Jane Street Market Data

March 20th, 2021

Marcos Arevalo	*Jasmine Guan*	*Cynthia Leung*	*Derek Leung*	*Dillen Padhiar*	*Timothy Tran*
Investigation, Theory	*Investigation, Theory*	*Investigation, Intro, Data, Background*	*Investigation, Theory*	*Investigation, Theory*	*Investigation, Theory*

1. Introduction

The act of stock trading involves buying and selling stocks to capitalize on market events by timing the market. Oftentimes, the rule of thumb is to sell stocks for a profit when they are priced at a high value, and buy stocks when they are low. Stock traders can have several transactions throughout the day, and the job requires much research and analysis on the market and market trends. Understanding time horizons, such as effects within short-term, intermediate-term, and long-term intervals, then become pivotal at understanding how to maximize one's payoff from a trade.

Stock trading can earn someone lots of money, and the flipside can occur as well. Because there are so many factors and past data to consider when trading, it becomes pivotal to understand market trends to successfully predict the return of a trade. The big question in this project is how can we understand trade predictions, as defined by returns. However, in this project, rather than jumping straight into machine learning models for stock forecasting, we aim to establish an understanding of stock market information and trends.

Firstly, there are many factors that affect stock performance prediction, and the dataset in this project 129 features per trade transaction. However it is unreasonable to create a model with 129 features. Therefore, we want to address this by figuring out which features are most distinct in returns per stock trade and understanding which is most vital in predicting return on stock trade.

Secondly, given historical data of trades over a 500 day period and returns on those trades, we aim to understand whether the returns or number of trades/transactions varied across this time period. The purpose of figuring out whether there are any patterns was to see whether there were specific time frames that may have yielded higher returns or contained more trades compared to others. If this were true, and seasonality is indeed a factor as represented in the data, this may be a component to consider when one is creating the actual predictive model. If there is no clear trend or pattern we can establish, the sole time factor may not make as much of an impact on the data.

Finally, we want to examine the volatility of returns given different time horizons. When trading, one may want to choose different stocks depending on whether the goal is long term or short

term; hence, understanding the differences among trades per different time horizons will create a better understanding of the nuances of stock trading in terms of the type of investment.

2. Data

The Jane Street Market dataset from Kaggle was originally created for Kaggle users to create predictive models that could forecast stock market returns for trades over a 500 day period. There are no company labels or any specific indicator to provide information whether a particular stock trade for a specific company was made over the time period; however, there is detailed, anonymized information about over 200,000 individual stock trade transactions. Each observation represents a single stock trade transaction, and it contains the date of transactions, return variables, and a set of 129 features about the trade. The return variable is represented by 'weight', 'resp_1', 'resp_2', 'resp_3', 'resp_4', and 'resp', and the set of features describing the trade transaction has no qualitative description about it.

Other datasets given within this Kaggle prompt included a features matrix and a training sample data for predictive modelling; however, these were not used and not included in this project. Our main focus is on the dataset that includes features, stock return variables, and date so that we can explore the trends and relationships between time, trades, and returns within the dataset.

3. Background

Much work has been conducted previously to understand market trends, price movement, and forecasting analysis. Because this project focuses on understanding the relationship between returns, trades, and features seen in the dataset over the 500 day period for prediction, understanding methods studies have utilized to forecast the stock market is critical.

In a study conducted by Ananthi and Vijayakumar published in 2020, the authors focused on predicting stock market prices using past data. The study began by breaking down stock market prediction into three sections: trendline identification, learning past price movements, and pattern formation. For trendline identification, the study represented price variation using OHLC (Open High Low Close) financial charts. Past price movements were pulled through online data, and k-NN regression was one of few to classify and test recent price movements. Technical indicators were also used to generate a signal between 0 and 100 for buying/shorting a stock. The implementation of the machine learning models were also curated based on the risk profile of the user, and this can account for adjustment based on the amount of units that should be held in each stock in a particular profile. The study also included sentiment analysis of each stock in the social media as well, but it was not thoroughly discussed in the paper.

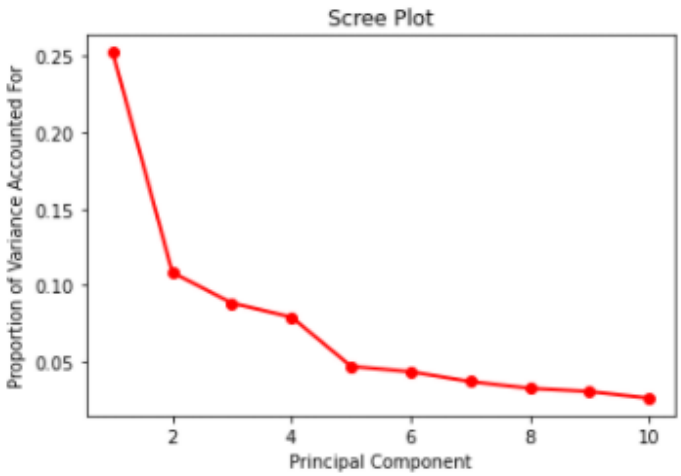
In a previous study conducted in 2001 by Abraham et al., the authors discuss the use of intelligent systems for stock market predictions, particularly the application of hybridized soft computing techniques for automated forecasting and trend analysis. Neural networks were a key method in creating predictive models in this study. Using stock data, the researchers created ANN-SCG algorithms and EFuNN training to forecast whether stock values were going up or down, represented by 1 and 0. The results demonstrated that the RMSE on test data is small and that the predictive model is reliable. The authors believe that the stock forecast errors could be reduced if individual neural networks were used rather than a single one.

In another study by Zhong and Enke in 2017, the study focused on forecasting daily stock market return using dimensionality reduction. The study used three methods for dimensionality reduction: PCA, FRPCAs, and KPCA. After the data was reduced, artificial neural networks (ANN) techniques were applied to accurately classify the daily stock price direction. Again, neural network techniques were leveraged to forecast stock market trends. The ANN classification results of 36 transformed datasets based on different PCAs, and the models were tested in a trading simulation to see if higher predictability implies higher profitability. Ultimately, the process using ANN-PCA models gives slightly higher prediction accuracy for stock price direction compared to the other PCA methods.

4. Investigations

Principal Component Analysis

Often, long term stock price changes are attributed to an overall bull or bear market and the specific features of the data are potentially dismissed as less important. In addition to long term results, our group was also interested in short term insights about the returns on a stock. Therefore, a primary question we asked when examining the data, concerned which features are the most distinct and telling about the return on investment, every day regardless of trends over long periods of time. In order to do this, we first had to find out what features best distinguish each trade from the group as a whole. In other words, we wanted to find the features that made up the principal components of the data.



Caption: This scree plot shows the proportion of variability that can be attributed to each principal component. While there are 129 features, we have found that about 74.65% of the variability can be explained by these 10 principal components.

