Telco Customer Churn Analysis

Project Overview

This project involves analyzing a fictional telecom dataset containing customer churn information. The dataset includes 7,043 observations and 33 variables, such as demographic details, services subscribed, and churn status. The objectives are to perform statistical analysis, explore churn behavior, and generate insights using R.

Dataset Overview

What was done:
The dataset was loaded and inspected to understand its structure and content. Key characteristics, such as the number of observations (7,043) and variables (33), were identified.

Why:
To get an overview of the data and ensure the required columns for analysis are present.

The dataset consists of:

- **Total Observations**: 7,043
- **Variables**: 33, including CustomerID, Tenure Months, Monthly Charges, Churn Value, and others.

Steps:

1. Loaded the dataset

```
telco <- read.csv("Telco.csv", stringsAsFactors = FALSE)
```

2. Inspected the data structure:

```
str(telco)
colnames(telco)
```

Data Cleaning

What was done:
Column names were standardized using the janitor::clean_names() function to make them consistent and easier to use in the analysis.

Why:
To eliminate potential issues arising from special characters or inconsistent formatting in column names.

Cleaned column names to standardize them for easier use:

```
library(janitor)
telco <- janitor::clean_names(telco)
```

Statistical Analysis

Discrete vs. Continuous Variables

What was done:
Categorized variables into discrete (e.g., gender, partner) and continuous (e.g., tenure_months, monthly_charges) based on their data type.

Why:
To determine the appropriate statistical techniques to apply to each variable type.

- **Discrete Variables**: Variables with distinct categories, e.g., gender, senior_citizen, partner, etc.
- **Continuous Variables**: Variables with numeric values, e.g., tenure_months, monthly_charges, total_charges.

Binomial Distribution

What was done:
Probabilities for various scenarios, such as customers staying or leaving, were calculated using the binomial distribution. For example, the probability of 350 customers staying out of 500.

Why:
To model customer churn as a binary event (stay/leave) and predict outcomes for different groups.

Probability Calculations:

Churn Probabilities

```
p_stay <- mean(telco$churn_value == 0)
p_leave <- 1 - p_stay
```

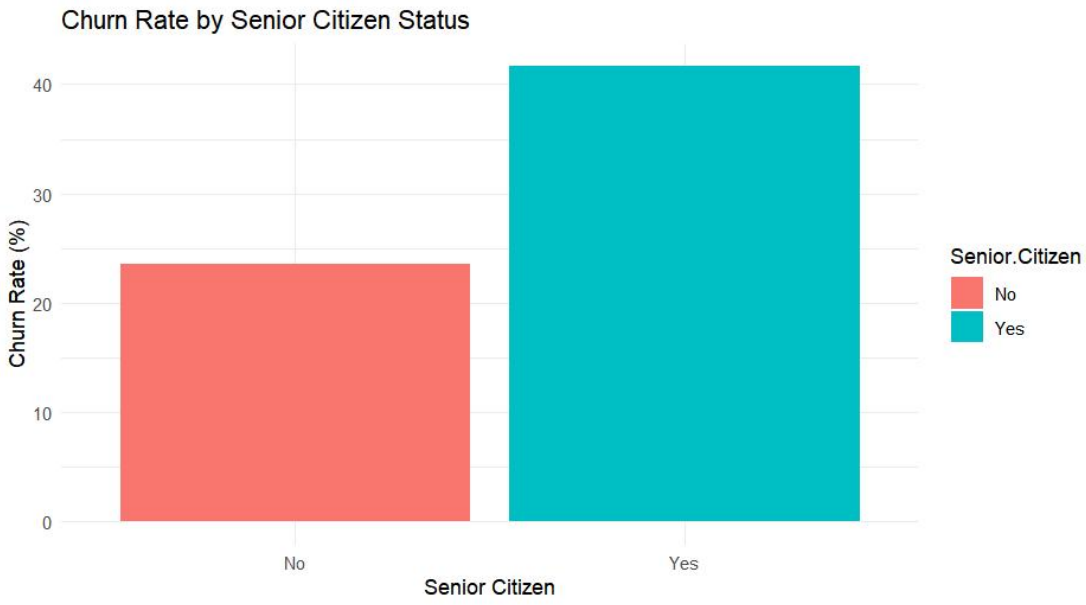
Scenarios

```
dbinom(350, size = 500, prob = p_stay)  # 350 customers staying out of 500
dbinom(160, size = 500, prob = p_leave)  # 160 customers leaving out of 500
pbinom(299, size = 1000, prob = p_leave)  # Less than 300 leaving out of 1000
```

Senior Citizens

```
senior_citizens <- telco %>% filter(senior_citizen == "Yes")
```

```
p_senior_stay <- mean(senior_citizens$churn_value == 0)
dbinom(200, size = 500, prob = p_senior_stay)
```



Normal Distribution

What was done:
The distribution of customer tenure was analyzed using the normal distribution. Probabilities were calculated for tenure falling within specified ranges, and a density plot was created.

