Applied data Project(or Assignment 1)

Pallavi Kollipara (s4015344)

2024-04-05

Problem Statement

1. Descriptive Statistics: Calculate, for example, the mean, median, mode, range, standard deviation, etc of both the S&P data and Bitcoin data.
2. Trend of S&P data and Bitcoin data over five years.
3. Compute the correlation coefficient between S&P data and Bitcoin data.
4. To assess whether the S&P data and Bitcoin data follow a normal distribution.

Loading Packages

```r
getwd()
```

```
## [1] "/Users/pallavikollipara/Desktop/rmit/sem2/Applied Analytics/Assignment1"
```

```r
library(readr)
library(dplyr)
library(ggplot2)
```

Importing the Data

After loading the necessary packages, I imported the S&P 500 and BTC-USD CSV files using the as.data.frame and read_csv() functions from readr. These functions convert the raw data into data frames, which are easier to work with and manipulate in R. Upon examining the BTC data frame, I noticed that it contained invalid columns from the 3rd position. To rectify this, I extracted columns 1 and 2 and stored them in a new variable named `Bitcoin`. Additionally, I renamed the second variable from "Close price adjusted" to "Price" to enhance clarity and usability for future analysis.

```r
SP500 <- as.data.frame(read_csv("S&P 500.csv"))
head(SP500)
```

```
##        Date   Price
## 1 2/04/2019 2867.24
## 2 3/04/2019 2873.40
## 3 4/04/2019 2879.39
## 4 5/04/2019 2892.74
## 5 8/04/2019 2895.77
## 6 9/04/2019 2878.20
```

```
BTC <- as.data.frame(read_csv("BTC-USD.csv"))
head(BTC)
```

```
##        Date Close price adjusted ...3 ...4 ...5 ...6 ...7 ...8 ...9 ...10 ...11
## 1 2/04/2019             4879.878   NA   NA   NA   NA   NA   NA   NA    NA    NA
## 2 3/04/2019             4973.022   NA   NA   NA   NA   NA   NA   NA    NA    NA
## 3 4/04/2019             4922.799   NA   NA   NA   NA   NA   NA   NA    NA    NA
## 4 5/04/2019             5036.681   NA   NA   NA   NA   NA   NA   NA    NA    NA
## 5 6/04/2019             5059.817   NA   NA   NA   NA   NA   NA   NA    NA    NA
## 6 7/04/2019             5198.897   NA   NA   NA   NA   NA   NA   NA    NA    NA
##   ...12
## 1    NA
## 2    NA
## 3    NA
## 4    NA
## 5    NA
## 6    NA
```

```
dim(SP500)
```

```
## [1] 1258    2
```

```
dim(BTC)
```

```
## [1] 1827   12
```

```
colnames(SP500)
```

```
## [1] "Date"  "Price"
```

```
colnames(BTC)
```

```
##  [1] "Date"                   "Close price adjusted" "...3"
##  [4] "...4"                   "...5"                 "...6"
##  [7] "...7"                   "...8"                 "...9"
## [10] "...10"                  "...11"                "...12"
```

```
Bitcoin<-BTC[1:2]
colnames(Bitcoin)[2]<- 'Price'
```

Task 1:

I calculated the descriptive statistics, including mean, median, mode, range, standard deviation, etc., for both the S&P 500 and Bitcoin data using the `summarise()` function to calculate all the necessary statistics separately.Comparing the descriptive statistics between the two datasets, we observed significant differences:

Central Tendency: The mean and median for Bitcoin prices are considerably higher than those for the S&P 500.

Variability: The range, standard deviation, and interquartile range (IQR) for Bitcoin prices are much larger compared to the S&P 500, indicating higher variability in Bitcoin prices.

Overall, these descriptive statistics provide insights into the differences in central tendency and variability between the S&P 500 and Bitcoin datasets, highlighting the distinct characteristics of each financial instrument.

```
SP500_stats<-
SP500 %>%summarise(Mean = round(mean(Price),2),
                   Median = round(median(Price),2),
                   Mode = round(as.numeric(names(sort(table(Price), decreasing = TRUE)[1])),2),
                   Min = round(min(Price),2),
                   Max = round(max(Price),2),
                   Range = round(max(Price) - min(Price),2),
                   SD = round(sd(Price),2),
                   Q1 = round(quantile(Price,probs = 0.25),2),
                   Q3 = round(quantile(Price,probs = 0.75),2),
                   IQR = round(Q3-Q1),2)

SP500_stats
```

```
##      Mean  Median    Mode    Min     Max   Range     SD      Q1     Q3  IQR 2
## 1 3867.88 3971.18 2926.46 2237.4 5254.35 3016.95 643.19 3273.73 4379.8 1106 2
```

