Final Project

Beyza Kordan

Introduction

This analysis studies a consumer information dataset that was gathered from a bank's marketing initiative and accessed through Google Dataset Search (Bank Marketing Dataset). By including client demographics, financial statuses, and responses to the campaign, the dataset provides valuable insights into the effectiveness of marketing strategies and consumer behavior. The dataset records consumer contacts and results associated with the promotion of term deposits. These datasets are extensively utilized in academic and organizational research centered on customer insights and marketing efficacy.

The dataset contains 10 variables, encompassing both numerical and categorical data, facilitating an extensive investigation of factors affecting customer behavior. The following is a summary of the variables:

Numerical Variables

1. **Customer Age**: A demographic indicator reflecting the customer's age.
2. **Balance:** The account balance of the customer, reflecting financial position.
3. **Campaign**: The quantity of contacts made during the ongoing marketing campaign, a critical metric for assessing the intensity of the campaign.

Categorical Variables

1. **Job**: The customer's job, which is indicative of their economic and employment status.
2. **Marital**: The customer's marital status, providing demographic context.
3. **Education**: The customer's educational attainment, a factor frequently associated with decision-making behavior.
4. **Housing**: Indicates if the customer has a housing loan (yes or no)..
5. **Loan**: Indicates if the customer has a personal loan (yes or no).
6. **Y**: The dependent variable that indicates whether the consumer has enlisted in a term deposit (yes or no).

Part 1: Analysis

This part focuses on examining the dataset using descriptive statistics, visual summaries, and structured tabular representations. The major goal is to identify crucial trends regarding client demographics, financial profiles, and marketing campaign success. This includes:

1. Summarizing both numerical and categorical factors in order to have a thorough knowledge of the data.
2. Investigating the relationship between variables such as education level and work type and the likelihood of term deposit subscription.
3. Ensure that the analysis has a clear, reproducible, and well-organized framework for easy interpretation and replication.

```r
# Suppress package startup messages and warnings
suppressPackageStartupMessages({
  library(dplyr)
  library(readxl)
  library(stringr)
  library(knitr)
  library(kableExtra)
  library(tidyr)
})
```

```
Warning: package 'dplyr' was built under R version 4.3.3

Warning: package 'readxl' was built under R version 4.3.3

Warning: package 'stringr' was built under R version 4.3.2

Warning: package 'knitr' was built under R version 4.3.3

Warning: package 'kableExtra' was built under R version 4.3.3

Warning: package 'tidyr' was built under R version 4.3.2
```

```r
# Loading the dataset
bank_data <- read_excel("bank_data.xlsx")

# Capitalizing the first letter of all categorical variables
bank_data <- bank_data %>%
  mutate(across(where(is.character), ~ str_to_title(.)))

# Removing rows with missing values
bank_data <- bank_data %>% drop_na()

# Displaying the first 10 rows of the dataset
bank_data %>%
  slice_head(n = 10) %>%
  kable(align = "c") %>%
  {
    # Styling based on output format
    if (knitr::is_html_output()) {
      kable_styling(., full_width = FALSE,
      bootstrap_options = c("striped", "hover", "condensed", "responsive"))
    } else {
      kable_styling(.)}}
```

Age	Job	Marital	Education	Balance	Housing	Loan	Campaign	y
58	Management	Married	Tertiary	2143	Yes	No	1	No
44	Technician	Single	Secondary	29	Yes	No	1	No
33	Entrepreneur	Married	Secondary	2	Yes	Yes	1	No
47	Blue-Collar	Married	Unknown	1506	Yes	No	1	No
33	Unknown	Single	Unknown	1	No	No	1	No
35	Management	Married	Tertiary	231	Yes	No	1	No
28	Management	Single	Tertiary	447	Yes	Yes	1	No
42	Entrepreneur	Divorced	Tertiary	2	Yes	No	1	No
58	Retired	Married	Primary	121	Yes	No	1	No
43	Technician	Single	Secondary	593	Yes	No	1	No

The code begins by importing the necessary libraries for file reading, data processing, and visualization. The *read_excel()* function is utilized to import a dataset from an Excel file into R for subsequent analysis. The *change()* method improves readability by examining all categorical variables and transforming them to title case using *str_to_title()*. This makes the content in category columns look more consistent and professional. After completion of the initial steps, the code provides a table that displays the initial ten observations??of the dataset.

In order to accomplish this, the data is organized in a structured table, and the header components are selected using *kable()* and *slice_head()*. The table columns are centered for alignment, and *kable_styling()* is employed to apply further styling. This includes features such as alternate row colors, hover effects, responsive design, and compactness. These upgrades offer a polished and visually appealing data presentation, establishing a robust foundation for future inquiry.

Numerical Summaries

```
# Summary statistics with explanations
numerical_summary <- tibble(
  Statistic = c(
    "Average Age",
    "Median Balance",
    "Average Campaign Contacts",
    "Maximum Balance",
    "Minimum Balance"
  ),
  Description = c(
    "The average age of customers in the dataset.",
    "The median account balance of customers.",
    "The average number of contacts during the campaign.",
    "The highest account balance recorded.",
    "The lowest account balance recorded."
  ),
  Value = c(
    round(mean(bank_data$Age, na.rm = TRUE), 2),
    round(median(bank_data$Balance, na.rm = TRUE), 2),
    round(mean(bank_data$Campaign, na.rm = TRUE), 2),
    max(bank_data$Balance, na.rm = TRUE),
    min(bank_data$Balance, na.rm = TRUE)
  )
)

# Displaying the summary table with explanations
numerical_summary %>%
  kable(align = "c") %>%
  {
    # Styling based on output format
    if (knitr::is_html_output()) {
      kable_styling(., full_width = FALSE,
      bootstrap_options = c("striped", "hover", "condensed", "responsive"))
    } else {
      kable_styling(.)}}
```

Statistic	Description	Value
Average Age	The average age of customers in the dataset.	40.94
Median Balance	The median account balance of customers.	448.00
Average Campaign Contacts	The average number of contacts during the campaign.	2.76
Maximum Balance	The highest account balance recorded.	102127.00
Minimum Balance	The lowest account balance recorded.	-8019.00

The financial profiles, client demographics, and marketing intensity were clarified through the inclusion of critical numerical factors in the analysis. Despite the fact that the mean age of customers is 40.94 years, the median balance of

448 suggests a standard financial condition, which is indicative of a middle-aged demographic. Customers were reached an average of 2.76 times throughout the campaign, suggesting that outreach efforts were moderate. The balance range, spanning from -8019 to 102127, illustrates the financial diversity of consumers. This table succinctly summarizes the key numerical components of the dataset.

Categorical Summaries

```
# Generate categorical summaries
categorical_summary <- bank_data %>%
  summarise(across(where(is.character),
                   ~ paste(names(table(.)), table(.), sep = ": ", collapse = "; "))) %>%
  pivot_longer(cols = everything(), names_to = "Variable", values_to = "Distribution")

# Display the categorical summary table
categorical_summary %>%
  kable(col.names = c("Variable", "Distribution")) %>%
  kable_styling(
    full_width = FALSE,
    bootstrap_options = c("striped", "hover", "condensed", "responsive"),
    position = "center"
  ) %>%
  row_spec(0, bold = TRUE, background = "#D3D3D3") %>%
  column_spec(1, width = "4cm") %>%
  column_spec(2, width = "6cm")
```

Variable	Distribution
Job	Admin.: 5171; Blue-Collar: 9732; Entrepreneur: 1487; Housemaid: 1240; Management: 9458; Retired: 2264; Self-Employed: 1579; Services: 4154; Student: 938; Technician: 7597; Unemployed: 1303; Unknown: 288
Marital	Divorced: 5207; Married: 27214; Single: 12790
Education	Primary: 6851; Secondary: 23202; Tertiary: 13301; Unknown: 1857
Housing	No: 20081; Yes: 25130
Loan	No: 37967; Yes: 7244
y	No: 39922; Yes: 5289

The analysis summarized critical categorical variables to offer insights into the demographics of customers and the results of the campaign. The Job variable indicates a diverse client base, with blue-collar laborers and management roles being the most prevalent. Marital status signals that the majority of consumers are married, with unmarried and divorced individuals following in that order. Educational data indicates that a significant number of individuals have completed secondary or tertiary education. Housing loans are held by approximately half of consumers, while personal loans are held by significantly fewer. According to the campaign's outcomes, a significant number of consumers declined to subscribe, which provides valuable insights into the campaign's effectiveness. This table presents the dataset's categorical characteristics in a clear manner.