Rather than try to use all 129 features in some kind of regression, we find the features that best distinguish the different stocks from each other. This allows us more flexibility in our regression model as well as a far easier interpretation of the results because there are fewer and more impactful variables to examine. The following is the output of principal components:

Table: Principal components

PC1	feature_56
PC10	feature_12
PC2	feature_24
PC3	feature_113
PC4	feature_126
PC5	feature_50
PC6	feature_56
PC7	feature_56
PC8	feature_77
PC9	feature_64

Caption: This small table shows the 10 principal components matched up with the feature that most strongly influenced them.

The features shown in the above table will be what we use to answer our question of which variables are distinct and telling about the return on investment for each day, regardless of the long term trends on returns.

Generalized Estimating Equations

After determining the most distinct variables, the next step is to find which of these variables is the most telling about the return on investment.

For the PCA and GEE sections, the data was cleaned to remove missing values. Because the data is complete and appears to be balanced, we picked an exchangeable correlation model. For the distribution family considered in the GEE, we picked poisson because there are a number of events that occur in a fixed period of time which impact the dependent variable.

GEE Regression Results

Dep. Variable:	resp	No. Observations:	2309487
Model:	GEE	No. clusters:	500
Method:	Generalized Estimating Equations	Min. cluster size:	29
		Max. cluster size:	18252
Family:	Poisson	Mean cluster size:	4619.0
Dependence structure:	Exchangeable	Num. iterations:	27
Date:	Fri, 19 Mar 2021	Scale:	1.000
Covariance type:	robust	Time:	23:52:41

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-6.5830	0.284	-23.195	0.000	-7.139	-6.027
feature_56	0.0021	0.001	3.130	0.002	0.001	0.003
feature_24	-0.0580	0.013	-4.602	0.000	-0.083	-0.033
feature_113	0.0444	0.010	4.539	0.000	0.025	0.064
feature_126	0.1197	0.013	9.019	0.000	0.094	0.146
feature_77	-0.0070	0.015	-0.457	0.648	-0.037	0.023
feature_64	-0.0238	0.015	-1.624	0.104	-0.052	0.005
feature_12	0.0566	0.017	3.422	0.001	0.024	0.089

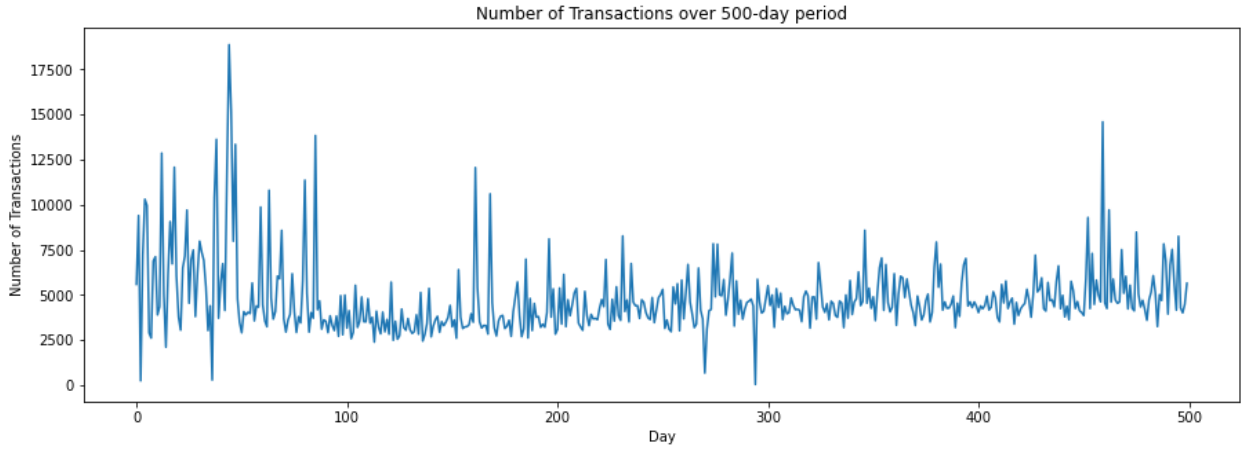
Skew:	-7.2087	Kurtosis:	1122.8056
Centered skew:	-7.2728	Centered kurtosis:	1134.4473

Caption: This table shows the results of the generalized estimating equations analysis. We used resp (return) as our dependent variable and regressed using the most distinct features across all of the days in the dataset. As a result, we should see the importance of these features across all of the time spanned in our data.

We found that features 56, 24, 113, 126, and 12 explain statistically significant differences in return at the 5% level. Of these features, 56, 113, 126, and 12 are positively correlated with the return while 24 is negatively correlated with the return. Feature 126 is notable in that it is a distinct feature with the highest positive correlation with return and feature 24 is a distinct feature with the highest negative correlation with return, even amongst the negatively correlated features with less confidence.

Transactions Over Time

Another part of the data we want to investigate is the number of transactions per day over the 500 day period. We want to understand whether the number of transactions follows a certain distribution or trend that we can quantitatively explain. To begin, the following demonstrates a simple line graph of the number of transactions per each day.

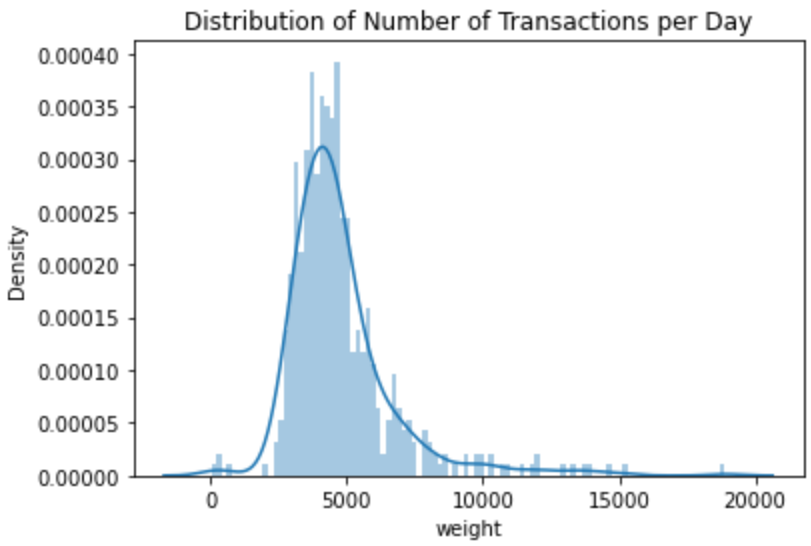


Caption: There seems to be more transactions made within the first 100 days, but after, it seems relatively stable, with the occasional spike. At the same time, volatility seems the greatest within the first 100 days as well, hitting both the maximum and minimum of the data.

From the above graph, no definitive conclusions can be drawn, as the graph, aside from the occasional spike and volatility in the first 100 days, seems relatively stable. To ensure that the number of transactions is not randomly scattered across the 500-day period, we simulated a random sample of 500 values using the data range. The following graph demonstrates the random scatter and observed data:

Caption: The orange line represents the randomly drawn data, and the blue line represents the observed data. The orange line clearly is very different from the blue line.