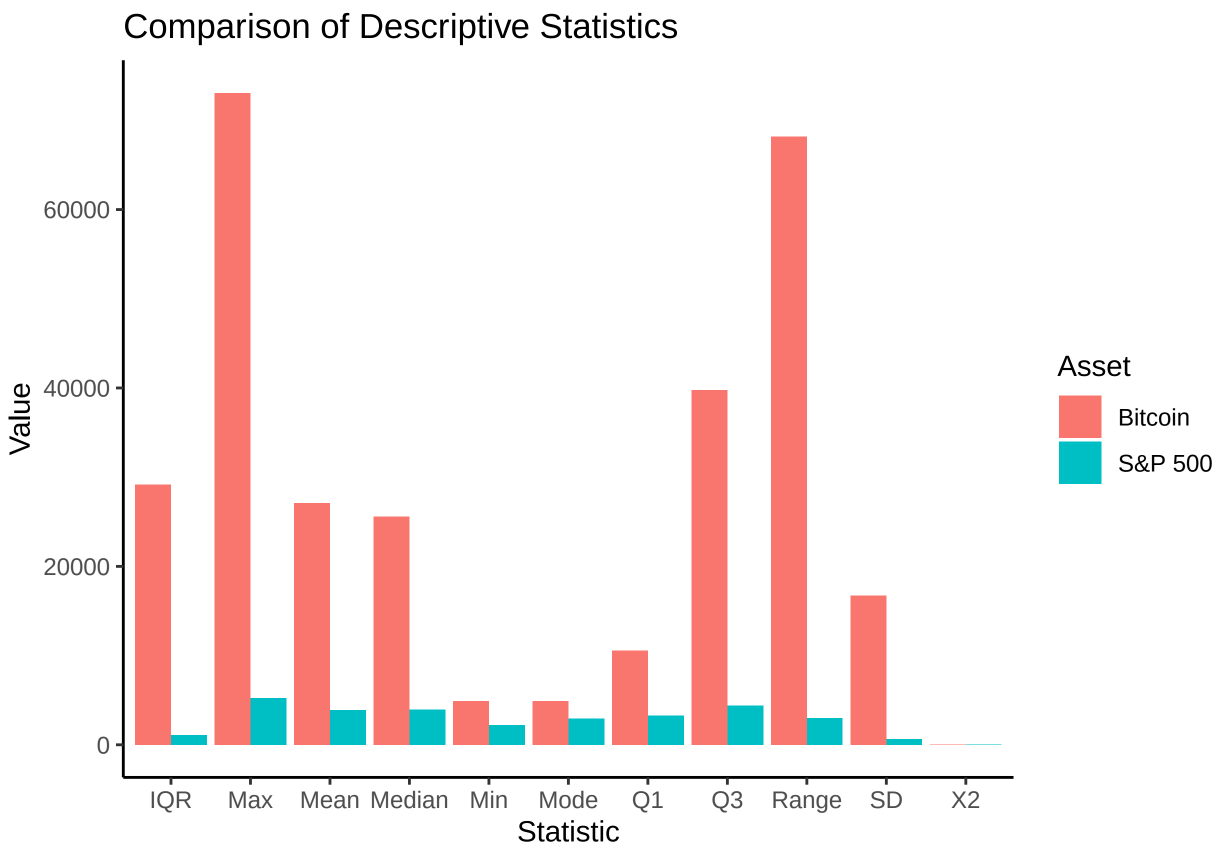
```
bitcoin_stats<-
Bitcoin %>%summarise(Mean =  round(mean(Price),2),
                     Median =  round(median(Price),2),
                     Mode = round(as.numeric(names(sort(table(Price), decreasing = TRUE)[1])),2),
                     Min = min(Price),
                     Max = max(Price),
                     Range =  round(max(Price) - min(Price),2),
                     SD =  round(sd(Price),2),
                     Q1 =  round(quantile(Price,probs = 0.25),2),
                     Q3 =  round(quantile(Price,probs = 0.75),2),
                     IQR = round(Q3-Q1),2)
bitcoin_stats
```

```
##       Mean   Median    Mode      Min     Max    Range       SD       Q1
## 1 27097.76 25576.39 4879.88 4879.878 73083.5 68203.62 16711.71 10579.55
##         Q3   IQR 2
## 1 39760.67 29181 2
```

```
summary_data <- rbind(data.frame(Asset = "S&P 500", SP500_stats),
                      data.frame(Asset = "Bitcoin", bitcoin_stats))

summary_data <-
  tidyr::pivot_longer(summary_data, cols = -Asset, names_to = "Statistic", values_to = "Value")

ggplot(summary_data, aes(x = Statistic, y = Value, fill = Asset)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Comparison of Descriptive Statistics",
       x = "Statistic",
       y = "Value") +
  theme_classic()
```



Task 2:

Below code generates two plots, one showing the trend of S&P 500 data over five years and the other showing the trend of Bitcoin data over the same period.

