Statistical analysis of Bitcoin and Gold Prices

1 Author Details

Tran Khanh An Dinh (Hayley Dinh)

2 Introduction & Problem Statement

In recent years, Bitcoin has immensely risen in popularity and is referred to by many investors as “digital gold” and believed to have the same benefit as gold in terms of hedging against inflation (Takinsoy 2021). In this report, the relationship and movement of gold and Bitcoin will be comprehensively analysed with data from 7 recent years to come to a conclusion if the claim is valid with relevant graphs where necessary to help with visualization and apprehension.

More specifically, even though gold and Bitcoin are perceived as similar due to value-preserving characteristics, how similar are they actually in terms of basic statistical values, trends, distribution and how close interrelated is their relationship? These questions are addressed through the following methods:

- Basic descriptive statistics
- Boxplot distribution analysis
- Individual trend line plots
- Correlation plots
- Histogram addressing normal distribution

3 Analysis and Findings

3.1 Load Packages

```r
# Loading necessary packages
library(readr)
library(dplyr)
library(ggplot2)
library(tidyr)
library(lubridate)
```

3.2 Data

```r
# Setting work directory
setwd("~/Documents/Masters/Applied Analytics")
# Importing data
BitcoinData = read_csv("data/Bitcoin Historical Data-1.csv")
GoldData = read_csv("data/Gold Futures Historical Data-1.csv")
```

3.3 Task 1 - Descriptive statistics

3.3.1 Basic statistics

```
# Basic descriptive statistics for Bitcoin
summary(BitcoinData)
```

```
##      Date               Price
##  Length:2557        Min.   :  3229
##  Class :character   1st Qu.:  8953
##  Mode  :character   Median : 23016
##                     Mean   : 29095
##                     3rd Qu.: 43580
##                     Max.   :106157
```

```
# Standard deviation of Bitcoin
sd(BitcoinData$Price)
```

```
## [1] 23404.75
```

```
# Basic descriptive statistics for Gold
summary(GoldData)
```

```
##      Date               Price
##  Length:1804        Min.   :1174
##  Class :character   1st Qu.:1499
##  Mode  :character   Median :1803
##                     Mean   :1785
##                     3rd Qu.:1949
##                     Max.   :2852
```

```
#Standard deviation of Gold
sd(GoldData$Price)
```

```
## [1] 371.1826
```

The price distribution of Gold and Bitcoin is massively different even through these simple statistics. Gold prices have a significantly smaller range, from just over 1000 to just under 3000. On the other hand, Bitcoin ranges from over 3000 to over 100000. Thus, Bitcoin has a considerably wider spread. Similarly, the standard deviation shows that Bitcoin has much higher variability, with gold's value being only 371 and Bitcoin's value being 23404. Regarding central tendency, Bitcoin's mean and median are on the very lower end of the dataset. Meanwhile, Gold's mean and median are quite central. To better illustrate the distribution of these financial instruments, let's examine the following graphs.

3.3.2 Boxplots

```
# Boxplot of Bitcoin Prices
ggplot(BitcoinData, aes(x="", y = Price),) +
  geom_boxplot(fill="cadetblue1") +
  stat_summary() +
  labs(x= "", y = "Bitcopin Price", title = "Bitcoin Price distribution")
```

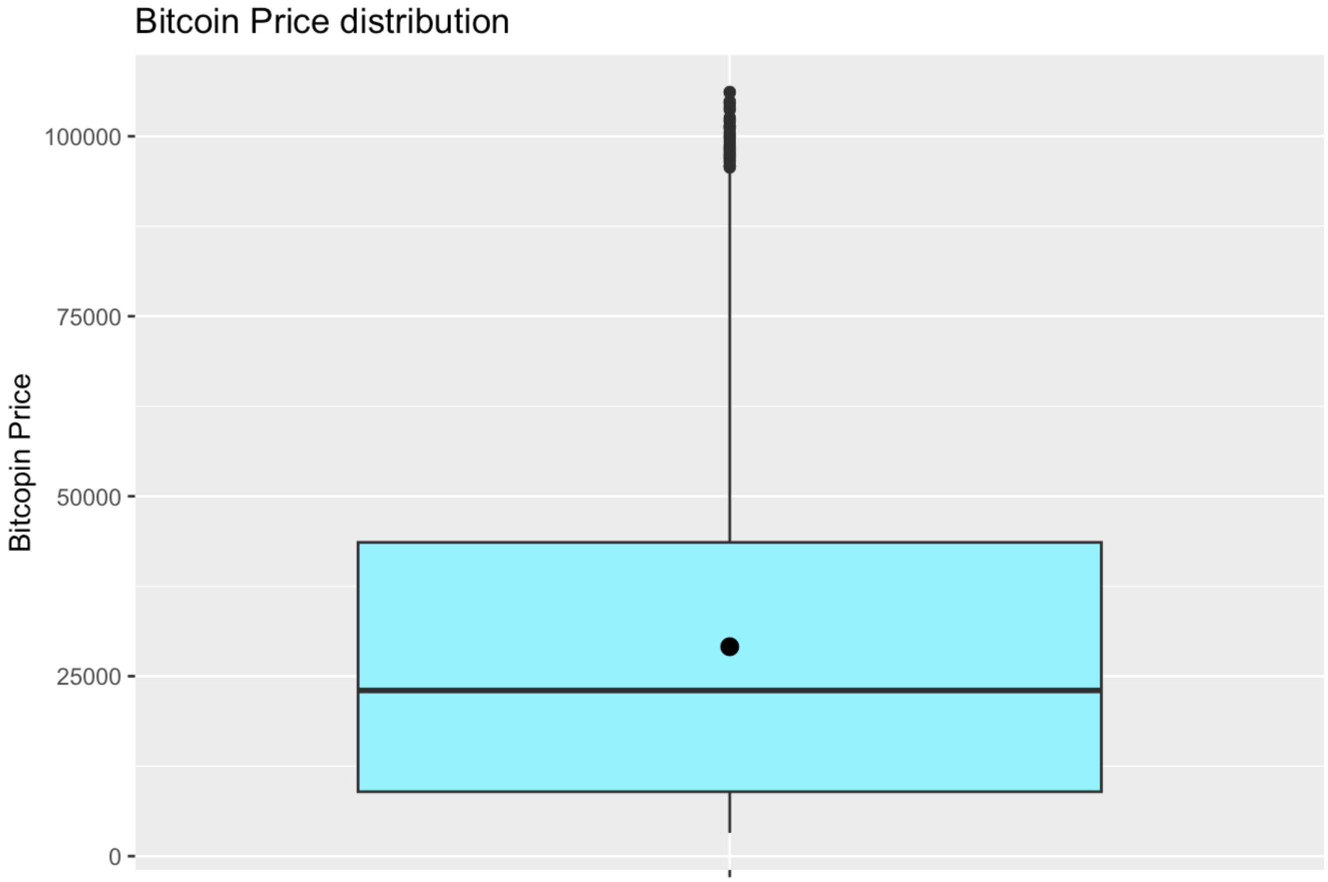

Figure 3.1: Boxplot of Bitcoin Prices

Referring to **Figure 3.1**, the distribution of Bitcoin prices is largely right skewed with a very high proportion of upper outliers.This is also shown by the fact that the mean is noticeably larger than the median. This proposes that most prices are on the lower end, however, there are a lot of unusually extreme high values. Considering real-world context, this may be because Bitcoin prices skyrocketed in more recent years, contrasting with past prices and creating a substantially skewed dataset.

```
# Boxplot of Gold Prices
ggplot(GoldData, aes(x="", y = Price)) +
  geom_boxplot(fill="gold1") +
  stat_summary() +
  labs(x= "", y = "Gold Price", title = "Gold Price distribution")
```