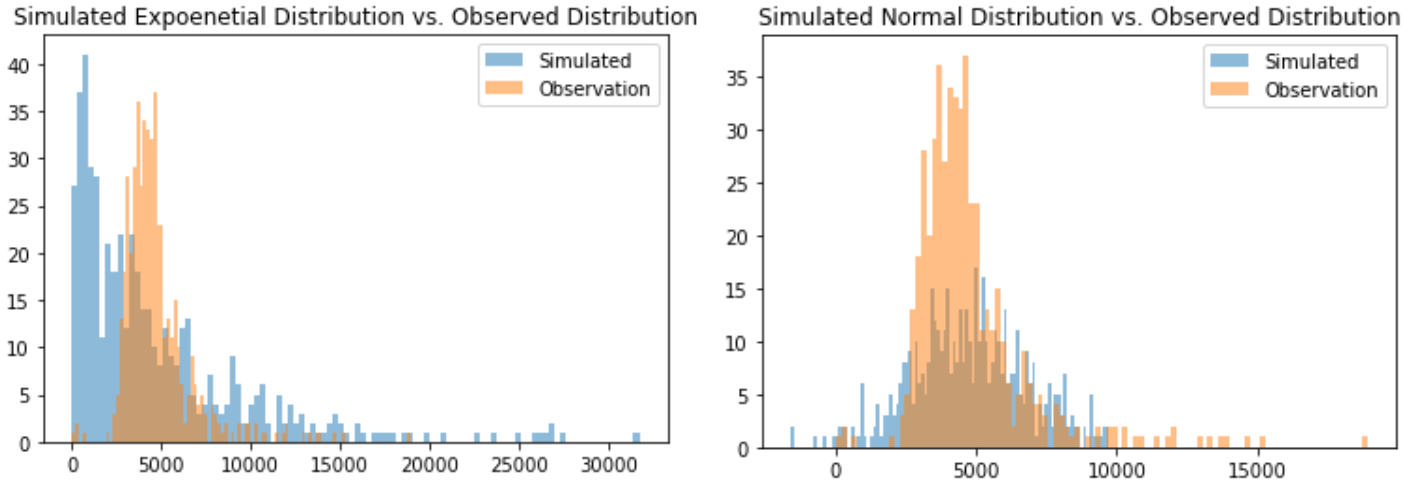
Now, we know that the observed number of transactions over time is not random, we want to see whether we can understand the distribution of the number of transactions per day to better understand the data. The following graph visualizes the distribution of the number of transactions per day.



Caption: The distribution above seems to have a steeper peak with a right-skewed tail. Most values seem clustered around 2500-5000.

Because we know the data is count data, we want to see if the histogram follows a Poisson distribution. Therefore, we use a chi-square test to see whether the observed distribution is similar to a simulated Poisson distribution using an estimated lambda-hat based on the observed mean number of transactions. From the results of the chi-square test, with a p-value of 0.0, we reject the null hypothesis that the observed data comes from the Poisson distribution, and we can conclude that the data does not come from the Poisson distribution.

In addition to a chi-square test to see whether it is a Poisson distribution, we also tested for an exponential distribution and a normal distribution. For the exponential distribution, we estimated lambda-hat using the mean of the observed data, and for the normal distribution, we used the mean and standard deviation of the observed data for the parameters. In both chi-square tests, the p-value was 0.0, so we can reject the null hypothesis in both these tests in favor of the alternative hypothesis. Therefore, the observed data does not come from either an exponential distribution or normal distribution as well. Graphs of the observed vs. simulated distribution of the two tests are visualized by the following graph:

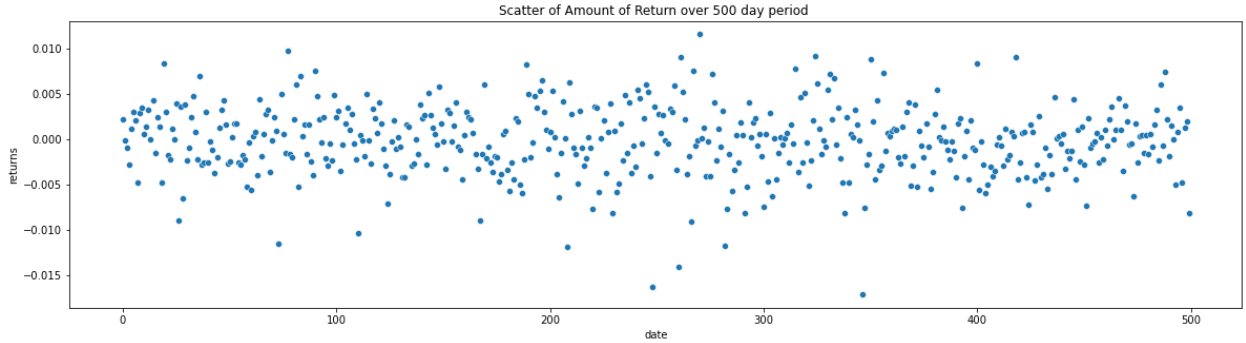


Caption: The left demonstrates the test with the exponential distribution and the right demonstrates the test with the normal distribution. We can see that the simulated vs observed in both cases are quite different from one another, as proved by the chi-square tests.

From all the above tests, we cannot conclude that the observed data comes from a specific distribution, but we can conclude that the observed distribution of number of transactions per day definitely does not come from a Poisson, exponential, nor normal distribution.

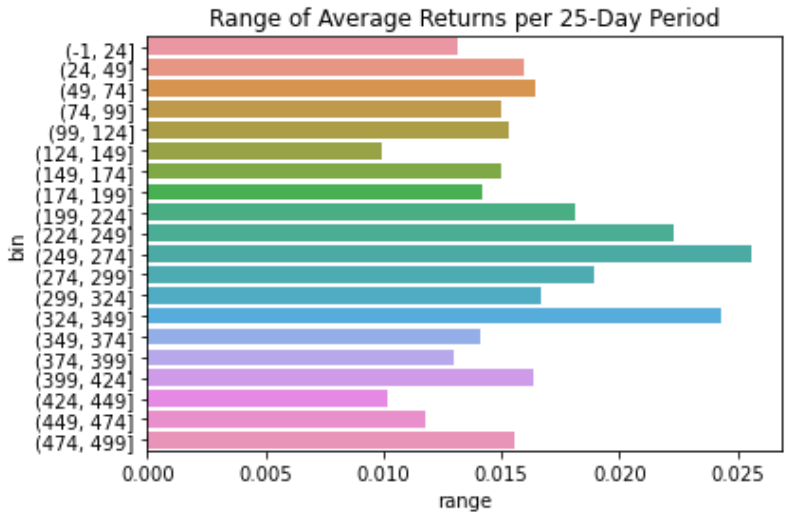
Returns by Day

Another key facet of the data we want to understand is whether the amount of returns per a transaction varied across the 500 day period. The goal is to understand whether there may be a certain time period that gave the most returns compared to other day periods. In this project, we defined returns as the product of weights and resp, and calculated a new 'returns' variable using these two variables.To understand the general trend over the 500 day period, we used a scatter plot to understand where the average amount of returns are per each day.



