Standard and Poor's 500 or

S&P 500 Stock Data Analysis using Statistical Analysis

DATA 602: Statistical Data Analysis

Group Project Report

Prepared By:

Andrii Voitkiv, 30199373

Prashant Mittal, 30192139

Raj Bhanvadia, 30157827

Table of Contents

Table of Figures

Introduction:

Nowadays, it's crucial to manage your finances or invest your money in a way that will help you be ready to combat inflation. Because if your money is not invested appropriately, inflation will reduce the value of your hard-earned dollars. The stock market is thus one option for investing your money. However, stock market investing calls for in-depth market understanding as well as the capital necessary to turn a profit. As of 2022, there are roughly 138 big stock exchanges around the world(List of Stock Markets, 2022).

We chose to analyze the Standards & Poor's 500 market index. The S&P 500 is a market index that tracks the progress of the top 500 performers across a range of industries. In this project, we'll look at a variety of tactics that a novice retail investor might employ to succeed in the stock market.

Guiding Questions:

Question 1:

How random are markets? How does the distribution of market returns compare to the normal distribution? So we framed our null hypothesis as follows,

- H(0): The sample comes from a normal distribution
- H(a): The sample comes from a non-normal distribution.

This analysis will help investors to set realistic expectations and subsequent risk management strategies. The S&P index for a given period will serve as our population.

Question 2:

How two different companies/populations from the S&P 500 index and the same sector behave in the market?

- The implication is to make the market an *efficient* place for investors by executing arbitrage strategies.
- In addition, eliminating similar stocks from the portfolio increases its *diversification*.

Two stocks from the S&P 500 list will act as our population.

Variables are Log Returns that are observed from stock data in terms of their recorded closed prices every day and subsequent changes.

Dataset Description:

It is about S&P 500 stock market data retrieved from Kaggle and collected from FRED and yfinance. It's a daily updated dataset that started from December 2009 but we have extracted it till the end of September 2022. There are 3 tables(Companies, Index, and Stocks) in a structured table format having 493 unique rows and 26 columns. Moreover, it has categorical columns including decimal, string, integer, and others. Here is the source: Remote source: https://www.kaggle.com/datasets/andrewmvd/sp-500-stocks.
It's an open source data available on Kaggle contributed by Larxel and under the license of CC0: Public Domain.

Data Analysis:

Question 1: Exploratory Task Analysis

Hypothesis:

H(0): The sample comes from a normal distribution.
H(a): The sample comes from a non-normal distribution.

```
# Read SP500 index data
df_index <- read.csv("/Users/berg/Projects/r projects/602_data/sp500_index_602_v2.csv")
# Peek at data
print(colSums(is.na(df_index)))
```

Figure 1 Data Load

Date	close	year	month	dom	day_name	month_name	dow	eom	days_to_eom	wom	returns	cum_returns
2012-09-11	1433.56	2012	9	11	Tuesday	September	1	2012-09-30	-19	2	0.003129980478	0.0031299805
2012-09-12	1436.56	2012	9	12	Wednesday	September	2	2012-09-30	-18	2	0.002090505688	0.0052204862
2012-09-13	1459.99	2012	9	13	Thursday	September	3	2012-09-30	-17	2	0.016178219630	0.0213987058

Figure 2 Data Display

In the above image, we load the data of the S&P 500 index which has various attributes includes such as Date, close, year, month, dom(day of month), day_name, month,_name dow(day of

week), eom(end of month), days_to_eom, wom(week of month), returns cum_returns, returns0.

We did data wrangling by using python and created a CSV for this project.

After this, working on descriptive statistics analysis using various plots can be seen below:

S&P 500 Time Series vs Daily Log Returns

By using ggplot, we are plotting time series for the period from 2012 to 2022.

In the below graph, the time series shows the index *price* over a period of time (2012-2022), and the ***uptrend*** can be observed.

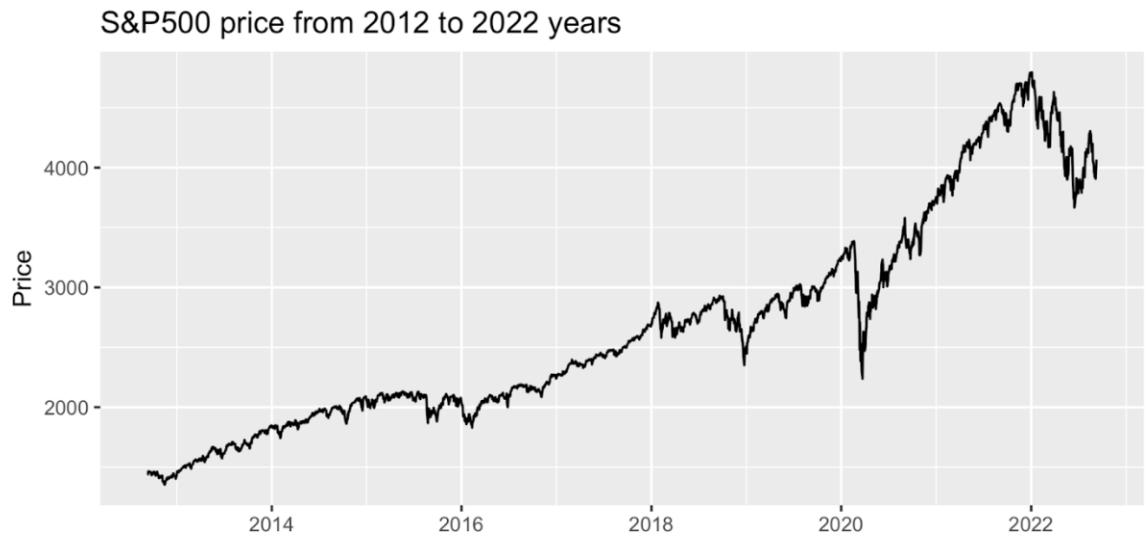

Figure 3 Time series for the period from 2012 to 2022

Now, in the below graph, we are plotting *daily log returns* over the same time period and it shows the volatility clustering phenomenon(Gonzalez-Rivera, et al., 2012). Also, can be seen high volatility in Feb-Mar 2020 due to the pandemic investors feared an economic slowdown.

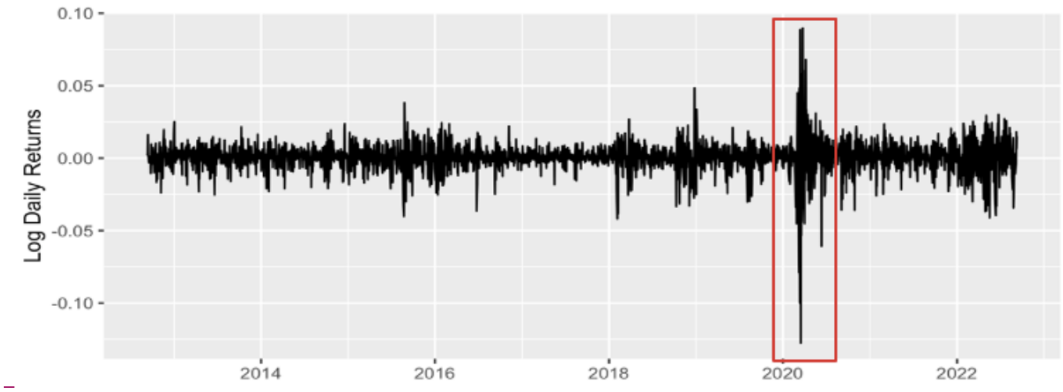

Figure 4 Daily Log returns for the period from 2012 to 2022

There are a number of reasons why not to work with the original variables which are prices, but rather with transformations of the variables such as logs, square roots, or other power transformations.

When the data are normally distributed, or at least symmetrically distributed, and have a constant variance, many statistical methods perform well.

The modified data frequently show less skewness and a more constant variable than the original variables. The log transformation of prices (ln(p1) - ln(p0) = ln(p1/p0)) is used in our analysis.

```
favstats(~returns, data = df_index)
```

```
##         min           Q1       median           Q3        max         mean
##  -0.1276521 -0.003507972 0.0006677019 0.005270995 0.08968316 0.0004157247
##          sd    n missing
##  0.0108538 2516       0
```

Figure 5 Favstats Function

Here, we have used the favstats() function to get all the statistics values such as median, mean, standard deviation, and so on. From this analysis, we have mean = 0.0004157, SD = 0.0108538. Also, values for min and max are -12.76% and 8.97% daily changes respectively. Even though the absolute value of min daily return is higher than the maximum daily return meaning some *extreme negative events*, the overall *trend is positive* because the mean is above zero and Q3 is higher than the absolute value of Q1.

Despite the large random fluctuations in SP500 index returns, we can see that series appears stationary, meaning that the nature of its random variation is constant over time. In particular, the series fluctuates about zero mean or nearly so.