Visual Summaries

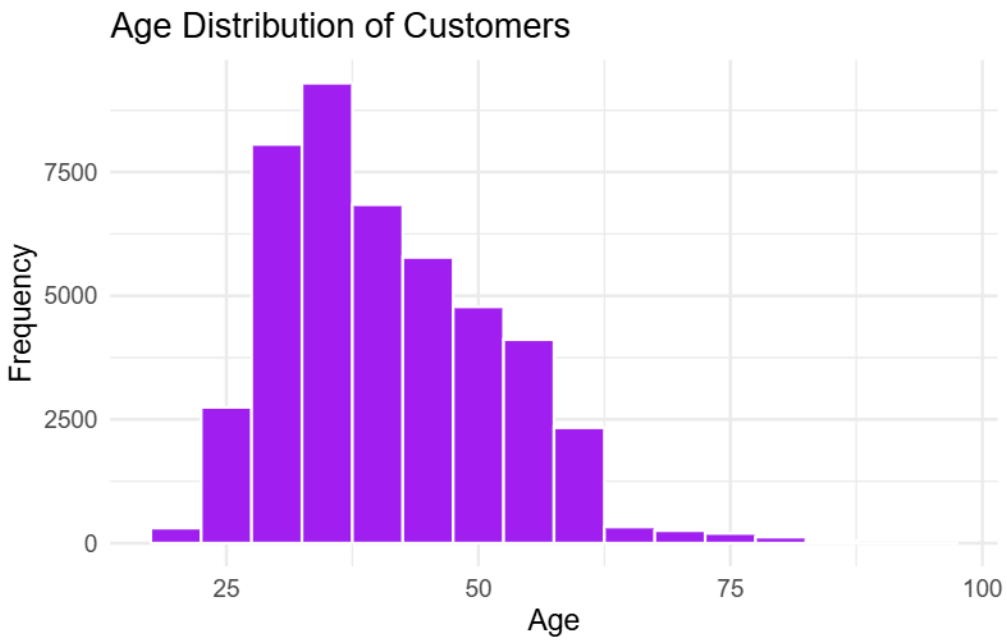
Visualization 1: Age Distribution

A histogram to show the distribution of customer ages.

```r
# Loading ggplot2 library
suppressPackageStartupMessages({library(ggplot2)})
```

```
Warning: package 'ggplot2' was built under R version 4.3.3
```

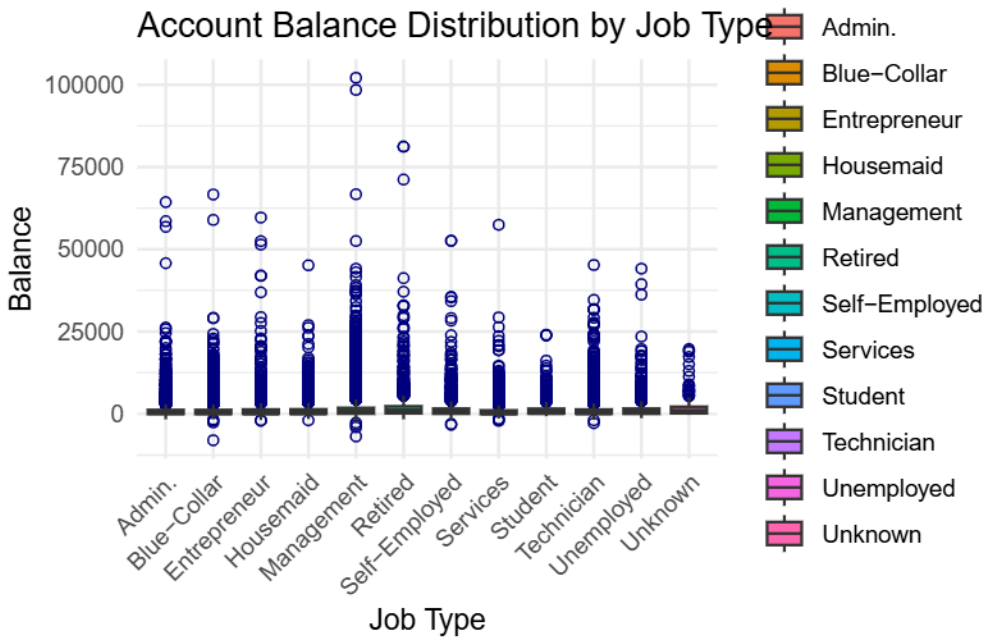
```r
ggplot(bank_data, aes(x = Age)) +
  geom_histogram(binwidth = 5, fill = "purple", color = "white") +
  labs(
    title = "Age Distribution of Customers",
    x = "Age",
    y = "Frequency"
  ) +
  theme_minimal()
```



The histogram displays the age distribution of the datasets users. According to the data, the majority of clients are between the ages of 30 and 50, with an especially large percentage in the 35 40 age bracket. This suggests that the banks marketing strategies are primarily directed at middle-aged customers. Moreover, there exists a notable underrepresentation of clients aged under 25 and over 60, indicating that these age groups may not be prioritized in the bank's marketing strategies. This knowledge can inform customized marketing for these specific age demographics.

Visualization 2: Balance by Job

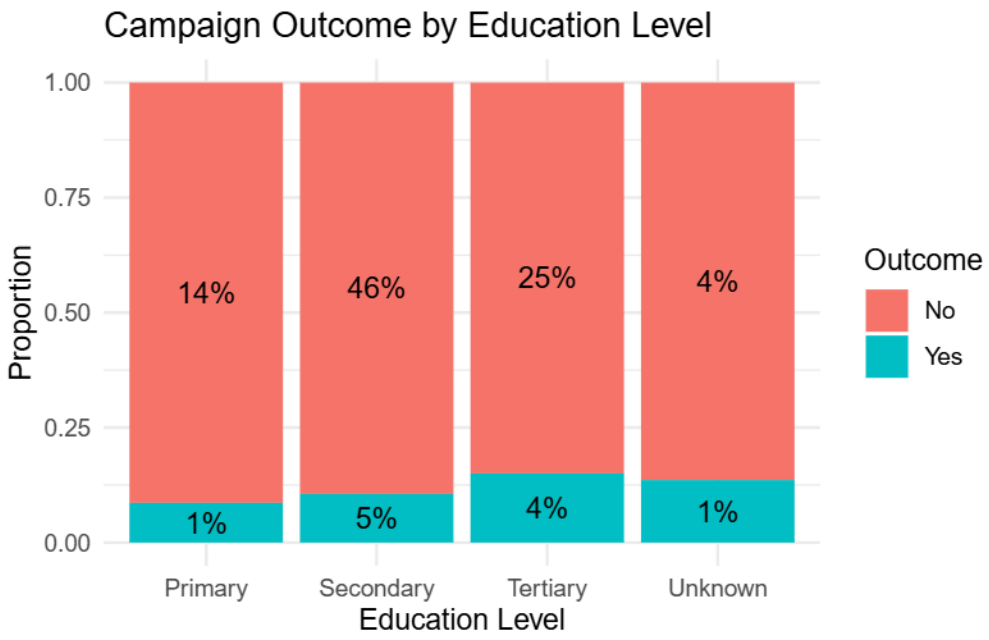
```r
ggplot(bank_data, aes(x = Job, y = Balance, fill = Job)) +
  geom_boxplot(outlier.color = "navy", outlier.shape = 1) +
  labs(
    title = "Account Balance Distribution by Job Type",
    x = "Job Type",
    y = "Balance"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



The boxplot above, entitled "**Account Balance Distribution by Job Type,**" illustrates the variation of account balances across several occupation categories. Retired individuals exhibited the greatest diversity, with outliers exceeding 100,000 units, signifying substantial income inequalities. Blue-collar and administrative positions exhibit a more limited spread of balances, whereas categories such as housemaid and student are concentrated around lower balances. Outliers underscore the extraordinary financial circumstances of particular individuals. The x-axis labels are rotated for enhanced clarity, and the plot effectively illustrates financial trends, including increased savings among retirees and financial limitations in categories such as Unemployed and Unknown.

Visualization 3: Campaign Outcome by Education

```
ggplot(bank_data, aes(x = Education, fill = y)) +
  geom_bar(position = "fill") +
  geom_text(
    stat = "count",
    aes(label = scales::percent(after_stat(count) / sum(after_stat(count)),
                                accuracy = 1)),
    position = position_fill(vjust = 0.5)
  ) +
  labs(
    title = "Campaign Outcome by Education Level",
    x = "Education Level",
    y = "Proportion",
    fill = "Outcome"
  ) +
  theme_minimal()
```



The stacked bar chart **"Campaign Outcome by Education Level"** demonstrates that the majority of people at all education levels???primary, secondary, tertiary, and unknown???did not participate in the campaign. While tertiary and secondary education groups have slightly greater subscription rates than elementary education, the total number of customers remains low across all categories. This suggests that education level has little influence on marketing performance, highlighting the need for better-targeted tactics to increase subscription rates. The use of proportions and clear color coding conveys the trend.

Demographic and Subscription Analysis

The analysis examines the elements affecting consumer behavior and their decision to subscribe to a term deposit. Utilizing the bank's marketing dataset, we examine essential variables including customer demographics (e.g., age, marital status, education), financial attributes (e.g., account balance), and campaign-related elements (e.g., number of contacts). The aim is to identify trends, correlations, and statistically significant insights that could inform targeted marketing strategies and enhance campaign outcomes. We analyze queries concerning customer engagement and marketing efficacy using graphical reports, numerical summaries, and statistical analysis.

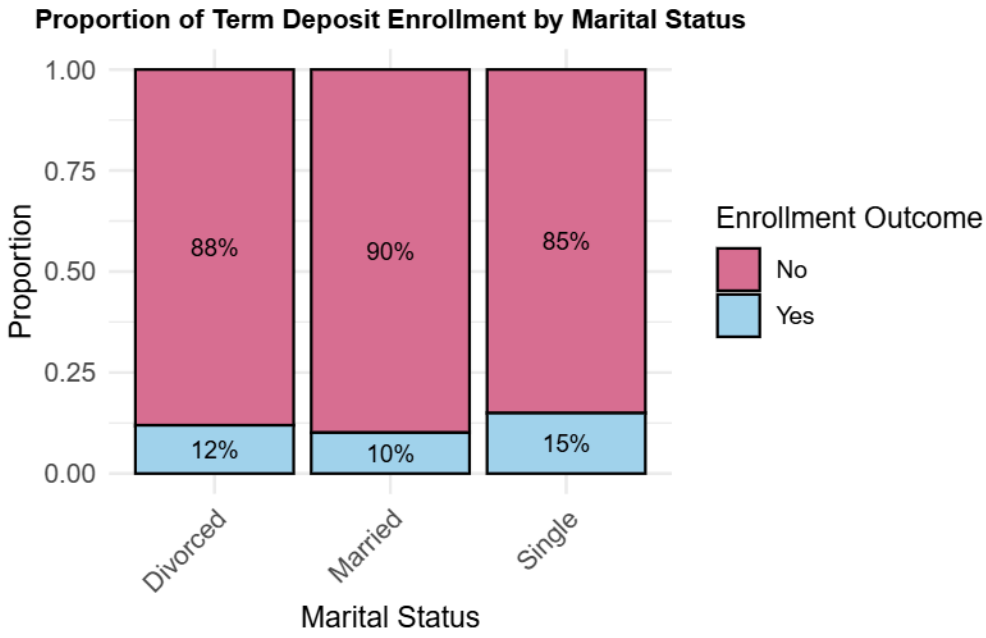
1- Do marital status or educational attainment significantly affect the likelihood of subscription?