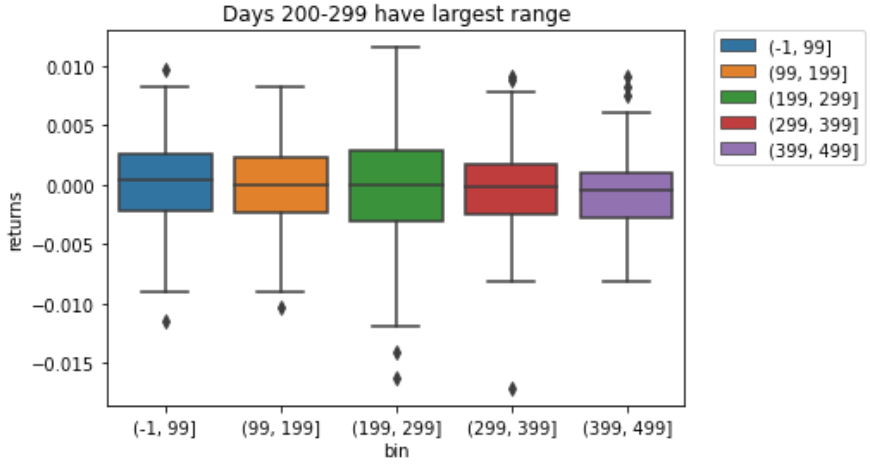
Caption: The average returns are evenly scattered throughout the 500 days; however, the average return points seem to be more scattered in the middle, between das 175-350.

The above graph demonstrates that there may be some variation in the average returns per day in the middle of the 500 day period. To initially understand the data, we binned the data into 25 day periods and calculated the mean, sum, min, max, and range. Among these summary statistics, there were not any particular values that stood out aside from the range. As demonstrated in the scatter plot above and in the data here, the range of the average returns seems to be the greatest among bins starting from 200-350. The following barplot depicts the ranges across the 25 day bins:



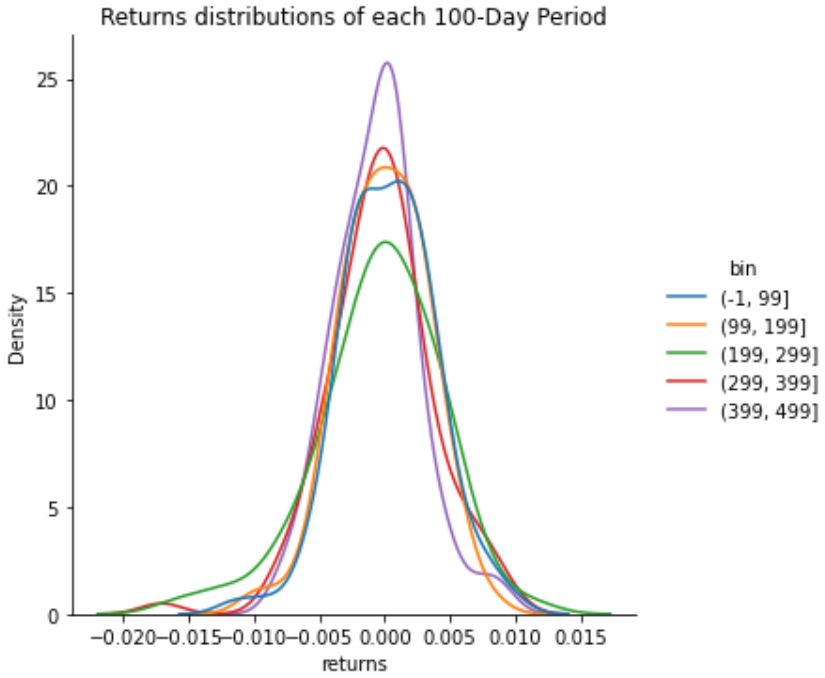
Caption: The bins in the middle of this bar plot have the greatest range compared to bins earlier or later in the 500-day period.

To further understand the data, we decided to separate the data into 5 100-day bins, and create boxplots for each to visualize the median, range, and outliers of the data a bit better. The following depicts the data:



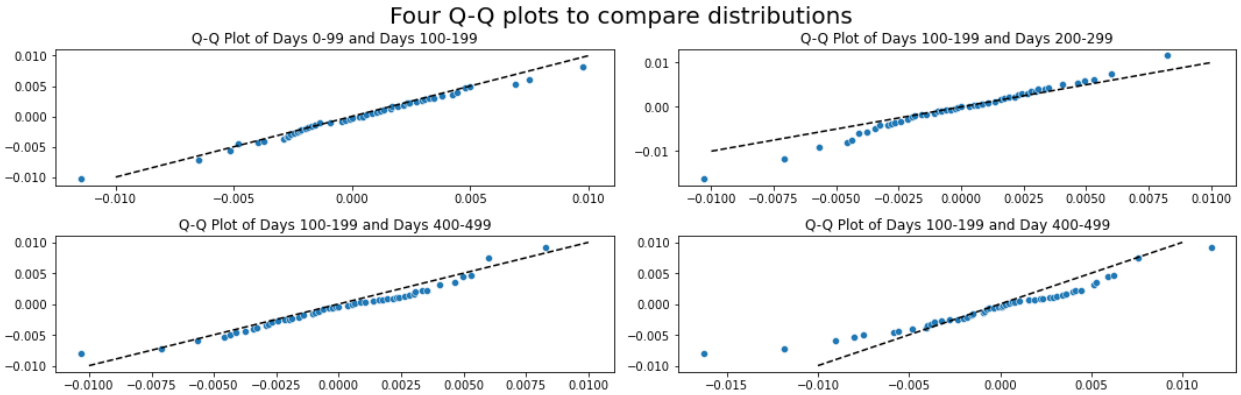
Caption: The above graph demonstrates that bin (199,299] has the greatest range compared to the other bins. Not only the 1st and 3rd quartile further from the median compared to the other bins, but the whiskers are also much further away as well. The other boxplots are more similar; however the last bin ((399,499]) has quite a few outliers on the higher end.

From the above box plot, we want to see whether the distribution of bin (199,299] is different from the other bins because the data seems more spread out. Additionally, we want to know whether there is any other variation we may have not been able to see in the box plot as well. We therefore plotted the distribution of the average returns for the days within each of the bins in the following plot:



Caption: The two plots that seem the most different compared to the rest are bin (199,299] and (399,499], represented by the green and purple line, respectively. The green line seems to be wider and lower compared to the other distributions, while the purple line seems to be narrower, with a steeper slope and higher peak.