S&P 500 Daily Returns by Month and Day of Week

In this part, we are plotting boxplots and violin plots integrating box plots inside which can be seen below:

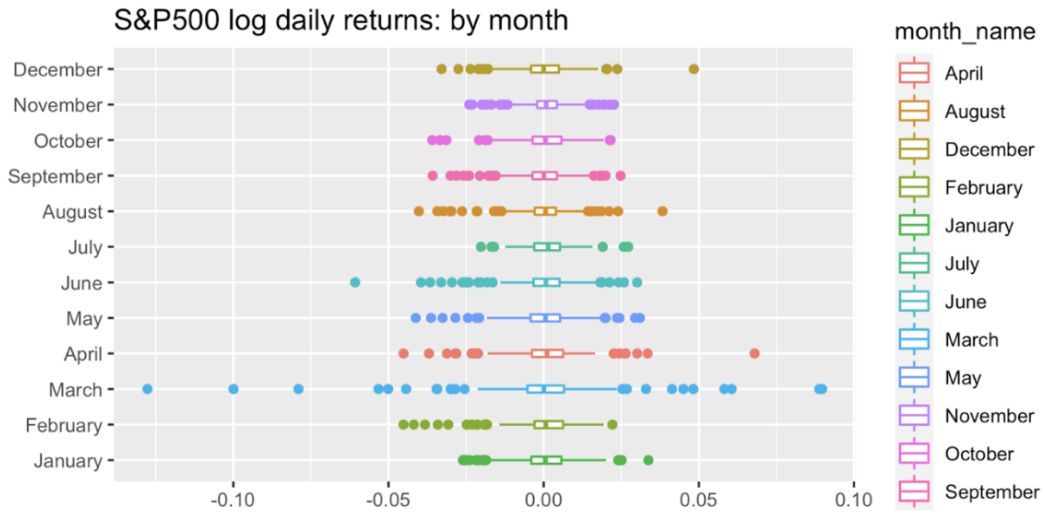

Figure 6 S&P 500 Daily Returns by Month

From the box plot, we are plotting log daily returns over the *months* can be seen in the above figure. Here, we can infer that March is a very turbulent month and July, November, and December are quiet months due to the holiday season.

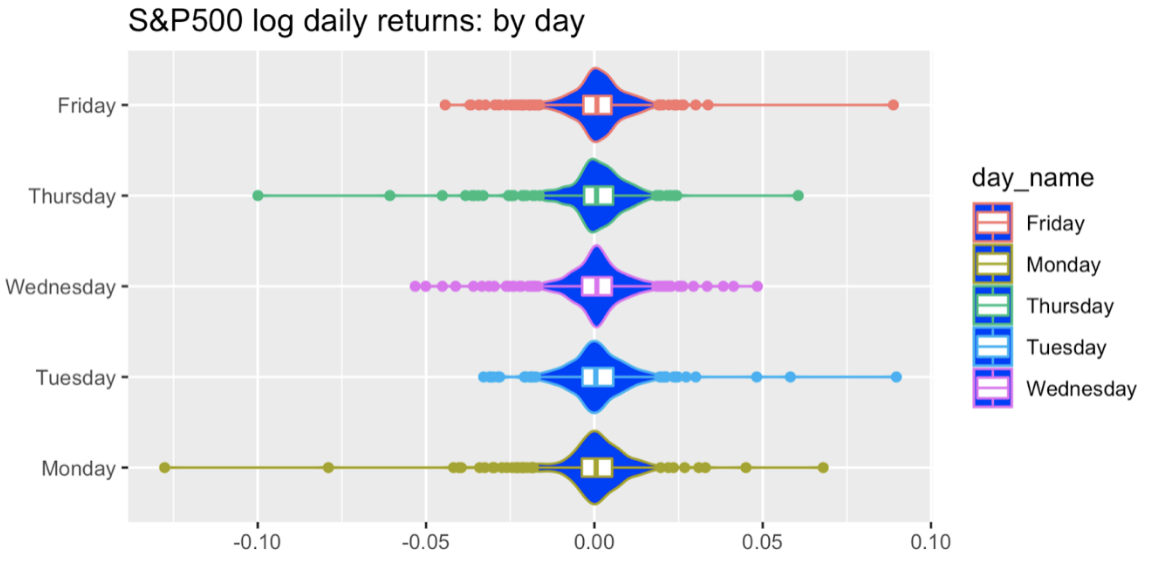

Figure 7 S&P 500 Daily Returns by Days

From the violin plot, we are plotting log daily returns over the *days of the week* can be seen in the above figure. Here, we can infer that Mondays have a negative tail, probably digesting negative weekend news. Mid of the week (Wednesdays) is quiet, and Thursdays are volatile due to the heavy economic calendar agenda.

Kernel Density Estimates:

Kernel density estimates of the daily log returns on the S&P 500 index.

```
e = ggplot(df_index, aes(x=returns)) +
  geom_histogram(binwidth=0.005) +
  ggtitle("Histogram")

f = ggplot(df_index, aes(x=returns)) +
  geom_density(color ="black", fill="steelblue", linetype = "dashed") +
  ggtitle("Density plot")

plot_grid(e, f, align="v", ncol = 2, nrow = 1)
```

Here is the snippet for plotting the Histogram and Density plot based on returns.

From this graph, we can determine the behaviour of the plot is *leptokurtic* means frequency distribution or its graphical representation having *greater kurtosis than the normal distribution*. Also, returns are more stacked around the mean that hints on the random nature of markets. These characteristics of the distribution allow analysts and investors to make statistical inferences about the expected return and risk of stocks.

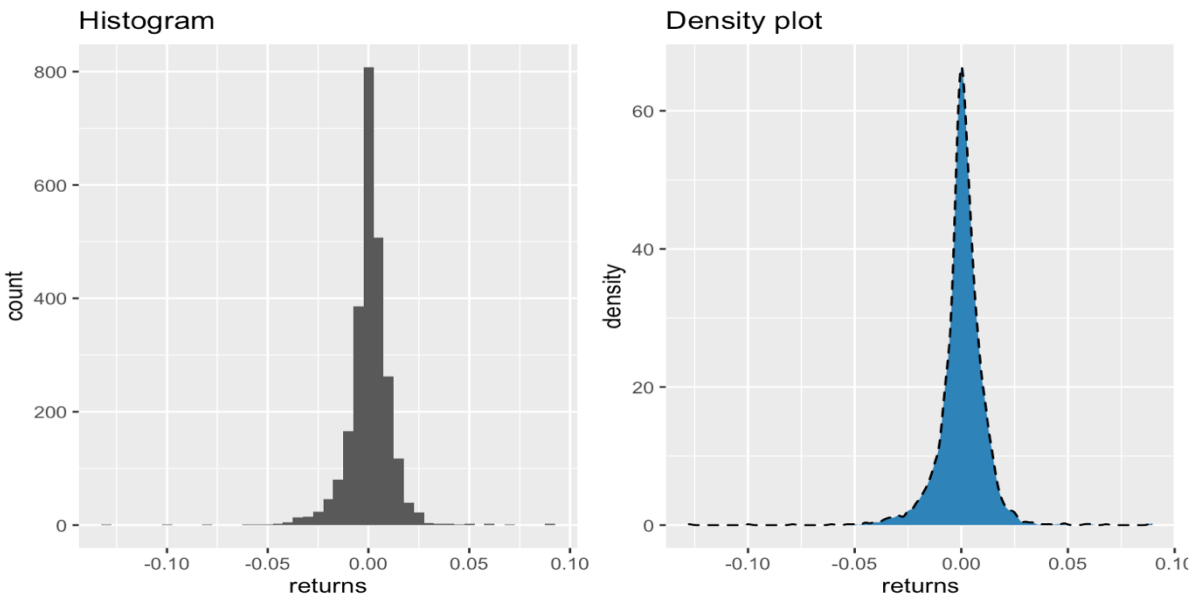

Figure 8 Histogram and Density plot based on Returns

```
g = ggplot(df_index, aes(x = returns, y = year, fill=year)) +
  geom_density_ridges() +
  ggtitle("S&P500 returns density plot: by year")

plot_grid(g, align="v", ncol = 1, nrow = 1)
```

We plotted the Density Estimates Curve based on Daily Log Returns and Yearly for each year to illustrate the distribution of the returns values for the S&P 500 during the duration of the

dataset, which is 2009-2022 can be seen in the below graph. The below graph shows a density plot from which we can see that there is skewness in the distribution of returns for all the years in the time period and for the year 2015, the distribution seems bimodal.

