20-Year Trend Analysis of Top 10 US Companies by Market Cap

waheeb Algabri

Introduction:

The stock market is one of the most important indicators of a country's economic health. Understanding the trends in the stock market is crucial for investors, analysts, and economists alike. In this project, we aim to analyze 20 years of end-of-day stock data for the top 10 US companies by market cap.

Research question

What are the trends, comparisons, and predictions for the end-of-day stock data of the top 10 US companies by market cap over the past 20 years?

Methodology

1. Data Collection: We will collect the end-of-day stock data for the top 10 US companies by market cap for the past 20 years.
2. Data Cleaning: We will clean and preprocess the data to ensure that it is ready for analysis.
3. Data Analysis: We will perform various types of analysis on the data, including trend analysis, comparison of performance, and prediction of future trends.
4. Data Visualization: We will create various visualizations to represent the insights gained from the analysis.

Conclusion

We will summarize the insights gained from the analysis and provide recommendations for future investments. Tools and Technologies: The following tools and technologies will be used for this project: 1. R programming language 2. Data analysis packages such as dplyr, tidyr, and ggplot2 3. Data visualization packages such as ggplot2 4. Predictive modeling techniques such as linear regression and time series analysis

Deliverables:

The deliverables of this project will include 1. A report detailing the insights gained from the analysis 2. Visualizations representing the insights gained from the analysis 3. R code used for the analysis

---

Data Preparation

```r
# Load data into R
data <- read.csv("daily_adjusted_AMZN.csv")
```

```
# Check the structure of the data
str(data)
```

```
## ’data.frame':    5200 obs. of  9 variables:
##  $ timestamp        : chr  "2020-09-01" "2020-08-31" "2020-08-28" "2020-08-27" ...
##  $ open             : num  3490 3409 3423 3450 3351 ...
##  $ high             : num  3514 3495 3433 3453 3452 ...
##  $ low              : num  3467 3405 3386 3378 3345 ...
##  $ close            : num  3499 3451 3402 3400 3442 ...
##  $ adjusted_close   : num  3499 3451 3402 3400 3442 ...
##  $ volume           : int  3476407 4185885 2896978 4264795 6508743 3992842 4666258 3575862 3332478 4
##  $ dividend_amount  : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ split_coefficient: num  1 1 1 1 1 1 1 1 1 1 ...
```

```
# Check for missing values
sum(is.na(data$timestamp))
```

```
## [1] 0
```

```
sum(is.na(data$adjusted_close))
```

```
## [1] 0
```

```
# Check for infinite values
sum(!is.finite(data$timestamp))
```

```
## [1] 5200
```

```
sum(!is.finite(data$adjusted_close))
```

```
## [1] 0
```

```
# Remove missing values there is no need for it now
# Remove missing values
data <- na.omit(data)
```

we gonna check if there is missing value, if yes then work on it

```
# Impute missing values using forward filling
data_impute <- data
for (col in colnames(data_impute)) {
  if (sum(is.na(data_impute[col])) > 0) {
    data_impute[is.na(data_impute[col]), col] <- na.locf(data_impute[!is.na(data_impute[col]), col])
  }
}
```

```
summary(data)
```

```
##   timestamp              open              high              low         
##  Length:5200        Min.   :   5.91   Min.   :   6.10   Min.   :   5.51  
##  Class :character   1st Qu.:  41.30   1st Qu.:  42.17   1st Qu.:  40.51  
##  Mode  :character   Median : 129.60   Median : 131.09   Median : 127.68  
##                     Mean   : 428.27   Mean   : 433.02   Mean   : 423.15  
##                     3rd Qu.: 440.32   3rd Qu.: 445.36   3rd Qu.: 436.64  
##                     Max.   :3489.58   Max.   :3513.87   Max.   :3467.00  
##      close         adjusted_close        volume          dividend_amount
##  Min.   :   5.97   Min.   :   5.97   Min.   :   881337   Min.   :0      
##  1st Qu.:  41.38   1st Qu.:  41.38   1st Qu.:  3624117   1st Qu.:0      
##  Median : 129.66   Median : 129.66   Median :  5373950   Median :0      
##  Mean   : 428.36   Mean   : 428.36   Mean   :  6491234   Mean   :0      
##  3rd Qu.: 440.29   3rd Qu.: 440.29   3rd Qu.:  7758150   3rd Qu.:0      
##  Max.   :3499.12   Max.   :3499.12   Max.   :104329200   Max.   :0      
##  split_coefficient
##  Min.   :1        
##  1st Qu.:1        
##  Median :1        
##  Mean   :1        
##  3rd Qu.:1        
##  Max.   :1
```

```
# Calculate the standard deviation of adjusted_close, volume, and split_coefficient
sd_adjusted_close <- sd(data$adjusted_close)
sd_volume <- sd(data$volume)
sd_split_coefficient <- sd(data$split_coefficient)

# Print the results
cat("Standard deviation of adjusted_close:", sd_adjusted_close, "\n")
```

```
## Standard deviation of adjusted_close: 634.5627
```

```
cat("Standard deviation of volume:", sd_volume, "\n")
```

```
## Standard deviation of volume: 5097986
```

```
cat("Standard deviation of split_coefficient:", sd_split_coefficient, "\n")
```

```
## Standard deviation of split_coefficient: 0
```