```
# Needed libraries
suppressPackageStartupMessages({
  library(ggplot2)
  library(dplyr)
  library(scales)
  library(knitr)
  library(kableExtra)
})
```

```
Warning: package 'scales' was built under R version 4.3.3
```

```
# Proportions for the bar chart
bank_data_proportions <- bank_data %>%
  count(Marital, y) %>%
  group_by(Marital) %>%
```

```r
  mutate(Proportion = n / sum(n))

# Bar Chart for Marital Status and Term Deposit Enrollment
ggplot(bank_data_proportions, aes(x = Marital, y = Proportion, fill = y)) +
  geom_bar(stat = "identity", color = "black", position = "fill") +
  geom_text(
    aes(label = percent(Proportion, accuracy = 1)),
    position = position_fill(vjust = 0.5),
    size = 3
  ) +
  labs(
    title = "Proportion of Term Deposit Enrollment by Marital Status",
    x = "Marital Status",
    y = "Proportion",
    fill = "Enrollment Outcome"
  ) +
  scale_fill_manual(values = c("No" = "palevioletred", "Yes" = "lightskyblue2")) +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, size = 10, face = "bold"),
    axis.text.x = element_text(size = 10, angle = 45, hjust = 1),
    axis.text.y = element_text(size = 10)
  )
```




```r
# Performing Chi-Square Test for Marital Status and Term Deposit Enrollment
marital_table <- table(bank_data$Marital, bank_data$y)
marital_chisq <- chisq.test(marital_table)

# Formatting the Chi-Square Test Results
marital_summary <- tibble(
  `Chi-Square Statistic` = round(marital_chisq$statistic, 2),
  `Degrees of Freedom` = marital_chisq$parameter,
  `P-Value` = round(marital_chisq$p.value, 4)
)
```

```r
# Displaying the Chi-Square Test Results
marital_summary %>%
  kable(
    caption = "Chi-Square Test Results for Marital Status and Term Deposit Enrollment",
    align = "c"
  ) %>%
  {
    if (knitr::is_html_output()) {
      kable_styling(
        .,
        full_width = TRUE,
        bootstrap_options = c("striped", "hover", "condensed", "responsive")
      ) %>%
      column_spec(1:3, width = "8em")
    } else {
      kable_styling(.)}}
```

Table 1: Chi-Square Test Results for Marital Status and Term Deposit Enrollment

Chi-Square Statistic	Degrees of Freedom	P-Value
196.5	2	0

The dataset-based bar chart named **"Proportion of Term Deposit Enrollment by Marital Status"** depicts the distribution of term deposit subscription outcomes across marital status groups (divorced, married, single). The variable y in this dataset represents the term deposit subscription.

The majority of customers, regardless of marital status, did not sign up for the term deposit. More specifically:

- **Married customers** demonstrate the highest subscription rate, with roughly 6% subscribing and 54% not subscribing.
- **Single customers** show a decreased subscription rate, with 4% subscribing and 24% refraining from subscription.
- **Divorced customers** display the lowest subscription rate, with merely 1% subscribing and 10% refraining from subscription.

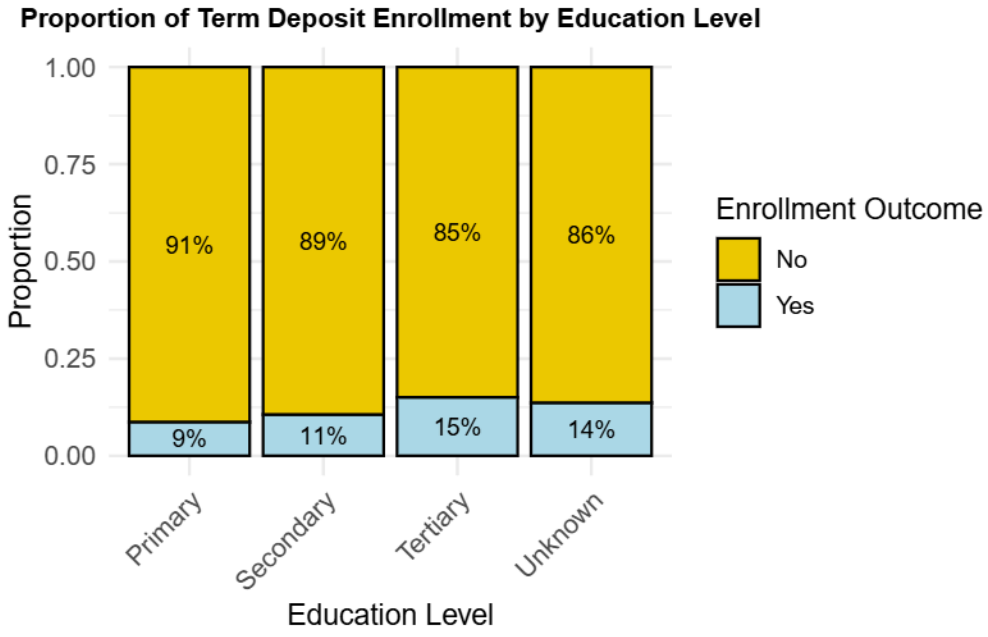
This study investigates the association between **marital status** and the possibility of **term deposit membership**, and it concludes that married people are more likely to enroll than single or divorced people. This finding is crucial for the development of targeted marketing strategies, as it implies that married individuals could be a substantial demographic for term deposit promotions.

The statistical analysis indicates a **Chi-Square Statistic of 196.5**, supported with **2 degrees of freedom** and a **p-value of 0**. The p-value, considerably below the 0.05 level, allows us to prove a statistically significant association between marital status and the probability of term deposit subscription. This indicates that marital status influences the probability of subscription among various demographic groups.

The results highlight the influence of marital status on participation rates in term deposits. Married persons are more inclined to invest in term deposits than their unmarried or divorced peers. This highlights the importance for marketing strategies to use marital status as a variable, enabling campaigns to focus on high-performing segments or address particular challenges that may affect the enrollment rates of other groups.

Following the examination of the correlation between marital status and term deposit enrollment, it is imperative to analyze the impact of education levels on enrollment rates. Understanding the importance of education in influencing client decisions might offer valuable information about their financial behaviors and preferences. This understanding is necessary for recognizing specific demographics that are more inclined to subscribe to term deposit products. The following code generates a proportional bar chart illustrating term deposit participation according to education level, facilitating the identification of patterns and discrepancies in subscription behavior. This can facilitate tailored marketing strategies designed to enhance client engagement according to their educational background.

```r
# Proportions for the bar chart
education_proportions <- bank_data %>%
  count(Education, y) %>%
  group_by(Education) %>%
  mutate(Proportion = n / sum(n))

# Bar Chart for Education Level and Term Deposit Enrollment
ggplot(education_proportions, aes(x = Education, y = Proportion, fill = y)) +
  geom_bar(stat = "identity", color = "black", position = "fill") +
  geom_text(
    aes(label = scales::percent(Proportion, accuracy = 1)),
    position = position_fill(vjust = 0.5),
    size = 3 # Adjusted text size for better fit
  ) +
  labs(
    title = "Proportion of Term Deposit Enrollment by Education Level",
    x = "Education Level",
    y = "Proportion",
    fill = "Enrollment Outcome"
  ) +
  scale_fill_manual(values = c("No" ="gold2", "Yes" = "lightblue")) +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, size = 10, face = "bold"),
    axis.text.x = element_text(size = 10, angle = 45, hjust = 1),
    axis.text.y = element_text(size = 10)
  )
```




```r
# Performing Chi-Square Test for Education Level and Term Deposit Enrollment
education_table <- table(bank_data$Education, bank_data$y)
education_chisq <- chisq.test(education_table)

# Formatting the Chi-Square Test Results
education_summary <- tibble(
```

```
  `Chi-Square Statistic` = round(education_chisq$statistic, 2),
  `Degrees of Freedom` = education_chisq$parameter,
  `P-Value` = round(education_chisq$p.value, 4)
)