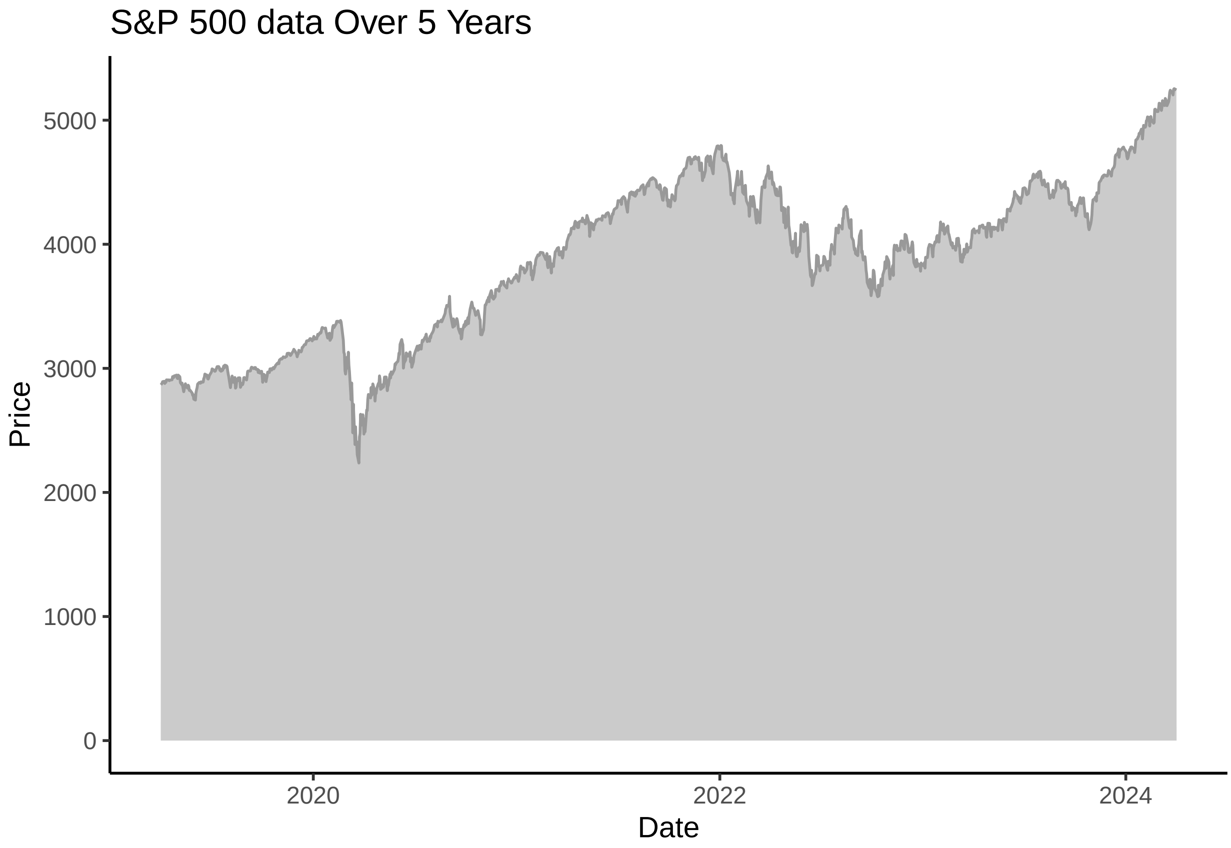
Interpretation of the Graphs:

1. **S&P 500 Data Trend:** The plot of S&P 500 data over five years shows a relatively steady and upward trend. There are fluctuations in the price, but overall, there seems to be a gradual increase in the S&P 500 index value over time. This suggests a positive growth trajectory in the stock market represented by the S&P 500 index.
2. **Bitcoin Data Trend:** In contrast, the plot of Bitcoin data over the same period exhibits a much more volatile pattern. There are significant fluctuations in Bitcoin prices, with periods of rapid increase followed by sharp declines. This volatility is characteristic of the cryptocurrency market, where prices can be highly unpredictable and subject to sudden shifts in demand and investor sentiment.

