Predicting the Resale Value of Sneakers

Winston Zhong

Problem Statement

The sneaker resale industry has been growing bigger and bigger over time. For a sneaker reseller, the hardest part of the job is buying sneakers at retail price. With the increasing usage of bots, it has become very competitive to buy sneakers from retail, but this is a different subject that I won't be getting into here. The resale price of a sneaker can vary depending on the sizing and time in which the sneaker is sold. So what size should you purchase, if you are lucky enough to get one? And should you sell them right when you get them or should you wait?

StockX is one of the biggest online resale market platforms for many different products. Sneakers and apparel are what it is most popularly used for. I was able to find a dataset from StockX containing sales for Yeezy 350 and Nike x Off-White sneakers that occurred from 09/01/2017 to 02/12/2019. Both of these are known to have very limited releases especially during that time span. Using this dataset, I developed a model to help predict sneaker resale prices and the effects of different variables on the resale price.

Data Wrangling

The dataset had 8 columns: Order Date, Brand, Sneaker Name, Sale Price, Retail Price, Release Date, Shoe Size, and Buyer Region. I had three string type columns, so I removed any leading or trailing spaces that were in those entries. For the two date columns, I changed them to datetime data types in order to find the difference between the two columns later. For the two price columns, I had to remove the dollar signs and commas and then change them into integer data types. This will help me later to create a return on investment column.

With almost every sneaker, there are always some select people that are able to obtain the sneaker earlier than the official release date. Because of this, there are people who resell them before the release date for a lot more money. For the purpose of this study, I decided to remove those entries which we can call outliers. To do that, I set my new dataset as the set where the Order Date is greater than the Release Date. This removed 5,601 of the original 99,956 data entries.

I want to create two new columns for the dataset in order to get a better understanding. I created a Days after release column and a Return on investment (ROI) column. Days after release was calculated by subtracting the release date column from the order date column. This gives us the number of days after the release of the sneaker that the sale was made. ROI was calculated by subtracting retail price from sale price and then dividing that by retail price, multiplied by 100.

Exploratory Data Analysis

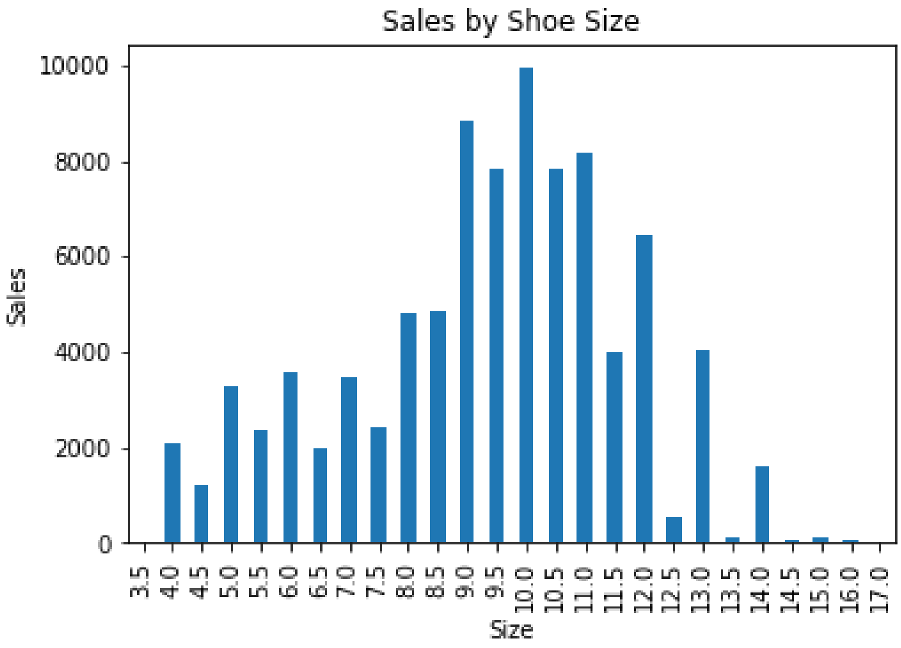

Figure 1

I plotted this graph in Figure 1 to show the number of sales for each shoe size. This gives us an idea of the demand for each shoe size. We can see that size 9 to size 11 have the highest demand. The lack of sales for the larger sizes could also be due to the lack of supply for those sizes.

