5 Word Cloud Analysis

```
# Create a Word Cloud for Future Financial Decision Impact
text_data = Corpus(VectorSource(df$Future_Financial_Decision_Impact))
text_data = text_data %>%
  tm_map(content_transformer(tolower)) %>%
  tm_map(removePunctuation) %>%
  tm_map(removeNumbers) %>%
  tm_map(removeWords, stopwords("english"))
```

1.4.5.1 Word Cloud for Future Financial Decision Impact

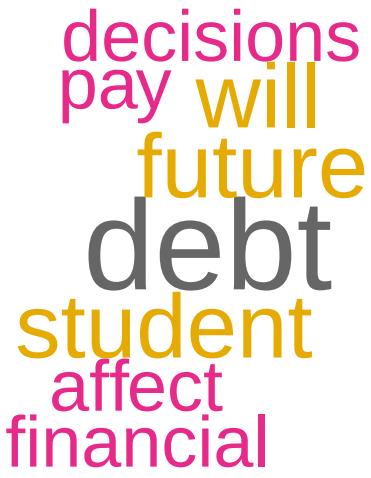
```
## Warning in tm_map.SimpleCorpus(., content_transformer(tolower)): transformation
## drops documents
```

```
## Warning in tm_map.SimpleCorpus(., removePunctuation): transformation drops
## documents
```

```
## Warning in tm_map.SimpleCorpus(., removeNumbers): transformation drops
## documents
```

```
## Warning in tm_map.SimpleCorpus(., removeWords, stopwords("english")):
## transformation drops documents
```

```
wordcloud(text_data, max.words = 100, random.order = FALSE, colors = brewer.pal(8, "Dark2"))
```



1.5 Inferential Analysis

1.5.1 Regression Analysis

```
# Linear Regression to Predict Financial Confidence
lm_model = lm(Financial_Confidence ~ Employment_Status + Student_Loans + Saving_For_Future_Expenses + Sa
summary(lm_model)
```

1.5.1.1 Predicting Financial Confidence

```
##
## Call:
## lm(formula = Financial_Confidence ~ Employment_Status + Student_Loans +
##     Saving_For_Future_Expenses + Saving_For_Retirement, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.2857 -0.4286 -0.1429  0.7143  0.8571
##
## Coefficients: (1 not defined because of singularities)
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 3.286e+00  4.860e-01   6.761 0.000262 ***
## Employment_Status                  NA         NA      NA       NA
## Student_Loans               2.857e-01  7.769e-01   0.368 0.723902
## Saving_For_Future_Expenses 2.012e-16  6.547e-01   0.000 1.000000
## Saving_For_Retirement       8.571e-01  5.915e-01   1.449 0.190573
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9258 on 7 degrees of freedom
## Multiple R-squared:  0.2667, Adjusted R-squared:  -0.04762
## F-statistic: 0.8485 on 3 and 7 DF,  p-value: 0.5099
```

1.5.2 Cross-Tabulations

```
# Cross-tabulation of Budgeting Method by Employment Status
budget_vs_employment = table(df$Budget_Management_Method, df$Employment_Status)
print(budget_vs_employment)
```

1.5.2.1 Cross-Tabulation: Budgeting Method by Employment Status

```
##
##                                0
##   I don't manage a budget      3
##   Manually (notebook or Excel) 3
##   No budgeting at all          3
##   Using a budgeting app        2
```

```
# Cross-tabulation of Financial Confidence by Saving for Future Expenses
confidence_vs_saving = table(df$Financial_Confidence, df$Saving_For_Future_Expenses)
print(confidence_vs_saving)
```

1.5.2.2 Cross-Tabulation: Financial Confidence by Saving Habits

```
##
##     0 1
##   2 1 0
##   3 0 3
##   4 3 2
##   5 0 2
```

