Analysis of real-world financial Datasets

Details

Name: Rohit Vasudev

Problem Statement

This report uses two real-world financial datasets: the S&P 500 and Bitcoin prices. In this report, statistical methods are applied to both of the datasets to analyse the Descriptive statistics of both the datasets, time series analysis for each dataset, correlation coefficient of both the datasets, and to assess if each of the datasets i.e. the S&P 500 data and Bitcoin data follow a normal distribution and uncover some insights or findings from these datasets.The purpose of this analysis is to uncover any insights, difference or similarities between the S&P 500 and Bitcoin historical data which spans over 7 years (from 2018 to 2025).

Load Packages

Firstly, we load the packages necessary for the analysis of the real-world financial datasets. We load packages such as readr, dplyr, ggplot2, knitr, and magrittr. These packages are used for loading the data, data manipulation, generating plots, and pipe operations.

```
# loading the necessary packages

library(readr)  # To read csv files
library(dplyr)  # For data wrangling
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)   # For generating plots
library(knitr)     # For dynamic report generation
library(magrittr)  # For pipe operators
```

Data Manipulation

This is the most important step before carrying out any data analysis of the financial data. The manipulation of data or data wrangling ensures that the data is consistent and is ready to use to uncover different trends, insights and conclusions. Here, we first load the into variables such as “sp” for S&P 500 Historical Data and “bitcoin” for Bitcoin Historical Data.

The glimpse() function from the dplyr package is used to see the structure of the S&P 500 and Bitcoin historical data. From observing the structure of both the financial datasets, it was found that the date column was originally a string or a character when the csv was first loaded. This was then corrected by converting it to the

Date type to a ensure that analysis could be carried out accurately. The conversion of the the type of the Date column was done with the help of the as.Date() function.

```
sp <- read_csv("S&P 500 Historical Data1825.csv")
```

```
## Rows: 1760 Columns: 2
## ── Column specification ────────────────────────────────────────────────────
## Delimiter: ","
## chr (1): Date
## num (1): Price
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
bitcoin <- read_csv("Bitcoin Historical Data1825.csv")
```

```
## Rows: 2558 Columns: 2
## ── Column specification ────────────────────────────────────────────────────
## Delimiter: ","
## chr (1): Date
## num (1): Price
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
glimpse(sp)
```

```
## Rows: 1,760
## Columns: 2
## $ Date  <chr> "1/08/2025", "31/07/2025", "30/07/2025", "29/07/2025", "28/07/20…
## $ Price <dbl> 6238.01, 6339.39, 6362.90, 6370.86, 6389.77, 6388.64, 6363.35, 6…
```

```
glimpse(bitcoin)
```

```
## Rows: 2,558
## Columns: 2
## $ Date  <chr> "1/08/2025", "31/07/2025", "30/07/2025", "29/07/2025", "28/07/20…
## $ Price <dbl> 113312.1, 115765.0, 117840.4, 117950.1, 118053.9, 119398.1, 1179…
```

```
# Converting Date datatype from a character to a Date type

sp$Date <- as.Date(sp$Date, "%d/%m/%Y")

bitcoin$Date <- as.Date(bitcoin$Date, "%d/%m/%Y")

glimpse(sp)
```

```
## Rows: 1,760
## Columns: 2
## $ Date  <date> 2025-08-01, 2025-07-31, 2025-07-30, 2025-07-29, 2025-07-28, 202…
## $ Price <dbl> 6238.01, 6339.39, 6362.90, 6370.86, 6389.77, 6388.64, 6363.35, 6…
```

```
glimpse(bitcoin)
```

```
## Rows: 2,558
## Columns: 2
## $ Date  <date> 2025-08-01, 2025-07-31, 2025-07-30, 2025-07-29, 2025-07-28, 202…
## $ Price <dbl> 113312.1, 115765.0, 117840.4, 117950.1, 118053.9, 119398.1, 1179…
```

```
dim(sp)
```

```
## [1] 1760    2
```

```
dim(bitcoin)
```

```
## [1] 2558    2
```