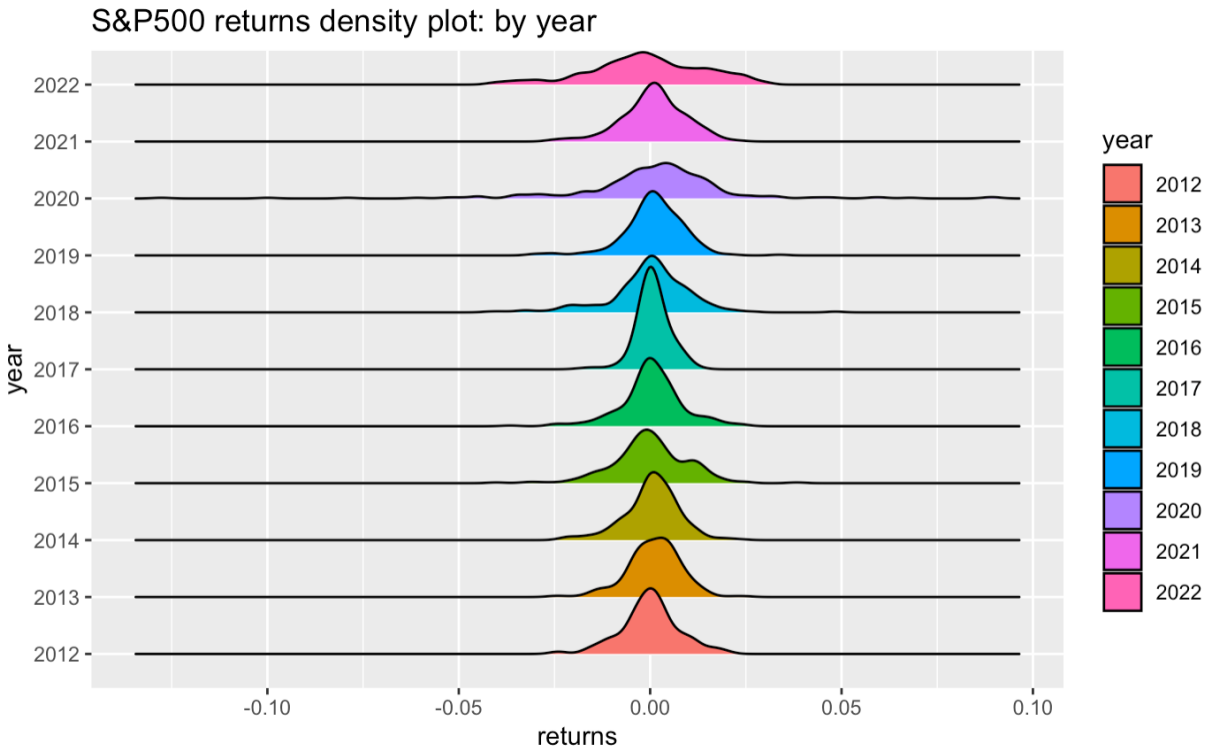

Figure 9 Density Plot by Year

Normality Curve over the Histogram of Returns:

```
h = ggplot(df_index, aes(x =returns)) +
  geom_histogram(aes(y =..density..), colour = "black", fill = "white", binwidth = 0.005) +
  stat_function(fun = dnorm, args = list(mean = mean(df_index$returns), sd = sd(df_index$returns))) +
  ggtitle("Normal curve over histogram of S&P500 returns")

plot_grid(h, ncol = 1, nrow = 1)
```

Using ggplot to plot the Normal curve over the histogram of returns can be seen in the below graph. This is done to check how the returns distribution compares to the normal distribution. For stock returns, the standard deviation is often called volatility. As we can see returns are not normally distributed because the left tail is longer than the right tail which means more negative events.

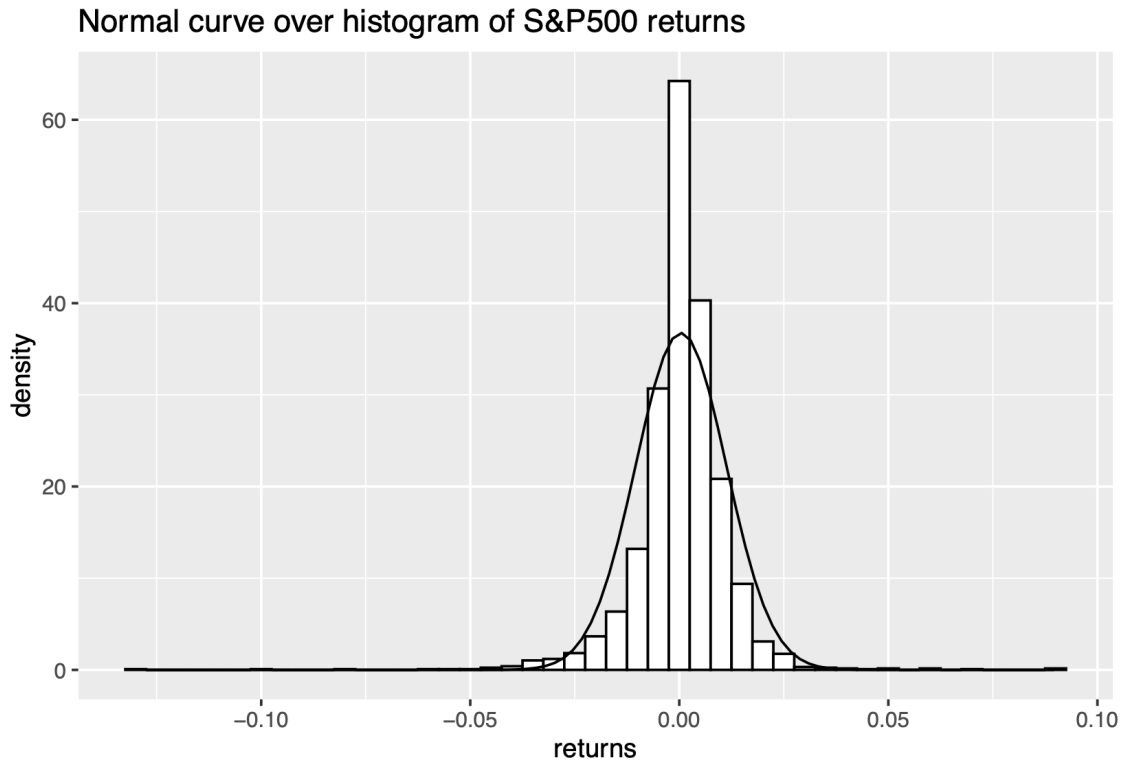

Figure 10 Normality Curve over the Histogram of Returns

Normality Test and Single Sample t-test:

Visual inspection is usually unreliable. It's possible to use a **significance test** comparing the sample distribution to a normal one in order to ascertain whether data show or not a serious deviation from normality before we can proceed with our statistical analysis.

There are several methods for **normality tests** such as **Kolmogorov-Smirnov (K-S) normality test, Jarque-Bera test,** and **Shapiro-Wilk's test**. We opted for the Shapiro-Wilks test for normality.

Shapiro–Wilk Test:

It is difficult to judge the Normality of returns when viewing a normal probability plot, due to sampling variation, so a statistical test of normality is useful. The null hypothesis (H0) is that the sample comes from a normal distribution and the alternative (Ha) is that the sample is from a non-normal distribution.

The Shapiro–Wilk test uses the normal probability plot to test these hypotheses. Specifically, the Shapiro–Wilk test is based on the correlation between i/n quantiles of the sample and the standard normal distribution, respectively. Under normality, the correlation should be close to

1 and the null hypothesis of normality is rejected for small values of the correlation coefficient. In R, the Shapiro–Wilk test can be implemented using the Shapiro.test function.

```
shapiro.test(df_index$returns)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df_index$returns
## W = 0.85991, p-value < 0.00000000000000022
```

If the test is significant, the distribution is non-normal. From the output, the p-value < 0.01 implies that the distribution of the data is *significantly different from a normal distribution*. In other words, we can not assume normality.

Now to find if the ***average of the returns is different from zero (statistically)***. This will help us prove or disprove our null hypothesis by performing a "t-test" as follows:

```
t.test(df_index$returns)
```

```
##
##  One Sample t-test
##
## data:  df_index$returns
## t = 1.9212, df = 2515, p-value = 0.05482
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.000008585553  0.000840034886
## sample estimates:
##    mean of x
## 0.0004157247
```

Here, we can observe that mean(x) = 0.0004 is different from zero not due to the randomness. Thus, we reject the null hypothesis based on P-value. In addition, we plotted the time series where we observed a strong uptrend during the period and the test's result is consistent with this plot.

Many statistical models assume that a random sample comes from a normal distribution. Normal probability plots are used to check this assumption, and, if the normality assumption seems false, to investigate how the distribution of the data differs from a normal distribution. Systematic deviation of the plot from a straight line is evidence of normality.

```
i = ggplot(df_index, aes(sample=returns)) +
  stat_qq(col="blue", distribution = qnorm) +
  stat_qqline(col="red", distribution = qnorm) +
  ggtitle("Normal Probability Plot of Log Daily Returns: SP500 Index")

plot_grid(i, ncol = 1, nrow = 1)
```

Here, we are plotting the Normal Probability Plot of the Log Daily Returns for the SP500 Index can be seen below. This graph shows that tails are heavier than normal which means more extreme events than normally expected.