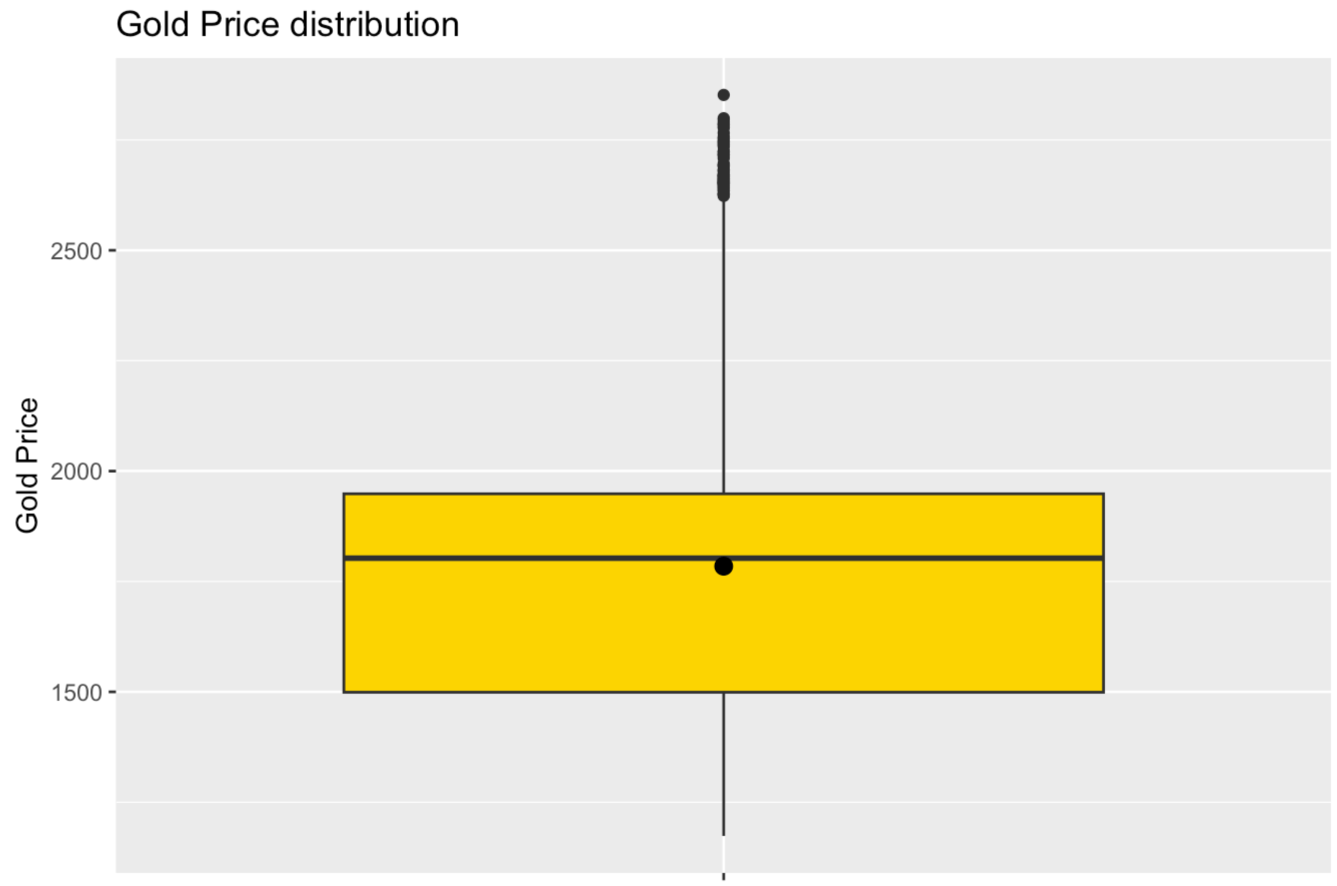

Figure 3.2: Boxplot of Gold Prices

Referring to **Figure 3.2**, the distribution of Gold prices is also but to a lesser extent right skewed, and again, with a very high proportion of upper outliers. Overall, the distribution is quite similar to that of Bitcoin. As opposed to Bitcoin, Gold’s skewness may be caused by the upper outliers, as the mean and median is relatively in the middle and are quite close together, suggesting that the overall data is more balanced around the midpoint.

```
# Joining both data sets by "Date" column to get the same number of observations in both
df = right_join(BitcoinData, GoldData, by = "Date")
colnames(df) = c("Date", "Bitcoin", "Gold")
# Reformating joined data set to long format
df_long = df %>%
  pivot_longer(cols = 2:3, names_to = "Type", values_to = "Price")
# Making boxplot of both Bitcoin and Gold
ggplot(df_long, aes(x = Type, y = Price, fill=Type)) +
  geom_boxplot() +
  scale_fill_manual(values = c("Bitcoin" = "cadetblue1", "Gold" = "gold1")) +
  labs(x = "Type", y = "Price", title = "Boxplot of Gold and Bitcoin prices") +
  theme_minimal()
```