Why:
To understand the spread and central tendencies of customer tenure and identify patterns.

Tenure Analysis:

Mean and SD

```
tenure_mean <- mean(telco$tenure_months, na.rm = TRUE)
tenure_sd <- sd(telco$tenure_months, na.rm = TRUE)
```

Probability Calculations

```
pnorm(30, mean = tenure_mean, sd = tenure_sd)  # Tenure < 30 months
pnorm(40, mean = tenure_mean, sd = tenure_sd) - pnorm(30, mean =
tenure_mean, sd = tenure_sd)  # 30 < Tenure < 40
```

Visualizing Tenure:

What was done:
A bar plot was created to compare churn rates between male and female customers.

Why:
To identify if gender influences churn behavior.

code

```
ggplot(telco, aes(x = tenure_months)) +
  geom_density(fill = "blue", alpha = 0.5) +
  labs(title = "Normal Distribution of Tenure", x = "Tenure (Months)", y =
"Density")
```

Confidence Intervals

What was done:
Confidence intervals were calculated for variables like `monthly_charges` to estimate the population mean within a specific range of confidence (e.g., 95%).

Why:
To provide insights into the expected range of values for these variables in the population.

- **Monthly Charges**

code

```
monthly_mean <- mean(telco$monthly_charges, na.rm = TRUE)
monthly_sd <- sd(telco$monthly_charges, na.rm = TRUE)
n <- nrow(telco)
error <- qt(0.975, df = n - 1) * (monthly_sd / sqrt(n))
CI_95 <- c(monthly_mean - error, monthly_mean + error)
```

Data Visualizations

Churn Rate by Gender

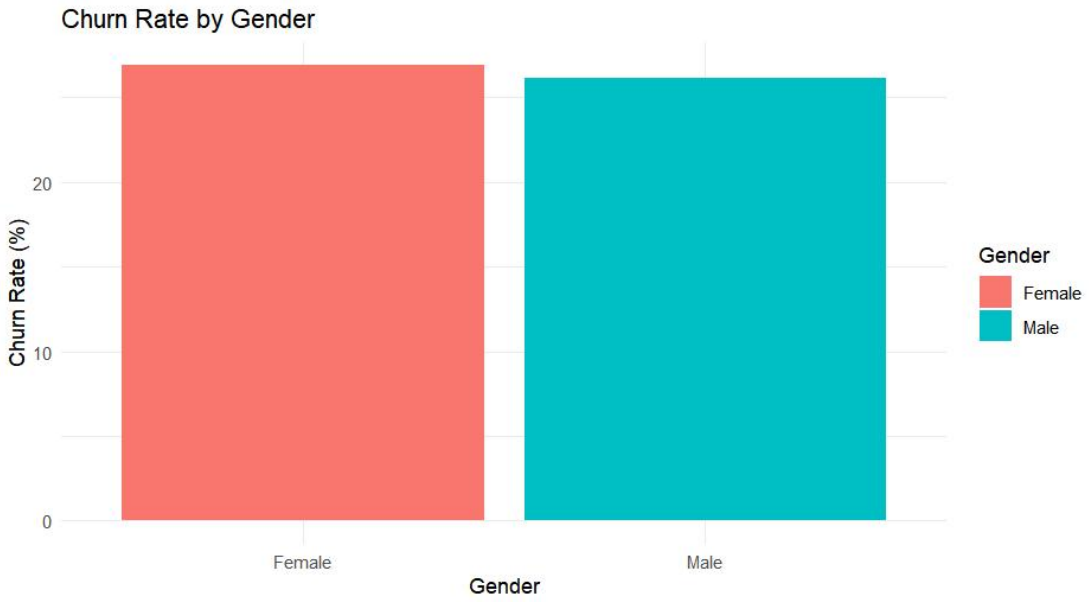
What was done:
A bar plot was created to compare churn rates between male and female customers.

Why:
To identify if gender influences churn behavior.