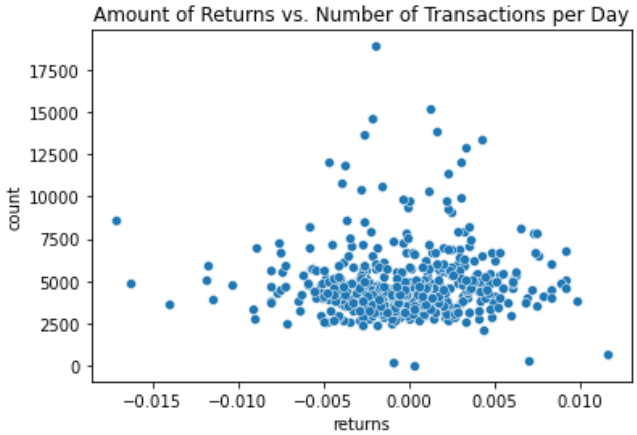
To test whether the distribution of the average returns are much different from each other, we selected a few distributions to compare to one another. We briefly tried to do a chi-square test to compare distributions, and it returned a very high p-value. Therefore, we believe that while the distributions vary slightly, the data comes from the same distribution, as expected. Rather than doing a chi-square test because we are unsure of the actual distribution, we chose to do a quantile-quantile plot to visualize how two distributions vary from one another. The following four plots compare the distributions of key representations from the distribution plot above.



Caption: As we predicted, the q-q plot between the first two bins was most similar, with bin (-1,99] slightly below bin (99, 199). However, the other q-q plots vary a bit more on the edges, while the area where the majority of the points remain relatively close to the y=x line.

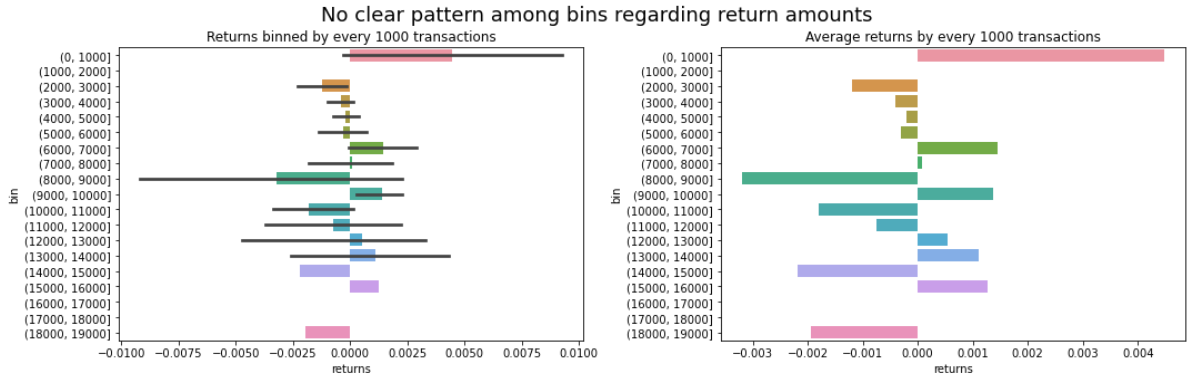
The q-q plot is key to understanding the relationship between the plots. The top right graph demonstrates that not only does period 200-299 have a greater range, but it begins with lower values compared to period 100-199 and ends with higher values compared to period 100-199. In the lower left plot, we also see that period 400-499 has lower values in the middle compared to period 100-199. In the lower right plot, we have compared period 100-199 and period 400-499, and this is where we see the most deviation from the line y=x, as expected. Initially, period 200-299 starts off with greater values than period 400-499; however, this switches when suddenly points go under the y=x line. The changes in the q-q plot can be depicted in the distribution plot of the 5 bins that was previously shown.

In addition to above, we tried to see whether there was any relationship between the amount of returns and number of transactions per day to see whether more transactions could correlate to higher returns; however, the following scatter plot shows that there isn't a correlation between the two.



Caption: There is no clear correlation between amount of returns and number of transactions per day. Points are scattered across the graph.

To be sure there is no correlation, we plotted the average returns based on bins of 1000 transactions, and again, found varied results.



Caption: Both graphs demonstrate that there is no clear pattern in terms of number of transactions and average return based on day, as the average return goes in both directions in consequent number of transactions, with no definitive pattern.

Volatility of Returns

Stock return is generally associated with the *Investment Time Horizon*, which is the period of time one plans to hold the investment until they cash out the money. Time horizon is generally dictated by investment goals and start time. When comparing long term and short term investment, high volatility (high variability) is usually associated with short term investment and vice versa. With this investment rule in mind, we wish to explore that, out of {*resp_1*, *resp_2*, *resp_3*, *resp_4*, *resp*}, which return time horizon is more volatile and can therefore be considered the shorter time horizon, and which return represents longer term investment.

We first begin with studying the distribution of the 5 resp variables. More specifically, it is key to first look at the density distribution to see what model the distributions take.

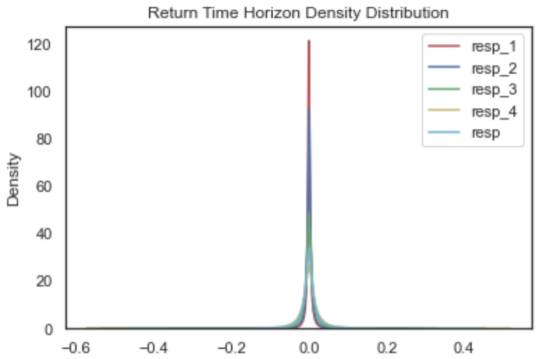


Figure: Density Distribution plot of the 5 return time horizons

Immediately, one can see that it is difficult to look at the distribution closely because of how the outliers cause the spread to become distorted and difficult to look at. However, it is clear that all the responses seem to be centered at 0, with different responses having a much higher peak than others. More specifically, *resp_1* seems to have the highest peak as opposed to *resp_4* which has the lowest. In order to study this more closely, we can redraw the graph to exclude outliers. For the sake of simplicity, we defined an outlier as any point that held a z-score higher or lower than 2.5. We then excluded these outliers and redrew the graph below.

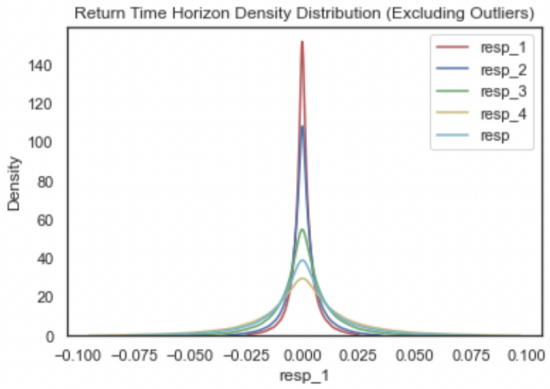


Figure: Density Distribution plot of the 5 return time horizons

Again, we can see a similar distribution (as expected) of all response variables centered at 0. We see that the order of height goes from *resp_4*, *resp*, *resp_3*, *resp_2*, and lastly *resp_1* with again *rest_1* being the highest peak. Because removing outliers provided a much clearer view of the data and said outliers will cause issues further into the investigation, all onward analysis was performed on data with outliers excluded. To further explore the distribution, we turn to a different viewpoint to see the spread on a clearer scale.