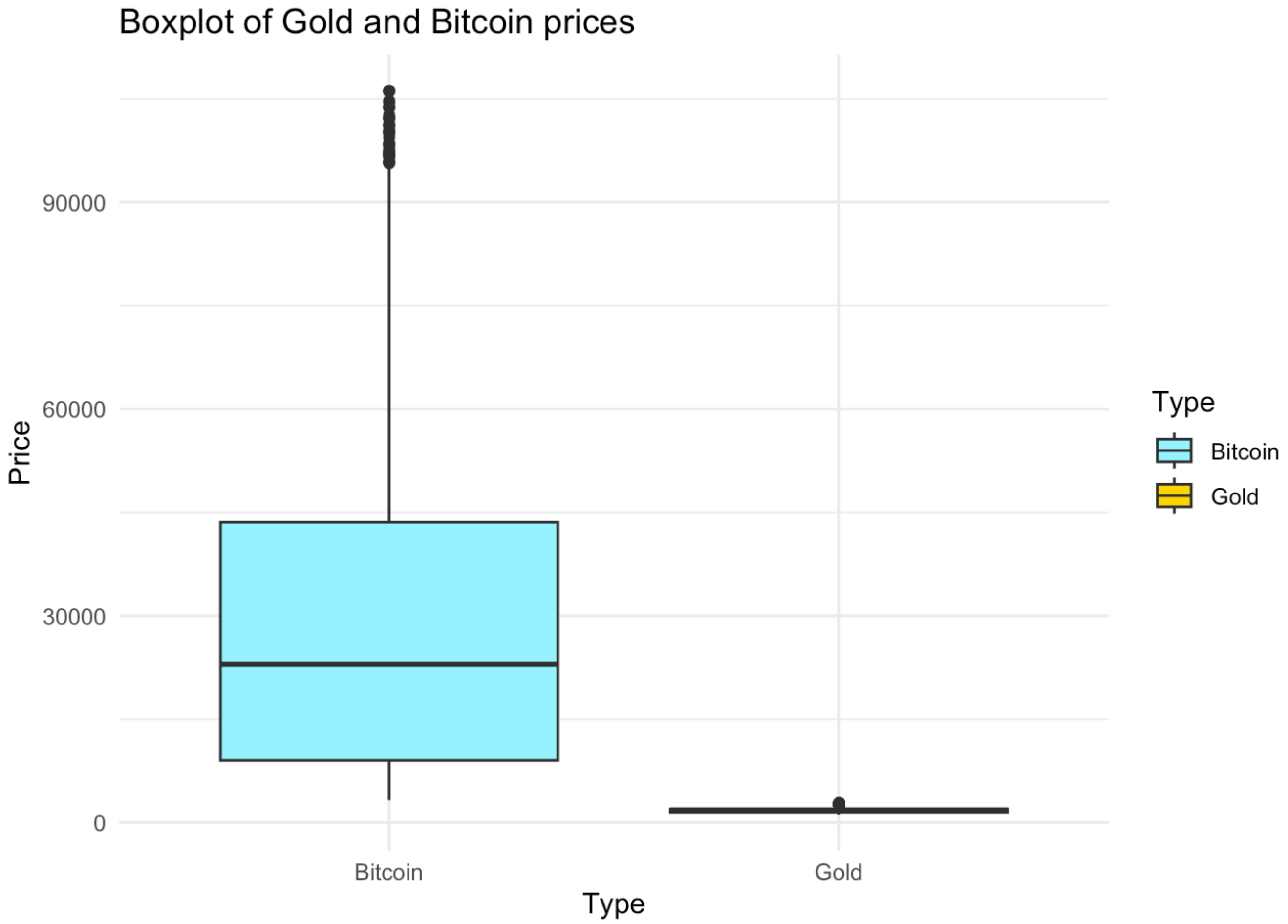

Figure 3.3: Boxplot of Bitcoin and Gold Prices

Nonetheless, when put together, it can clearly be seen how the spread and variability of Bitcoin prices completely dominates Gold prices as Gold prices can barely show up in the graph. To summarise, even though both datasets are right skewed and include upper outliers, Gold prices are much more consistent and symmetric compared to Bitcoin prices, which seems to be extraordinarily volatile.

3.4 Task 2 - Trend of Bitcoin and Gold price movement and their correlation by period

3.4.1 Bitcoin

```
# Set date to the correct format
df$Date = as.Date(df$Date, format = "%d/%m/%Y")

# Creating line plot for Bitcoin
ggplot(df, aes(x = Date, y = Bitcoin)) +
  geom_line(color="cadetblue2") +
  geom_smooth(method = "lm", color = "tomato1", se=FALSE) +
  labs(title = "Price", x = "Date", y = "Price") +
  theme_minimal()
```

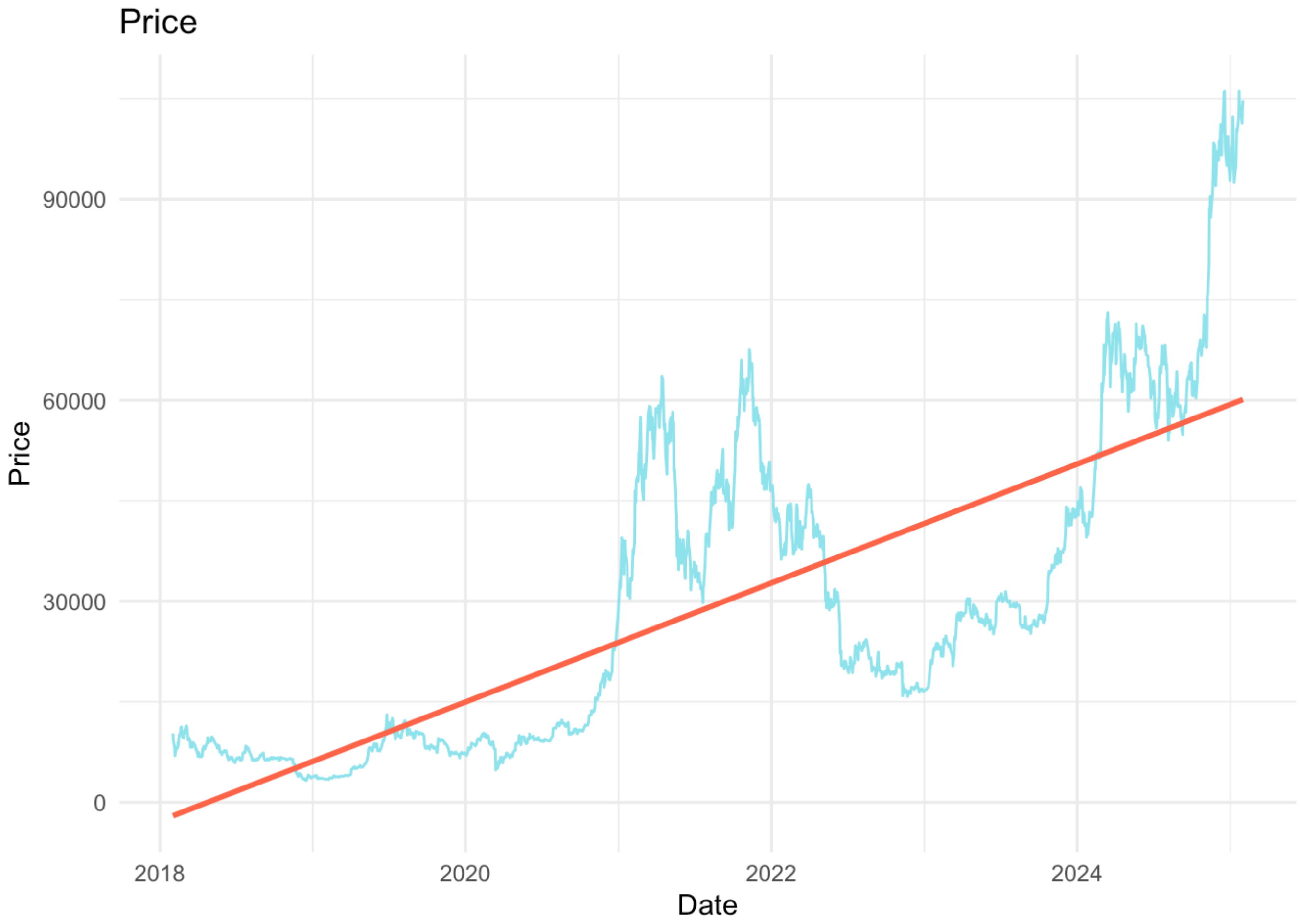

Figure 3.4: Line Chart of Bitcoin Prices through 7 years

Corroborating with previous evidence, the price of Bitcoin moves up and down very rapidly and drastically. This is distinct through several points, such as the price soaring in 2021 as well as in 2024-2025. On the other hand, there is a very extreme dip around 2023. Overall, there is a very prevalent upwards trend.

3.4.2 Gold

```
# Creating line plot for Gold
ggplot(df, aes(x = Date, y = Gold)) +
  geom_line(color="gold1") +
  geom_smooth(method = "lm", color = "tomato1", se=FALSE) +
  labs(title = "Price", x = "Date", y = "Price") +
  theme_minimal()
```