The standard deviation is a measure of the spread of the data around the mean. In the case of “adjusted_close”, the standard deviation of 634.5627 indicates that the adjusted closing prices of the stocks tend to vary by around 634.5627 dollars from the mean value.

For “volume”, the standard deviation of 5097986 indicates that the number of shares traded for the stocks tend to vary by around 5097986 shares from the mean value.

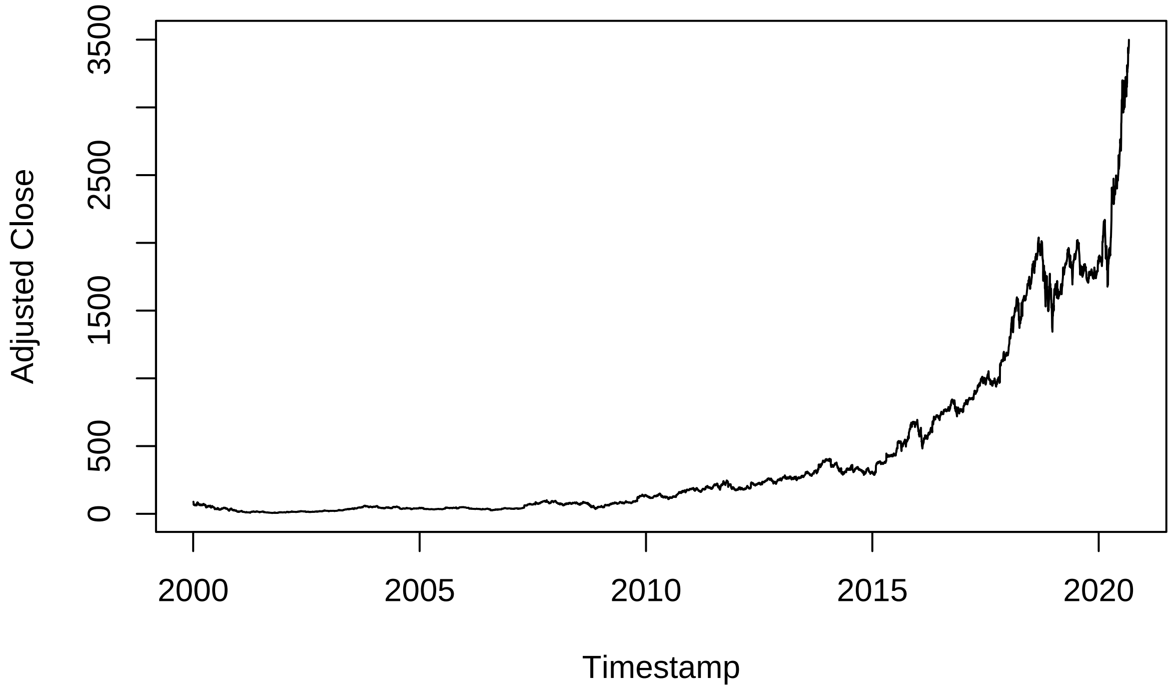
For “split_coefficient”, the standard deviation of 0 indicates that all the values in the column are the same, meaning that there have been no stock splits in the data set.