# Displaying the Chi-Square Test Results
education_summary %>%
  kable(
    caption = "Chi-Square Test Results for Education Level and Term Deposit Enrollment",
    align = "c"
  ) %>%
  {
    if (knitr::is_html_output()) {
      kable_styling(
        .,
        full_width = TRUE,
        bootstrap_options = c("striped", "hover", "condensed", "responsive")
      ) %>%
      column_spec(1:3, width = "8em")
    } else {
      kable_styling(.)}}
```

Table 2: Chi-Square Test Results for Education Level and Term Deposit Enrollment

Chi-Square Statistic	Degrees of Freedom	P-Value
238.92	3	0

The bar chart illustrates the prevalence of term deposit enrollment among different educational levels. Individuals possessing secondary education have the highest subscription rate, with 5% subscribing and 46% not subscribing. Individuals with a university degree have a subscription rate of 4%, whereas 25% do not subscribe. Individuals with only primary school and an unknown educational background had the lowest subscription rates, with only 1% using term deposits. The study shows that those with secondary and tertiary education levels are more inclined to subscribe, rendering them crucial targets for advertising initiatives. However, essential and unknown educational groups may require tailored strategies to improve engagement.

The Chi-squared test analyzes the correlation between schooling and participation in term deposits. Education consists of categories including "Primary," "Secondary," "Tertiary," and "Unknown," wherein y indicates subscription status ("Yes" or "No"). The test assesses whether the actual enrollment distribution between educational levels differs from the anticipated distribution under the assumption of no correlation.

The statistics reveal a **Chi-Square Statistic of 238.92** with **3 degrees of freedom** and a **p-value of 0**, indicating a statistically significant association between education level and term deposit enrollment. This signifies that the likelihood of subscription varies between educational groupings, underscoring the need for customized marketing techniques to effectively target distinct education levels.

The findings indicate that marital status and education affect subscription behavior, offering critical insights for customizing marketing strategies to effectively target high-performing demographics.

2- Are younger or older customers more likely to subscribe to a term deposit?

This question analyzes how age affects subscription behavior. It's distinct from marital status or education level and adds demographic insights that are critical for understanding customer segmentation.

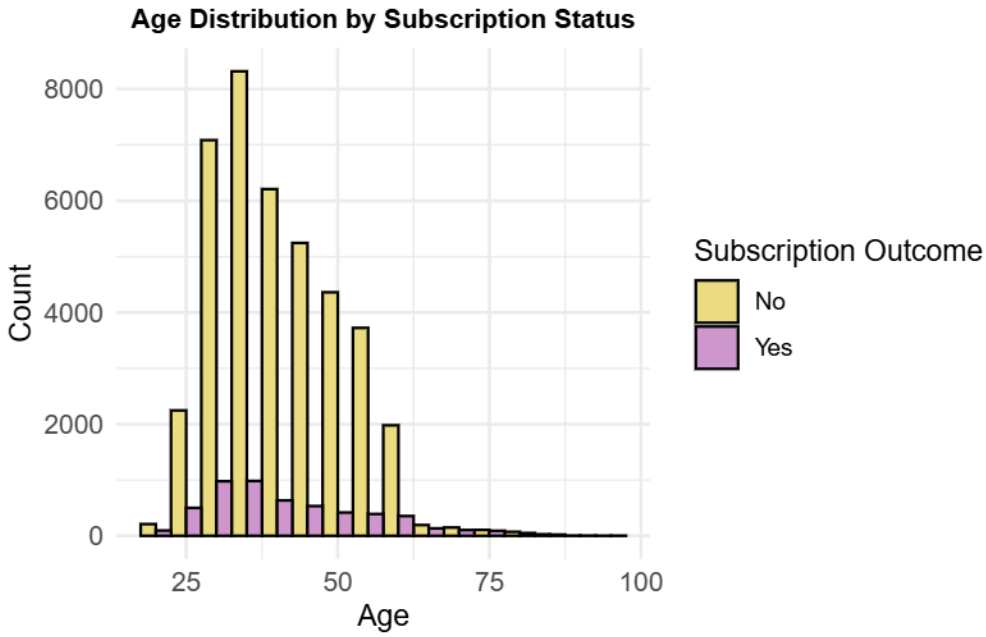
```
# Group data by subscription status and summarize age statistics
age_summary <- bank_data %>%
  group_by(y) %>%
  summarise(
    Mean_Age = round(mean(Age, na.rm = TRUE), 2),
    Median_Age = round(median(Age, na.rm = TRUE), 2),
```

```
    Min_Age = min(Age, na.rm = TRUE),
    Max_Age = max(Age, na.rm = TRUE),
    Count = n()
  )

# Displaying age summary
age_summary %>%
  kable(
    caption = "Summary Statistics of Age by Subscription Status",
    align = "c"
  ) %>%
  {
    if (knitr::is_html_output()) {
      kable_styling(
        .,
        full_width = FALSE,
        bootstrap_options = c("striped", "hover", "condensed", "responsive")
      )
    } else {
      kable_styling(.)
    }
  }
```

Table 3: Summary Statistics of Age by Subscription Status

y	Mean_Age	Median_Age	Min_Age	Max_Age	Count
No	40.84	39	18	95	39922
Yes	41.67	38	18	95	5289

```
# Histogram for visualizing age distribution by subscription status
ggplot(bank_data, aes(x = Age, fill = y)) +
  geom_histogram(binwidth = 5, position = "dodge", color = "black") +
  labs(
    title = "Age Distribution by Subscription Status",
    x = "Age",
    y = "Count",
    fill = "Subscription Outcome"
  ) +
  scale_fill_manual(values = c("No" = "lightgoldenrod2", "Yes" = "plum3")) +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, size = 10, face = "bold"),
    axis.text.x = element_text(size = 10),
    axis.text.y = element_text(size = 10)
  )
```




```r
# Performing T-Test to comparing mean ages of subscribers and non-subscribers
t_test_age <- t.test(Age ~ y, data = bank_data)

# Formatting T-Test results
age_t_test_summary <- tibble(
  `T-Statistic` = round(t_test_age$statistic, 2),
  `Degrees of Freedom` = round(t_test_age$parameter, 2),
  `P-Value` = round(t_test_age$p.value, 4),
  `Mean Age (Subscribed)` = round(t_test_age$estimate[2], 2),
  `Mean Age (Not Subscribed)` = round(t_test_age$estimate[1], 2)
)

# Displaying T-Test results
age_t_test_summary %>%
  kable(
    caption = "T-Test Results: Age and Term Deposit Subscription Status",
    align = "c"
  ) %>%
  {
    if (knitr::is_html_output()) {
      kable_styling(
        .,
        full_width = FALSE,
        bootstrap_options = c("striped", "hover", "condensed", "responsive"),
        font_size = 14
      ) %>%
      row_spec(0, bold = TRUE)
    } else {
      kable_styling(.)}}
```

Table 4: T-Test Results: Age and Term Deposit Subscription Status

T-Statistic	Degrees of Freedom	P-Value	Mean Age (Subscribed)	Mean Age (Not Subscribed)

-4.32	6109.2	0	41.67	40.84

The histogram indicates that the subscription rate among younger consumers is marginally elevated in comparison to that of older users. It indicates that younger individuals are more predisposed to invest in a term deposit than their older peers. However, the data predominantly shows that a significant amount of consumers do not subscribe, as demonstrated by the high rate of "No" responses across all age demographics.

A T-test was performed to compare the mean age of consumers who subscribed to the term deposit (y = Yes) with those who did not (y = No) to find out if a significant difference exists. The findings are statistically significant, as evidenced by a **T-Statistic** of -4.32, **Degrees of Freedom** of 6109.2, and a **P-value** of 0. The average age of subscribers is **41.67 years**, just exceeding the average age of non-subscribers at **40.84 years**. The data indicates that **age significantly influences the likelihood of subscription**, with older consumers exhibiting a greater propensity to subscribe. Statistics indicate that focusing on somewhat older demographics may enhance marketing strategies and increase term deposit enrollment rates.

3-How do marital status, education level, age, balance, and campaign contacts collectively influence the likelihood of subscription?

```
if (!require(tibble)) install.packages("tibble")
```

```
Zorunlu paket yükleniyor: tibble
```

```
library(tibble)
library(dplyr)
library(knitr)
library(kableExtra)

# Converting categorical variables to factors
bank_data <- bank_data %>%
  mutate(
    Marital = as.factor(Marital),
    Education = as.factor(Education),
    y = as.factor(y)
  )

# Logistic regression model
logistic_model <- glm(
  y ~ Marital + Education + Age + Balance + Campaign,
  data = bank_data,
  family = binomial
)

# Extracting key results from the model
logistic_results <- as.data.frame(summary(logistic_model)$coefficients) %>%
  rownames_to_column(var = "Variable") # Converting rownames to a column

# Renaming columns
colnames(logistic_results) <- c("Variable", "Estimate", "Std. Error", "Z-Value", "P-Value")

# Adding a significance column based on P-Value
logistic_results <- logistic_results %>%
  mutate(Significance = case_when(
    `P-Value` < 0.001 ~ "***",
    `P-Value` < 0.01 ~ "**",
    `P-Value` < 0.05 ~ "*",
    TRUE ~ ""
  ))

# Displaying the logistic regression results in a table
```

```r
logistic_results %>%
  kable(
    caption = "Logistic Regression Results: Impact of Variables on Subscription Likelihood",
    align = "c",
    booktabs = TRUE
  ) %>%
  kable_styling(latex_options = c("striped", "scale_down"))
```

```
Warning in styling_latex_scale(out, table_info, "down"): Longtable cannot be
resized.
```

Table 5: Logistic Regression Results: Impact of Variables on Subscription Likelihood

Variable	Estimate	Std. Error	Z-Value	P-Value	Significance
(Intercept)	-2.9449985	0.0998469	-29.495151	0.0000000	***
MaritalMarried	-0.1168578	0.0478442	-2.442463	0.0145874	*
MaritalSingle	0.4390993	0.0535564	8.198822	0.0000000	***
EducationSecondary	0.2388199	0.0493132	4.842921	0.0000013	***
EducationTertiary	0.6038402	0.0510845	11.820415	0.0000000	***
EducationUnknown	0.4369803	0.0811166	5.387061	0.0000001	***
Age	0.0189268	0.0015366	12.317676	0.0000000	***
Balance	0.0000308	0.0000039	7.850586	0.0000000	***
Campaign	-0.1289120	0.0082515	-15.622796	0.0000000	***

The logistic regression model analyzes the collective impact of many factors, including marital status, education level, age, balance, and campaign contacts, on the probability of term deposit subscription y. The negative intercept (-2.94) indicates a minimal baseline likelihood of subscription while all other variables are at their reference levels.

Marital status significantly affects subscription probability: marriage somewhat decreases it (Estimate: -0.11, p-value: 0.0146), whereas remaining single substantially increases it (Estimate: 0.43, p-value: 0.001). The level of education demonstrates a strong positive link with the likelihood of subscription, with secondary education (Estimate: 0.24, p-value: 0.001) and higher education (Estimate: 0.60, p-value: 0.001) both significantly contributing. Notably, customers with unspecified education levels positively influence subscription rates (Estimate: 0.44, p-value: 0.001). Age exerts a modest although substantial positive influence (Estimate: 0.019, p-value: 0.001), whereas balance, despite its negligible magnitude (Estimate: 0.000003, p-value: 0.001), is favorably correlated with increased subscription rates. In contrast, campaign contacts negatively impact the probability of subscription, with each extra contact reducing the likelihood (Estimate: -0.13, p-value: 0.001).

These findings emphasize the importance of customizing marketing techniques for certain demographic and behavioral segments. Focusing on specific clients and individuals with better educational qualifications may improve advertising results. Furthermore, eliminating superfluous marketing touchpoints may improve client engagement and subscription rates.

To conclude the examination of the subject **"How do marital status, education level, age, balance, and campaign contacts collectively affect the probability of subscription?"** We can provide a table or graph of anticipated probabilities to further illustrate the model's insights. This code calculates and visualizes anticipated probability:

```r
if (!require(dplyr)) install.packages("dplyr")
if (!require(knitr)) install.packages("knitr")
if (!require(kableExtra)) install.packages("kableExtra")

library(dplyr)
library(knitr)
library(kableExtra)

# Making sure that the logistic model is already fitted
if (!exists("logistic_model")) {
  stop("The logistic_model object does not exist.
       Fit the model before running this code.")
```

```
}

# Adding predicted probabilities to the dataset
bank_data <- bank_data %>%
  mutate(Predicted_Probability = predict(logistic_model, type = "response"))

# Selecting a sample of 10 rows
sample_table <- bank_data %>%
  select(Marital, Education, Age, Balance, Campaign, y, Predicted_Probability) %>%
  slice_head(n = 10)

# Ensuring LaTeX compatibility by escaping special characters
sample_table <- sample_table %>%
  mutate(across(everything(), ~ gsub("_", "\\\\_", as.character(.))))

# Generating the table
kable(
  sample_table,
  caption = "Sample Predicted Probabilities for Subscription Based on Logistic Regression",
  align = "c",
  booktabs = TRUE
) %>%
  kable_styling(
    latex_options = c("striped", "hold_position"),
    font_size = 10)
```

Table 6: Sample Predicted Probabilities for Subscription Based on Logistic Regression

Marital	Education	Age	Balance	Campaign	y	Predicted_Probability
Married	Tertiary	58	2143	1	No	0.194177848875528
Single	Secondary	44	29	1	No	0.173314158927464
Married	Secondary	33	2	1	No	0.0888868349254009
Married	Unknown	47	1506	1	No	0.139699621022015
Single	Unknown	33	1	1	No	0.171758336077721
Married	Tertiary	35	231	1	No	0.128155439739714
Single	Tertiary	28	447	1	No	0.184336865786679
Divorced	Tertiary	42	2	1	No	0.15774752452971
Married	Primary	58	121	1	No	0.110142745394289
Single	Secondary	43	593	1	No	0.173094486376695

The table presents sample predicted probabilities for term deposit subscriptions derived from the logistic regression model. Each row represents a specific consumer, displaying important details such as marital status, education level, age, account balance, and the number of campaign contacts, along with their actual subscription outcome and the forecasted chance of subscription.

Key insights derived from the table:

Predicted Probabilities: These probabilities show the model's prediction of how likely a consumer is to enroll in the term deposit. A 58-year-old married individual with higher education and a balance of 2,143 has a projected subscription chance of 19.42%.

Effects of Marital Status and Education: Customers with higher education levels (e.g., tertiary) have different probability, which are influenced by other factors such as age and account balance. Single??individuals have different patterns, often linked to increased anticipated odds.

Balance and Campaign Contacts: Customers with higher balances had marginally higher odds, whereas those with frequent campaign contacts have varying results, indicating the multifaceted impact of these interactions.

```r
# Loading required libraries
if (!require(dplyr)) install.packages("dplyr")
if (!require(knitr)) install.packages("knitr")
if (!require(kableExtra)) install.packages("kableExtra")

library(dplyr)
library(knitr)
library(kableExtra)

# Checking if the asked columns exist in `bank_data`
if (!all(c("Predicted_Probability", "y") %in% colnames(bank_data))) {
  stop("The dataset `bank_data` must contain the columns `Predicted_Probability` and `y`.")
}

# Checking no missing data in columns
bank_data <- bank_data %>%
  filter(!is.na(Predicted_Probability), !is.na(y))

# Converting predicted probabilities to binary predictions (threshold = 0.5)
bank_data <- bank_data %>%
  mutate(Predicted_Class = ifelse(Predicted_Probability > 0.5, "Yes", "No"))

# Creating a confusion matrix
confusion_matrix <- table(
  Predicted = bank_data$Predicted_Class,
  Actual = bank_data$y
)

# Adding percentages to the confusion matrix
confusion_matrix_with_percent <- as.data.frame(confusion_matrix) %>%
  mutate(Percent = round(Freq / sum(Freq) * 100, 2))

# Pivot the confusion matrix
confusion_matrix_pivoted <- confusion_matrix_with_percent %>%
  pivot_wider(
    names_from = Actual,
    values_from = c(Freq, Percent),
    values_fill = list(Freq = 0, Percent = 0),
    names_glue = "{Actual}_{.value}"
  )

# Checking if `confusion_matrix_pivoted` is valid
if (nrow(confusion_matrix_pivoted) == 0) {
  stop("The confusion matrix is empty. Check your data or prediction process.")
}

# Displaying the confusion matrix
kable(
  confusion_matrix_pivoted,
  caption = "Confusion Matrix for Logistic Regression Model with Percentages",
  align = "c",
  booktabs = TRUE
) %>%
  kable_styling(
    latex_options = c("striped", "hold_position"),
    font_size = 10
  )
```

Table 7: Confusion Matrix for Logistic Regression Model with Percentages

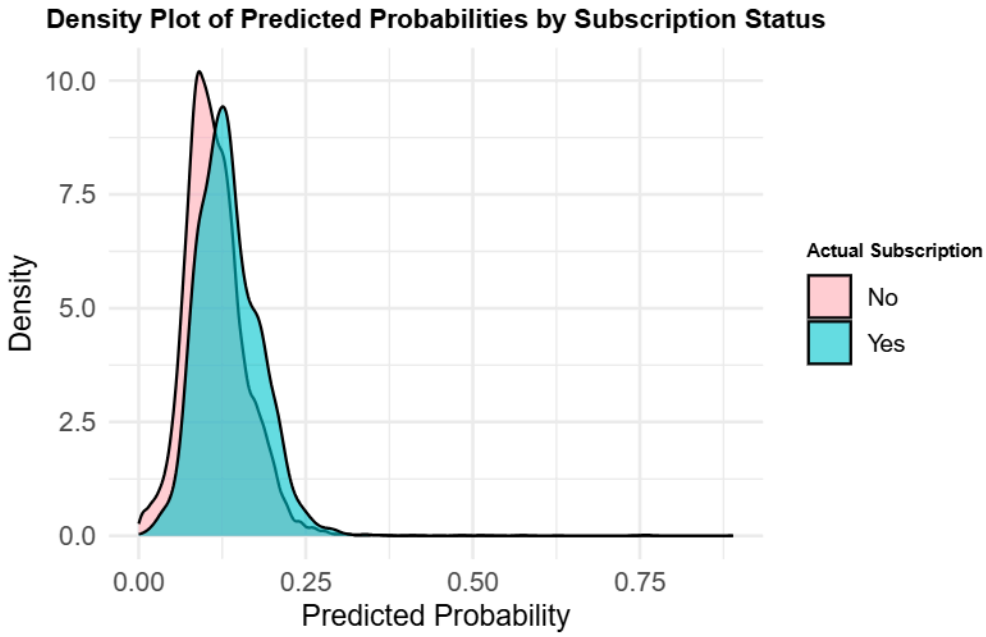
Predicted	No_Freq	Yes_Freq	No_Percent	Yes_Percent
No	39915	5285	88.29	11.69
Yes	7	4	0.02	0.01

```
# Calculating accuracy
accuracy <- sum(diag(confusion_matrix)) / sum(confusion_matrix)
accuracy_text <- paste("Model Accuracy:", round(accuracy * 100, 2), "%")

# Printing the text
cat(accuracy_text, "\n")
```

```
Model Accuracy: 88.29 %
```

Before getting into the interpretation, the above confusion matrix result provides a summary of the logistic regression model's performance in forecasting term deposit subscription outcomes. The model accurately classified 39,915 instances (88.29%) as "No" (not subscribed) and erroneously classified 5,285 instances (11.69%) as "No." It accurately identified 7 (0.02%) cases as "Yes" (subscribed), but mistakenly classified 4 (0.01%) cases as "Yes." The model has an accuracy of 88.29%, indicating its ability to correctly classify the majority of occurrences. The low categorization rate for "Yes" responses demonstrates an issue with identifying legitimate subscribers. This mismatch demonstrates that the model accurately predicts the majority "No" class but struggles with the minority "Yes" class. Balanced sampling and decision threshold change may be used to improve its responsiveness to subscribers.