Task 1

For Task 1 of this project, the financial datasets were used to calculate different statistics such as the mean, median, mode, range, standard deviation, etc. Insights such as the measures of central tendency and variability were compared between both the datasets i.e. the S&P 500 and Bitcoin Historical Data over a period of 7 years from 2018-2025.

```
desc_stats_comparison_table <- data.frame(
  Stats = c("Mean", "Median", "SD", "Variance", "Min", "Max", "Range", "IQR"),

  SP500 = c(mean(sp$Price), median(sp$Price), sd(sp$Price), var(sp$Price), min(sp$Price), max
(sp$Price), max(sp$Price) - min(sp$Price), IQR(sp$Price)),

  Bitcoin = c(mean(bitcoin$Price), median(bitcoin$Price), sd(bitcoin$Price), var(bitcoin$Pric
e), min(bitcoin$Price), max(bitcoin$Price), max(bitcoin$Price) - min(bitcoin$Price), IQR(bitco
in$Price))
)


# Displaying with the help of knitr package to display it on a pdf

knitr::kable(desc_stats_comparison_table, caption = "Comparison of Descriptive stats between S
&P 500 and Bitcoin datasets", digits = 2)
```

Comparison of Descriptive stats between S&P 500 and Bitcoin datasets

Stats	SP500	Bitcoin
Mean	4112.21	35568.82
Median	4110.74	28002.00
SD	1010.76	28828.00
Variance	1021629.89	831053563.98
Min	2237.40	3228.70
Max	6389.77	119965.50

Stats	SP500	Bitcoin
Range	4152.37	116736.80
IQR	1481.13	43790.20

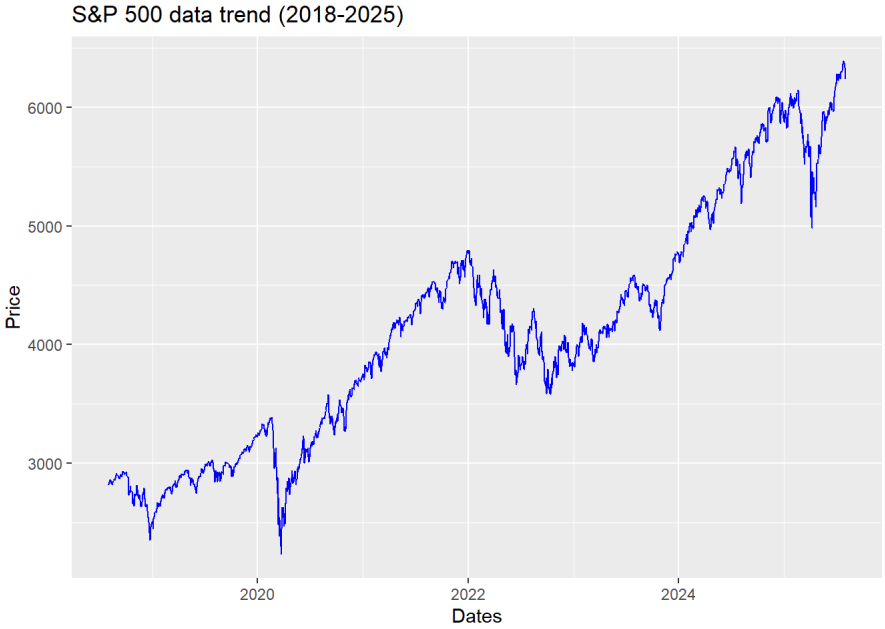
The table above was generated with the help of the knitr package, which helps in dynamic report generation. This is a cross table which shows the comparison between the different descriptive statistics measures of central tendency and variability between the two datasets.

From the table, we can see that the bitcoin historical data has a higher mean and median price compared to that of the S&P500 historical data. The Bitcoin data also has a much larger range and IQR. From the difference between the statistics of the datasets we can see that the Bitcoin data saw some dramatic changes from low to very high values compared to the S&P 500 historical data. It is clear that the Bitcoin experienced huge drastic changes as shown by it's higher range, IQR and standard deviation. The prices in the S&P 500 historical data fluctuate less dramatically compared to that of the Bitcoin historical data.