Something else I wanted to look at was whether or not there was a significant difference in the ROI between the two brands, Yeezy and Nike x Off-White. I did this by doing a hypothesis test on the difference between the two means.

Null Hypothesis: Mean ROI for Yeezy is equal to mean ROI for Nike x Off-White
Alternative Hypothesis: Mean ROI for Yeezy is not equal to mean ROI for Nike x Off-White

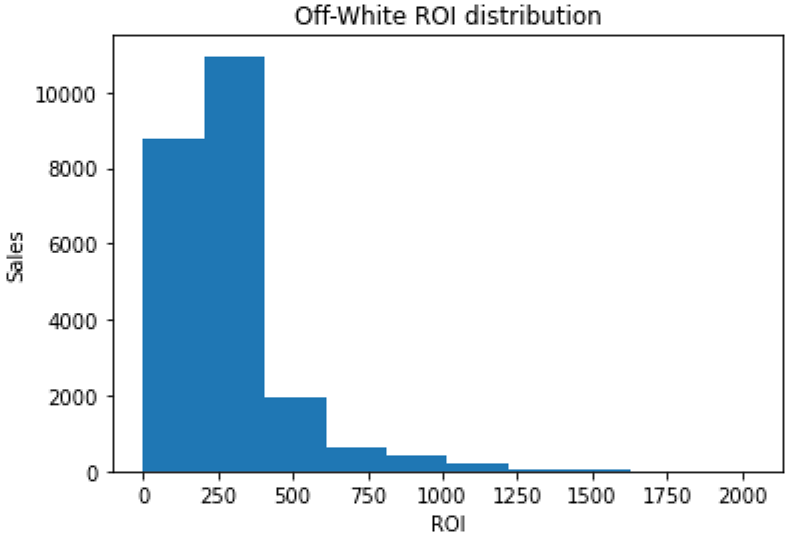

Figure 2

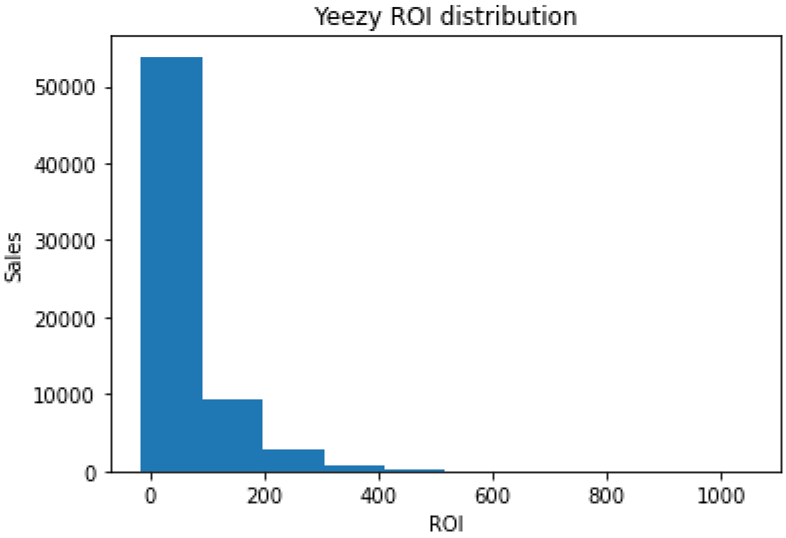

Figure 3

Graphing out the ROI distribution for the two brands seen on Figure 2 and Figure 3 shows that the distributions are not normal. This led me to perform a permutation test, which gave me the graph below of the distribution of difference in means of ROI between the two brands.