R code:

```
gender_churn <- telco %>%
  group_by(gender) %>%
  summarise(churn_rate = mean(churn_value == 1) * 100)

ggplot(gender_churn, aes(x = gender, y = churn_rate, fill = gender)) +
  geom_bar(stat = "identity") +
  labs(title = "Churn Rate by Gender", x = "Gender", y = "Churn Rate (%)")
```



Churn Rate by Contract Type

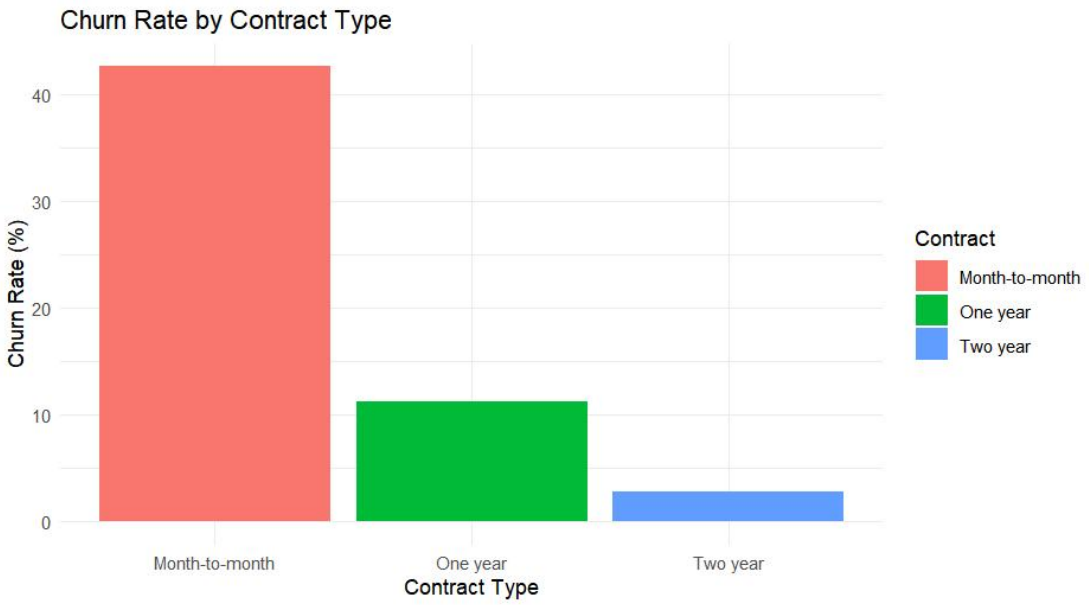
What was done:
A bar plot was created to visualize churn rates for different contract types (e.g., month-to-month, one-year).

Why:
To assess whether contract type impacts churn rates.

R code:

```
contract_churn <- telco %>%
  group_by(contract) %>%
  summarise(churn_rate = mean(churn_value == 1) * 100)

ggplot(contract_churn, aes(x = contract, y = churn_rate, fill = contract)) +
  geom_bar(stat = "identity") +
  labs(title = "Churn Rate by Contract Type", x = "Contract Type", y = "Churn
Rate (%)")
```