Task 2

For Task 2 of this data project, the focus was on showing the time series analysis or trends or patterns of the financial data over seven years (2018-2025). This task is about covering insights like for instance if there is an increasing or a decreasing pattern from the time series analysis for each of the datasets respectively.

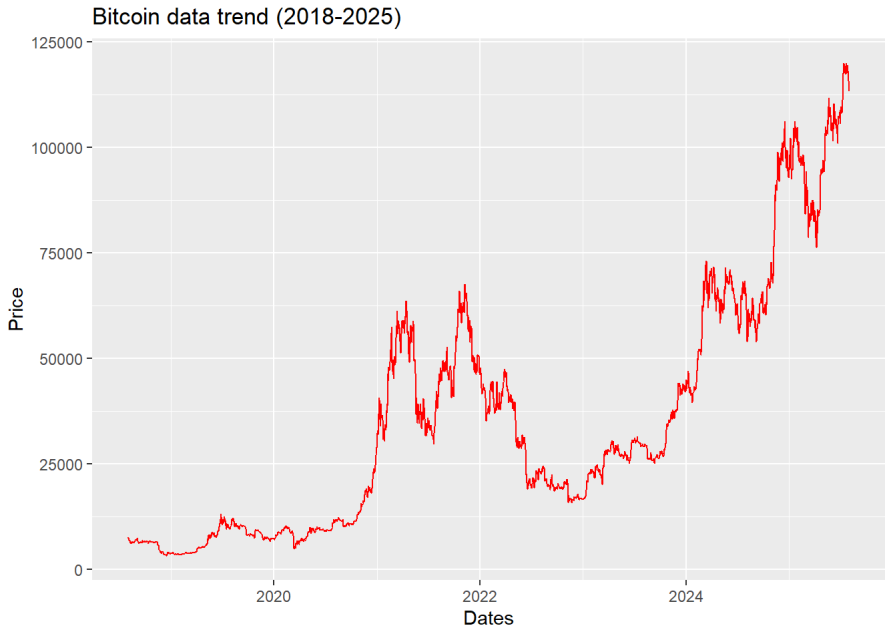
```
ggplot(sp, aes(x = Date, y = Price)) +
  geom_line(color = "blue") +
  labs(title = "S&P 500 data trend (2018-2025)", x = "Dates", y = "Price")
```



Overall, from the time series plot generated above we can see that the S&P 500 data shows an upwards trend over 7 years from 2018 to 2025. However, there are some declines in the trend in the mid 2020s and the beginning of 2022. Despite the shortcomings in the trend, there is a steady increase in the trends of the S&P

500 prices. This shows that the S&P 500 had a steady growth over the last 7 years. The steady growth of the S&P500 data or the stock market index is denoted by a blue color trend line in the plot.

```
ggplot(bitcoin, aes(x = Date, y = Price)) +
  geom_line(color = "red") +
  labs(title = "Bitcoin data trend (2018-2025)", x = "Dates", y = "Price")
```



From the Bitcoin trend plot or time series plot generated above we can see that the trends are more drastic and not as stable compared to the S&P 500 trend. There are sudden spikes and declines observed from the time series analysis of the Bitcoin data over the last 7 years. There was a sudden spike in the Bitcoin prices in early 2021 and a drastic drop in the prices towards the end of 2021, which then saw a spike in the prices of Bitcoin again from the year 2023. These drastic changes in the prices of Bitcoin shows that it is not a very stable asset compared to the S&P 500. This shows that Bitcoin is a highly volatile asset as compared to the S&P500. The trend or pattern of the changes in the prices of Bitcoin is denoted by a red trend line in the plot.

The two financial datasets are merged together with the help of an inner join on the Date column from both the datasets. This allows for the comparison of the prices between S&P500 and Bitcoin. This also helps in the calculation of the correlation calculated every 6 months between the datasets over seven years.

```
merged_data <- inner_join(sp, bitcoin, by = "Date", suffix = c("_sp500", "_bitcoin"))
head(merged_data)
```

Date <date>	Price_sp500 <dbl>	Price_bitcoin <dbl>
2025-08-01	6238.01	113312.1
2025-07-31	6339.39	115765.0
2025-07-30	6362.90	117840.4

Date <date>	Price_sp500 <dbl>	Price_bitcoin <dbl>
2025-07-29	6370.86	117950.1
2025-07-28	6389.77	118053.9
2025-07-25	6388.64	117631.9

6 rows

```
dim(merged_data)  # Shows the dimensions of the merged dataset
```

```
## [1] 1760    3
```

Correlation shows us how the relationship between the 2 financial datasets are. It is a measure that helps show the strength of the relationship between the datasets.