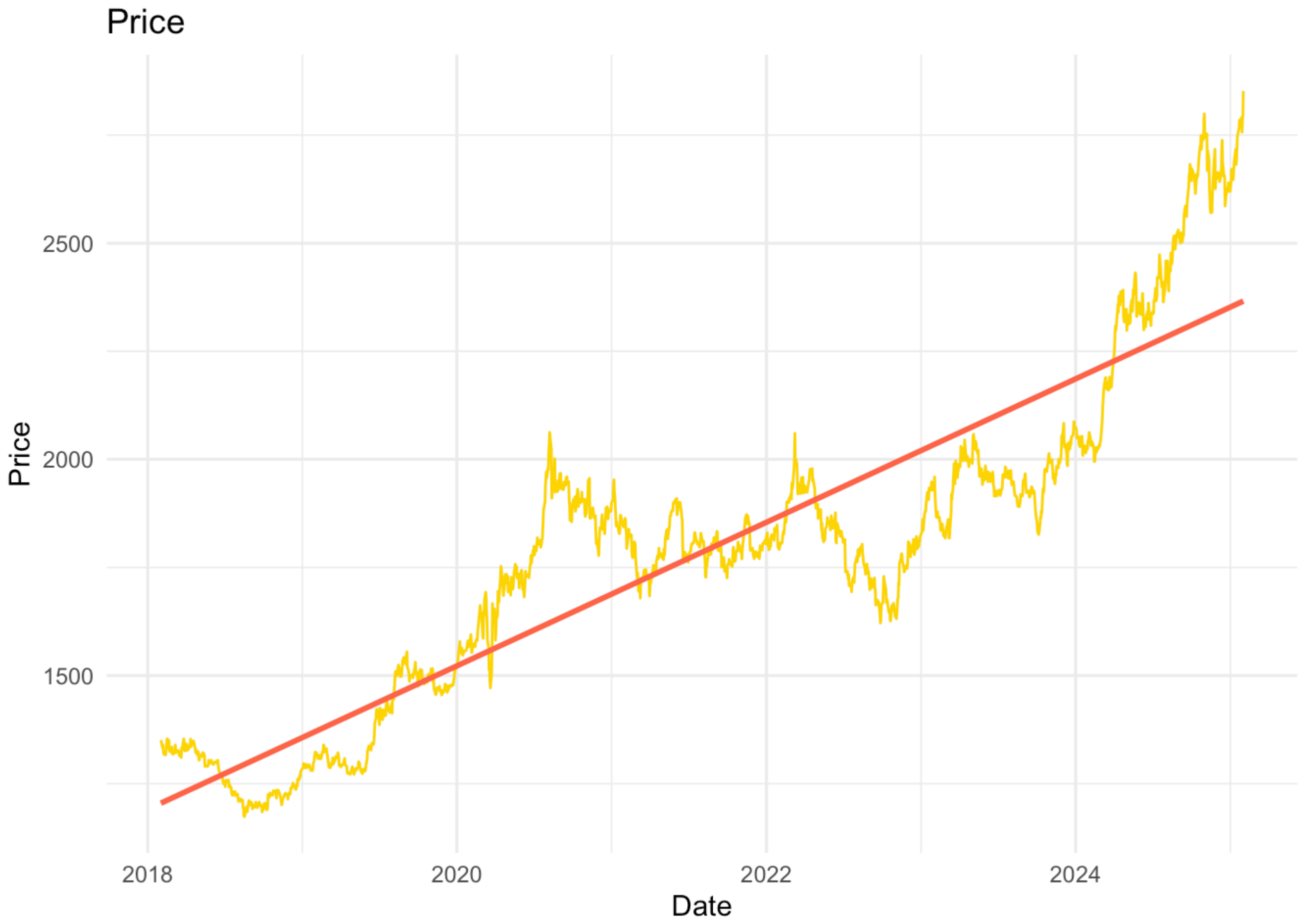

Figure 3.5: Line Chart of Gold Prices through 7 years

Again, in line with presented calculations, Gold prices form a much straighter line through 7 recent years. Interestingly, there is also a noticeable peak at around 2021 and strong upwards movement around 2024-2025. Furthermore, there is also a small dip around 2023. Generally, Gold prices also show a powerful and firm rising movement.

3.4.3 Correlation by period

```
# Grouping data by 6 months chunks
price_correlation = df %>% mutate(Period = floor_date(Date, "6 months"))
# Isolate grouped data to show only period and correlation
price_correlation = price_correlation %>% group_by(Period) %>% summarise(Correlation
= cor(Bitcoin, Gold))
# Making correlation plot
ggplot(price_correlation, aes(x = Period, y = Correlation)) +
  geom_line(color="mediumpurple1", size = 1) +
  geom_smooth(method = "lm", color = "tomato1", se=FALSE) +
  geom_point() +
  labs(title = "Correlation by price", x = "period", y = "correlation") +
  theme_minimal()
```

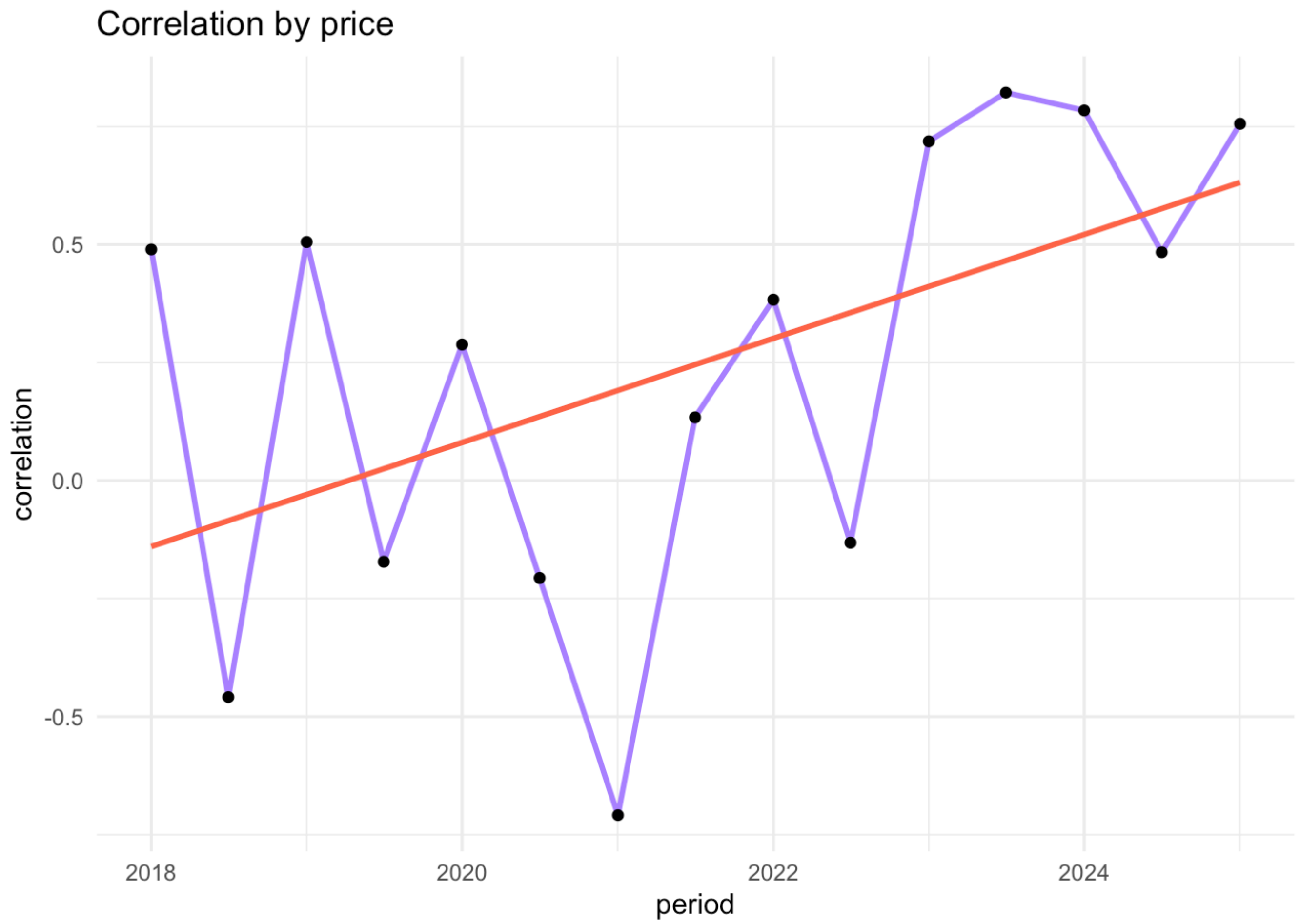

Figure 3.6: Correlation between Bitcoin and Gold by 6 months intervals

Correlation between the prices of Gold and Bitcoin by 6 month periods show high volatility. More specifically, 5 of 15 data points suggest negative correlation, with one datapoint reaching below -0.5. The upward sloping trend line indicates that the two assets are behaving more similarly as time passes. Despite that, the erratic spread of the data points hint at a not very consistent or strong correlation.