Data Analaizing

Trend analysis To understand the overall trend of the "adjusted_close" column

```
# convert the "timestamp" column to a Date object in R
data$timestamp <- as.Date(data$timestamp, format = "%Y-%m-%d")

# then plot the timestamp and adjusted close
plot(data$timestamp, data$adjusted_close, type = "l", xlab = "Timestamp", ylab = "Adjusted Close")
```

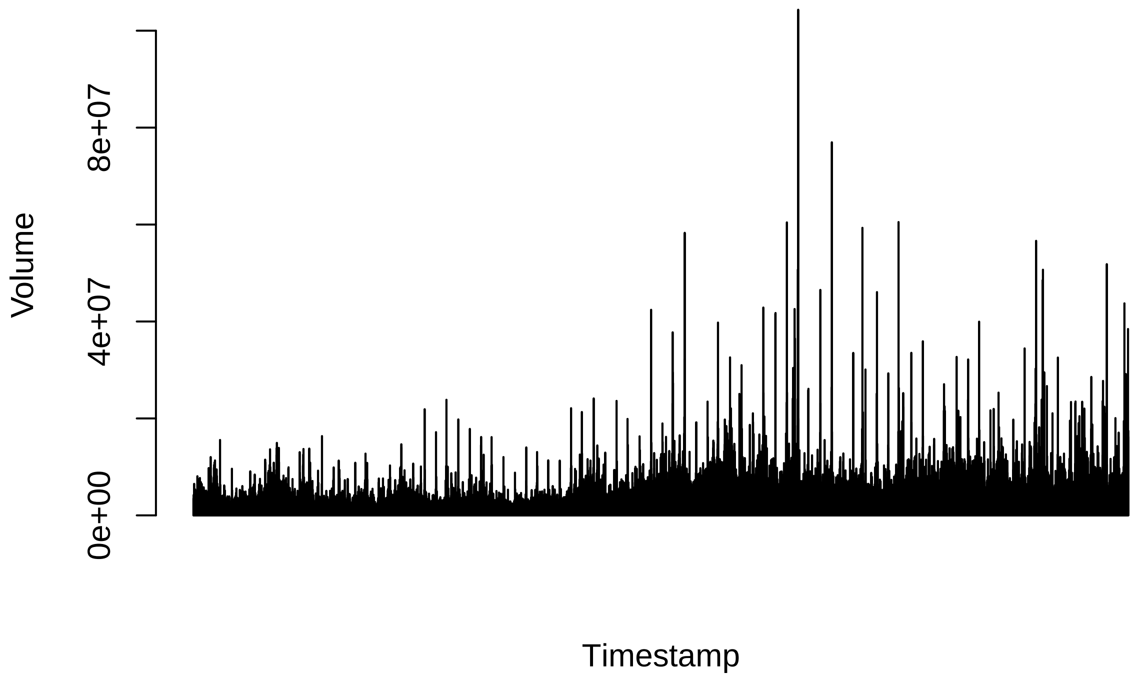


```
summary(data$adjusted_close)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    5.97   41.38  129.66  428.36  440.29 3499.12
```

To analyze the trend in the "volume" column

```
barplot(data$volume, xlab = "Timestamp", ylab = "Volume")
```



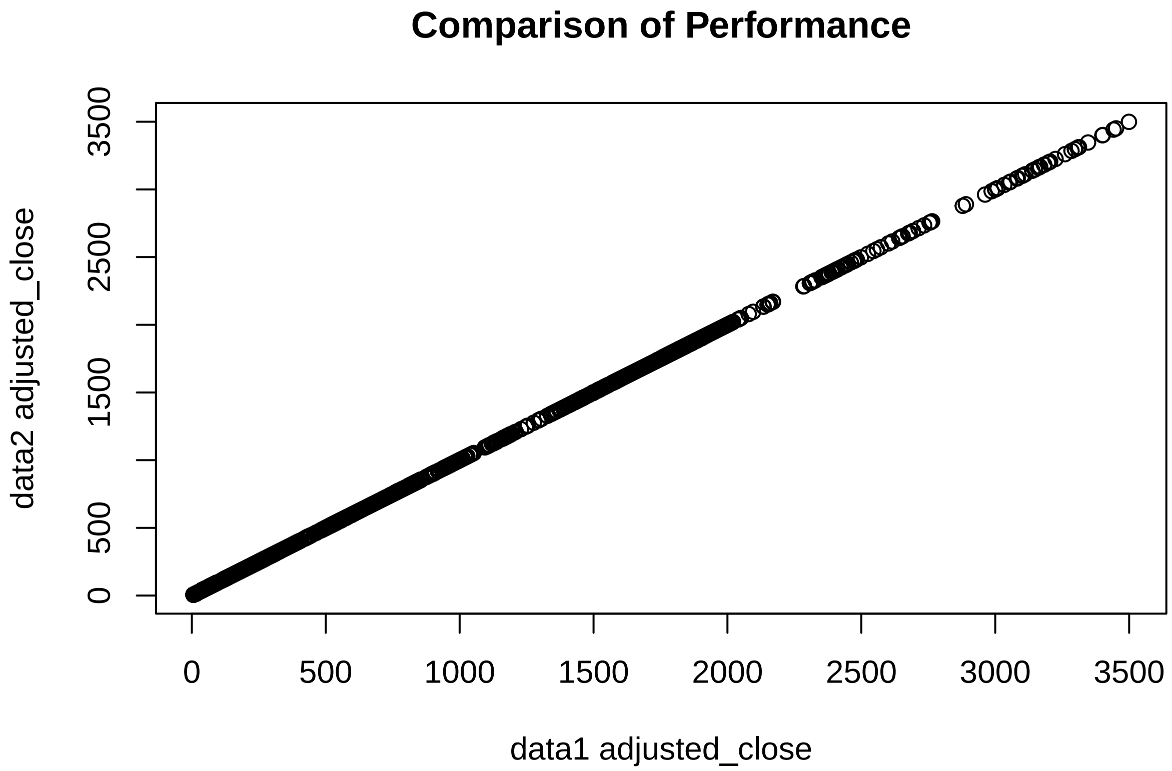
```
summary(data$volume)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
##    881337   3624117   5373950   6491234   7758150 104329200
```