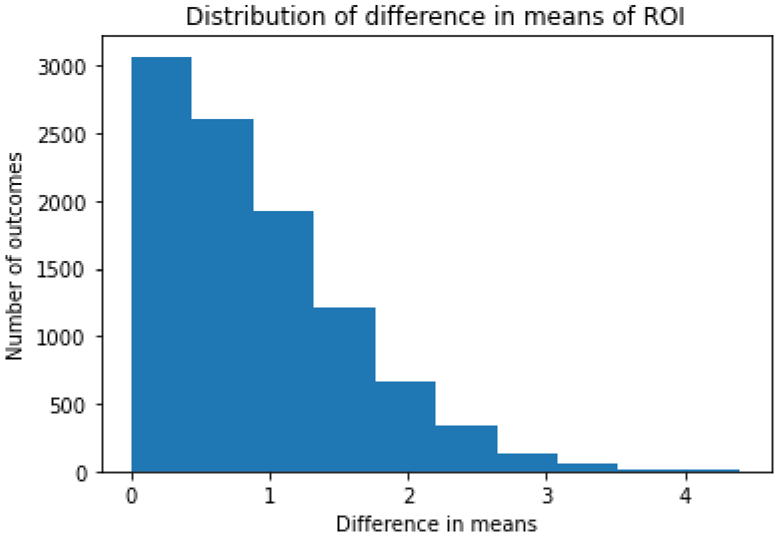

Figure 4

I found the p-value to be 0, which means we reject the null hypothesis. We conclude that the two brands have significantly different ROI, with Nike x Off-White having significantly higher ROI than Yeezy.

Next, I wanted to make more changes to my dataset. I changed the brand column into a binary variable, with Yeezy represented by 0 and Nike x Off-White represented by 1. Making this a numerical column will allow me to use it later in the modeling step. The days after release column also needed to be a numerical column. I removed “ days” from each entry and converted the entries to an integer type. Then I dropped the rest of them non-numerical columns, as well as the sale price column. This leaves us with just 5 columns: Brand, Retail Price, Shoe Size, Days after release, and ROI.

Pre-processing and Training Data Development

Here I want to pre-process my data and get it ready for the modeling step. I first round the ROI column to two decimal places just to make it easier to read. Then I binned the days after release column into four categories: 2 weeks, 2 months, 1 year, and over 1 year. Selling a sneaker 2 weeks after release is like selling it right when you get it because it takes 1 to 2 weeks for it to be delivered. The time period between 2 weeks and 2 months is to see what resale prices are like after the immediate resales have been processed. After 2 months, I wanted to see what resale prices were like within 1 year and also over 1 year. In order to do a train test split on my data, I need all my columns to be numerical so I did one-hot encoding on the days after release column. This created four new columns as binary variables while removing the categorical days after release column.

Now, I do a train test split on my dataset with the test size equal to 0.2. The ROI column is our dependent variable and the rest are our independent variables. I used standard scaler and min max scaler on the X train and test datasets to create 2 additional datasets to test our model on. I also did a log transform on the y train and test datasets to see if we would get better modeling results.

Modeling

To find the best model for my dataset, I tested two different modeling methods: linear regression and random forest regressor. I determined each model’s accuracy by calculating their R2 score, mean absolute error, and mean squared error. Using the original train test split datasets, I found out that the random forest regressor was the better modeling method for my data. I got a R2 score of 0.46, mean absolute error of 63, and mean squared error of 11781 for the linear regression model, while for the random forest regressor model, I got a R2 score of 0.67, mean absolute error of 50, and a mean square error of 7116. This shows a huge improvement going from the linear regression model to the random forest regressor model.

With some outliers in my y dataset, I wanted to try modeling my data with the y dataset log transformed. I got a worse result from this, with a R2 score of 0.57. I then tried using the scaled X datasets for modeling on random forest and I achieved similar results as my original random forest model. So I used cross validation to assess the R2 scores of each model. This also gave me R2 scores with very similar ranges.

I wanted to optimize the hyperparameters of my model, so I used randomized search cross validation on the random forest regressor model. As a result, I found these to be the optimal hyperparameters:

```
{'n_estimators': 26,
 'min_samples_split': 5,
 'min_samples_leaf': 4,
 'max_features': 'auto',
 'max_depth': None,
 'bootstrap': True}
```

These hyperparameters give me my best model with a mean R2 score of 0.68 and a R2 score standard deviation of 0.0083. It also gives me a mean absolute error of 50, meaning that the model can predict a sneaker’s ROI and be off by about 50 percentage points on average.