3.4.4 Correlation of returns by period

```
# Adding returns to original data
df_returns = df %>% mutate(BitcoinReturns = (Bitcoin/lag(Bitcoin)), GoldReturns = (Gold/lag(Gold)))
# Grouping returns data by 6 months chunks
df_returns = df_returns %>% mutate(Period = floor_date(Date, "6 months"))
# Isolate grouped data to show only period and correlation
returns_correlation = df_returns %>% group_by(Period) %>% summarise( Correlation = cor(BitcoinReturns, GoldReturns))
returns_correlation = returns_correlation[1:14,]

# Making correlation plot in regards to returns
ggplot(returns_correlation, aes(x = Period, y = Correlation)) +
  geom_line(color="mediumpurple1", size = 1) +
  geom_smooth(method = "lm", color = "tomato1", se=FALSE) +
  geom_point() +
  labs(title = "Correlation by returns", x = "period", y = "correlation") +
  theme_minimal()
```

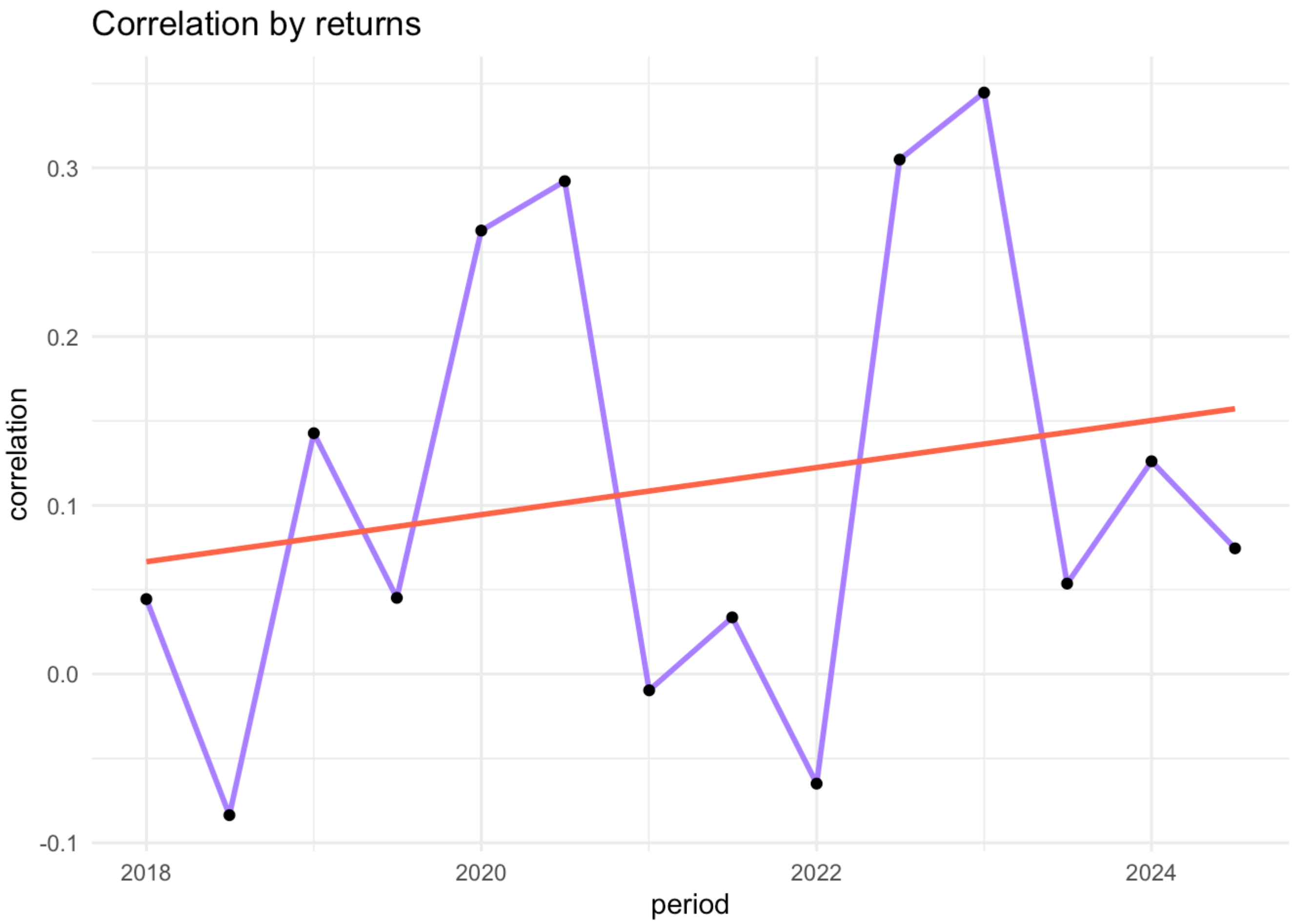

Figure 3.7: Correlation between Bitcoin and Gold by 6 months intervals

With the purpose to more accurately investigate the correlation between the prices of Bitcoin and Gold, research suggests testing correlation using returns as values instead (Kamalov et al. 2021). While the graph still appears to be volatile, it is a lot more stable compared to **Figure 3.7**, there are now only 3 of the 15 data points diving below 0, and not by much. The trend line also follows an upward slope, but much more gentle, implying that the returns of the two are becoming ever so slightly more similar through time. Whilst the increase in correlation is weaker, there is a much more consistent correlation relationship when using returns.

3.5 Task 3 - Correlation and relationship between Bitcoin and Gold data

3.5.1 Relationship by price

```
# Calculating correlation coefficient
cor(df$Bitcoin,df$Gold)
```

```
## [1] 0.8328287
```

```
# Reformat table to long format for graph
df_long = df %>%
  pivot_longer(cols = 2:3, names_to = "Type", values_to = "Price")

# Making price scatter plot
ggplot(df_long, aes(x = Date, y = Price, color = Type)) +
  geom_point() +
  geom_smooth(method = "lm", color = "tomato1") +
  scale_color_manual(values = c("cadetblue1","gold1")) +
  labs(x = "Date", y = "Price", title = "Bitcoin and Gold Prices scatterplot") +
  theme_minimal()
```