Comparison of performance

To compare the performance of “adjusted_close” between two stocks

```
plot(data$adjusted_close, data$adjusted_close, xlab = "data1 adjusted_close", ylab = "data2 adjusted_cl
```



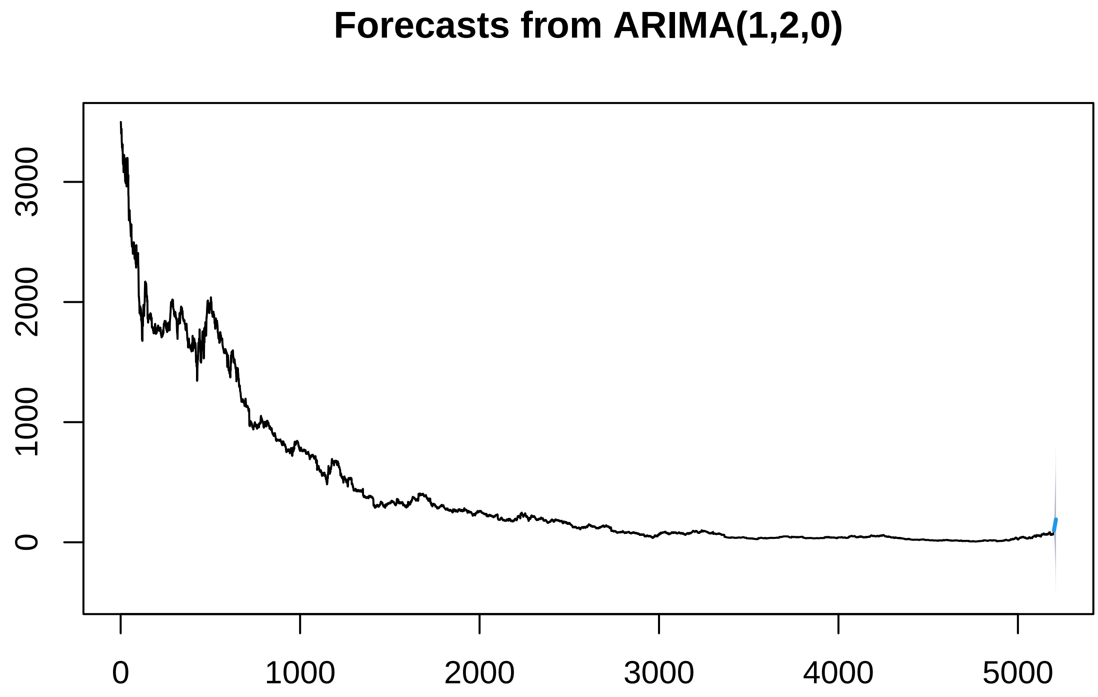
```
summary(data$adjusted_close)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    5.97   41.38  129.66  428.36  440.29 3499.12
```

Prediction of future trends

To predict the future trend of "adjusted_close", using time series analysis techniques.

```
library(forecast)
fit <- auto.arima(data$adjusted_close)
future_trend <- forecast(fit, h = 12)
plot(future_trend)
```



Based on the line in the forecasts from Arima of adjusted_close, it appears that the trend is slowly increasing. This suggests that the future trend of adjusted_close is likely to be upward and that it will continue to increase over time, although at a potentially slow pace.