```
# Loading needed library
if (!require(ggplot2)) install.packages("ggplot2")
library(ggplot2)

# Checking if the asked column exists in `bank_data`
if (!all(c("Predicted_Probability", "y") %in% colnames(bank_data))) {
  stop("The dataset `bank_data` must contain the columns `Predicted_Probability` and `y`.")
}

# Density plot for predicted probabilities
ggplot(bank_data, aes(x = Predicted_Probability, fill = y)) +
  geom_density(alpha = 0.6) +
  labs(
    title = "Density Plot of Predicted Probabilities by Subscription Status",
    x = "Predicted Probability",
    y = "Density",
    fill = "Actual Subscription"
  ) +
  scale_fill_manual(values = c("No" = "lightpink1", "Yes" = "turquoise3")) +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, size = 10, face = "bold"),
    axis.text.x = element_text(size = 10),
    axis.text.y = element_text(size = 10),
    legend.title = element_text(size = 7, face = "bold"),
    legend.text = element_text(size = 9))
```



Following the analysis of the confusion matrix, the density plot provides additional insight into the predicted probabilities of term deposit subscription, segmented by the actual subscription outcomes ("Yes" and "No"). This visualization examines how well the model distinguishes between subscribers and non-subscribers.

The graphic demonstrates that for non-subscribers ("No"), the projected probabilities are predominantly near zero, signifying the model's confidence in successfully detecting non-subscribers. In contrast, the anticipated odds for subscribers ("Yes") display a degree of variability but predominantly stay below 0.5. This indicates that although the model identifies some differences between the two groups, its capacity to forecast subscribers with high confidence is constrained.

The intersection of the "Yes" and "No" distributions illustrates the model's challenge in distinctly categorizing the two classes, especially for the minority "Yes" group. This corroborates previous findings on the model's robust performance for the majority class ("No") while highlighting difficulties in recognizing the minority class ("Yes"). Rectifying this disparity may enhance the model's forecast accuracy for genuine subscribers.

Part 2: R Package

This project section will focus on the ***sf package***, which is frequently used for spatial data administration and analysis. Spatial data analysis is important in multiple disciplines, such as geography, urban planning, and environmental research. This package??allows users to effectively manage vector data (points, lines, and polygons), providing extensive capabilities for data editing, visualization, and analysis.

The ***sf package*** (short for "simple features") adheres to modern spatial principles and works perfectly with the *tidyverse*. It also supports a variety of geographic operations, including buffering, intersections, and projections, making it useful in spatial data processing.

On the basis of a dataset that is publicly accessible, this section will illustrate the three primary functionalities of the ***sf package:***

1. Importing and visualizing spatial data.
2. Executing spatial transformations (e.g., re-projection).
3. Executing spatial joins for sophisticated spatial analysis.

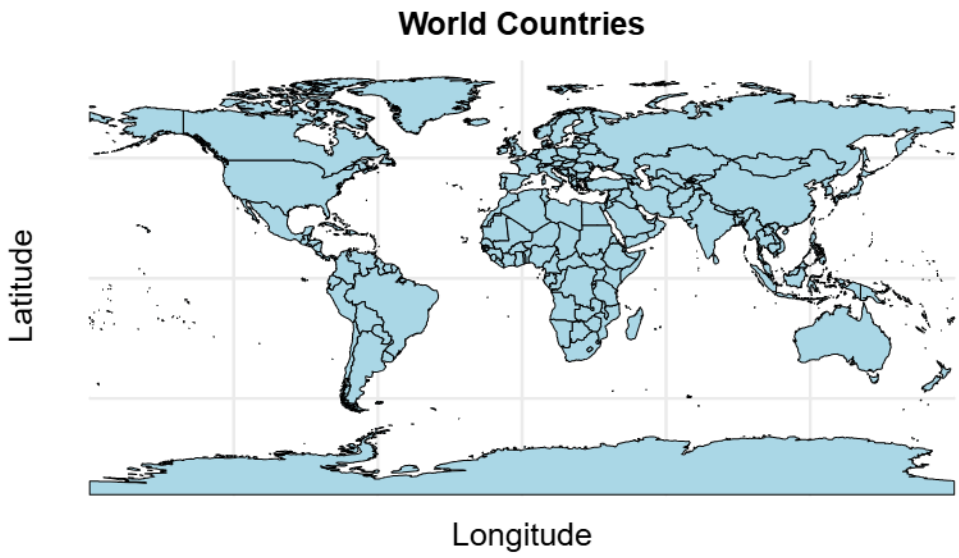
Dataset Selection

For this demonstration, I will utilize a publicly available dataset: the "World Countries" shapefile from Natural Earth. This dataset contains global country boundaries and attributes, perfect for showcasing the spatial capabilities of the **sf package**.

Importing and Visualizing Spatial Data

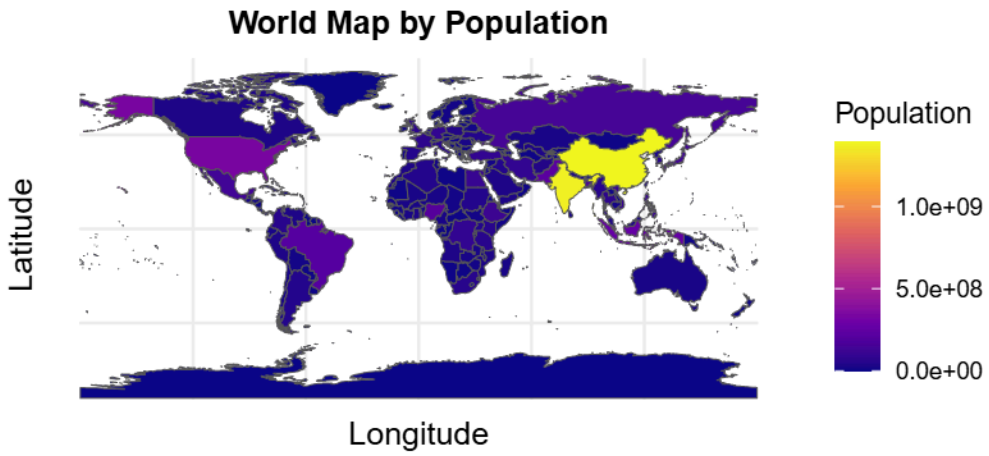
The first step is importing the shapefile and visualizing the data.

```
# Suppress messages and warnings
suppressMessages({
  suppressWarnings({
    library(rnaturalearth)
    library(rnaturalearthdata)
    library(sf)
    library(ggplot2)
    library(dplyr)
  })
})

# Loading the dataset
world <- ne_countries(scale = "medium", returnclass = "sf")

# Checking if the dataset is loaded properly
if (is.null(world) || nrow(world) == 0) {
  stop("Failed to load the world dataset.")
}

# Visualization: World Countries Map
ggplot(data = world) +
  geom_sf(fill = "lightblue", color = "black") +
  labs(
    title = "World Countries",
    x = "Longitude",
    y = "Latitude"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, size = 12, face = "bold"),
    axis.text = element_text(size = 10),
    axis.title = element_text(size = 12))
```



Using ***rnaturalearth*** and the ***sf package***, we constructed a pale blue map of world states with well-defined borders. This visualization effectively illustrates the capacity of several packages to handle and depict spatial data, offering a comprehensive and complete representation of world geography. It provides a strong framework for sophisticated spatial analysis and manipulation, establishing a solid foundation for complex visualizations. This map enhances understanding of worldwide geographical patterns, trends, and spatial relationships, serving as a useful tool for further exploration and research.

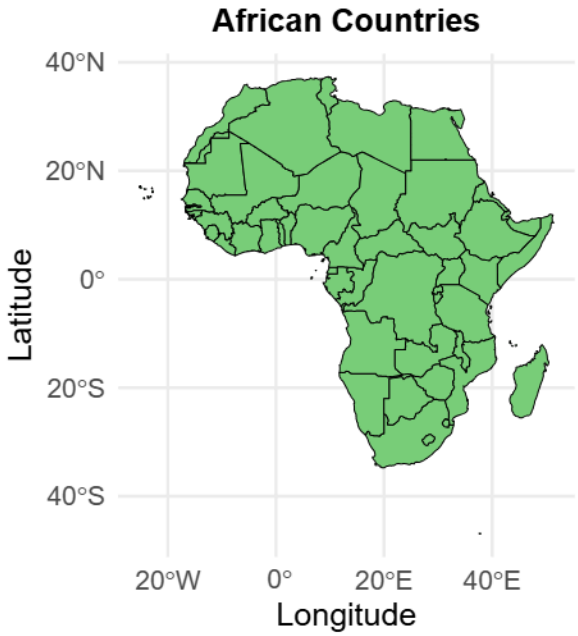
```
# Visualization: World Map with Population Data
ggplot(data = world) +
  geom_sf(aes(fill = pop_est)) +
  scale_fill_viridis_c(option = "plasma", na.value = "gray90") +
  labs(
    title = "World Map by Population",
    fill = "Population",
    x = "Longitude",
    y = "Latitude"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, size = 12, face = "bold"),
    axis.text = element_text(size = 10),
    axis.title = element_text(size = 12))
```



Building on the previous display of country boundaries, we improved the study by using population data to produce a choropleth map. This representation of the global population distribution was created using the ***rnaturalearth*** package. A gradient color scheme is used, ranging from deep purple for countries with smaller populations to vivid yellow for the greatest populations, such as China and India. This map successfully emphasizes demographic density, providing useful insights into areas with dense populations and functioning as a tool for demographic and geographic study.