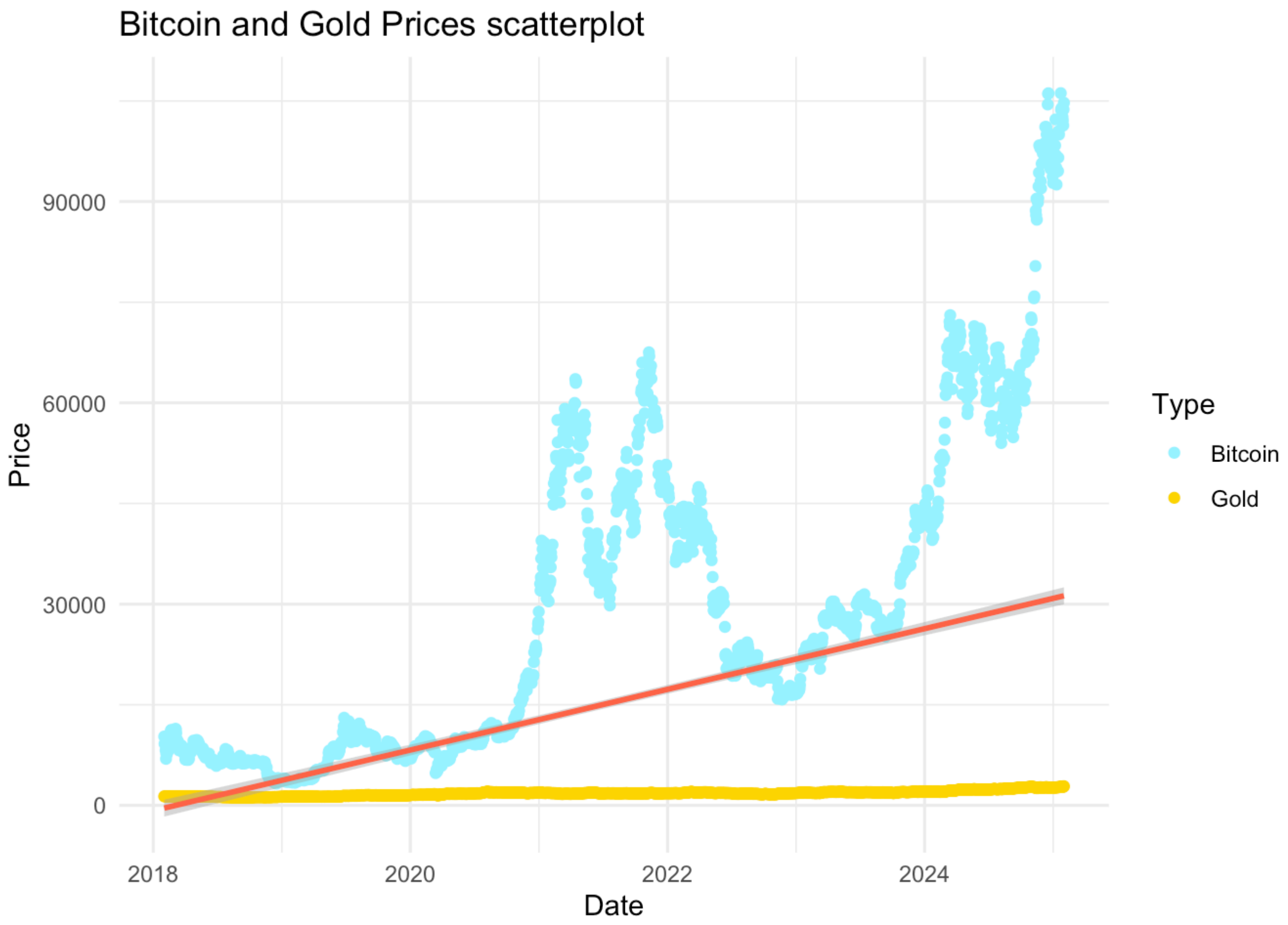

Figure 3.8: Scatterplot of Bitcoin and Gold prices

Purely using the prices of Bitcoin and Gold throughout the 7 years, there is an impressive correlation of 0.83, a strong positive relationship. Referring to **Figure 3.8**, it can be seen that both financial data sets are moving upwards in price, though, the increase in Gold prices is much more insubstantial. Furthermore, the data points are not clustered around the trend line. Because the spread of data is so widely different between the two variables, the trend is not quite visually clear. Also, because these calculations use raw prices, there could be misleadingly positive correlation because of overall financial trends that impact all investment assets. For that reason, let’s once again examine the relationship between their returns.

3.5.2 Relationship by returns

```
# Making separate returns table
df_returns_clean = df_returns[2:1804,c("Date","BitcoinReturns","GoldReturns")]
cor(df_returns_clean$BitcoinReturns,df_returns_clean$GoldReturns)
```

```
## [1] 0.1254079
```

```
# Reformatting returns table
df_long_returns = df_returns_clean %>%
  pivot_longer(cols = 2:3, names_to = "Type", values_to = "Price")
# Making returns scatter plot
ggplot(df_long_returns, aes(x = Date, y = Price, color = Type)) +
  geom_point() +
  geom_smooth(method = "lm", color = "tomato1") +
  scale_color_manual(values = c("cadetblue1","gold1")) +
  labs(x = "Date", y = "Returns", title = "Bitcoin and Gold Returns scatterplot") +
  theme_minimal()
```

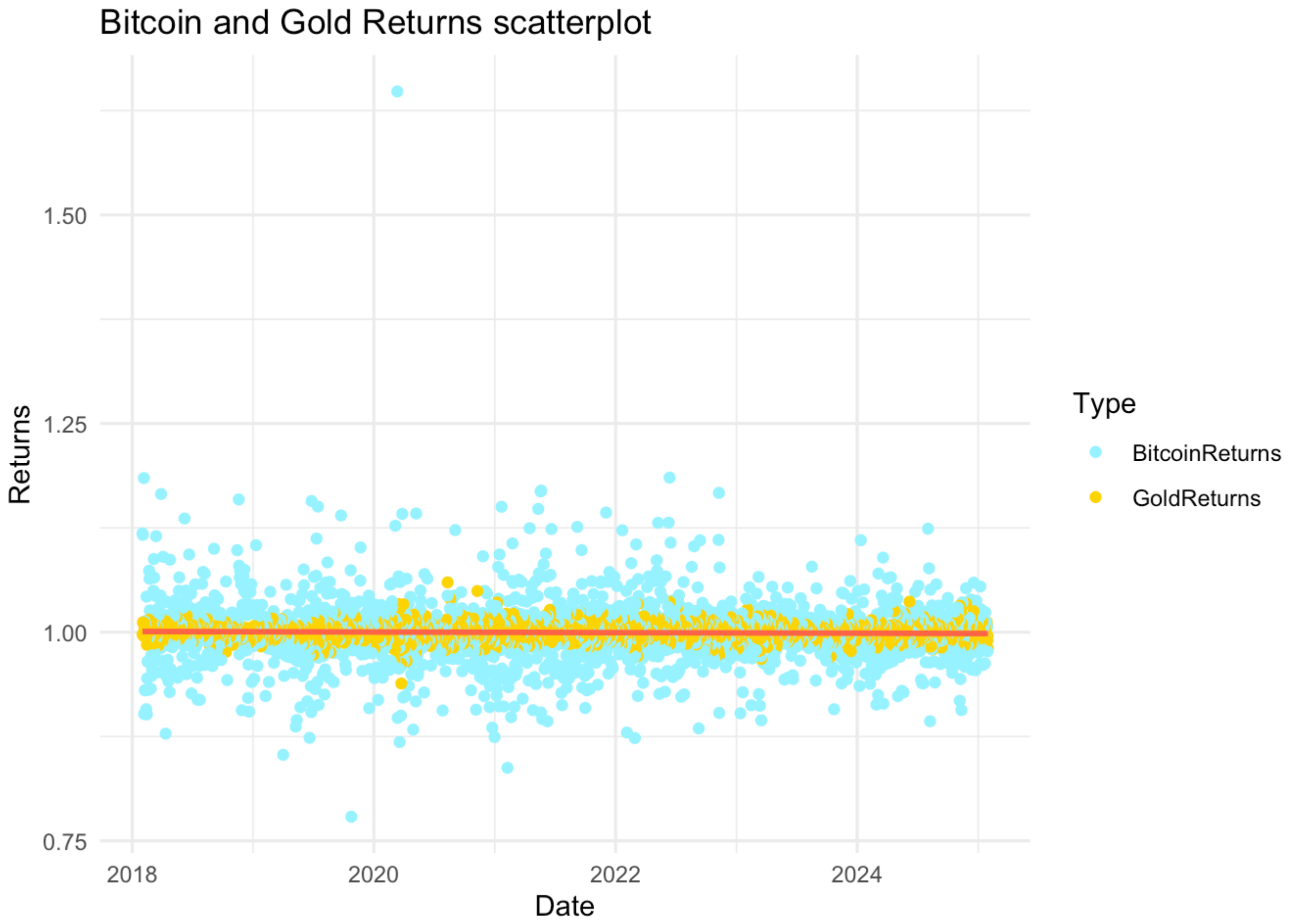

Figure 3.9: Scatterplot of Bitcoin and Gold returns

Results changed extensively when values using returns. First off, the correlation coefficient dropped to a mere 0.12, displaying a negligible correlation. Similarly, **Figure 3.9** shows a completely flat trend line, indicating no relationship or trend. However, it can still be very clearly seen how Bitcoin returns are much more unstable while Gold returns are much more stable as they are packed more tightly. This new graph suggests that the previously calculated correlation between raw prices are largely motivated by overall financial trends and the two financial instruments are not actually meaningfully connected.