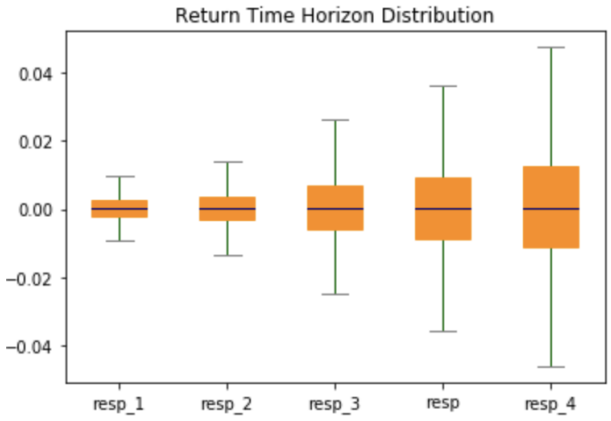


Figure: Box-plot of the 5 return time horizons with outliers removed

Distribution Description

Time Horizon	*resp_1*	*resp_2*	*resp_3*	*resp*	*resp_4*
Return Average	0.000338847	0.000412703	0.000641281	0.000842295	0.001003522
Standard Deviation	0.009791262	0.013603640	0.021041706	0.029945465	0.036390541

In the box plot above, we continue to see the increased variability in the time horizons. More specifically, we see that all returns of the five time horizons are centered around 0, which goes in accordance with the stochastic nature of the stock market. We again see the spread of the 5 returns increasing in order (with resp between *resp_3* and *resp_4*) which again reflects the volatility of the specific stock. We can then deduce the range of each return time horizon.

Since *resp_1* shows the least volatility in returns, we can deduce that *resp_1* indicates the longest investment time horizon. When an asset is held for longer periods of time, the fluctuation in the market due to random events can be evened out to reflect the overall return rate of the specific stock. Although long term investors sometimes don't receive the high gains from buying at lowest and selling at highest, they also don't face the risk of large scale loss. On the other hand, *resp_4* shows the highest variability in distribution. The high gain and high loss indicates that the investment period is relatively short and is more directly influenced by market fluctuations.

In order to continue analyzing the comparison between the resp variables, we select notable resp variables and start comparing specific variables against one another. This allows us to narrow down our search and formulate more specific graphs to analyze the time horizons and how this information can be interpreted.

Compare Distribution of Returns of Different Time Horizon

Comparison #1 --- Shortest vs. Longest Time Horizon

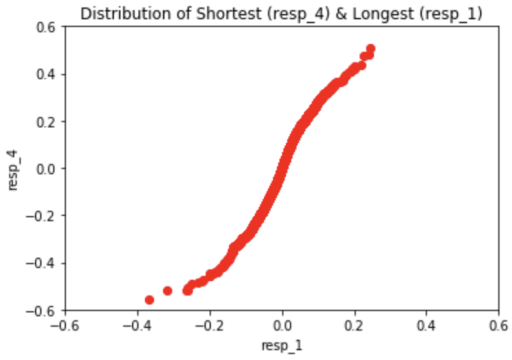


Figure: Return of shortest time horizon (resp_4) plotted against return of longest time horizon (resp_1)

We first turn to a comparison between the two resp variables with the largest difference in volatility: *resp_4* and *resp_1*. We do this through the creation of a Q-Q plot, which shows us if both of the variables come from the same distribution. From the plot, it is clear that the comparison distribution goes from convex to concave at around the 0 return area for both. We can divide this observation into negative return and positive return to further understand the effect of time horizon on investment.

For the negative return, since the two axes are the same, the graph below (0,0) shows convex trend, which means that the *resp_4* values are lower than the *resp_1* values. When in negative returns, this indicates that a shorter time horizon is associated with bigger loss in comparison to longer investment. Short-term investment is associated with higher risk and the larger negative returns reflect that. For the positive returns, the graph above (0,0) shows concave trend, which means that the resp_4 values are higher than the *resp_1* values. When in positive returns, this indicates that a shorter time horizon is associated with larger gains. Short-term investors pay close attention to the low and high points of the market and make frequent decisions that can increase profit in contrast to long term investment that sacrifices high returns for stability.

Comparison #2 --- Second Shortest vs. Second Longest Time Horizon

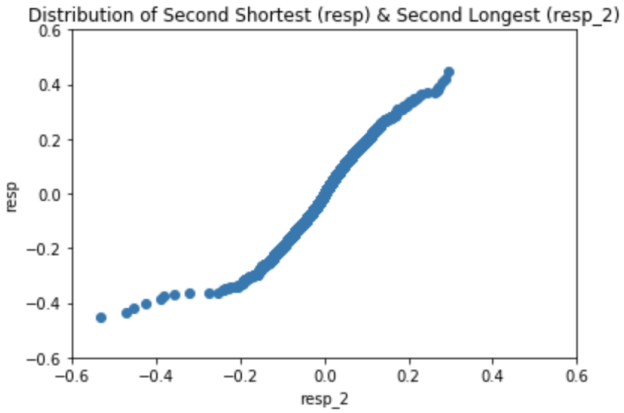


Figure: Return of second shortest time horizon (resp) plotted against return of second longest time horizon (resp_2)

We continue to do a similar comparison between the resp variables. We now look at the distribution of the second shortest (*resp*) and second longest (*resp_2*) time horizons. The comparison distribution of *resp* and *resp_2* shows a similar trend as that of *resp_4* and *resp_1* except the distribution appears flatter than the previous one on the same axis scaling. This indicates that higher loss and higher gain is still more associated with the shorter time horizon, further proving our point that a shorter time horizon is indicated by larger volatility.

Comparison #3 --- Longest vs. Second Longest & Shortest Time Horizon

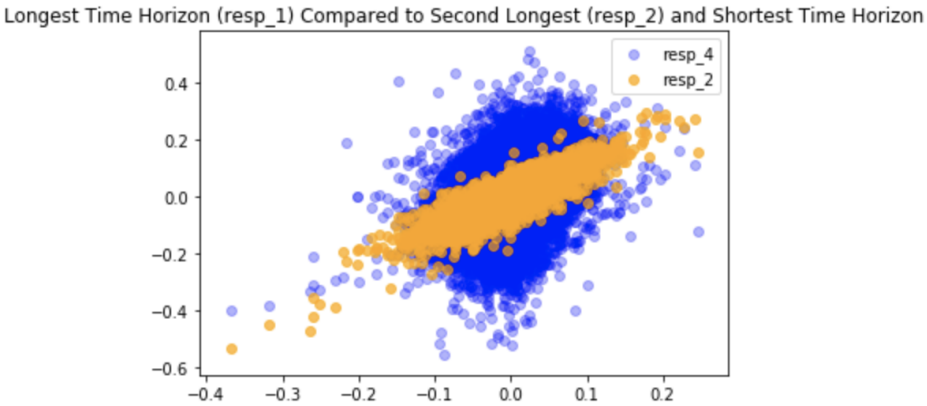


Figure: Second longest time horizon and shortest time horizon plotted against the longest time horizon

Next, we look at a comparison between *res_1*, the longest time horizon, resp_2, the second longest, and the shortest time horizon . As seen from the plot, when plotted against the longest time horizon (*resp_1*), the shortest (*resp_4*) shows significantly greater variability than that of the second longest (resp_2) time horizon. This variability is an indicator for volatility and is correlated to the risk and returns of stock investment.