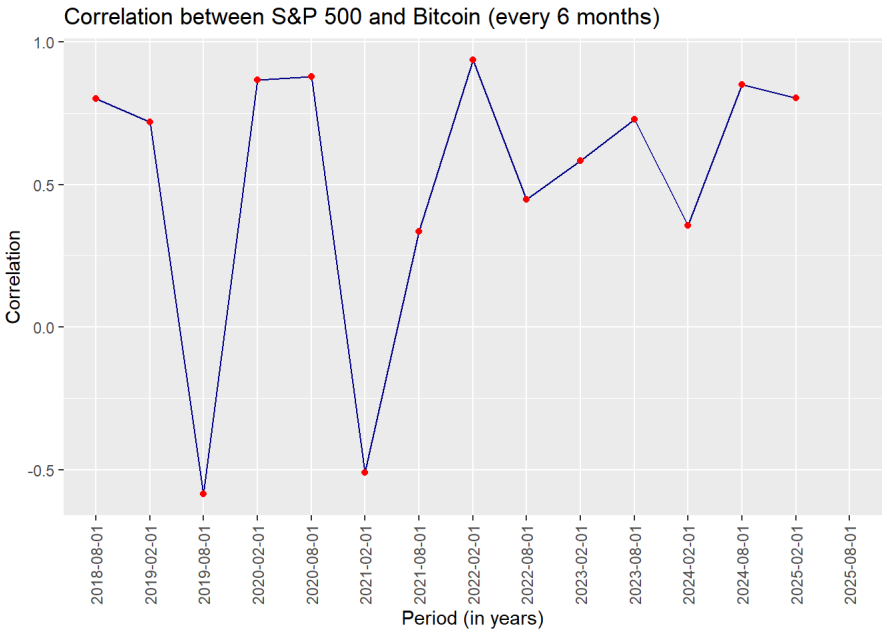
```
merged_data$Period <- cut(merged_data$Date, breaks = "6 months")  # grouping the dates by 6 mo
nths


correlation_6months <- merged_data %>% group_by(Period) %>% summarise(Correlation = cor(Price_
sp500, Price_bitcoin, use = "complete.obs"))

ggplot(correlation_6months, aes(x = Period, y = Correlation, group = 1)) +
  geom_line(color = "darkblue") +
  geom_point(color = "red") +
  labs(title = "Correlation between S&P 500 and Bitcoin (every 6 months)", x = "Period (in yea
rs)", y = "Correlation") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))
```

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

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



A correlation plot was generated to show the relationship between the Bitcoin and S&P 500 prices every 6 months over the last 7 years from 2018-2025. From the plot, we can see that the correlation points are scattered across and are not consistent through the years. At some points through the years, there are stronger relationships between the prices of Bitcoin and S&P500, and sometimes have weaker correlations which suggests that the Bitcoin prices sometimes changes independently regardless of the movements of the stock market.