3.6 Task 4 - Assess whether the Gold data and Bitcoin data follow a normal distribution

A histogram of a normally distributed data set will follow the standard bell curve shape. Its basic characteristics include symmetry, with one peak centered around the mean and median. Therefore, plotting a histogram for each financial instrument can provide a general visual idea if the data set follows a normal distribution. After that, for a more in-depth analysis, each asset will also get a Q-Q plot to more certainly support the analysis.

3.6.1 Bitcoin

```
# Making histogram for Bitcoin
ggplot(df, aes(x = Bitcoin)) +
  geom_histogram(fill="cadetblue2",binwidth = 350) +
  labs(title = "Histogram of Bitcoin", x = "Price", y = "Frequency") +
  theme_minimal()
```

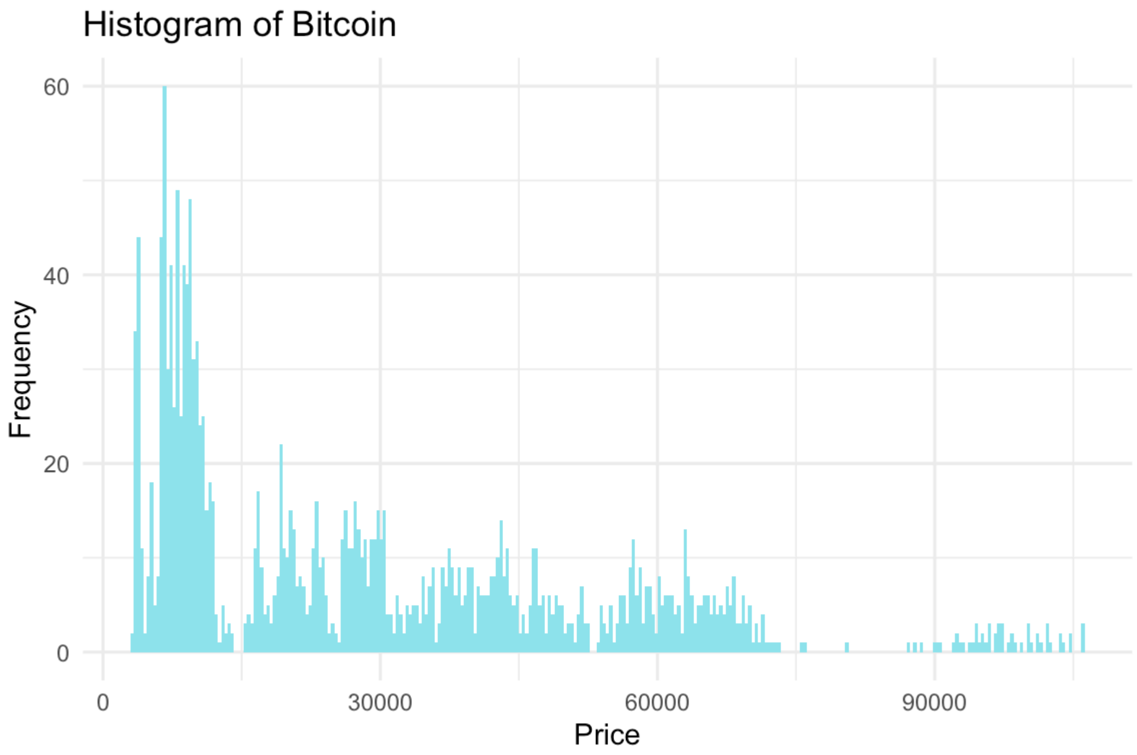

Figure 3.10: Histogram of Bitcoin prices

The histogram for Bitcoin already quite obviously represented the fact that it does not follow a normal distribution. It is completely skewed and asymmetric. The largest peak appears on the very lower tail and not in the middle. Also, the long tail to the right suggests a high value of upper outliers, which is not characteristical of a normal distribution.

```
# Making QQ plot for Bitcoin
qqnorm(df$Bitcoin)
qqline(df$Bitcoin, col = "blue")
```

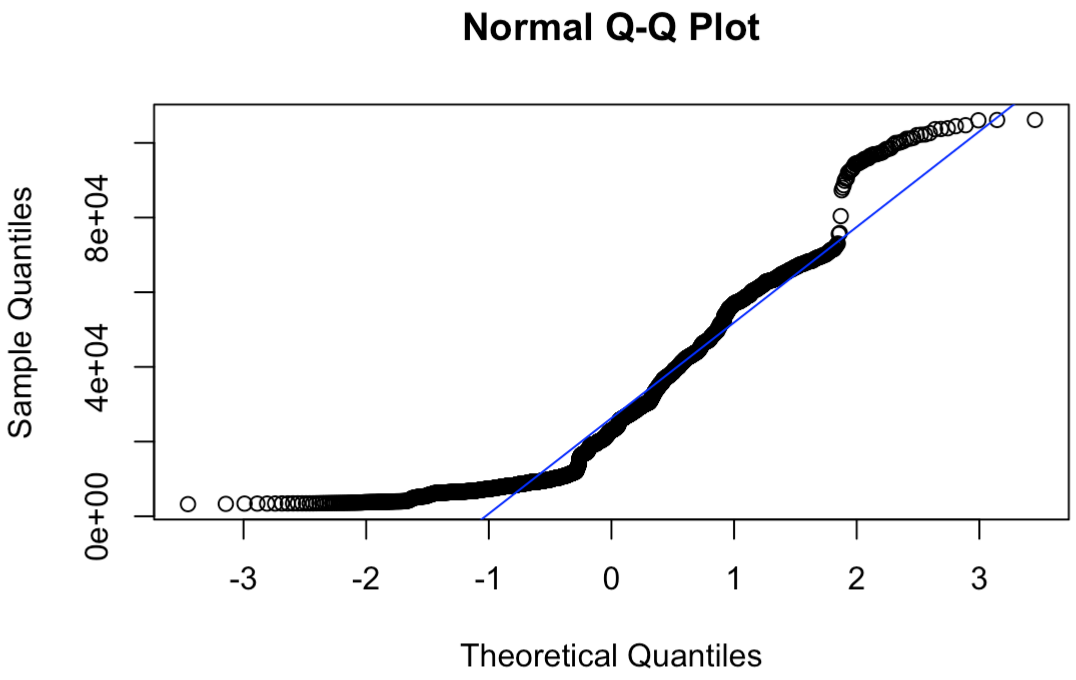

Figure 3.11: QQ plot of Bitcoin
A normal Q-Q plot would have values tightly adhering to the theoretical line in all sections of the data. Thus, this Q-Q plot indicates what was already shown through the histogram, there are no sections of the data that is clustered tightly around the line, especially the two tails.