Conclusion

Now I can use the model to make predictions and see how sneaker resellers can maximize their profits. First, I used the model to predict when is the best time to sell. I calculated the median ROI for each time period, and the time period with the highest median ROI was between 2 months and 1 year with a median ROI of 89.35%. The next highest median ROI was 59.01%, which belonged to sneakers sold within the first two weeks of the release. This shows a 30% increase in profit if a sneaker reseller waits a few months to sell. I created violin plots for each time period shown in Figure 5 and you can see that the one year plot shows more entries with higher ROI’s. Note that the scale on each plot is different.

Next, I wanted to find out which shoe size would get the highest return. I first categorized the shoe sizes into three classes: small, medium, and large. Small sizes were less than 8, medium sizes were from 8 to 11.5, and large were 12 and higher. I found that the highest median ROI was 63.98% which were for the large sizes. Medium sizes came in a close second with a median ROI of 61.85% and small sizes were much less with a 53.85% median ROI. I also created violin plots for this shown in Figure 6 for visualization.

In Figure 7, I created a bar chart to show the median ROI for each shoe size and we can see that the two largest sizes, 15 and 16, give a significantly higher ROI. This is most likely due to the supply of sneakers in those sizes. So for sneaker resellers it will be very difficult to get those sizes, but if they do, they will make significantly more profit. The rest of the figures, Figure 8 to 11, just show the median ROI for each size based on the time period the sneaker was sold. These are good visualizations to see how the ROI changes over time for each size.