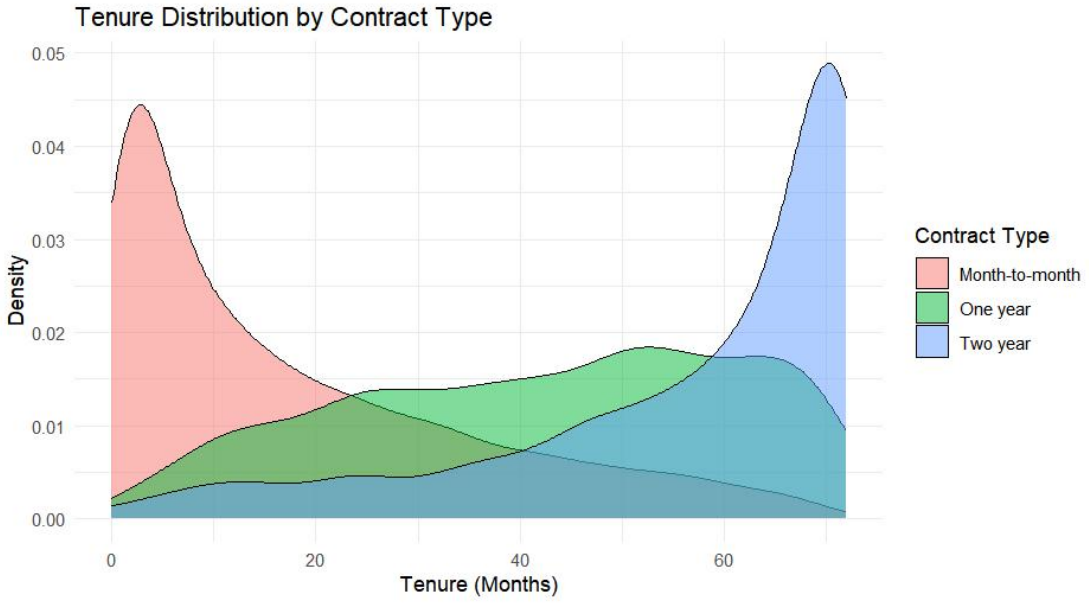


Tenure distribution by contract type

To analyze how customer tenure varies based on the type of contract they have (e.g., month-to-month, one-year, or two-year contracts). This can help identify patterns, such as whether longer contracts are associated with higher retention.

R code

```
ggplot(Telco, aes(x = Tenure.Months, fill = Contract)) +
  geom_density(alpha = 0.5) +
  labs(title = "Tenure Distribution by Contract Type",
       x = "Tenure (Months)", y = "Density", fill = "Contract Type") +
  theme_minimal()
```



Correlation Heatmap

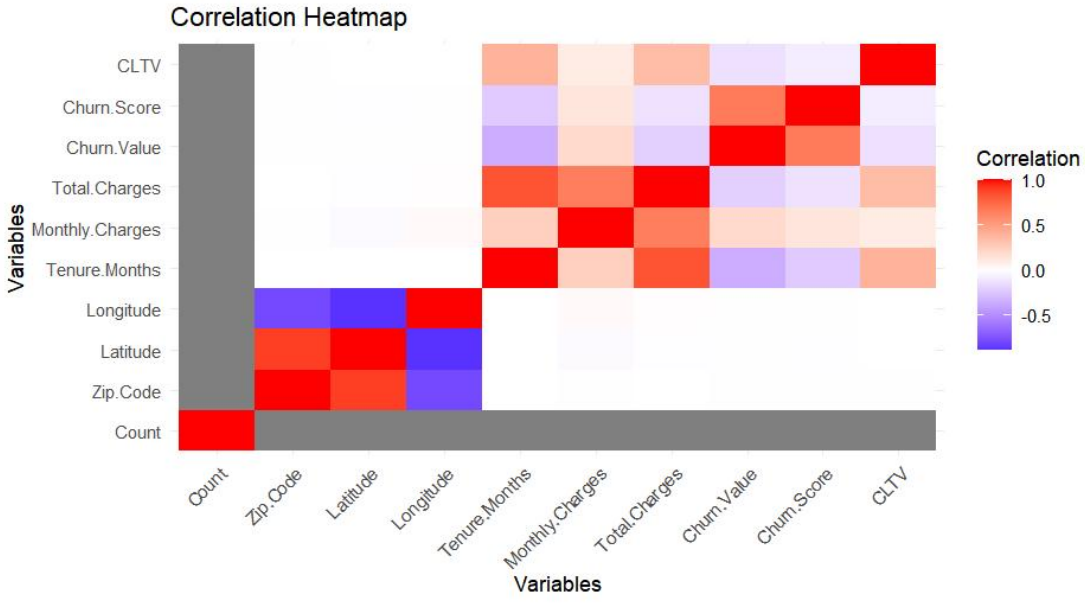
What was done:
A heatmap was generated to show correlations between numeric variables like tenure_months, monthly_charges, and total_charges.

Why:
To uncover relationships between variables that may affect churn.

R code:

```
numeric_data <- telco %>% select_if(is.numeric)
cor_matrix <- cor(numeric_data, use = "complete.obs")
library(reshape2)
cor_melted <- melt(cor_matrix)

ggplot(cor_melted, aes(Var1, Var2, fill = value)) +
  geom_tile() +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white", midpoint = 0) +
  labs(title = "Correlation Heatmap", x = "Variables", y = "Variables", fill =
"Correlation")
```



Key Findings

- **Churn Rates**:
  - Gender: Higher churn rate among specific genders (to be specified from results).
  - Contract Type: Month-to-month contracts have the highest churn rate.
- **Financial Insights**:
  - Customers with higher monthly charges are more likely to churn.

- **Tenure**:
    - Longer tenure customers are less likely to churn.

Conclusion

This project provided insights into customer behavior, churn tendencies, and key factors affecting churn. The analysis demonstrated proficiency in:

- Data cleaning and preparation.
- Statistical calculations and probability distributions.
- Creating insightful visualizations using R.