3.6.2 Gold

```
# Making histogram for Gold
ggplot(df, aes(x = Gold)) +
  geom_histogram(fill="gold1",binwidth = 5) +
  labs(title = "Histogram of Gold", x = "Price", y = "Frequency") +
  theme_minimal()
```

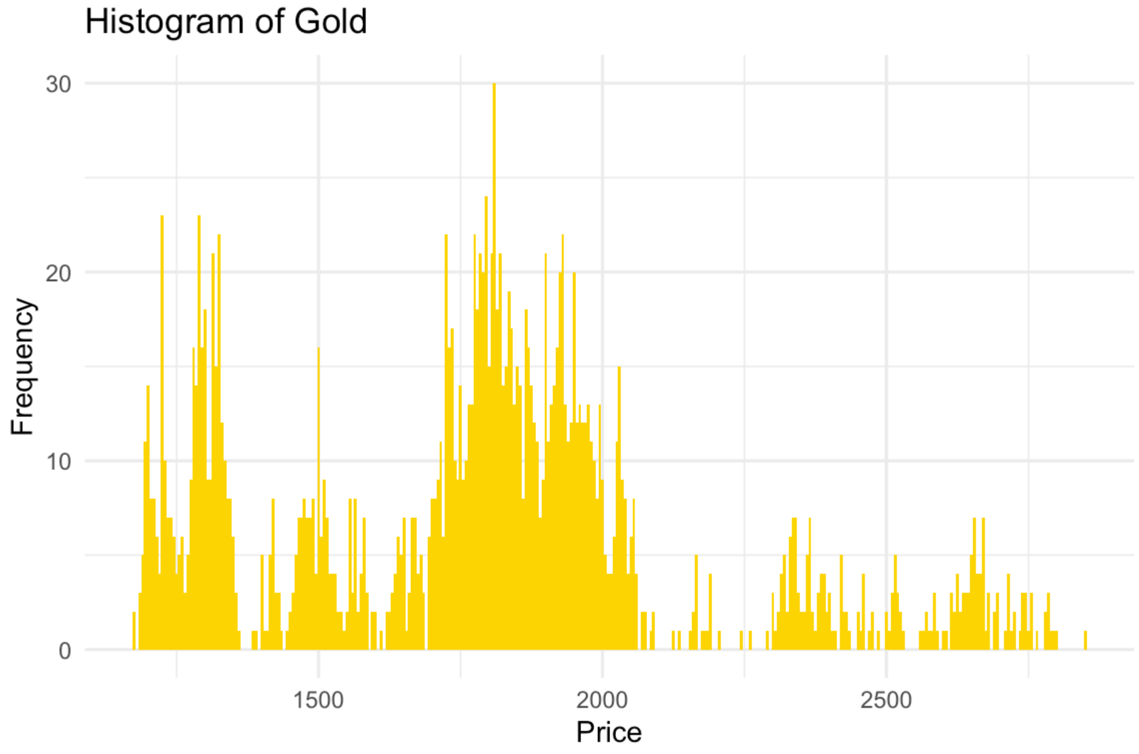

Figure 3.12: Histogram of Gold Prices

The distribution of Gold data more closely follows a normal distribution, which has the highest peak close to the middle of the data and better symmetry compared to Bitcoin. Despite that, it is multi-modal which multiple peaks, which already makes it apparent that it does not follow a normal distribution. Moreover, similarly to Bitcoin, the data is still largely right skewed, lacking needed symmetry with more data points on the lower end and a long right tail of extreme upper outliers.

```
# Making QQ plot for Gold
qqnorm(df$Gold)
qqline(df$Gold, col = "blue")
```

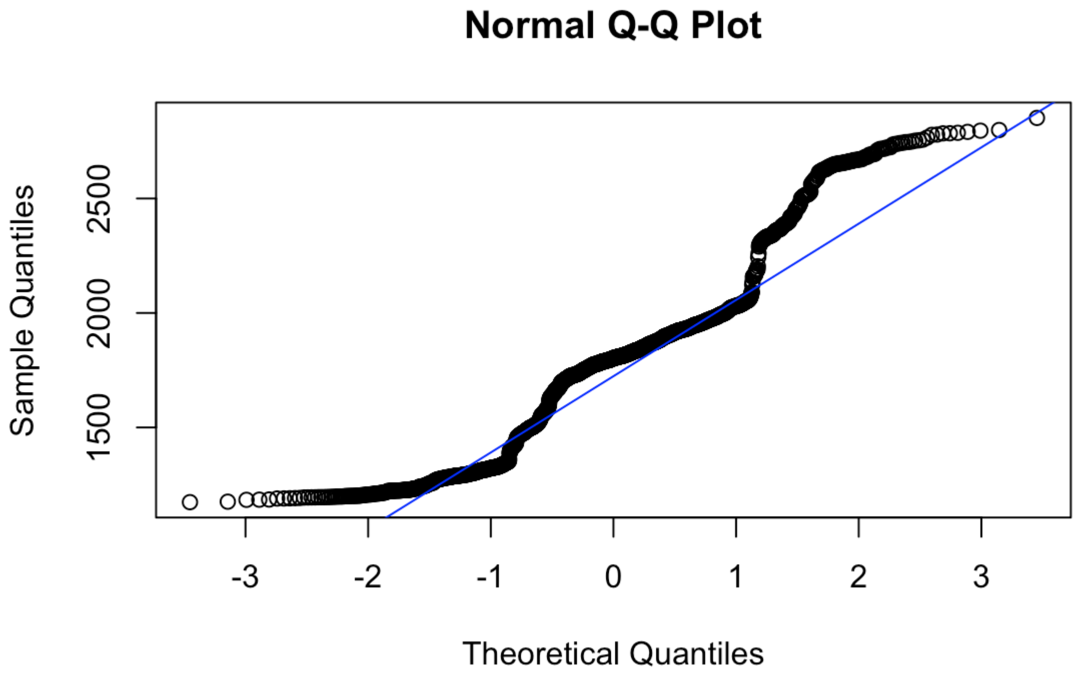

Figure 3.13: QQ plot of Gold

The Q-Q plot for gold corroborates previous information, as it very clearly still does not follow theoretical values. It is slightly more normally distributed than Bitcoin, but largely deviates from that once again especially at the head and tails of the data.

4 Conclusion

This report explored whether Bitcoin and gold share similar statistical behavior over the past seven years, focusing on trends, variability, correlation, and distribution.

Through descriptive statistics and boxplots, results show that Bitcoin is far more volatile than gold, with a wider price range and much greater variability. Both assets are right-skewed with upper outliers, but Bitcoin's distribution is more extreme and less symmetrical. Gold, by contrast, displays more stability and central tendency.

Correlation by period both with raw prices and returns suggest an increase in correlation over time. Although not by as much with returns. This indicates that the movement of the two assets is becoming more similar through time. Not calculating by period, a strong correlation (r = 0.83) was found using raw prices. However, this relationship nearly vanished (r = 0.12) when analyzing returns. This suggests that the apparent connection between the two is more reflective of shared market trends than a true relationship.

Finally, through both histograms and Q-Q plots, neither dataset followed a normal distribution. Bitcoin's returns are especially skewed, while gold is somewhat closer but still deviates from normality.

Overall, despite some perceived similarities, Bitcoin and gold behave differently in key statistical ways. Bitcoin is highly more risky and unstable compared to traditional gold, which hints that it is not a reliable investment choice. The two assets should not be assumed to serve the same role in investment strategy without deeper analysis.

5 References

Kamalov F, Gurrib I and Rajab K (2021) ‘Financial Forecasting with Machine Learning: Price Vs Return’, Journal of Computer Science, 17(3): 251-264, doi:10.3844/jcssp.2021.251.264 (doi:10.3844/jcssp.2021.251.264).

Takinsoy J (2021) ‘Bitcoin: A New Digital Gold Standard in the 21st Century?’, doi:10.2139/ssrn.3941857 (doi:10.2139/ssrn.3941857).