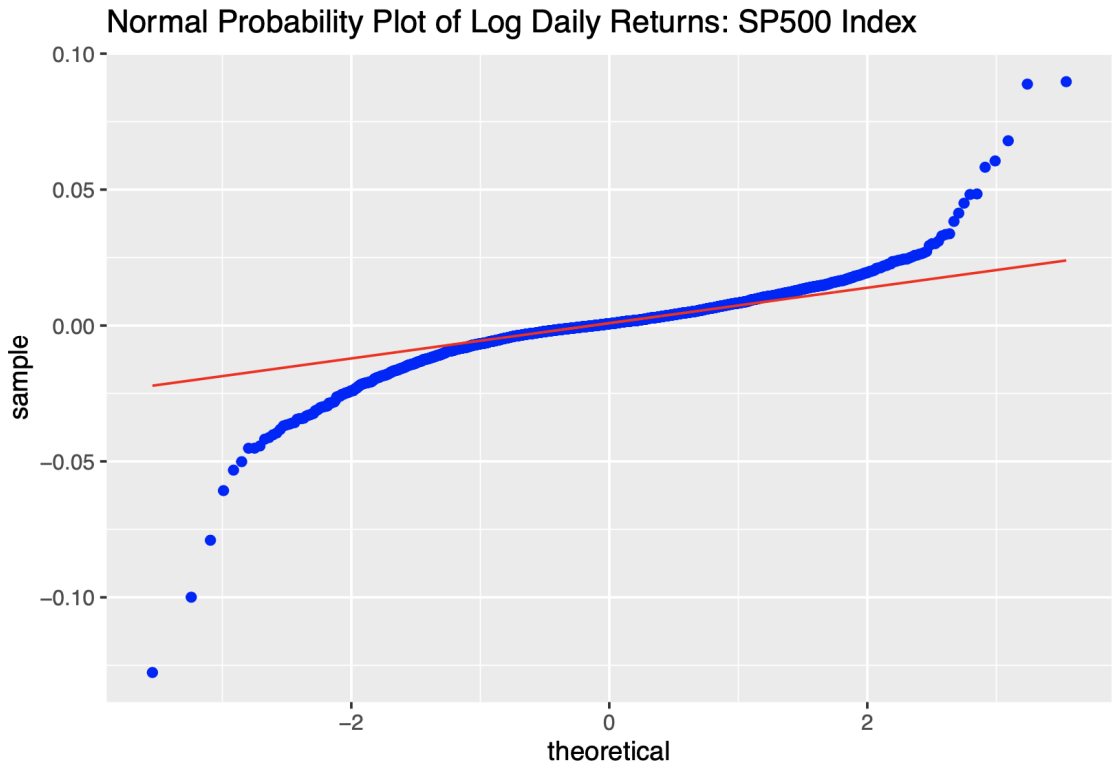

Figure 11 Normal Probability Plot of the Log Daily Returns

The curve is concave on the left and convex on the right. QQ-plot indicates, respectively, left skewness, right skewness, and heavy tails (compared to the normal distribution). By the tail of the distribution is meant the regions are far from the normal line which can be seen in the above graph.

The distribution is outlier-prone, meaning that the extreme observations on both the left and right sides are significantly more extreme than they would be for a normal distribution. It is a

general property of the t-distribution that the tails become heavier as the degrees-of-freedom parameter decreases and the distribution approaches the normal distribution as the degrees of freedom approach infinity. Heavy-tailed distributions with little or no skewness are common in finance and, the *t-distribution is a reasonable model for index returns*.

T-Plots:

We are now using the t-distribution to model the returns. Degree of freedom has longer tails on which data values can lie over the normal line and with an increase in the number of degree-of-freedom; decrement can be seen in the distribution of normality. But in our case, with df = 3, almost all data points lie over the normal line, and can we say that it shows normalization in distribution. Now, we plotted the qqplot of t-distribution over different degrees of freedom can be seen below. From the plot, we can infer that the t-distribution with 3 degrees of freedom fits much better than the normal distribution.

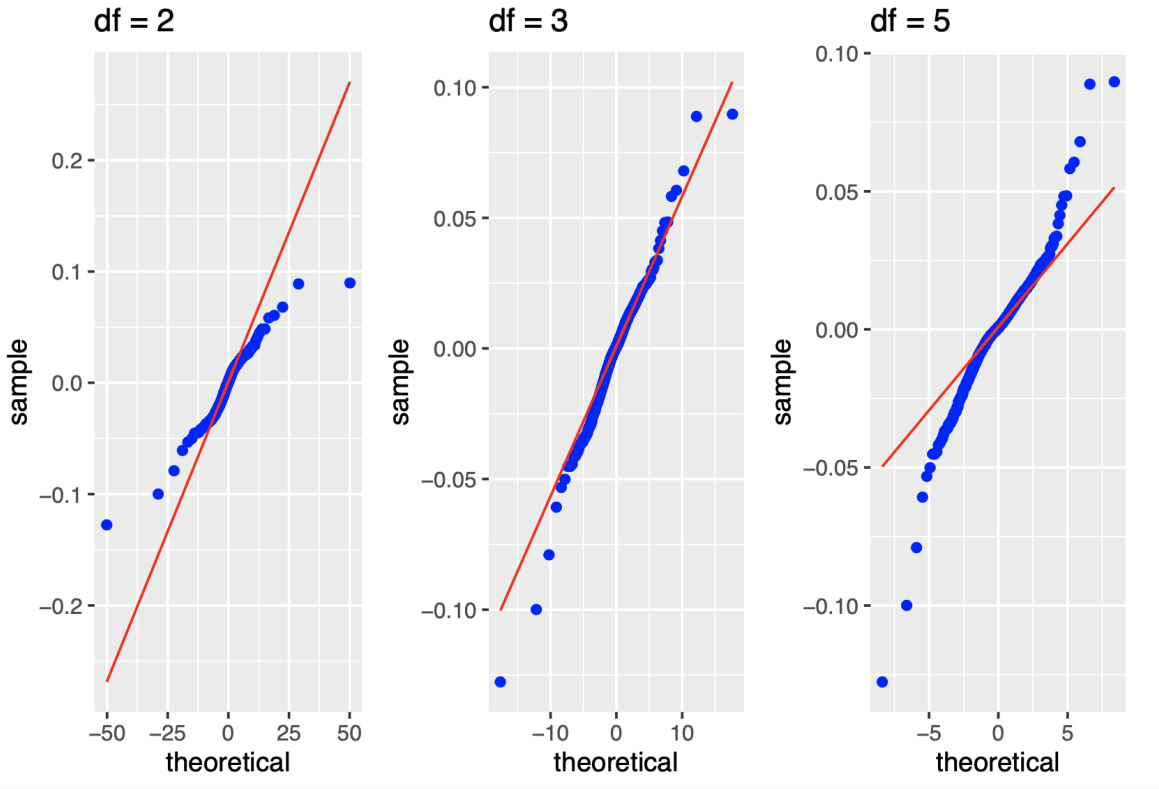

Figure 12 T-distribution to Model the Returns

It is worthwhile to keep in mind that the historical data have more extreme outliers than a t-distribution. There are two reasons why the t-model does not give a credible probability of a negative return as extreme as in March 2020. First, the t-model is symmetric, but the return

distribution appears to have some *skewness in the extreme left tail*, which makes extreme negative returns more likely than under the t-model. Second, the t-model assumes constant conditional volatility, but volatility changes and was high in March 2020 (see plot returns).

Question 1: Conclusion

For this guiding question, we have observed a P-value significantly lower than 0.05 by using the Shapiro test that ***rejects normality*.** In addition to this, the histogram and density plots don't follow a normal curve; the distribution represents a leptokurtic shape with greater kurtosis than normal distribution and also fatter tails: negative values on the left tail more than on the right tail - asymmetrical. The *mean of returns* equal to 0.0004 is **statistically different from 0** proved by the one sample t-test. And the time series plot shows the uptrend to confirm this. Moreover, the family of **t-distributions models the returns** much better by counting for fat tails which is standard for financial data.

Question 2: Statistical Analysis

In this question, we are going to use linear regression and two population testing to find out: How do two different companies/populations from the S&P 500 index and the same sector behave in the market?

The implication is to make the market an efficient place for investors by executing arbitrage strategies.

```
# Read SP500 index data
df_stock_returns <- read.csv("/Users/berg/Projects/r projects/602_data/sp500_stocks_returns_602.csv")
df_stock_cum_returns <- read.csv("/Users/berg/Projects/r projects/602_data/sp500_stocks_cumreturns_602.
```

Date	A	AAL	AAP	AAPL
2009-12-31	*NA*	*NA*	*NA*	*NA*
2010-01-04	*NA*	*NA*	*NA*	*NA*
2010-01-05	-0.010921690	0.10724564	-0.005961516	0.001727402
2010-01-06	-0.014481528	0.06493182	0.002720492	-0.014306505

The above snippet of code is reading the data files in which data wrangling is done in python.

Visual Representations:

Time series of Pairs of companies from different Sectors:

Based on cumulative returns for some of the stock and looking for our two populations depicted as:

The below graph shows the cumulative returns over the years from 2010 to 2020 with an interval of 5 years. Also, we can see the pairs from various industries including Food and Beverages, Gas & Oils, Technology, and Financial Services. From this visualization, we can clearly determine that *stocks from the same sector share a strong bond* between them whether it be PEP & KO from Food and Beverage or MSFT & AAPL from the Technology sector (Hofert, et al., 2018).