Comparison #4 --- Similar Time Horizons

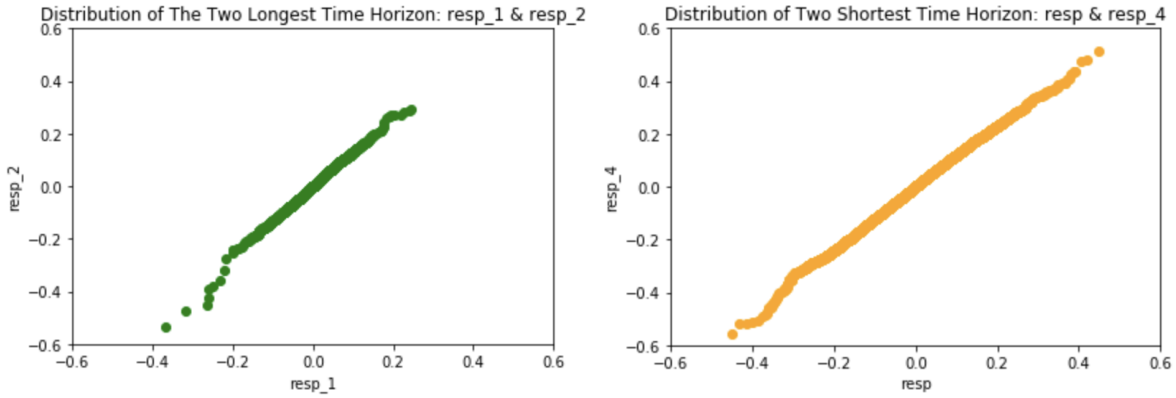


Figure:
Left: return of two of the longest time horizons
Right: return of two of the shortest time horizons

Lastly, we examine the distribution between the two shortest and two largest time horizons. When comparing the two longest time horizons (*resp_1*, *resp_2*) and the two shortest time horizons (*resp*, *resp_4*), the trend appears to be much more linear with slight convexity at the negative tail and concavity at the positive tail. This indicates that similar time horizons correspond to similar returns except at the tail of negative and positive return. Since the volatility

between similar time horizons are similar, their corresponding returns also have similar distributions.

5. Theory

1. Quantile-Quantile Plot

A Quantile-quantile plot (often shortened as Q-Q plot) is used to determine if two distributions are the same or not. This is done by pairing order statistics from each of the two distributions into coordinates. We check the line formed by these points, and conclude depending on the line created. If a straight line is created, then we can conclude that both distributions are actually the same, and vice versa.

- Null Hypothesis: $Dist_x \sim Dist_y$ (In which X and Y come from the same distribution)
- Alternative Hypothesis: $Dist_x \neq Dist_y$ (In which X and Y are different distributions)

Shape of Q-Q Plot:

- Convexity: if data appears convex relative to the 45 degree line, it means that the y-quantiles are lower than the x-quantiles and therefore y-values have a tendency to be lower than x-values.
- Concavity: on the other hand, concave data indicates that x-values have a tendency to be lower than y-values.

2. Box Plot

Box plots numerically displays the center, spread and quantiles of data. It has lines extending from the edge of the box to indicate further variability in the data. Box plots are nonparametric, which means that it displays the center and spread of data without any assumptions of the actual distribution. The middle 50% of data are included in the box and the outer whiskers indicate the maximum and minimum without outliers. From the shape of the box and whiskers, we can infer the center, degree of dispersion and skewness from the data.

Multiple box plots are displayed side by side and give a clear picture of how the overall distributions differ from each other through the shape of the box. Longer box and whiskers indicate higher variability while smaller box and shorter whiskers indicate more concentrated data.

3. Standard Deviation

Standard Deviation is calculated by: $SD = \sqrt{\frac{\Sigma(x - \overline{x})}{n}}$

It is the statistical measure of market volatility, measuring how widely trading values are dispersed from the average price. If a stock is trading at narrow price range, it will have low

standard deviation and therefore low volatility. Prices fluctuating with increasing standard deviation show higher than average strength and weakness, causing high yields and high losses. In terms of time horizon, low volatility is generally associated with long term investments.

4. Outlier Identification

In order to obtain high order inference about real life data, outliers are usually discarded when exploring data. In order to filter outliers, we decided to remove data that are 3 standard deviations away from the center. The Empirical Rule states that 99.7% of data are within standard deviation from the center, therefore discarding low probability events to increase generalizability is needed.

- Z-score: number of standard deviations from the center
  - $z = \frac{x - \mu}{\sigma}$
    - In the case where μand σare unknown, we use sample average to estimate and square root of sample variance to estimate
- Empirical Rule

Number of SD Away from Data	**1**	**2**	**3**
Percentage of Data from Both Ends	65%	95%	99.7%

  -

5. Poisson Distribution

Let X be a discrete random variable. We can claim X is approximately distributed by the Poisson distribution with rate $\lambda > 0$ such that $E[X] = \lambda$ and that Var $[X] = \lambda$.The probability mass function (PMF) follows such that $P[X = k] = \frac{\lambda^k}{k!} * e^{-\lambda}$ where $k \in Z_{\geq 0}$. For our research however, λ is unknown thus we instead estimate it using λ̂, which we define to be the average of the discrete random variable X within each unit interval where the average is defined to be $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. The process of estimating λ̂ is called a method of movement but we also use maximum likelihood estimation to reach a similar conclusion.

6. Chi-Square Goodness-Of-Fit Test

The Chi-Square test is a parametric test which tests that a distribution follows a certain distribution such as normal, poisson etc. We always then define H_0, the null hypothesis, that it follows the distribution specific to the test. The $\chi\chi\chi\chi\chi\chi^2$ test relies on the probabilistic result stemming from a theorem: begin with the summation $\sum_{j=1}^{r} \frac{(v_j - np_j)^2}{np_j}$ where v_j is a random

dependent variable with binomial distribution (n, p_j) and np_j is its expectation. This summation then converges in distribution to a χ^2 random variable with r-1 degrees of freedom. When performing the actual test, you first must start with H_0 which claims that independent observations $W_1 ... W_n$ have a certain distribution. Depending on what this certain distribution is, you will calculate the unknown p_j by choosing the amount of intervals that occur in the distribution which we define as $B_1 ... B_r$. When choosing r, in practice it is done such that either $np_j \geq 5$ for all j or $r = 2 * n^{2/5}$.

7. Chi-Square Test for Poisson Distribution

If we assume H_0 is true, we then estimate lambda using a maximum likelihood estimator. Each p_j value can be calculated then by $P(Poisson(\bar{\lambda}) \in B_j)$. Our test statistic is modeled by the $\chi^2\ r - 1$ distribution and we follow the summation given in the previous section. For example, if we are given a $\hat{p}$ composted of $k = \{0, 1, 2\}$, we separate this as:

$$P((Poisson(\lambda) = 0) \cup (Poisson(\lambda) = 0)) \cup Poisson(\lambda) = 0))$$
$$= P(Poisson(\lambda) = 0) + P(Poisson(\lambda) = 1) + P(Poisson(\lambda) = 2)$$
$$= \frac{(\lambda)^0}{0!} * e^{-\lambda 0} + \frac{(\lambda)^1}{1!} * e^{-\lambda} + \frac{(\lambda)^2}{2!} * e^{-\lambda} = e^{-\lambda}[1 + \lambda + \lambda^2] = H_0$$

If the test statistic we find is larger than the test percentile of the χ^2 distribution, we fail to reject the null hypothesis. One could also the p-values to determine if the hypothesis fails or not. The p-value for this test is defined as the P[Theoretical distribution of T under the H_0 > observed value of the test statistic]. We fail to reject the H_0 if the p-value is above this certain threshold.