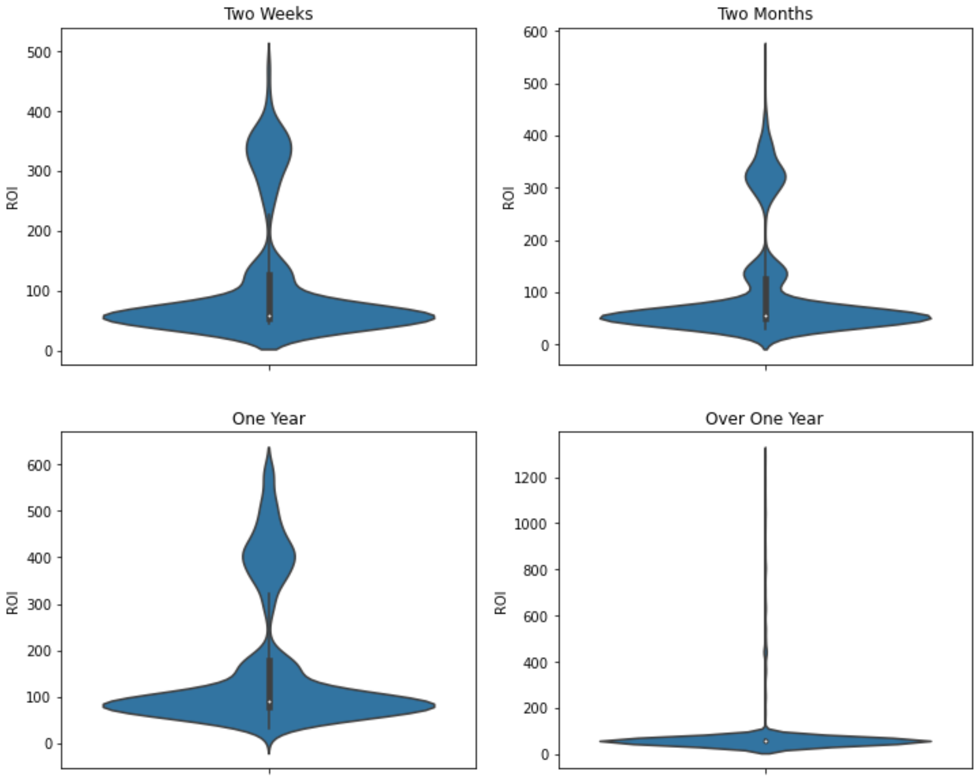

Figure 5

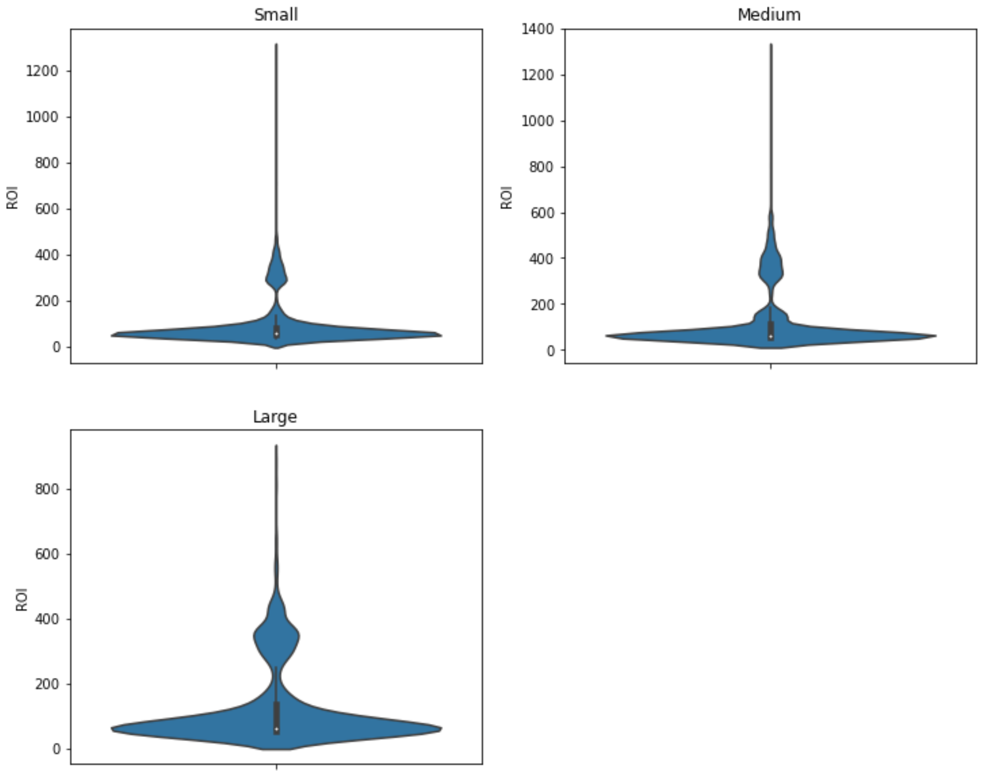

Figure 6

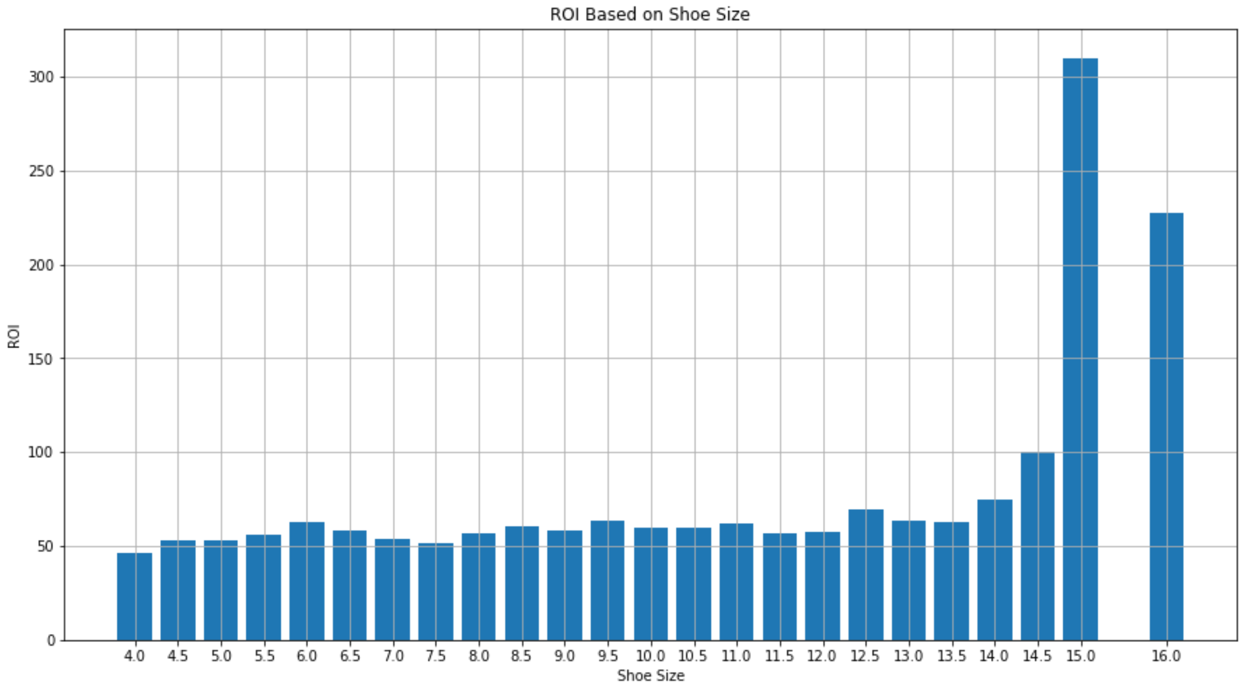

Figure 7

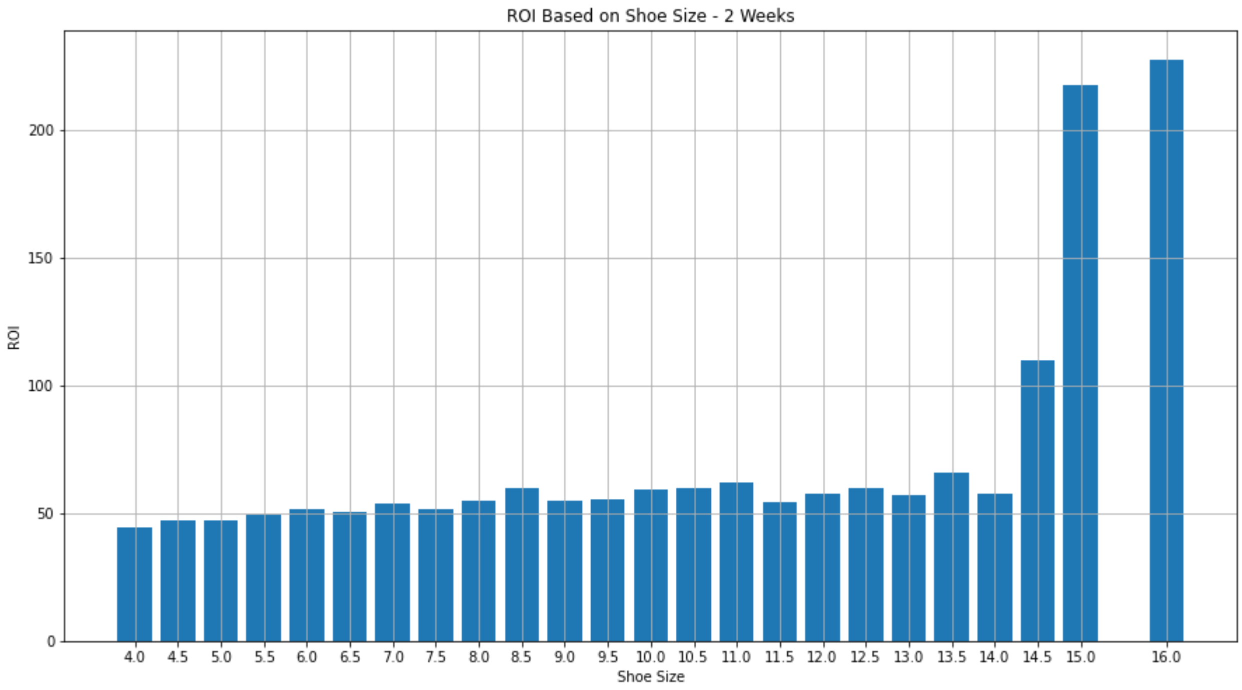

Figure 8

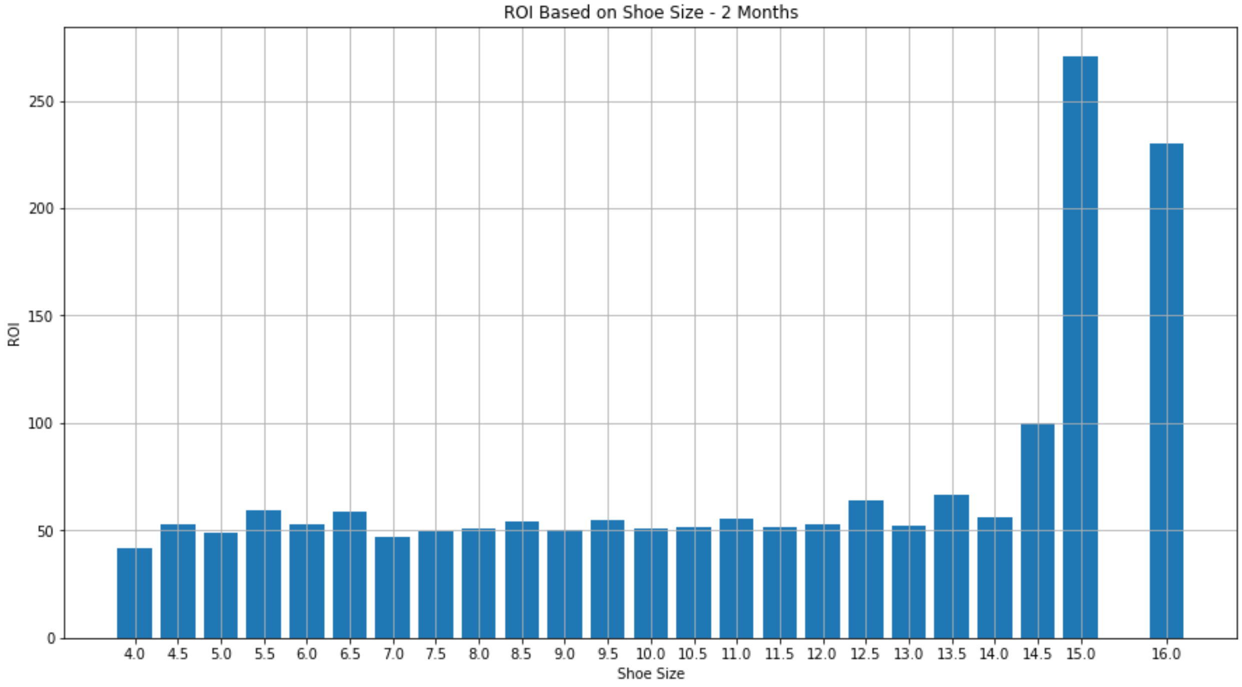

Figure 9