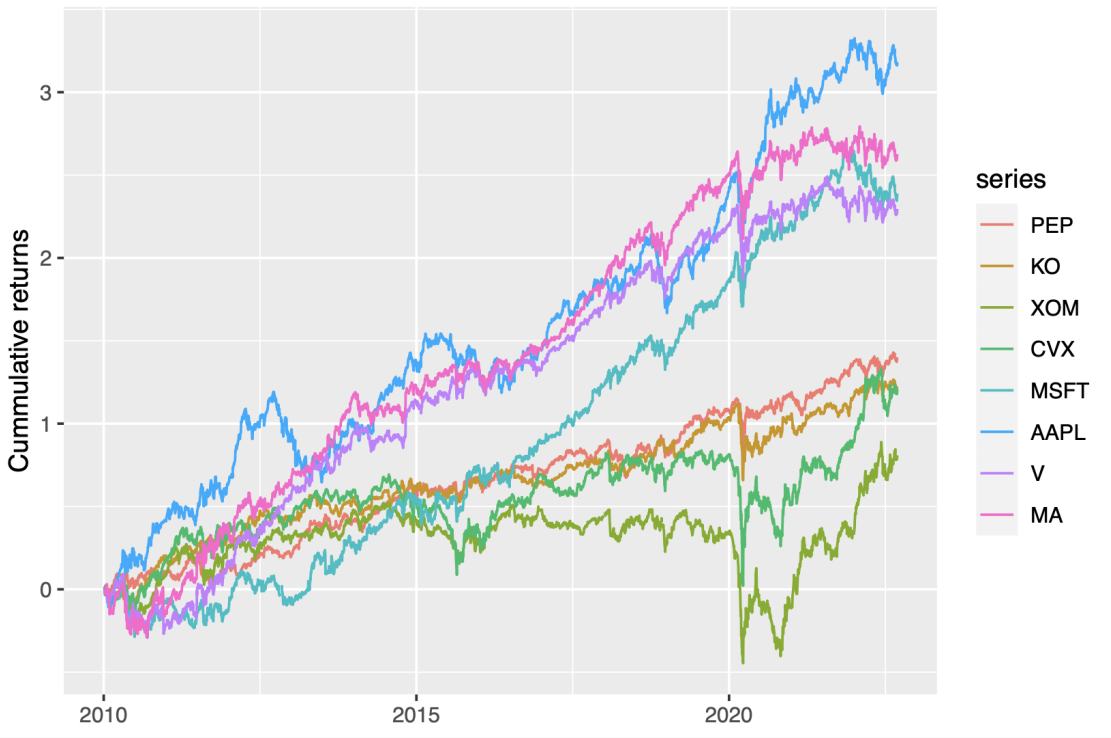

Figure 13 Time series of Pairs of companies from different Sectors

Next, we want to check how correlated these returns for the stocks in the above graphs are on a daily basis. To analyse this, we use the correlation function of R to find out all the possible *pairwise correlations* between these stocks. The result is a matrix that lists all the possible pairwise correlations.

	PEP	KO	XOM	CVX	MSFT	AAPL	V	MA
PEP	1.0000000	0.7102439	0.3923810	0.3951931	0.4889682	0.3992099	0.4451504	0.4485192
KO	0.7102439	1.0000000	0.4555481	0.4601172	0.4266665	0.3500878	0.4490066	0.4686299
XOM	0.3923810	0.4555481	1.0000000	0.8337852	0.3770320	0.3310620	0.4420956	0.4636053
CVX	0.3951931	0.4601172	0.8337852	1.0000000	0.4185843	0.3568863	0.4673546	0.4819508
MSFT	0.4889682	0.4266665	0.3770320	0.4185843	1.0000000	0.5883075	0.5583753	0.5702065
AAPL	0.3992099	0.3500878	0.3310620	0.3568863	0.5883075	1.0000000	0.4838209	0.5170416
V	0.4451504	0.4490066	0.4420956	0.4673546	0.5583753	0.4838209	1.0000000	0.8486198
MA	0.4485192	0.4686299	0.4636053	0.4819508	0.5702065	0.5170416	0.8486198	1.0000000

Figure 14 Matrix for pairwise correlations

Pearson Correlation:

It's hard to read the correlation matrix, so we decided to plot a *heatmap* to better visualize these correlations between the different stocks. The Pearson correlation **measures the strength of the linear relationship between two variables**. It has a value between -1 to 1, with a value of -1 meaning a total negative linear correlation, 0 being no correlation, and + 1 meaning a total positive correlation.

In this graph, we are using color pallet to understand the relationship between values as RED:

1.0 and BLUE: 0.0. We can infer from the below graph that companies from the same sector share a strong correlation between them but when we make the comparison between the two stocks each belonging to a different sector, shows weaker relation.

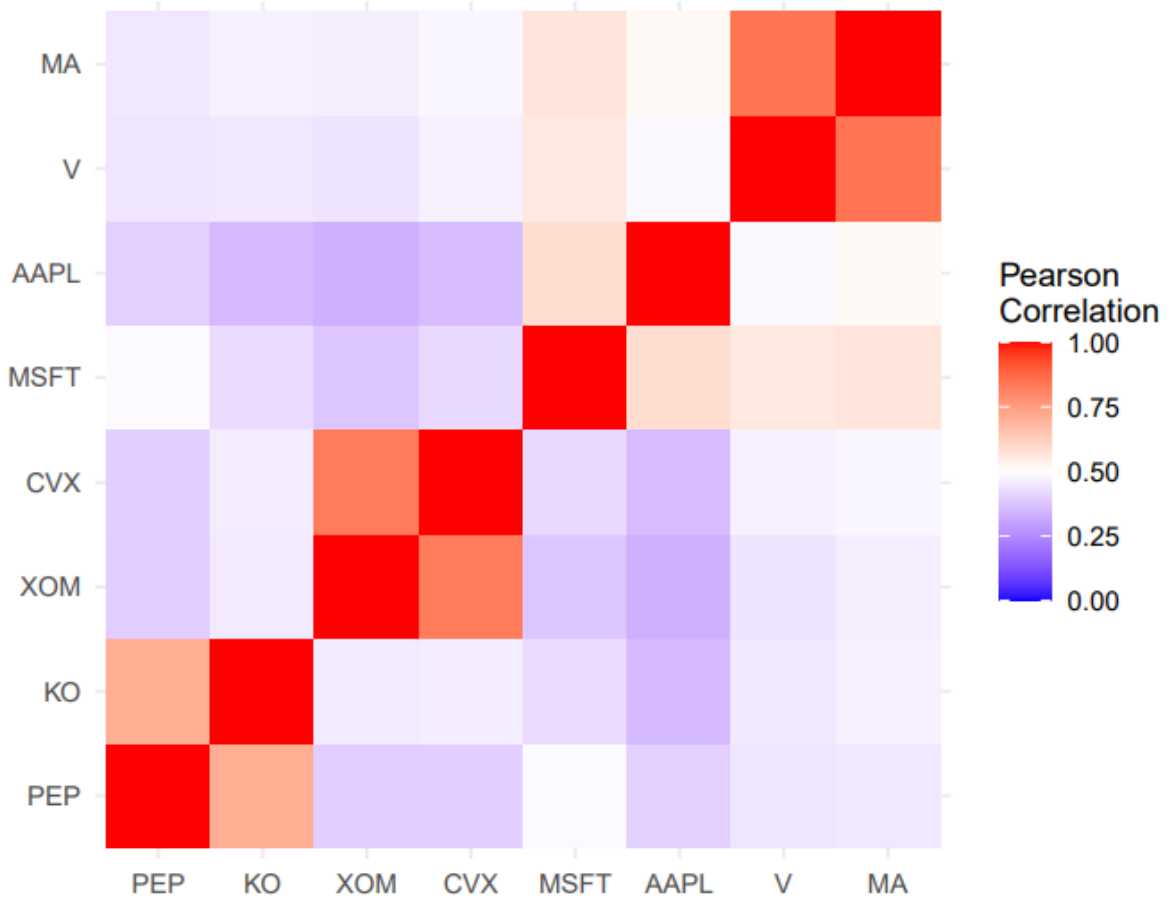

Figure 15 Pearson Correlation Heatmap

This clearly shows that there's a strong relationship between Pepsi and Coca-Cola stocks. We can infer from this graph about the correlation between PEP and KO.

PEP and AAPL show a weak relationship and can extract the correlation value for these two i.e., 0.3992.

From our analysis, we can see that PEP and KO are positively correlated which is 0.71. We'll confirm these statistics from our further analysis.

Henceforth, we chose our ***two populations*** as:

PEP: PepsiCo returns

KO: Coca-Cola returns

Hypothesis:

H0: They are statistically the same i.e., Mean of returns PEP - Mean returns KO = 0.

Ha: Mean of returns PEP - Mean returns KO != 0.