Task 3

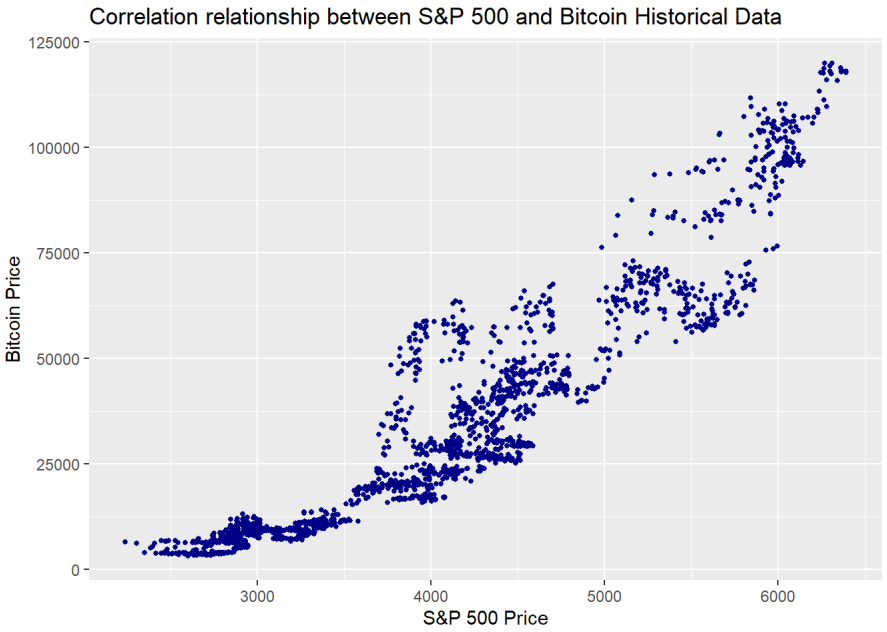
This task is about computing the correlation coefficient. The correlation coefficient is a statistical measure which is use to measure the strength of the relationship between two variables and the direction in which they move.

```
correlation_coeff <- cor(merged_data$Price_sp500, merged_data$Price_bitcoin, use = "complete.o
bs")
correlation_coeff
```

```
## [1] 0.9323243
```

The correlation coefficient was calculated based on the merged dataset of the financial data created from the previous task. The correlation coefficient calculated above came to about 0.93 which represents a strong relationship between the financial data. i.e. Bitcoin and S&P 500.

```
# Plotting of the correlation coefficient relationship of the merged financial historical data
ggplot(merged_data, aes(x = Price_sp500, y = Price_bitcoin)) +
  geom_point(size = 1, color = "darkblue") +
  labs(title = "Correlation relationship between S&P 500 and Bitcoin Historical Data", x = "S&
P 500 Price", y = "Bitcoin Price")
```



From the scatter plot generated above, we can see that the correlation coefficient or the relationship between the financial data is very strong, moving in the same direction upwards in the plot from 2018 to 2025. This plot strongly depicts the strong correlation coefficient calculated which came to about 0.93. This shows that bitcoin and S&P 500 prices made similar movements over the last 7 years or observed periods. [12]

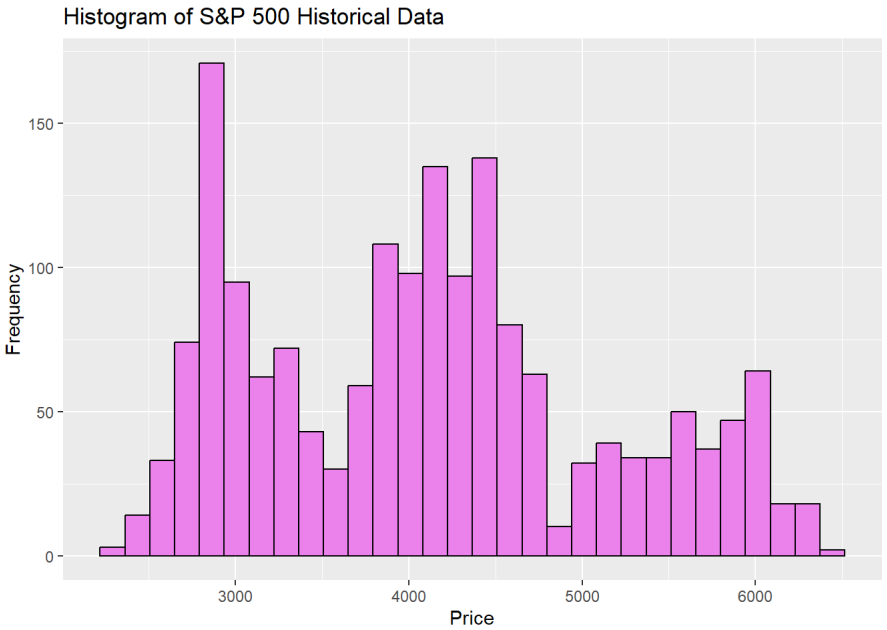
Task 4

Histograms

This task includes the analysis of datasets separately to assess whether it follows a normal distribution with the help of histograms, Q-Q plots, and Shapiro-Wilk tests of normality.

```
# Histogram for S&P Historical Data
ggplot(sp, aes(x = Price)) +
  geom_histogram(fill = "violet" ,color = "black") +
  labs(title = "Histogram of S&P 500 Historical Data", x = "Price", y = "Frequency")
```