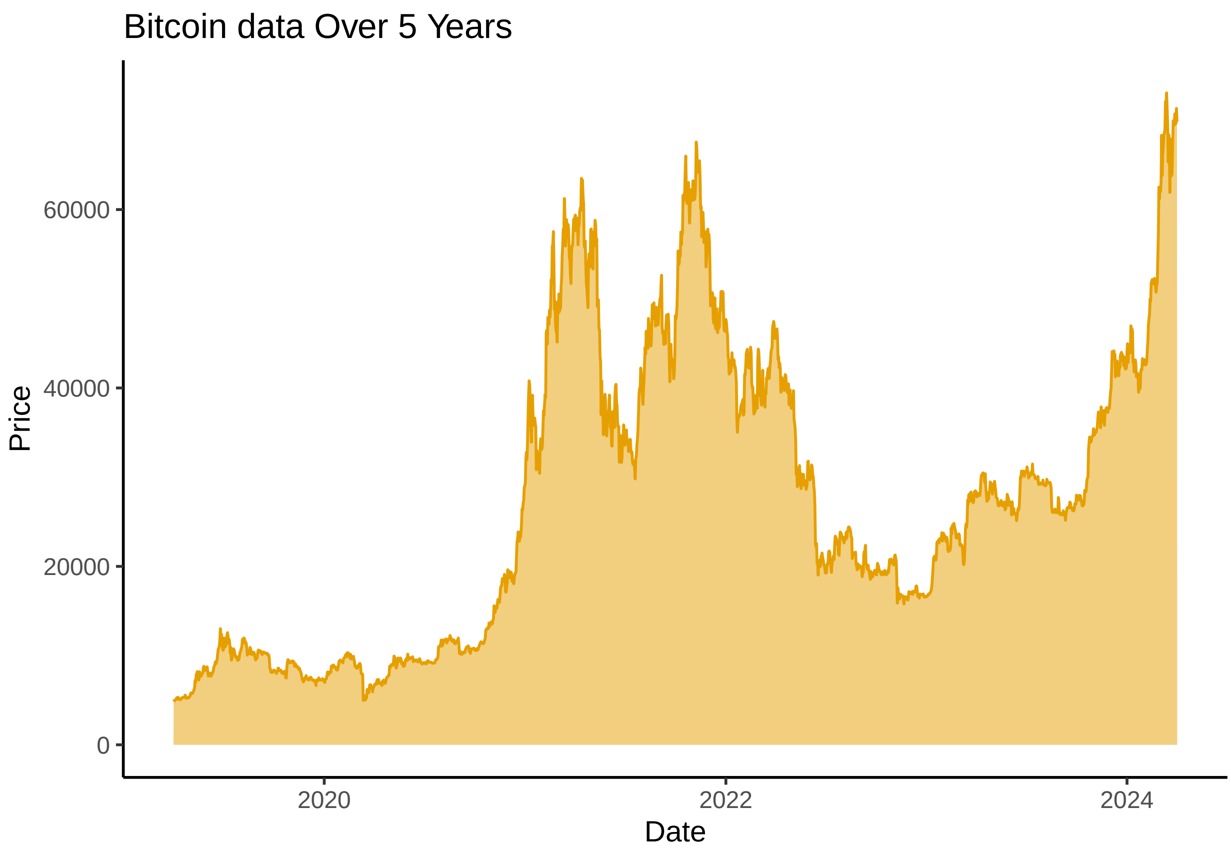
Overall, the graphs illustrate the distinct trends and patterns observed in the S&P 500 and Bitcoin data over five years. While the S&P 500 shows a relatively steady upward trend, Bitcoin prices demonstrate much greater volatility, reflecting the speculative nature of the cryptocurrency market.

```r
# Convert date strings to Date objects
SP500$Date <- as.Date(SP500$Date, format="%d/%m/%Y")
Bitcoin$Date <- as.Date(Bitcoin$Date, format="%d/%m/%Y")
```

```r
# Plot for S&P 500 trend
ggplot(SP500, aes(x = Date, y = Price)) +
  geom_area(colour = "#999999",fill="#999999",alpha=0.5) +
  labs(title = "S&P 500 data Over 5 Years",
       x = "Date",
       y = "Price")+
  theme_classic()
```



```r
# Plot for Bitcoin trend
ggplot(Bitcoin, aes(x = Date, y = Price)) +
  geom_area(colour = "#E69F00",fill = "#E69F00",alpha=0.5) +
  labs(title = "Bitcoin data Over 5 Years",
       x = "Date",
       y = "Price")+
  theme_classic()
```



For the below code first I merged the S&P 500 and Bitcoin datasets by the "Date" column to facilitate a comparison of prices across both datasets. The resulting merged dataset was assigned to the variable `merge_data`. Subsequently, missing values were handled to ensure data integrity.Next, a new variable named `merge_six_months` was created, and a new column named "Six_Months" was introduced in the `merge_six_months` dataframe. The `cut()` function was used to bin the dates into six-month intervals. Each row in the dataset was assigned a corresponding six-month interval based on its date. Afterward, the correlation between the price columns (`Price.x` and `Price.y`) of the `merge_six_months` dataframe was calculated for each six-month interval. The data was grouped by the "Six_Months" column, and the correlation was calculated using the `cor()` function. Finally, a `ggplot` was created to visualize the correlation using a line graph.