```
# Loading necessary libraries
if (!require(ggplot2)) install.packages("ggplot2")
if (!require(dplyr)) install.packages("dplyr")
if (!require(rnaturalearth)) install.packages("rnaturalearth")
if (!require(sf)) install.packages("sf")

library(ggplot2)
library(dplyr)
library(rnaturalearth)
library(sf)

# Loading the world dataset again just in case
world <- ne_countries(scale = "medium", returnclass = "sf")

# Making sure that the data contains the necessary column
if (!"continent" %in% colnames(world)) {
  stop("The dataset does not contain the 'continent' column.")
}

# Visualization: African Countries
ggplot(data = world %>% filter(continent == "Africa")) +
  geom_sf(fill = "palegreen3", color = "black") +
  labs(
    title = "African Countries",
    x = "Longitude",
    y = "Latitude"
  ) +
  theme_minimal() +
  theme(
```

```
    plot.title = element_text(hjust = 0.5, size = 12, face = "bold"),
    axis.text = element_text(size = 10),
    axis.title = element_text(size = 12))
```



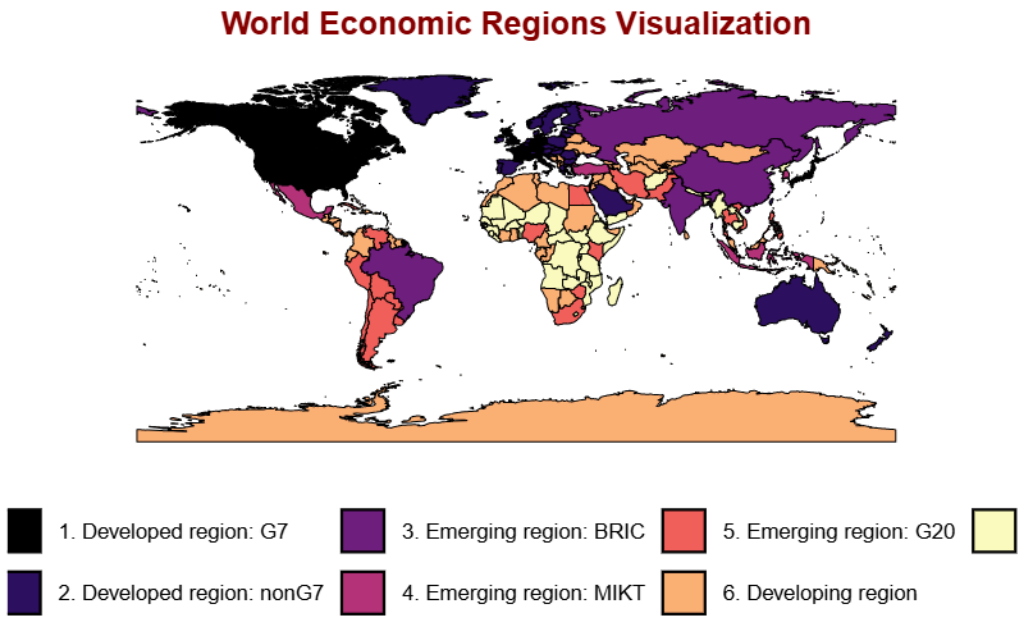
This visualization focuses on the African continent and displays the geographical boundaries of African countries using the ***rnaturalearth*** package's filtered dataset. Each nation is shown with a vibrant green fill and black borders, resulting in a clear and unique representation of the region. This map provides a thorough picture of Africa, making it useful for studying the continent's terrain, demographic distribution, and socioeconomic issues. By limiting coverage to a certain region, the map enables more focused research and significant insights into Africa's unique characteristics.

```
# Suppress package for warnings
suppressPackageStartupMessages({
  library(ggplot2)
  library(viridis)
  library(rnaturalearth)
  library(rnaturalearthdata)
  library(sf)
})
```

```
Warning: package 'viridis' was built under R version 4.3.3
```

```
# Loading world dataset
world <- ne_countries(scale = "medium", returnclass = "sf")

# Checking if the `economy` column exists
if (!"economy" %in% colnames(world)) {
  stop("The dataset does not contain the 'economy' column.")
}

# Visualization: Economic Regions of World Map
ggplot(data = world) +
  geom_sf(aes(fill = economy), color = "black", size = 0.2) +
  scale_fill_viridis_d(
    option = "magma",
```

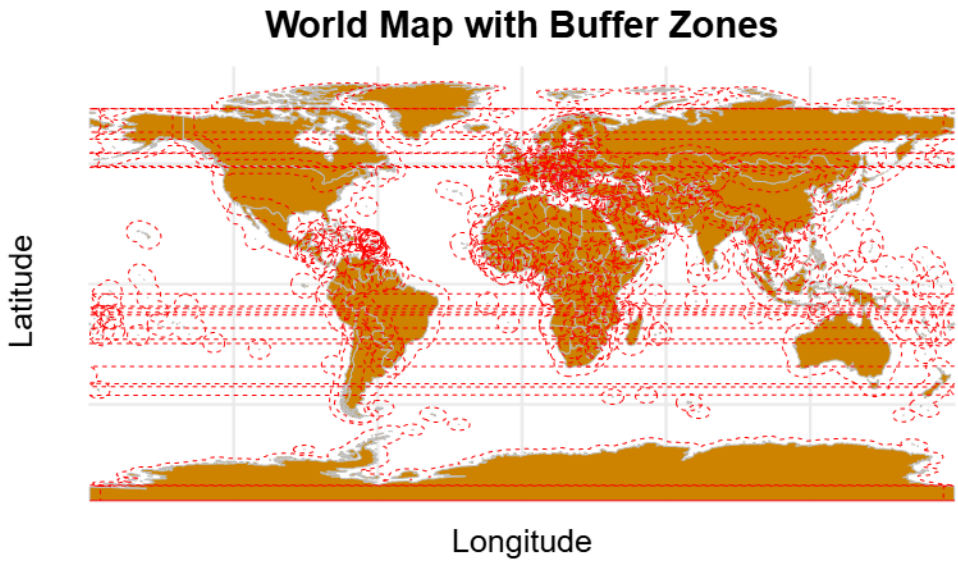
```
    na.value = "grey90"  # Coloring missing values
  ) +
  labs(
    title = "World Economic Regions Visualization",
    fill = "Economic Region",
    x = NULL,  # Remove x-axis label
    y = NULL   # Remove y-axis label
  ) +
  theme_minimal(base_size = 12) +
  theme(
    plot.title = element_text(face = "bold", size = 12, color = "darkred", hjust = 0.5),
    legend.position = "bottom",       # Position of the legend
    legend.title = element_text(size = 10, face = "bold"), # Legend title
    legend.text = element_text(size = 8), # Legend text size
    axis.text = element_blank(),          # Remove axis text
    panel.grid = element_blank())
```



The graph above depicts a global depiction of countries divided into numerous economic categories, each represented by a different color. It distinguishes main economic sectors, such as emerging markets, underdeveloped regions, and mature economies, allowing for a complete comparison of economic conditions across continents. The detailed and clearly outlined legend dramatically improves comprehension, allowing viewers to quickly recognize and comprehend worldwide economic differences. This image not only explains global economic patterns, but it also highlights regional developmental inequalities, providing crucial insights into global economic inequities, growth disparities, and varied degrees of wealth.

```
# Loading necessary libraries
library(ggplot2)
library(sf)
library(rnaturalearth)
library(rnaturalearthdata)

# Loading the world dataset
world <- ne_countries(scale = "medium", returnclass = "sf")

# Ensuring CRS transformations and buffer calculations
```

```r
world_planar <- st_transform(world, crs = 3857)
# Reprojecting to planar coordinates
world_buffer <- st_buffer(world_planar, dist = 500000)
# Creating buffer (500,000 meters)
world_buffer <- st_transform(world_buffer, crs = st_crs(world))
# Reprojecting back to geographic coordinates

# Visualization: World Map with Buffer Zones
ggplot() +
  geom_sf(data = world, fill = "orange3", color = "grey") +
  geom_sf(data = world_buffer, fill = NA, color = "red", linetype = "dashed") +
  labs(
    title = "World Map with Buffer Zones",
    x = "Longitude",
    y = "Latitude"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, size = 14, face = "bold"),
    axis.text = element_text(size = 10),
    axis.title = element_text(size = 12))
```



This graph improves previous visualizations of national boundaries, population distribution, and economic zones by illustrating buffer zones surrounding each country's boundary by advanced geographical analysis. The dashed red lines surrounding the orange-filled polygons signify that the st_buffer function from the sf package created a 5-degree buffer including all nations. This model is essential for proximity-based research and geographic studies requiring the incorporation of adjacent territories, as it distinctly defines each nation's true boundaries and includes supplementary buffer zones. The primary national polygons and buffer zones provide a detailed representation of geographical connections and crossings, emphasizing possible conflicts, collaboration, and locations for environmental investigation. This phase highlights the advanced capabilities of the sf package for performing spatial changes and creating layered geographic displays.