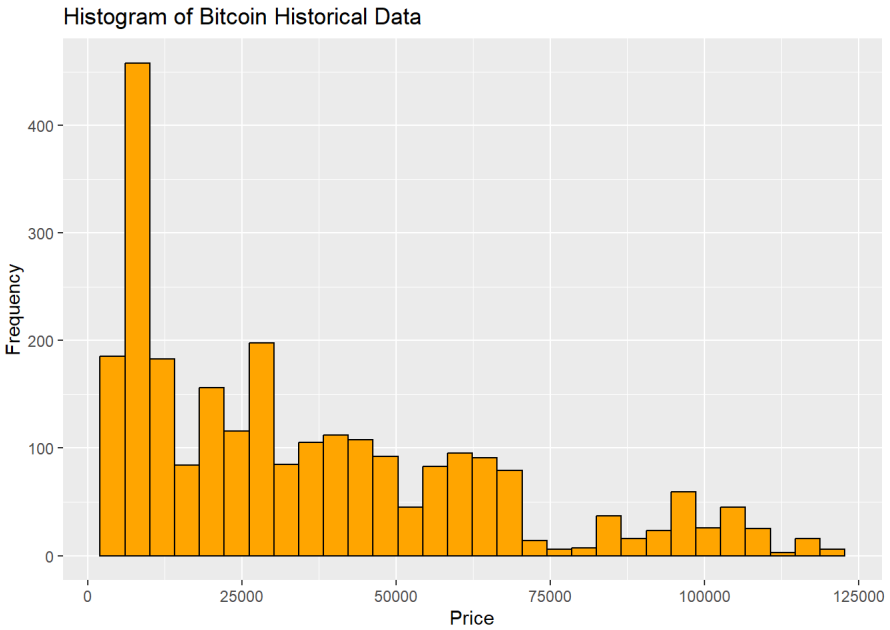
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Histograms are useful to check for normality. if the histogram has a bell-shaped curve in the middle of the plot, it follows normality. The histogram generated above shows a distribution that is not a perfectly shaped bell-curve. This shows that the S&P 500 data does not follow a normal distribution. [11]

```
# Histogram for Bitcoin Historical Data
ggplot(bitcoin, aes(x = Price)) +
  geom_histogram(fill = "orange" ,color = "black") +
  labs(title = "Histogram of Bitcoin Historical Data", x = "Price", y = "Frequency")
```

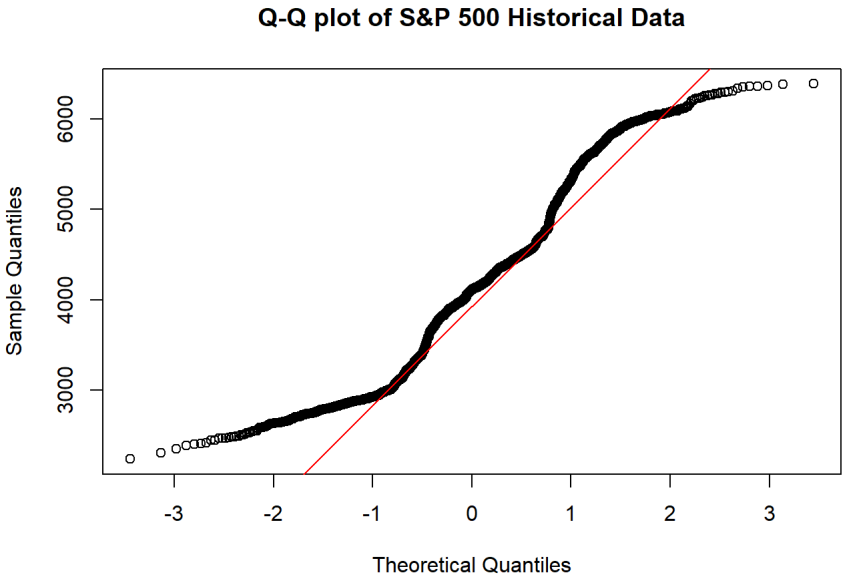
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