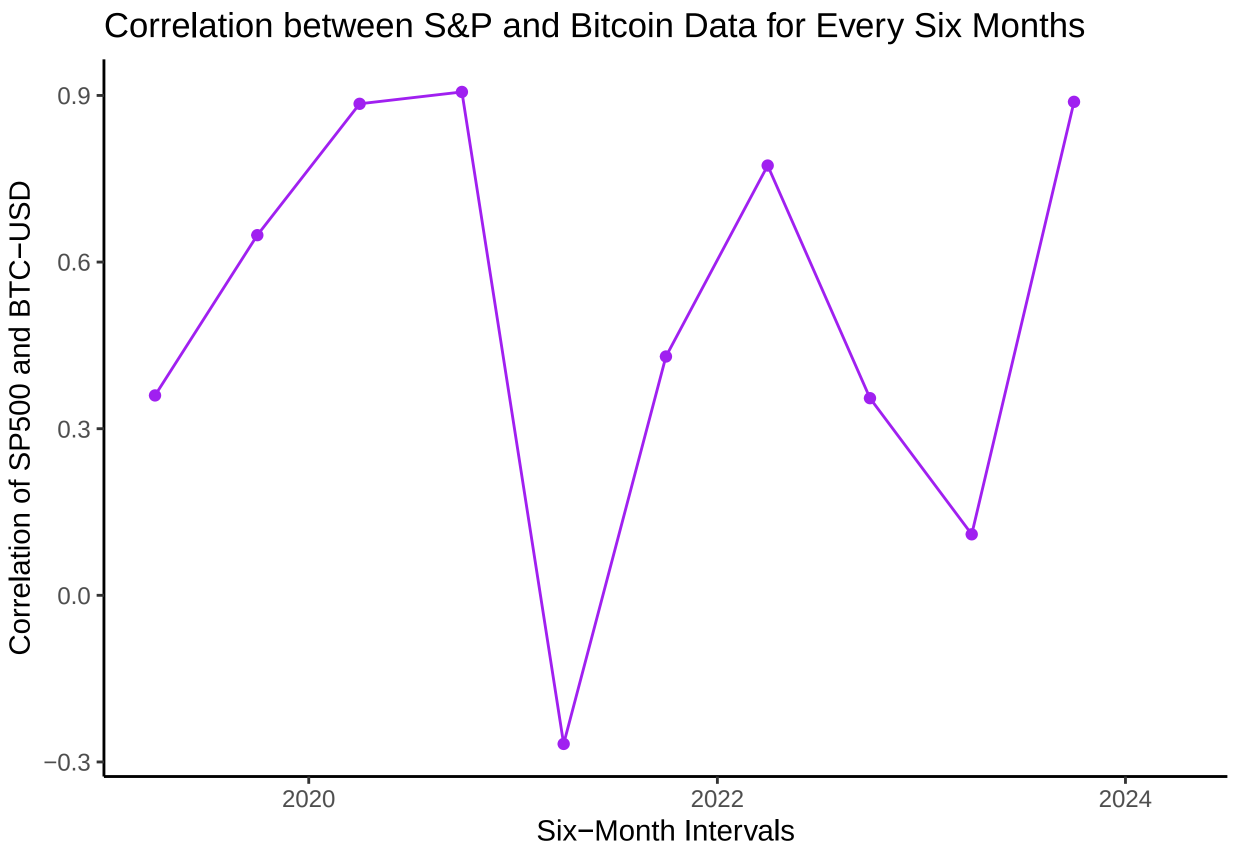
```r
# Merge the datasets by date
merged_data <- merge(SP500, Bitcoin, by = "Date", all = TRUE)

# Remove rows with missing or non-finite values
merged_data <- na.omit(merged_data)

merge_six_months <- merged_data %>%
  mutate(Six_Months = as.Date(cut(Date, breaks = "6 months")))

correlation_data <- merge_six_months %>%
  group_by(Six_Months) %>%
  summarise(Six_Months_Correlation = cor(Price.x, Price.y, use = "complete.obs"))

ggplot(correlation_data, aes(x = Six_Months, y = Six_Months_Correlation)) +
  geom_line(color = "purple" ) +
  geom_point(color = "purple") +
  labs(title = "Correlation between S&P and Bitcoin Data for Every Six Months",
       x = "Six-Month Intervals",
       y = "Correlation of SP500 and BTC-USD") +
   theme_classic()
```



TASK 3:

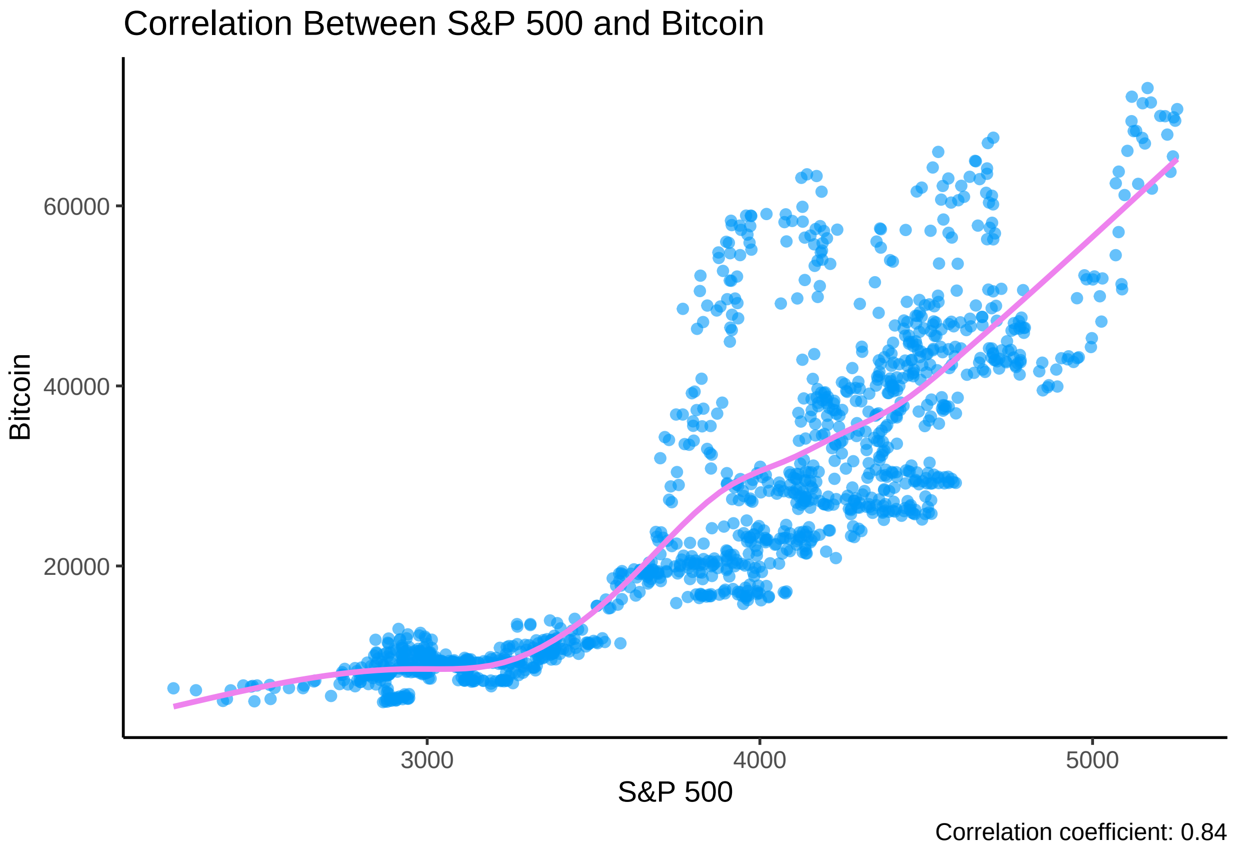
The `cor()` function was utilized to compute the correlation coefficient between S&P 500 prices and Bitcoin prices. A scatter plot was created using `ggplot()` and `geom_point()`, with S&P 500 prices on the x-axis and Bitcoin prices on the y-axis. Each point in the scatter plot represents a pair of corresponding prices from the merged dataset. Additionally, a smooth line was incorporated into the plot using `geom_smooth()` with a specific smoothing method ("gam"). The scatter plot, along with the smooth line, offers a visual representation of the relationship between S&P 500 and Bitcoin prices. If the points align to form a straight line and the smooth line closely tracks the points, it suggests a strong linear relationship between the two variables. The correlation coefficient displayed in the plot's caption quantifies the strength of this relationship, with a correlation value of 0.84 indicating a strong positive correlation between S&P 500 and Bitcoin prices.

```
# Compute correlation coefficient
correlation <- cor(merged_data$Price.x, merged_data$Price.y)

# Plot scatter plot with specific smoothing method
```

```
ggplot(merged_data, aes(x = Price.x, y = Price.y)) +
  geom_point( color = "#0099f9",alpha = 0.6) +
  geom_smooth(color = "violet" ,method = "gam",se = FALSE) +
  labs(title = "Correlation Between S&P 500 and Bitcoin",
       x = "S&P 500",
       y = "Bitcoin",
       caption = paste("Correlation coefficient:", round(correlation, 2)))+
   theme_classic()
```

```
## 'geom_smooth()' using formula = 'y ~ s(x, bs = "cs")'
```