8. Hypothesis Test

In hypothesis testing, you test two hypotheses, the null hypothesis (H_0) and the alternative hypothesis (H_1). After testing, you either can conclude to accept or fail to reject the null hypothesis. The H_0 typically contains the answer to the scientific question of topic, and the purpose of this test is to see if the answer is true or not. Depending on the context of the scientific question, the test can have 4 different types. If one is testing towards a large positive/negative value, it is classified as a one-sided test. Furthermore, if one tests for both, then it is considered a two-sided test. The other two types of hypothesis tests come into play when testing for distributions. Concluding the null hypothesis is false classifies the test as complex, and specifying an alternative type of distribution classifies it as a simple test.

When performing hypothesis tests, one must consider Decision Errors, a flaw one can make when arriving at a conclusion. Failing to reject the H_0 and accepting it instead results in a Type II error. On the other hand, rejecting the null hypothesis when one was supposed to fail to reject results in a Type I error. Hypothesis tests are designed so that type I errors could be controlled through various options such as levels of a test and confidence level with dependency on the null hypothesis. This is due to the fact that society relies on the Type I error not occurring

during testing. Type II errors are represented by the power of a test, where power is defined by 1 - Type II error. Unlike the Type I error, Type II errors are dependent on the alternative hypothesis.

9. PCA

When analyzing data with a high dimension of features, we often need a dimension reduction strategy in order to make faster and more easily interpreted, simpler models. Principal component analysis is used to find which variables account for the most variability amongst each subject, keeping as much important information as possible while making our models and other exploration more simple.

The first step in this process is to standardize the variables. Because PCA works off of variances, we need to ensure that all variables are on comparable scales. This is done by taking each value and subtracting the mean for that feature, then dividing it by the standard deviation, as shown below.

$$Z = \frac{Value - Mean}{Standard\ Deviation}$$

The next step is to create a covariance matrix. This is done to see how the data varies from the mean relative to the other data. By finding information like highly correlated data, we can find characteristics such as redundant information, which can be rectified in our dimension reduction. The matrix will be symmetric and n x n where n is the number of features and it will contain all possible covariance pairs. For instance, if there were 3 features, the matrix would have the following structure.

$$\begin{pmatrix} x,x & x,y & x,z \\ y,x & y,y & y,z \\ z,x & z,y & z,z \end{pmatrix}$$

Once we have this matrix, we can compute the eigenvectors and eigenvalues which will give us our principal components. Eigenvectors and eigenvalues are calculated by solving for this equation.

$$AX = \lambda X$$

We solve this equation where A is our matrix, X is a non-zero eigenvector and λ is the eigenvalue. The number of eigenvectors and eigenvalues we acquire will be equal to the amount of features we have. Each eigenvector represents a direction, or location across the data, where variance is high. And each eigenvalue represents the weight of the corresponding eigenvector which tells us exactly how much variance there is in the vector's location. By keeping only the eigenvector and eigenvalue pairs that account for the most variability in the data and identifying the features that best correspond to these eigenvectors, we find the most important features in our dataset.

10. Generalized Estimating Equation

When analyzing longitudinal data, a regression approach we can consider using is the generalized estimated equation (GEE). There is the regression model used for the mean response of a general linear model and a within subject correlation model. With the within subject model, we obtain the covariance inverse used to obtain our coefficient estimates and the model-based standard error for these estimated coefficients.

The mean response is given by $\mu_{ij} = E[Y_{ij} \mid X_{ij}]$ which is given by a regression model. This is a function of the covariates in the data X_{ij}. The generalized linear model is chosen based on what type the outcome Y is:

Continuous Outcome	Count Outcome	Binary Outcome
$E[Y_{ij} \mid X_{ij}] = \mu_{ij}$	$E[Y_{ij} \mid X_{ij}] = \mu_{ij}$	$E[Y_{ij} = 1 \mid X_{ij}] = \mu_{ij}$
$\mu_{ij} = X_{ij}^T\beta$	$log(\mu_{ij}) = X_{ij}^T\beta$	$logit(\mu_{ij}) = X_{ij}^T\beta$

We must note that μ_{ij} does not condition on anything besides X_{ij}.

An alternative model we can use for a GEE is the covariance model. To do so, we first assume the form for variance that depends on μ_{ij} and outcomes:

Continuous Outcome	Count Outcome	Binary Outcome
$var[Y_{ij} \mid X_{ij}] = \sigma^2$	$var[Y_{ij} \mid X_{ij}] = \mu_{ij}$	$var[Y_{ij} \mid X_{ij}] = \mu_{ij}(1 - \mu_{ij})$

Based on the correlation within subject, a covariance model must be selected as well given by $corr(Y_{ij}, Y_{ik} \mid X_i)$.

Independence	Exchangeable	Auto-regressive	Unstructured
$corr(Y_{ij}, Y_{ik} \mid X_i) = 0$	$corr(Y_{ij}, Y_{ik} \mid X_i) = \rho$	$corr(Y_{ij}, Y_{ik} \mid X_i) = \rho^{\lvert j-k \rvert}$	$corr(Y_{ij}, Y_{ik} \mid X_i) = \rho_{jk}$

These models are semi-parametric as they only specify mean and variance which are the first two moments. Due to this, we must use $\widehat{\beta}$ to solve a GEE as we can't assume the distribution on ϵ which alters how we estimate β. We solve the equation:

$$\widehat{\beta} = \beta: \sum_{i=1}^{N} D_i^T * V_i^{-1}(Y_i - \mu_i) = 0$$

We estimate β using the defined variable $\mu_i = g^{-1}(X_i\beta)$ where g is a link function.

Works Cited

1. Ananthi, M., Vijayakumar, K. Stock market analysis using candlestick regression and market trend prediction (CKRM). J Ambient Intell Human Comput (2020). https://doi.org/10.1007/s12652-020-01892-5

2. X. Zhong, D. Enke. Forecasting daily stock market return using dimensionality reduction. Expert Systems with Applications, 67 (2017), pp. 126-139 https://doi.org/10.1016/j.eswa.2016.09.027

3. Abraham A., Nath B., Mahanti P.K. (2001) Hybrid Intelligent Systems for Stock Market Analysis. In: Alexandrov V.N., Dongarra J.J., Juliano B.A., Renner R.S., Tan C.J.K. (eds) Computational Science - ICCS 2001. ICCS 2001. Lecture Notes in Computer Science, vol 2074. Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-45718-6_38