The histogram of the Bitcoin historical data generated above also does not a normal distribution as it clearly does not follow a slope like structure and does not have a bell-shaped curve. Hence, we can conclude that the Bitcoin historical data also does not follow normality. [11]

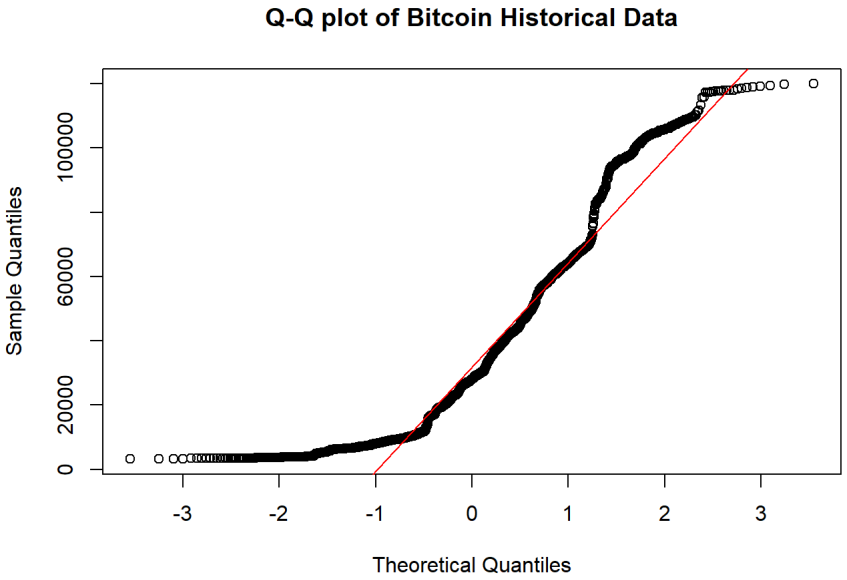
Q-Q plots

```
# Q-Q Plot for S&P Historical Data
qqnorm(sp$Price, pch = 1, frame = TRUE, main = "Q-Q plot of S&P 500 Historical Data")
qqline(sp$Price, col = "red")
```



Q-Q plots are another useful way of checking if the financial data follows normality. If the data points follow the straight line on the plot it indicates normality. From the Q-Q plot generated above, we can see that there are deviations in the data points following the straight inclined line on the plot. Hence, we can conclude that S&P 500 does not follow normality as the data points are deviating from the line on the plot. [8]

```
# Q-Q Plot for Bitcoin Historical Data
qqnorm(bitcoin$Price, pch = 1, frame = TRUE, main = "Q-Q plot of Bitcoin Historical Data")
qqline(bitcoin$Price, col = "red")
```



The Q-Q plot of the Bitcoin data generated from above also does not follow normality as there are deviations of the data points from the red straight line on the plot. Hence, we can conclude that Bitcoin historical data does not follow normality.

Shapiro-wilk normality tests

Shapiro-Wilk normality tests were applied to both the datasets separately to check for normality.

```
# Normality test using Shapiro-Wilk test of normality for S&P 500 historical data
shapiro.test(sp$Price)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  sp$Price
## W = 0.95469, p-value < 2.2e-16
```

The Shapiro-Wilk normality test outcome of the S&P 500 data from above indicates that W = 0.95469. The value of W is fairly close to 1, which means it's somewhat closer to normal. But, the value of p (p-value) is lesser than 0.05 ($p < 0.05$) which means it does not follow normality. For the data to follow normality the p-value must be greater than 0.05 ($p > 0.05$) [7]. But, this is not the case for the S&P 500 historical data. Therefore, S&P 500 does not follow normality.

```
# Normality test using Shapiro-Wilk test of normality for Bitcoin historical data
shapiro.test(bitcoin$Price)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  bitcoin$Price
## W = 0.88853, p-value < 2.2e-16
```

The Shapiro-Wilk normality test outcome of the Bitcoin historical data from above indicates that W = 0.88853. The value is further away from 1 than that of the S&P 500 historical data, a bit further away from being normal. The value of p (p-value) is lesser than 0.05 ($p < 0.05$) [7]. This indicates that the Bitcoin historical data also does not follow normality.