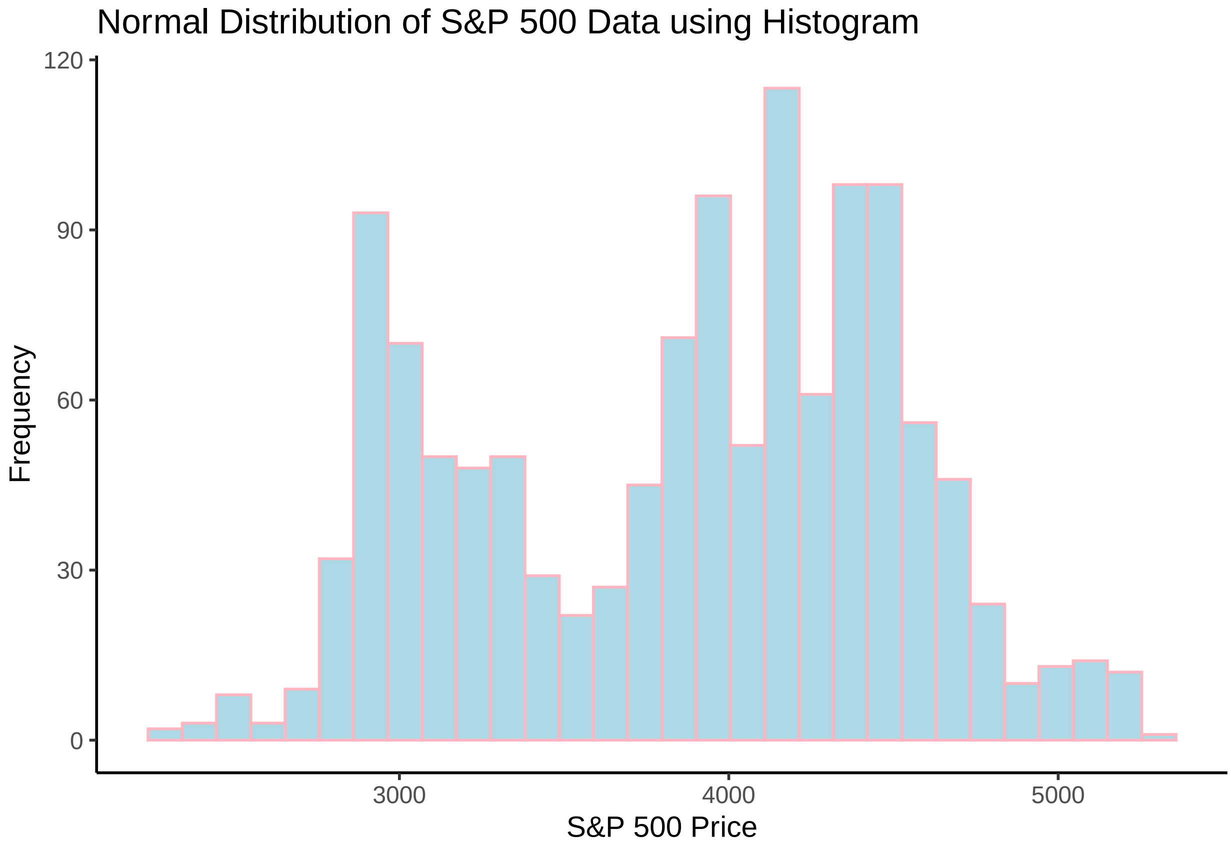


TASK 4:

Histograms are plotted for both the S&P 500 data and Bitcoin data using ggplot() and geom_histogram(). Histograms represent the distribution of data by displaying the frequency of observations within specified intervals, or "bins", along the x-axis. In these plots, the x-axis represents the price of the respective financial instrument, and the y-axis represents the frequency of occurrence. For the S&P 500 data histogram, the distribution appears approximately symmetric and bell-shaped, which is characteristic of a normal distribution. The data seems to cluster around the center, with fewer observations towards the extremes. Similarly, for the Bitcoin data histogram, the distribution is slightly right-skewed but also exhibits some characteristics of a normal distribution. However, there are noticeable outliers on the right side, indicating some deviation from normality.

The Shapiro-Wilk test is performed to formally assess whether the data follows a normal distribution. This statistical test calculates a p-value, which is used to determine whether the null hypothesis of normality can be rejected or not. A low p-value (typically less than 0.05) suggests that the data significantly deviates from a normal distribution. For the S&P 500 and Bitcoin data, if the p-value obtained from the Shapiro-Wilk test is greater than 0.05, it indicates that there is no significant evidence to reject the null hypothesis of normality, suggesting that the data follows a normal distribution.

```r
# Plot histogram for S&P 500 data
ggplot(SP500, aes(x = Price)) +
  geom_histogram(fill = "lightblue", color = "lightpink", bins = 30) +
  labs(title = "Normal Distribution of S&P 500 Data using Histogram",
       x = "S&P 500 Price",
       y = "Frequency")+
  theme_classic()
```