1.5.3 Clustering Analysis

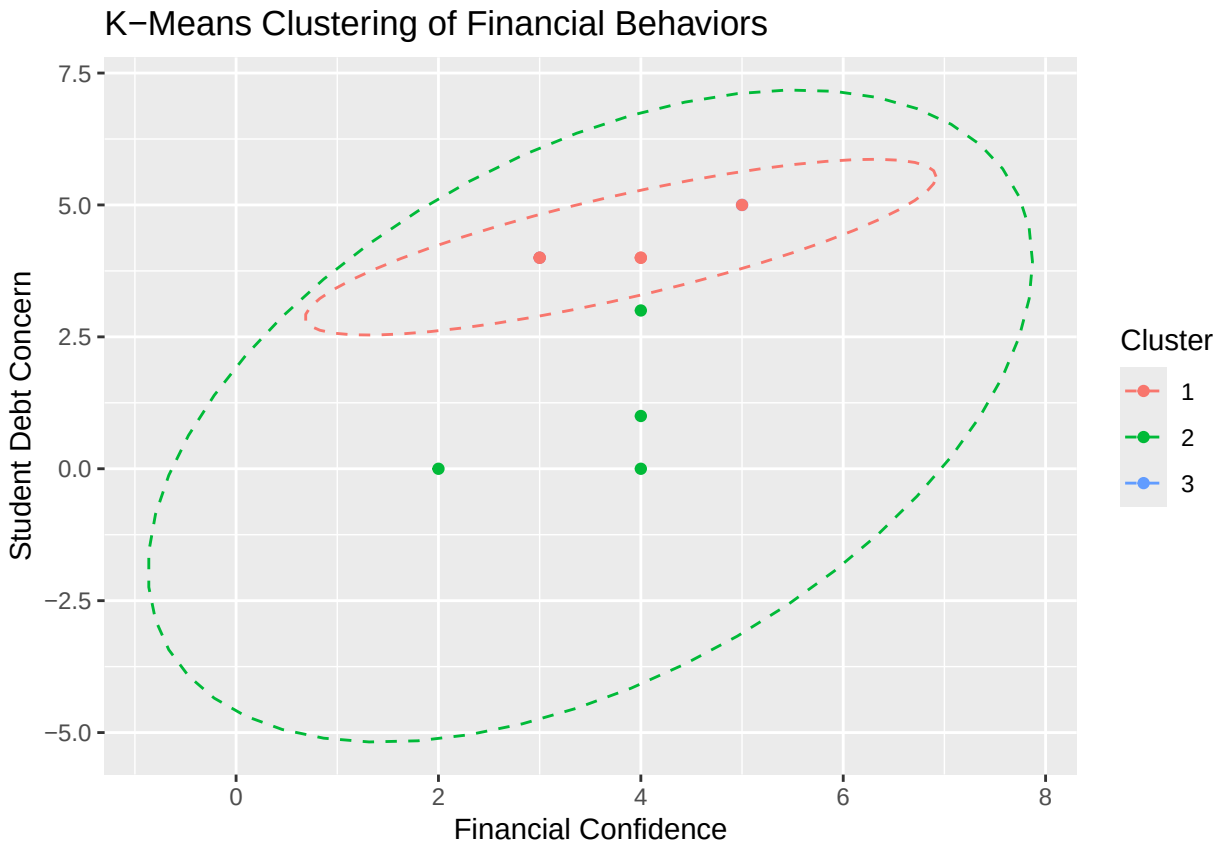
```
# K-Means Clustering
df_numeric = df %>% select_if(is.numeric)
kmeans_result = kmeans(df_numeric, centers = 3)
df$Cluster = as.factor(kmeans_result$cluster)

# Plot Clustering Results with Circles
ggplot(df, aes(x = Financial_Confidence, y = Student_Debt_Concern, color = Cluster)) +
  geom_point() +
  stat_ellipse(aes(group = Cluster), type = "norm", linetype = 2) +
  labs(title = "K-Means Clustering of Financial Behaviors", x = "Financial Confidence", y = "Student Del
```

1.5.3.1 K-Means Clustering of Financial Behaviors

```
## Too few points to calculate an ellipse
```

```
## Warning: Removed 1 row containing missing values or values outside the scale range
## (`geom_path()`).
```