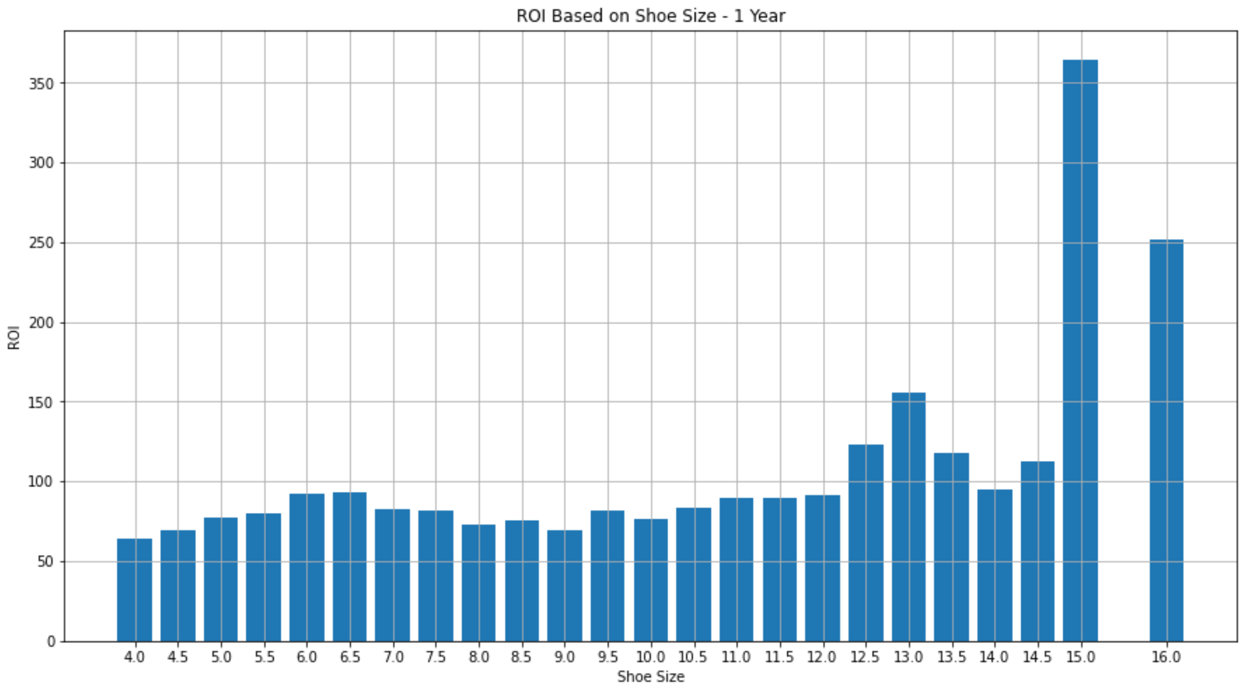

Figure 10

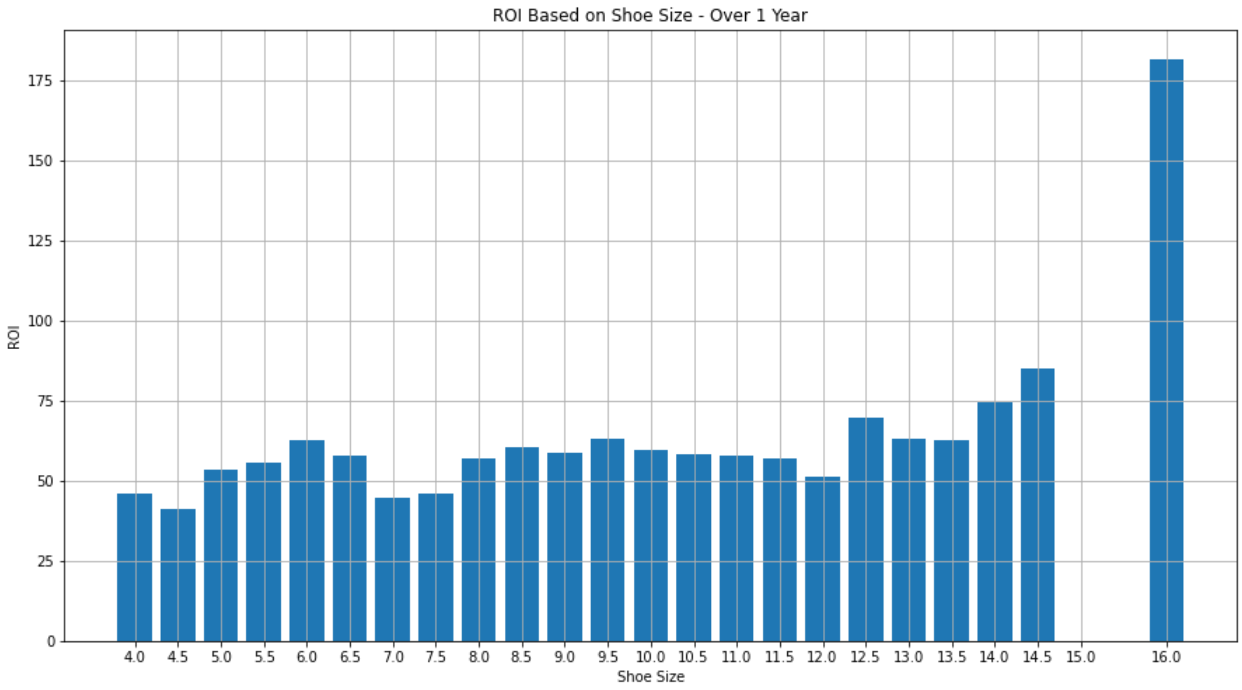

Figure 11

Further Research

The supply of a sneaker has a lot to do with the resale value of a sneaker. We know that all these sneakers are limited, but some are more limited than others. From my experience, I know that the largest sizes have the lowest supply, which we can see clearly affected the resale value. This kind of information would greatly improve our model and help us create better predictions. Another thing I would like to add is the resale value of the same sneakers on other platforms. There are many other ways to resell sneakers, StockX just happens to be the most popular platform for sneaker reselling. It would be interesting to see how different resale prices are based on the platform the transaction is made.