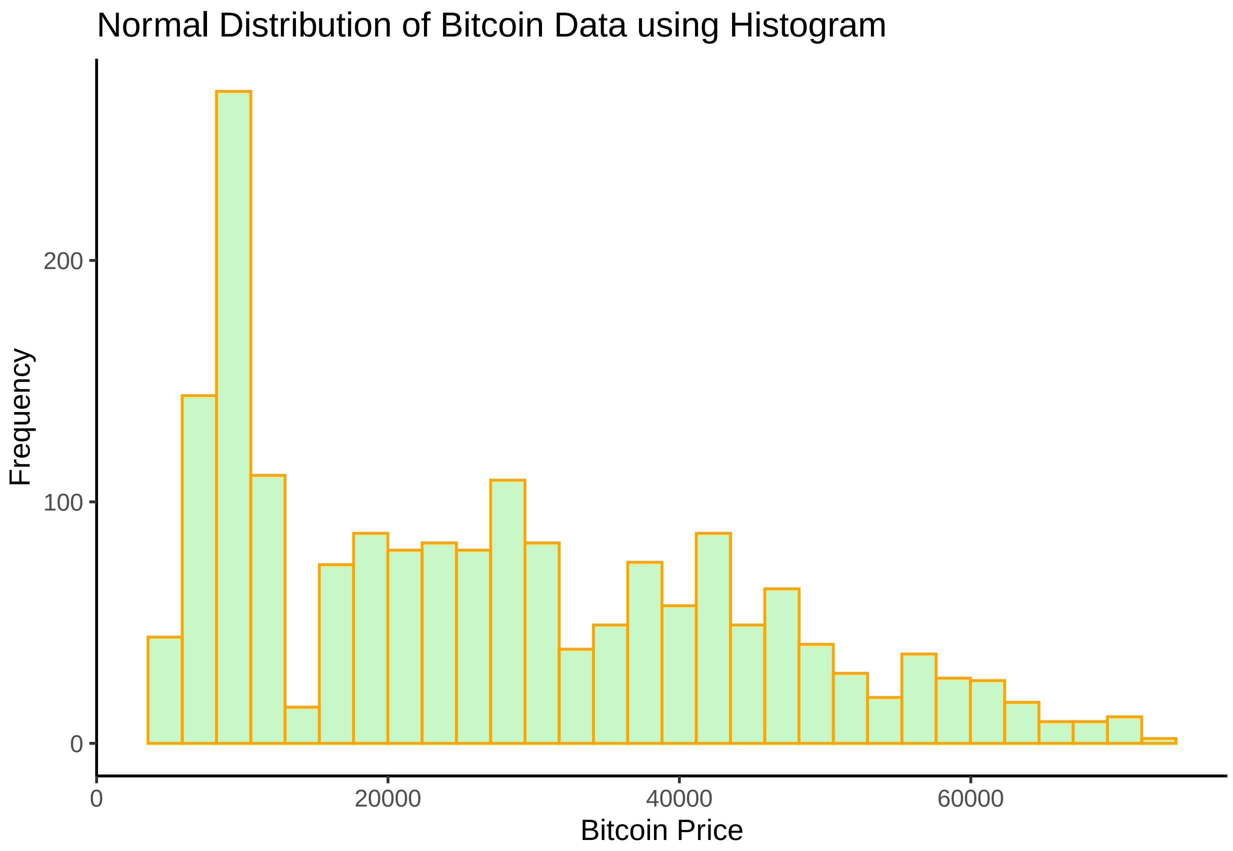


```r
# Shapiro-Wilk test for normality
shapiro.test(SP500$Price)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  SP500$Price
## W = 0.96053, p-value < 2.2e-16
```

```r
# Plot histogram for Bitcoin data

ggplot(Bitcoin, aes(x = Price)) +
  geom_histogram(bins = 30, alpha=0.5,fill = "lightgreen", color = "orange") +
  labs(title = "Normal Distribution of Bitcoin Data using Histogram",
       x = "Bitcoin Price",
       y = "Frequency")+
  theme_classic()
```



```r
# Shapiro-Wilk test for normality
shapiro.test(Bitcoin$Price)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Bitcoin$Price
## W = 0.93079, p-value < 2.2e-16
```

References:

1. Stats and R. (2020 January 22). Descriptive statistics in R. https://statsandr.com/blog/descriptive-statistics-in-r/#mode
2. stack overflow (2019 November 17). How to show code but hide output in RMarkdown?. https://stackoverflow.com/questions/47710427/how-to-show-code-but-hide-output-in-rmarkdown

3. APPLIED ANALYTICS (2016). Module 2 Descriptive Statistics through Visualisation. https://astral-theory-157510.appspot.com/secured/MATH1324_Module_02.html

4. APPLIED ANALYTICS (2016). R Bootcamp - Course 3: Basic Statistics in R. https://astral-theory-157510.appspot.com/secured/RBootcamp_Course_03.html