1.5.4 Percentage Analysis of Survey Responses

```
# Percentage of Responses for Key Questions

# Percentage of Students with Student Loans
total_responses = nrow(df)
students_with_loans = sum(df$Student_Loans == 1)
percentage_loans = (students_with_loans / total_responses) * 100
cat("Percentage of students with student loans:", percentage_loans, "%\n")
```

```
## Percentage of students with student loans: 18.18182 %
```

```
# Percentage of Students Saving for Future Expenses
saving_for_expenses = sum(df$Saving_For_Future_Expenses == 1)
percentage_saving_expenses = (saving_for_expenses / total_responses) * 100
cat("Percentage of students saving for future expenses:", percentage_saving_expenses, "%\n")
```

```
## Percentage of students saving for future expenses: 63.63636 %
```

```
# Percentage of Students Saving for Retirement
saving_for_retirement = sum(df$Saving_For_Retirement == 1)
percentage_saving_retirement = (saving_for_retirement / total_responses) * 100
cat("Percentage of students saving for retirement:", percentage_saving_retirement, "%\n")
```

```
## Percentage of students saving for retirement: 45.45455 %
```

```
# Percentage of Employed Students
employed_students = sum(df$Employment_Status == 1)
percentage_employed = (employed_students / total_responses) * 100
```

```r
cat("Percentage of students employed while attending school:", percentage_employed, "%\n")
```

```
## Percentage of students employed while attending school: 0 %
```

```r
# Percentage of Students Confident in Financial Independence
confident_financial_independence = sum(df$Financial_Independence_Confidence == 1)
percentage_confident_independence = (confident_financial_independence / total_responses) * 100
cat("Percentage of students confident in achieving financial independence:", percentage_confident_indepe
```

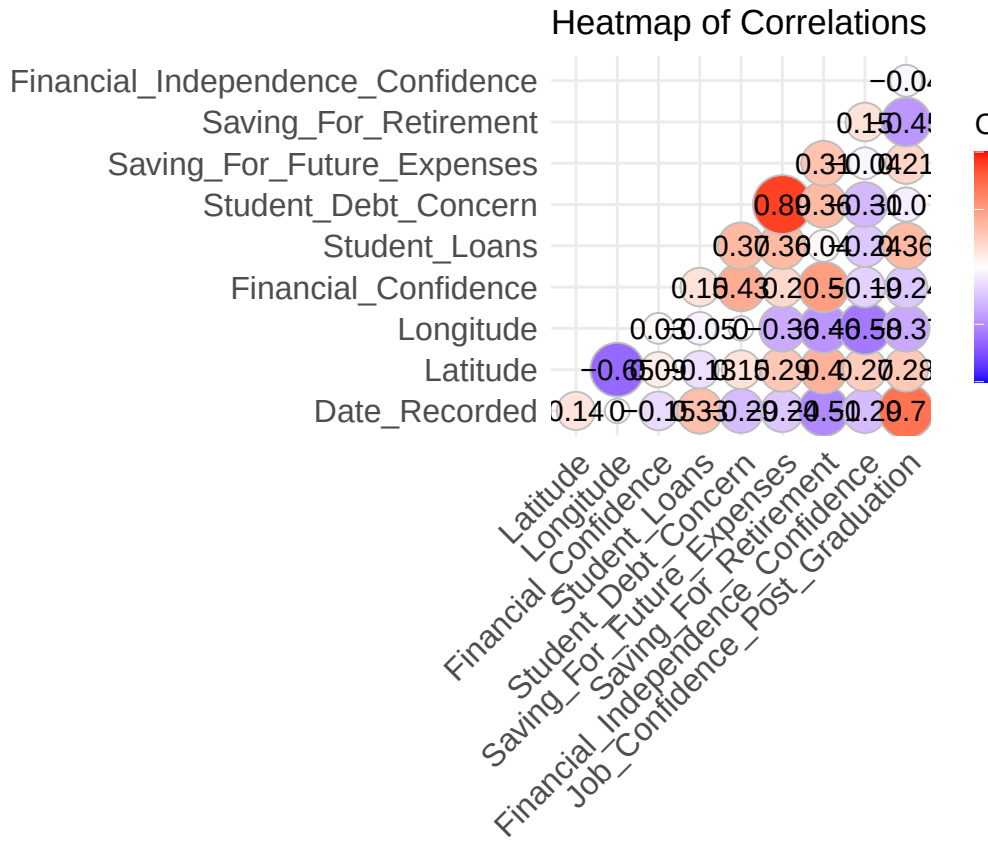
```
## Percentage of students confident in achieving financial independence: 72.72727 %
```

```r
# Percentage of Students Confident in Job Post Graduation
confident_job_post_graduation = sum(df$Job_Confidence_Post_Graduation == 1)
percentage_confident_job = (confident_job_post_graduation / total_responses) * 100
cat("Percentage of students confident in finding a job post-graduation:", percentage_confident_job, "%\n
```

```
## Percentage of students confident in finding a job post-graduation: 63.63636 %
```