Similarities: PepsiCo vs Coca-Coca

```
symbol1 = "KO"
symbol2 = "PEP"

pepsicola = na.omit(df_stock_returns[,c("Date", symbol1, symbol2)])
```

Here, we are plotting a line graph for PEP and KO stocks to determine the performance of each share.

```
# Plot series
o = ggplot(df_stock_cum_returns_pepco_melt, aes(Date, value), na.rm=TRUE) +
  geom_line(aes(colour = symbols)) +
  ylab("Cummulative returns") +
  xlab("")
```

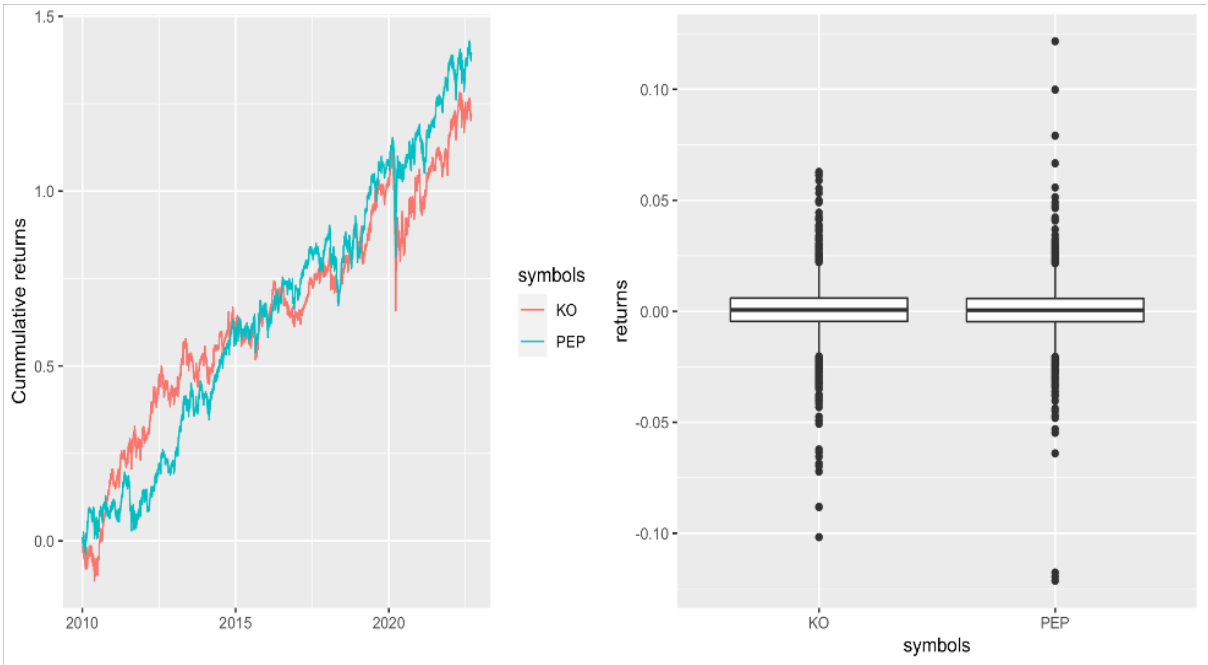

Figure 16 Line graph and Box Plot for PEP and KO stocks

The line graph plotted above depicts the relation between two stocks that are from the same industry over the period of the S&P 500. The box plot shows more outliers in the returns for these two stocks but we can see *more outliers* in PEP stock returns.

Bootstrap Test Plots for Mean Difference:

Here, we are performing Bootstrap testing to find out the difference of mean between two populations.

Tests if there is a difference between the two populations:

```
diff_obs = mean(pepsicola[,symbol1]) - mean(pepsicola[,symbol2])
diff_obs
```

```
## [1] -0.00005449945
```

The observed difference is slightly negative but is this due to randomness or there is a difference between the two populations.

We are plotting the distribution of the difference of means of two populations by executing 2000 number of simulations.

In the first graph below, we plot the results of the bootstrap test over the absolute mean line of returns. The distribution looks good with values symmetrically distributed on either side of the mean line. Whereas in the second figure below, we visualized the variance ratio of the returns over the mean line.

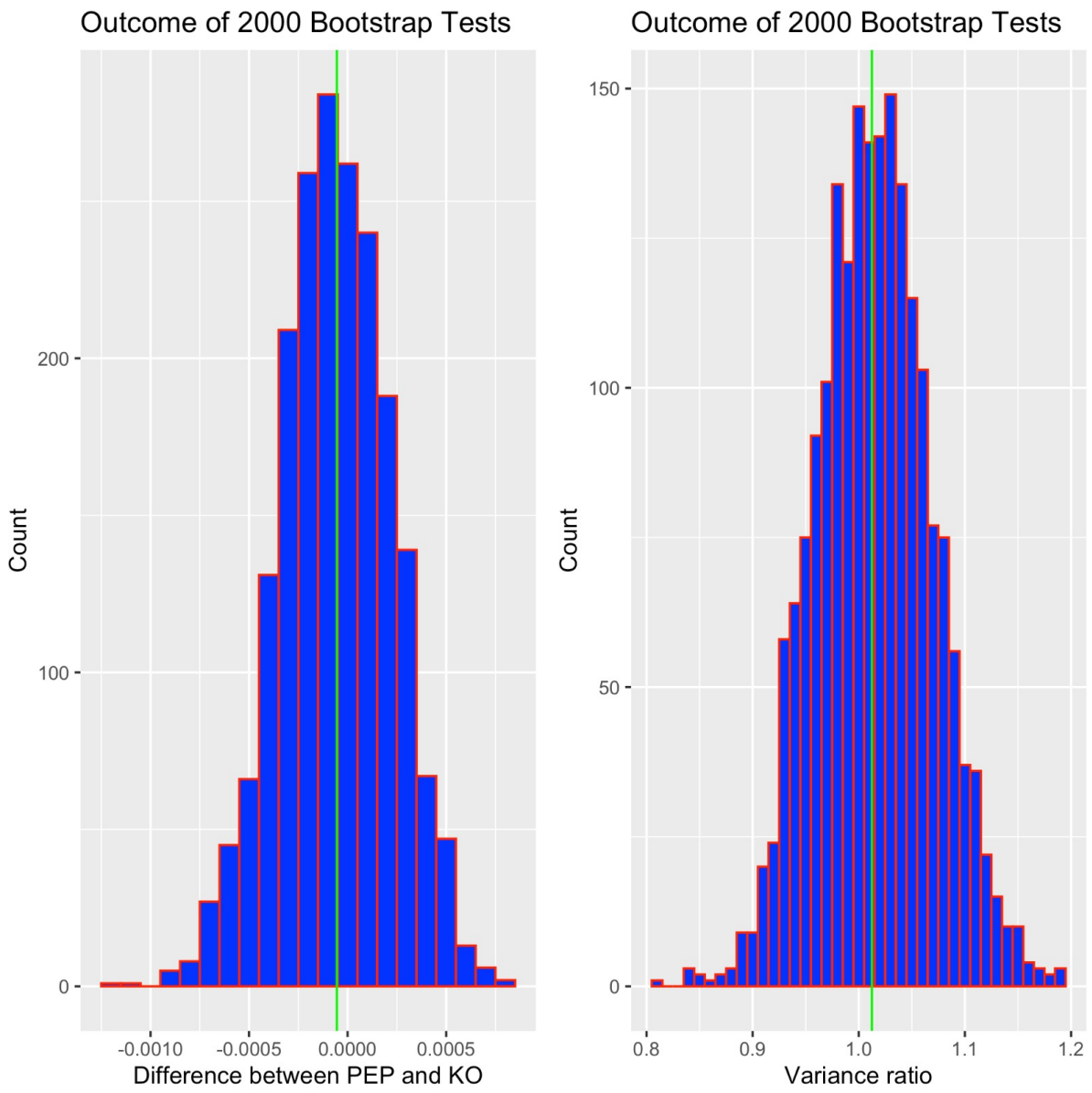

Figure 17 Bootstrap Tests for Mean and Variance Ratio

95% boot confidence interval for variance ratio to find a difference of means of two populations:

```
qdata(~division_sd, c(0.025, 0.975), data=boot_df)
```

```
##       2.5%     97.5%
## 0.9125129 1.1303962
```

Inference:

A confidence interval indicates that the 95% interval of variance ratio for the population mean

is between 0.9 and 1.1. Since the difference of mean for the confidence interval of PEP and KO is greater than zero, we can conclude that the mean of values of PEP is greater than KO, and with a factor that is in the range of the confidence interval of variance ratio.

Linear Regression

Stock market forecasting is an attractive application of linear regression. One of these assumptions is that variables in the data are independent. Namely, this dictates that the residuals (difference between the predicted value and observed value) for any single variable aren't related(Gharehchopogh, et al., 2013).