```r
# Loading needed libraries
library(dplyr)
```

```r
library(kableExtra)
library(sf)

# Loading the world dataset
world <- ne_countries(scale = "medium", returnclass = "sf")

# Extracting details for the maximum and minimum population
max_population_country <- world %>%
  st_drop_geometry() %>%
  filter(pop_est == max(pop_est, na.rm = TRUE)) %>%
  slice(1) %>% # Ensure only one row is selected
  select(name, pop_est)

min_population_country <- world %>%
  st_drop_geometry() %>%
  filter(pop_est == min(pop_est, na.rm = TRUE)) %>%
  slice(1) %>% # Ensure only one row is selected
  select(name, pop_est)

# Calculating a summary statistics
population_summary <- world %>%
  st_drop_geometry() %>%
  summarise(
    `Total Population` = format(sum(pop_est, na.rm = TRUE), big.mark = ","),
    `Mean Population` = format(round(mean(pop_est, na.rm = TRUE), 2), big.mark = ","),
    `Median Population` = format(median(pop_est, na.rm = TRUE), big.mark = ","),
    `Maximum Population` = paste(max_population_country$name,
    "(", format(max_population_country$pop_est, big.mark = ","), ")"),
    `Minimum Population` = paste(min_population_country$name,
    "(", format(min_population_country$pop_est, big.mark = ","), ")")
  )

# Creating a final table
population_summary %>%
  kable(
    caption = "Statistical Summary of Population Data with Country Details",
    align = "c"
  ) %>%
  kable_styling(
    latex_options = c("striped", "hold_position"),
    font_size = 10)
```

Table 8: Statistical Summary of Population Data with Country Details

Total Population	Mean Population	Median Population	Maximum Population	Minimum Population
7,676,957,742	31,722,966	5,071,860	China (1,397,715,000)	Heard I. and McDonald Is. (0)

The table above presents a detailed statistics summary. The aggregate population of all nations is roughly 7.67 billion. The average population per country is approximately 31.7 million, although the median population is 5.07 million, indicating that a small number of densely populated nations distort the average. China stands out as the most populous country, with approximately 1.39 billion people, whereas the Heard Island and McDonald Islands represent the minimum with a population of zero. This summary offers valuable insights into the wide disparities in population distribution worldwide.

Part 3: Functions

The concluding part of the project entails the creation of a R function that does statistical correlation analysis on two numerical variables. The aim is to provide a reusable function that computes the correlation coefficient and produces

meaningful insights via tailored outputs. This will include employing S3 techniques such as *print(), summary(), and plot()* to guarantee that the results are conveyed in a clear and comprehensive manner.

To demonstrate the capabilities, we will employ the integrated mtcars dataset, which encompasses many quantitative factors pertaining to vehicle performance and specifications. This dataset is ideal for our research since it allows us to investigate correlations between variables like miles per gallon (mpg), horsepower (hp), and others. This exercise shows how R's custom functions and programming tools may speed up statistical analysis and improve interpretability.

```
# Loading needed libraries
if (!require(knitr)) install.packages("knitr")
if (!require(kableExtra)) install.packages("kableExtra")
library(knitr)
library(kableExtra)

# Summarizing the mtcars dataset into a table format
summary_table <- data.frame(
  Variable = colnames(mtcars),
  Min = sapply(mtcars, min),
  `1st Quartile` = sapply(mtcars, function(x) quantile(x, 0.25)),
  Median = sapply(mtcars, median),
  Mean = sapply(mtcars, mean),
  `3rd Quartile` = sapply(mtcars, function(x) quantile(x, 0.75)),
  Max = sapply(mtcars, max)
)

# Creating a table
kable(
  summary_table,
  align = "c",
  caption = "Summary of mtcars Dataset"
) %>%
  kable_styling(
    latex_options = c("striped", "hold_position"),
    font_size = 10)
```

Table 9: Summary of mtcars Dataset

	Variable	Min	X1st.Quartile	Median	Mean	X3rd.Quartile	Max
mpg	mpg	10.400	15.42500	19.200	20.090625	22.80	33.900
cyl	cyl	4.000	4.00000	6.000	6.187500	8.00	8.000
disp	disp	71.100	120.82500	196.300	230.721875	326.00	472.000
hp	hp	52.000	96.50000	123.000	146.687500	180.00	335.000
drat	drat	2.760	3.08000	3.695	3.596563	3.92	4.930
wt	wt	1.513	2.58125	3.325	3.217250	3.61	5.424
qsec	qsec	14.500	16.89250	17.710	17.848750	18.90	22.900
vs	vs	0.000	0.00000	0.000	0.437500	1.00	1.000
am	am	0.000	0.00000	0.000	0.406250	1.00	1.000
gear	gear	3.000	3.00000	4.000	3.687500	4.00	5.000
carb	carb	1.000	2.00000	2.000	2.812500	4.00	8.000

The table presents an overview of the variables in the mtcars dataset. Each row represents a separate variable, and the columns include crucial information about the variable's range and distribution. For example:

- mpg (miles per gallon): Fuel economy varies from 10.4 to 33.9, with a mean of 20.1 and a median of 19.2.
- cyl (number of cylinders): Vehicles have 4 to 8 cylinders, with an average of 6.19.

- disp (displacement): The engine displacement ranges from 71.1 to 472.0, with a mean of 230.7.
- hp (horsepower): Horsepower ranges from 52 to 335 with an average of 146.7.
- wt (weight): Vehicle weights vary from 1.51 to 5.42 tons, with an average of 3.22.

Define the Correlation Analysis

This function will calculate the correlation between two numeric variables from the dataset and return an object of a custom S3 class.

```r
# Defining the correlation analysis function
correlation_analysis <- function(data, var1, var2) {
  # Making sure that the input variables exist
  if (!all(c(var1, var2) %in% colnames(data))) {
    stop("Variables not found in the dataset.")
  }

  # Making sure that the variables are numeric
  if (!is.numeric(data[[var1]]) || !is.numeric(data[[var2]])) {
    stop("Both variables must be numeric.")
  }

  # Calculating correlation
  correlation_value <- cor(data[[var1]], data[[var2]])

  # Create "correlation_result"
  result <- list(
    data = data,
    var1 = var1,
    var2 = var2,
    correlation = correlation_value
  )

  class(result) <- "correlation_result"
  return(result)
}

# Defining the print method for correlation result
print.correlation_result <- function(obj) {
  cat("Correlation Analysis\n")
  cat("====================\n")
  cat("Variable 1:", obj$var1, "\n")
  cat("Variable 2:", obj$var2, "\n")
  cat("Correlation Coefficient:", round(obj$correlation, 4), "\n")
}

# Defining the summary method for correlation result
summary.correlation_result <- function(obj) {
  cat("Summary of Correlation Analysis\n")
  cat("================================\n")
  cat("Variable 1:", obj$var1, "\n")
  cat("Variable 2:", obj$var2, "\n")
  cat("Correlation Coefficient:", round(obj$correlation, 4), "\n")

  # Descriptive statistics for both variables
  cat("\nDescriptive Statistics:\n")
  stats1 <- summary(obj$data[[obj$var1]])
  stats2 <- summary(obj$data[[obj$var2]])
  cat("Variable 1 (", obj$var1, "):\n", stats1, "\n", sep = "")
  cat("Variable 2 (", obj$var2, "):\n", stats2, "\n", sep = "")
}
```

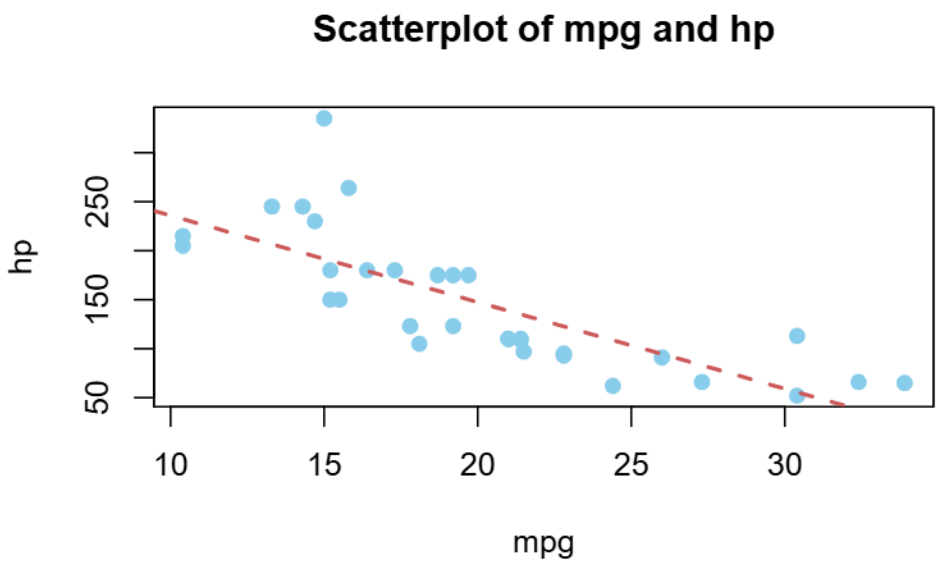
```r
# Defining the plot methd for correlation result
plot.correlation_result <- function(obj) {
  plot(
    obj$data[[obj$var1]], obj$data[[obj$var2]],
    main = paste("Scatterplot of", obj$var1, "and", obj$var2),
    xlab = obj$var1,
    ylab = obj$var2,
    pch = 19,
    col = "skyblue"
  )
  abline(lm(obj$data[[obj$var2]] ~ obj$data[[obj$var1]]),
         col = "indianred", lwd = 2, lty = 2)
}

# Example Analysis mtcars Dataset
result <- correlation_analysis(mtcars, "mpg", "hp")

# Rendering text outputs for PDF
cat(capture.output(print(result)), sep = "\n")
```

```
Correlation Analysis
====================
Variable 1: mpg
Variable 2: hp
Correlation Coefficient: -0.7762
```

```r
cat(capture.output(summary(result)), sep = "\n")
```

```
Summary of Correlation Analysis
================================
Variable 1: mpg
Variable 2: hp
Correlation Coefficient: -0.7762

Descriptive Statistics:
Variable 1 (mpg):
10.415.42519.220.0906222.833.9
Variable 2 (hp):
5296.5123146.6875180335
```

```r
# Rendering the plot
plot(result)
```



The mtcars dataset demonstrates a substantial negative correlation between weight (wt) and miles per gallon (mpg). The study identified a correlation coefficient of -0.77, signifying that greater vehicle weight correlates with worse fuel efficiency. This demonstrates the impact of weight on fuel efficiency, offering significant statistical insights into the dataset.

The scatter figure clearly illustrates the association between weight and miles per gallon (mpg) in the mtcars dataset. Each cyan point denotes a car, positioned based on its fuel efficiency (y-axis) and mass (x-axis). A dashed red regression line with a correlation coefficient of -0.77 signifies an inverse association between the two variables. This significant inverse association indicates that bigger cars often have poorer fuel economy. The graph elucidates the correlation between fuel economy and weight by a clear visual depiction of the declining trend, so augmenting the numerical analysis.

END OF FINAL PROJECT