Conclusion

This report compared the S&P 500 and Bitcoin using descriptive statistics, time series analysis with the help of patterns or trends, correlations, and Shapiro-Wilk normality tests. The results of different insights and analysis show that Bitcoin is a highly volatile asset as compared to S&P 500 which is more stable and steady. Both the datasets did not follow normality or a normal distribution and the correlation of the datasets varied over a period of time. Overall, this analysis shows the difference between the stability of S&P 500 which is more traditional and the volatility of Bitcoin.

References

[1] “Date values in R”, datacamp.com. [Online]. Available: https://www.datacamp.com/doc/r/dates (https://www.datacamp.com/doc/r/dates)

[2] “R Data Frames”, w3schools.com. [Online]. Available: https://www.w3schools.com/r/r_data_frames.asp (https://www.w3schools.com/r/r_data_frames.asp)

[3] “The function knitr::kable()”, bookdown.org. [Online]. Available: https://bookdown.org/yihui/rmarkdown-cookbook/kable.html (https://bookdown.org/yihui/rmarkdown-cookbook/kable.html)

[4] “kable: Create tables in Markdown”, rdocumentaton.org. [Online]. Available: https://www.rdocumentation.org/packages/knitr/versions/1.50/topics/kable (https://www.rdocumentation.org/packages/knitr/versions/1.50/topics/kable)

[5] “ggplot2 line plot: Quick start guide”, sthda.com. [Online]. Available: https://www.sthda.com/english/wiki/ggplot2-line-plot-quick-start-guide-r-software-and-data-visualization#data (https://www.sthda.com/english/wiki/ggplot2-line-plot-quick-start-guide-r-software-and-data-visualization#data)

[6] “Correlation Test Between Two Variables in R”, sthda.com. [Online]. Available: https://www.sthda.com/english/wiki/correlation-test-between-two-variables-in-r (https://www.sthda.com/english/wiki/correlation-test-between-two-variables-in-r)

[7] “Normality Test in R”, sthda.com. [Online]. Available: https://www.sthda.com/english/wiki/normality-test-in-r (https://www.sthda.com/english/wiki/normality-test-in-r)

[8] “QQ-plots: Quantile-Quantile plots - R-Base Graphs”, sthda.com. [Online]. Available: https://www.sthda.com/english/wiki/qq-plots-quantile-quantile-plots-r-base-graphs (https://www.sthda.com/english/wiki/qq-plots-quantile-quantile-plots-r-base-graphs)

[9] “How to merge data in R using R merge, dplyr or data.table”, geeksforgeeks.org. [Online]. Available: https://www.geeksforgeeks.org/r-language/how-to-merge-data-in-r-using-r-merge-dplyr-or-data-table/ (https://www.geeksforgeeks.org/r-language/how-to-merge-data-in-r-using-r-merge-dplyr-or-data-table/)

[10] “Why is Bitcoin Volatile?”, investopedia.com. [Online]. Available: https://www.investopedia.com/articles/investing/052014/why-bitcoins-value-so-volatile.asp# (https://www.investopedia.com/articles/investing/052014/why-bitcoins-value-so-

volatile.asp#):~:text=If%20you're%20looking%20to,are%20to%20make%20any%20gains.

[11] "How to Test for Normality in R", statology.org. [Online]. Available: https://www.statology.org/test-for-normality-in-r/ (https://www.statology.org/test-for-normality-in-r/)

[12] "Scatterplot & Correlation", study.com. [Online]. Available: https://study.com/academy/lesson/scatter-plot-and-correlation-definition-example-analysis.html# (https://study.com/academy/lesson/scatter-plot-and-correlation-definition-example-analysis.html#):~:text=A%20scatterplot%20with%20a%20positive,upwards%20from%20left%20to%20right.&text=T