Hypothesis:

H0: The null hypothesis about the slope. (beta = 0)
It means that KO stock returns doesn't tend to change linearly when PEP changes - there is no linear association between the two variables.
Ha: There is a linear association: linear association between the two variables. (beta != 0)

Estimation:

y = a+b*x

```
s = ggplot(data=pepsicola, aes(x = PEP, y = KO)) +
  geom_point(size=2, position="jitter", color="blue") +
  geom_smooth(method="lm", col="red") +
  xlab("PEP returns") +
  ylab("KO returns") +
  ggtitle("")

plot_grid(s, ncol = 1, nrow = 1)
```

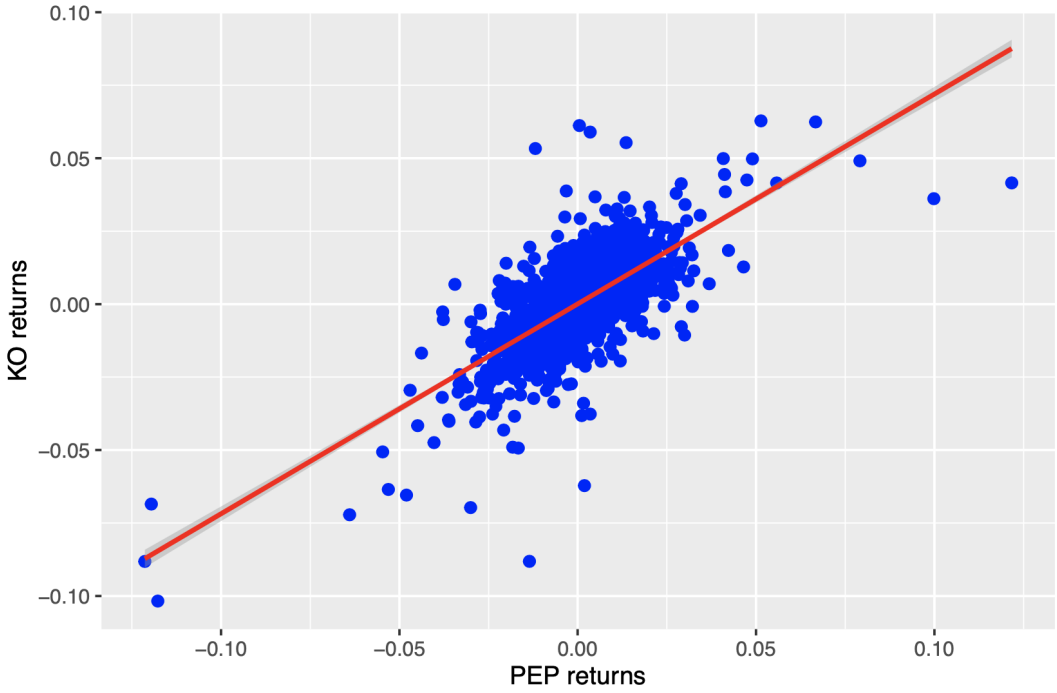

Figure 18 Linear Relationship through Scatter Plot

- From this scatterplot we can see a pattern that runs from lower-left to upper-right, meaning a positive relationship.
- The form appears to be a cloud of points stretched out in a generally consistent, straight form, although some points stray away from it.
- Regarding the strength of the relationship, the points cluster somewhat tightly, occasionally vague. Also, there are some outliers standing away from the overall pattern.

```
predict_stock = lm(KO ~ PEP, data=pepsicola)
predict_stock$coefficients
```

```
##   (Intercept)           PEP
## 0.00006837328 0.71906802557
```

Expected(KO Return) = 0.000068 + 0.72 * PEP Return

A positive coefficient beta (slope) that is approximately equal to 0.72 indicates that as the return of PepsiCo increases let's say 1%, the mean return of KO also tends to increase by 0.72%.

ANOVA and Normality of Residuals Condition:

Using this **Analysis Of Variance** approach, the value of FObs serves as the test statistic to test(Wang, 2008).

H0: Slope = 0

Ha: Slope not equal to 0

```
summary(aov(predict_stock))
```

```
##               Df Sum Sq Mean Sq F value              Pr(>F)
## PEP            1 0.1978 0.19779    3243 <0.0000000000000002 ***
## Residuals   3186 0.1943 0.00006
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Conditions:

There are two conditions upon which the model building that we have begun are built upon.

1. The y-variable, or commonly known as the response variable, is Normally distributed with a mean mu and standard deviation of sigma - normality of the residuals

2. For each distinct value of the x-variable (the predictor variable), the y-variable has the same standard deviation sigma - homoscedasticity

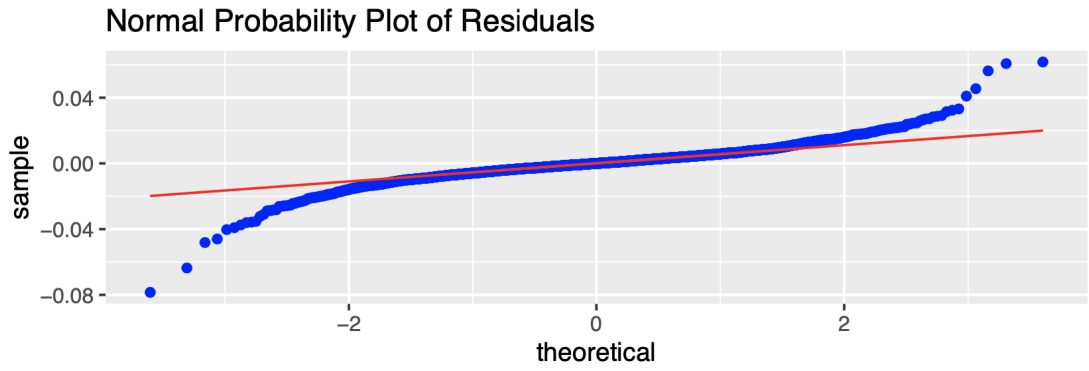

Figure 19 Normal Probability Plot for Residuals

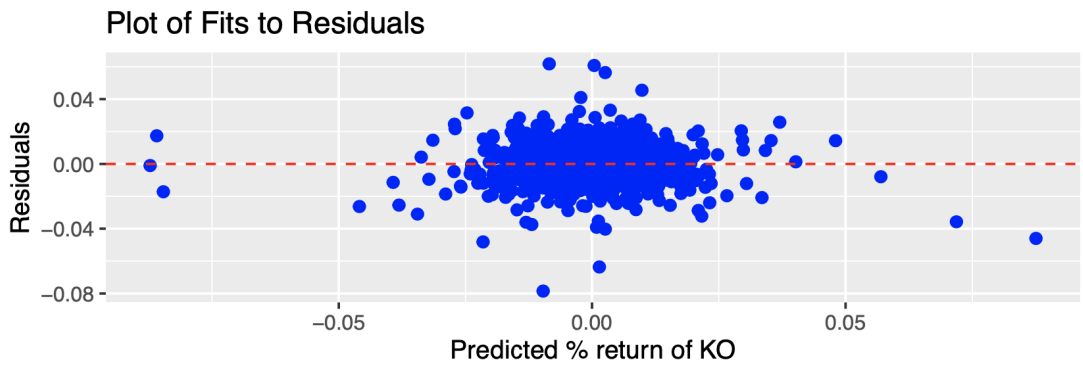

Figure 20 Fits to Residuals Plot

From the qq-plot, the points appear in a straight line and follow the diagonal line. This suggests normality.

The scatterplot of residuals doesn't appear to have any shape, pattern, or direction. The points appear fairly uniformly scattered about the flat dotted line (zero line). This suggests homoscedasticity.

The null hypothesis about the slope (beta) is that it equals to 0. It means that KO stock returns doesn't tend to change linearly when PEP changes - there is no linear association between the two variables. Ha: there is a linear association. However, conditions doesn't appear to hold but we can remove outliers and make a separate analysis as we can consider this for future research due to the out-of-scope from this project.

```
coef(summary(predict_stock))
```

```
##                 Estimate   Std. Error    t value  Pr(>|t|)


## (Intercept) 0.00006837328 0.0001384206  0.4939531 0.6213733
## PEP         0.71906802557 0.0126265543 56.9488719 0.0000000
```

Hypothesis and ANOVA Inference:

We have evaluated the p-value for the linear model based on two populations that is less than 0.05. Therefore, P-value = 0.0000 < 0.05(alpha) determines rejecting Null Hypothesis that KO stock returns doesn't tend to change linearly when PEP changes - there is no linear association between the two variables. Hence, beta is significant and ***Ha(beta != 0) failed to reject or***

Accept means *there's a linear relationship between PEP and KO.*

Nearly 57 standard errors from the hypothesized value (beta = 0) certainly seem big. The P-value 0 confirms that a t-ratio this large would be unlikely to occur if the true slope were zero. Reject the null hypothesis and conclude that there is a positive linear relationship between stocks.

Confidence Interval:

```
confint(predict_stock, conf.level=0.95)
```

```
##                    2.5 %       97.5 %
## (Intercept) -0.0002030292 0.0003397757
## PEP          0.6943110287 0.7438250225
```

As beta (slope) may vary for another sample of data, so we are 95% confident to say that the true slope varies somewhere between 0.6943 to 0.7438. That suggests if PEP will increase by 1%, we expect KO stock to increase somewhere between 0.69% to 0.74%.

Bootstrap Test Plot for Correlation:

The below graph determines the correlation coefficient by performing Bootstrap Statistics.
We can infer from this graph, the data is normalized and from this data, we are 95% confident, that the true value of correlation is between 0.67 and 0.74.

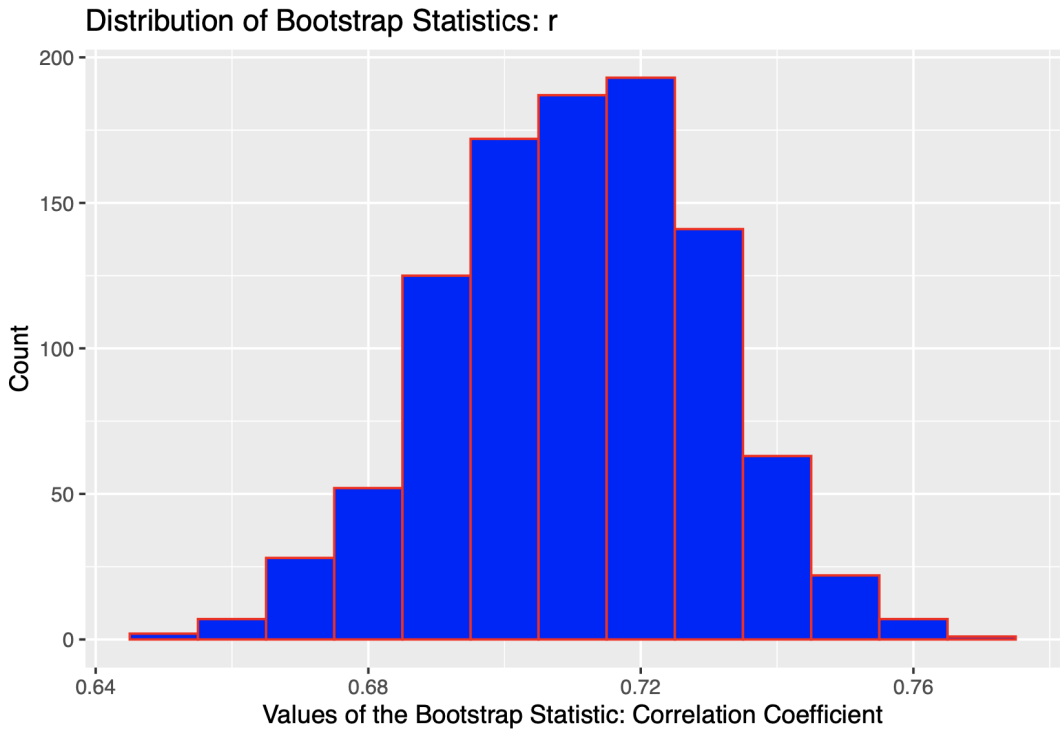

Figure 21 Correlation Coefficient

```
qdata(~cor, c(0.025, 0.975), data=boot_df)
```

```
##      2.5%     97.5%
## 0.6716284 0.7457278
```

From this Confidence Interval for beta, we are 95% confident, that the true value of correlation is between 0.67 and 0.74.

BONUS:

Calculate the Spread:

```
slope = as.numeric(predict_stock$coefficients[2])
pepsicola$spread = pepsicola$KO - slope * pepsicola$PEP
pepsicola$cumspread = cumsum(pepsicola$spread)
```

The above chunk of code calculates the spread which means the mean difference between PepsiCo and Coca-Cola stocks returns.

The below graph depicts the difference between PEP and KO stocks. We found that the difference stays as Stationary Mean. Hence, the mean of the spread doesn't change as much. It

is calculated based on Predicted KO return - Slope * Original PEP return can be seen below:

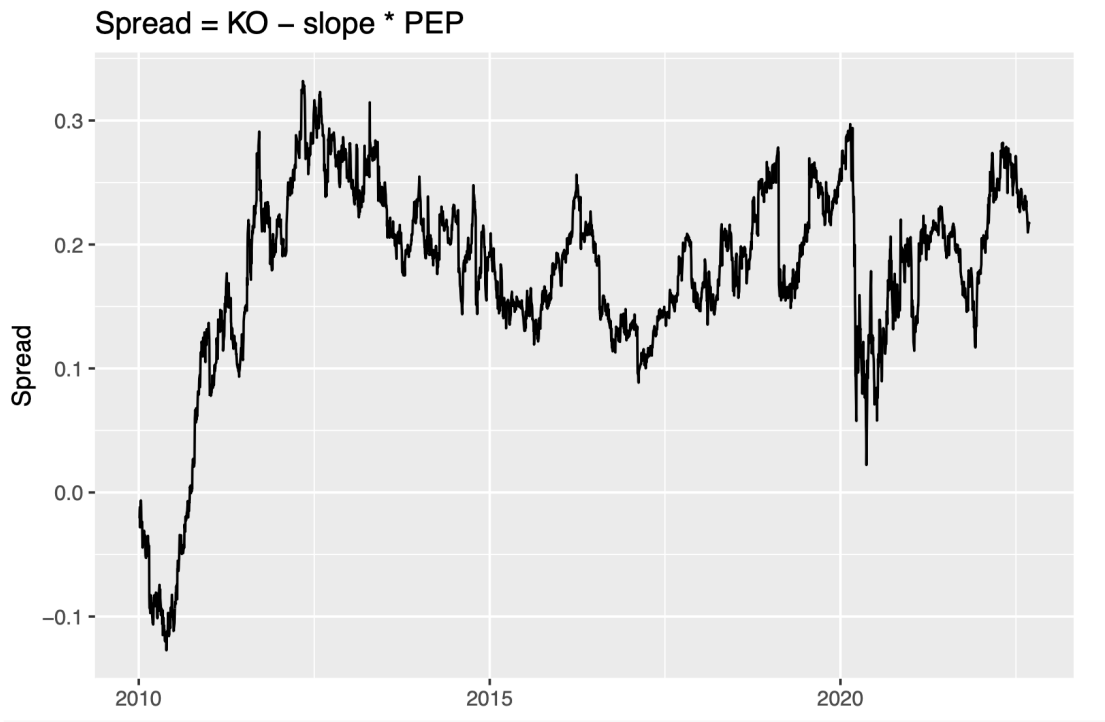

Figure 22 Spread: Means Difference (PEP and KO)

Applying Model:

```
pepco_model = function(pepreturns){
  ko_returns = 0.00006837328 + 0.71906802557 * pepreturns
  return(ko_returns)}

KO_predicted = pepco_model(pepco_test$PEP) # apply model
pepco_test = data.frame(pepco_test, KO_predicted) # add to df
```

Performance of Model:

Calculating the R squared by using a model and got the value R-squared: 0.807193.

Also, Mean Squared Errors = 0.00005320894

R-squared is a regression error metric that justifies the performance of the model. It represents the value of *how much the independent variable is able to describe the value for the response/target variable*. In our case, it is approx 81% - very high.

Scatter Plot to Evaluate the Performance of the Model:

In the below graph, we are comparing the actual KO returns with predicted *on unseen data*.

Figure 23 Performance of Model

Question 2: Conclusion

From our statistical analysis, we can infer that two stocks that belong to the same industry do behave statistically similarly. In this guiding question, we did this analysis for Pepsi and Coca - Cola which belong to the Food and Beverage industry. From our statistical analysis, we can see that Pepsi and Coca-Cola share a strong linear relationship based on linear regression with an R-squared value of 0.807 and beta approximately equal to 0.72 with a confidence interval of beta between 0.67 and 0.74. Also, from the scatter plot for the predicted KO and original KO data values, we can justify the performance of the Linear Model.

Conclusion:

Guiding Question 1:

1. Shapiro test performed to check normality: P-value significantly lower than 0.05, ***rejects normality.***

2. One-Sample t-test performed which gives a *mean of returns* equal to 0.0004 is statistically different from 0.

3. T-distributions with 3 *degrees of freedom* fit much better than the normal distribution.

Guiding Question 2:

1. From our statistical analysis, we can infer that two stocks that belong to the same industry do behave similarly and on the contrary, if we were to choose the two stocks each from different sectors they exhibit a weak relationship.

2. Linear regression on the two populations (**PEP and KO**) helps to identify the linear relations between these two groups with the *R-squared value of 81%* which is very high. Hence, H(a) adapted for Pepsi and Coca-Cola share a strong linear relationship (beta != 0).

References

- Gonzalez-Rivera, G. and Arroyo, J., 2012. Time series modeling of histogram-valued data: The daily histogram time series of S&P500 intradaily returns. International Journal of Forecasting, 28(1), pp.20-33. Doi: https://doi.org/10.1016/j.ijforecast.2011.02.007
- Hofert, M. and Oldford, W., 2018. Visualizing dependence in high-dimensional data: An application to S&P 500 constituent data. Econometrics and statistics, 8, pp.161-183. Doi: https://doi.org/10.1016/j.ecosta.2017.03.007
- Gharehchopogh, F.S., Bonab, T.H. and Khaze, S.R., 2013. A linear regression approach to prediction of stock market trading volume: a case study. *International Journal of Managing Value and Supply Chains*, *4*(3), p.25.
- Wang, J.C., 2008. Investigating market value and intellectual capital for S&P 500. *Journal of intellectual capital*. Doi: https://doi.org/10.1108/14691930810913159