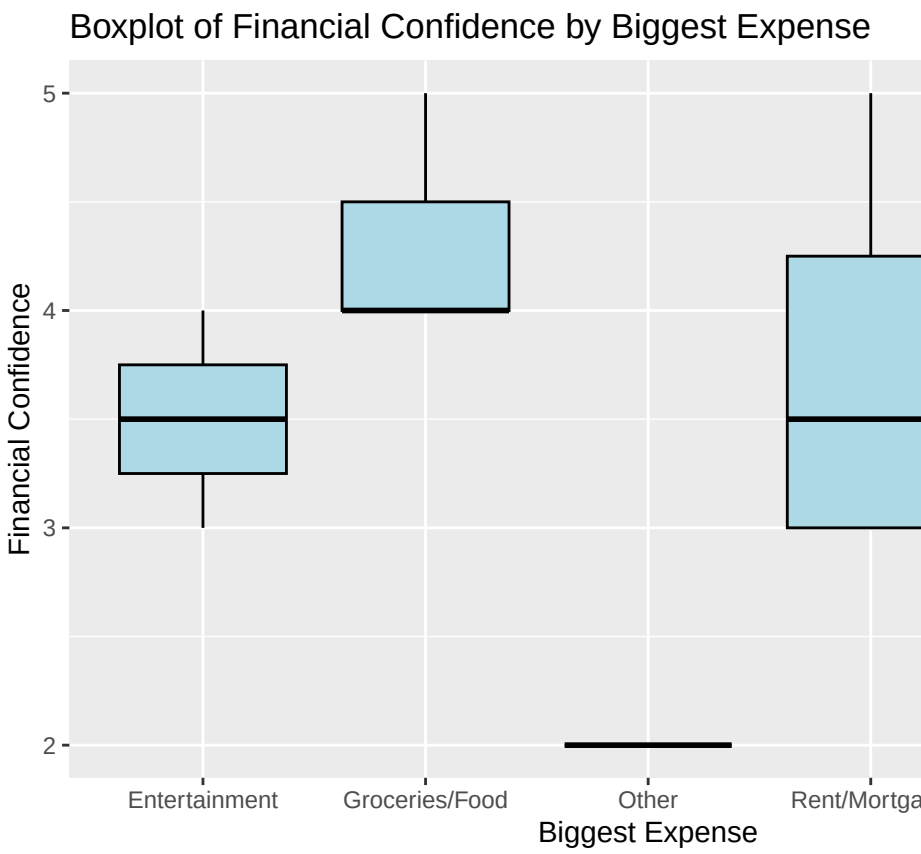
1.5.5 Additional EDA Suggestions

```r
# Heatmap of Correlations to visualize relationships between variables
ggcorrplot(correlation_matrix, method = "circle", type = "lower", lab = TRUE, title = "Heatmap of Correl
```



1.5.5.1 Heatmap of Correlations

```r
# Boxplot for Biggest Expense Categories
ggplot(df, aes(x = Biggest_Expense, y = Financial_Confidence)) +
  geom_boxplot(fill = "lightblue", color = "black") +
  labs(title = "Boxplot of Financial Confidence by Biggest Expense", x = "Biggest Expense", y = "Financ
```



1.5.5.2 Boxplots for Spending Habits

1.5.6 Geographic Analysis Using Mapview

#{r} #library(mapview) # Plot the locations using mapview #mapview(df, xcol = "Longitude", ycol = "Latitude", crs = 4269, grid = FALSE) #