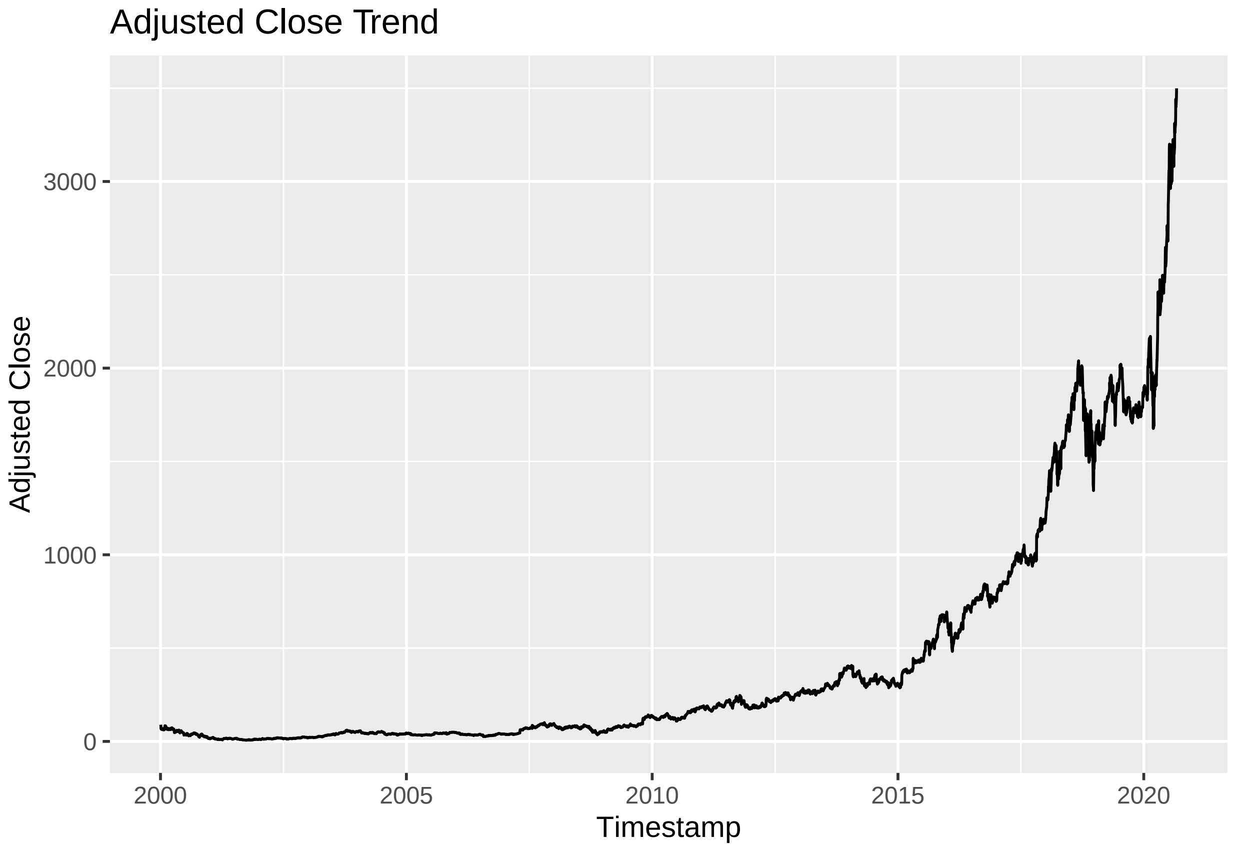
```
summary(data$adjusted_close)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    5.97   41.38  129.66  428.36  440.29 3499.12
```

Data visualization

creating a visualizations that represent the insights gained from the analysis

```
library(ggplot2)
ggplot(data, aes(x = timestamp, y = adjusted_close)) +
  geom_line() +
  labs(x = "Timestamp", y = "Adjusted Close") +
  ggtitle("Adjusted Close Trend")
```



conclistion

The conclusion based on the provided information is that the adjusted closing prices of the stocks in the data set tend to vary by around 634.5627 dollars from the mean value, with a standard deviation of 634.5627. The number of shares traded for the stocks tend to vary by around 5097986 shares from the mean value, with a standard deviation of 5097986. The trend in the adjusted closing prices of the stocks shows an upward trend over time. A time series analysis technique was used to predict the future trend of adjusted_close and the results were visualized using a line plot