Starbucks Portal

Things Starbucks's Internal Users Should Know

Hassan Mortagy HAM2149
Aion Feehan AF2981
Jidapa Limpiwattakee JL4751

Introduction:

Starbucks is an international coffeeshop that is arguably the most famous worldwide. It started as a roaster and retailer of whole bean and ground coffee, tea and spices with a single store in Seattle's Pike Place Market in 1971. The company now operates more than 24,000 retail stores in 70 countries. Starbucks recently started leveraging analytics and data science to improve and optimize their operations. The goal of our project is to contribute to Starbucks's analytics by developing tools providing various pieces of information and services, that would be of benefit for Starbucks's internal users.

Part I: Starbucks Geographical Analysis

Using the Starbucks Stores Data (we provide a diagram of all datasets used in our submittals), we were able to determine that the city that has the highest number of the stores is in Shanghai City. On the other hand, the city with the least number of Stores is 's-Hertogenbosch'. As for countries, the United States of America has the most Starbucks Stores in the world; whereas, Andorra has only one store. We created a function where you can input the name of the city or the country and get the number of the stores at that place. Furthermore, we wanted to further analyze the locations of all stores to get a grasp of Starbucks' strategy. We plotted the Starbucks worldwide store concentration (see appendix), which got us wondering what are the two Starbucks locations that are closest in proximity to one another. The results are astounding. There are 9,398 Starbucks stores that are almost 0 km apart. In addition, just for the fun of it we also wanted to know where the furthest Starbucks from where we are right now (our classroom) is. We created a function that does exactly that, and the results are placed in the appendix. To complement our geographical analysis, we developed a function that gives directions to the nearest Starbucks for a given user, and in New York City completes that with data on pedestrian traffic in that area, relative to the average in New York in that timeframe (morning or afternoon). It turns out that around campus is relatively quiet (unsurprisingly the vast majority of traffic is concentrated downtown). We lastly wanted to see just how far away from a Starbucks store you can get on the planet, and wrote a script to estimate the point on the earth that's furthest away from their locations. This point is located on the artic circle north of Russia (see appendix).

Part II: Worldwide Store analysis:

We wanted to take our geographical analysis further and look at how Starbucks has expansed worldwide, and what factors determine how many stores they have in a given location. We started with our kaggle base dataset which is a directory of all Starbucks stores worldwide. We then enhanced this dataset by web scraping country specific socio-economic and demographic factors like GDP, median age, land size, gross national income (GNI), and others, and combined them with our original data set. We then ran a regression, using the number of stores as our dependant variable, and the country specific variables as our predictors, to determine which factors affect Starbucks's expansion worldwide. We found that the number of stores could be explained, relatively well, using a linear model and the country specific variables, as the model R^2 was 0.78. We also found that the country with the highest Starbucks density, given by Starbucks per million inhabitants, is the USA as expected (See appendix). We therefore wanted to do a deeper analysis and looked at the state of California and all its cities to know how does Starbucks determine its stores in the US specifically. California is a very large state with a lot of different cities and it would therefore be representative, but maybe a possible extension for this project would be analyzing other states as well. The results are provided below:

Table 2: Important factors in determining No. of stores worldwide

Number	Predictor
1	Population
2	Yearly Change
3	Migrants (net)
4	World Share
5	GDP per capita (US$)
6	GNI per capita (US$)

Table 2: Important factors that determine No. of stores in California cities

Number	Predictor
1	Median Household Income
2	Population
3	Percentage of white population
4	Land Area (Sq. miles)

Part III: Starbucks Stock Price and Revenue Predictions:

Starbucks stock (SBUX) has one of the highest earnings per share (EPS) and trading volumes in the food and beverage industry. Since our tools are intended for internal use, it is therefore very important to have accurate stock predictions for business development and investment purposes. We have thus developed a function that utilizes various machine learning models to predict the SBUX closing price for the next business day. The function first web-scraps the closing stock price of Starbucks for the past two years from the date the function is run, and then transform the data into a time-series using the price of the previous 10 days as predictors. We then perform LASSO regression, and support vector regression with cross validation to obtain the best performing model. The best model is then used to make the price prediction for the following business day taking weekends and holidays into account. Our algorithm is performing really well and on average obtains a mean-squared error of 0.4, as shown in the figure below (see appendix for more details on cross validation and model evaluation results).

Furthermore, we also wanted to develop a tool that determines the projected revenue for the next quarter. This would be very helpful because it would help the executive team set future realistic goals, and also determine their ability to achieve future goals and whether they are currently following historical trends. We approach this problem by web scraping revenue data from the web and then feeding this data into a model that projects the revenue for the next quarter. We were able to obtain revenue data from 2005 only, and the challenge here is that even at the quarter level the revenue data is not large. This reduced number of data points did not enable us to fit advanced predictive models, so we just used a simple linear regression, using the revenue of the previous four quarters as our predictors. We used the previous four specifically because found four to be the optimal number using cross validation. The model is actually working well and has an $R^2 = 0.93$, and this was because the revenue data has a strong linear trend which played in our favor (see figure below).

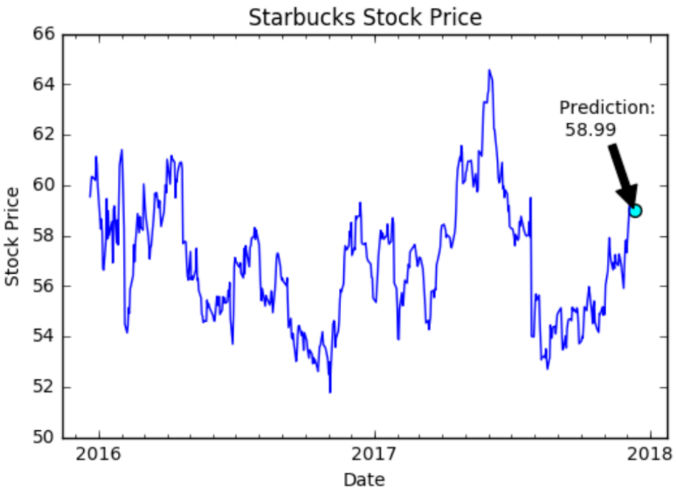


Figure 1: Starbucks closing stock price prediction for 12/11/2017

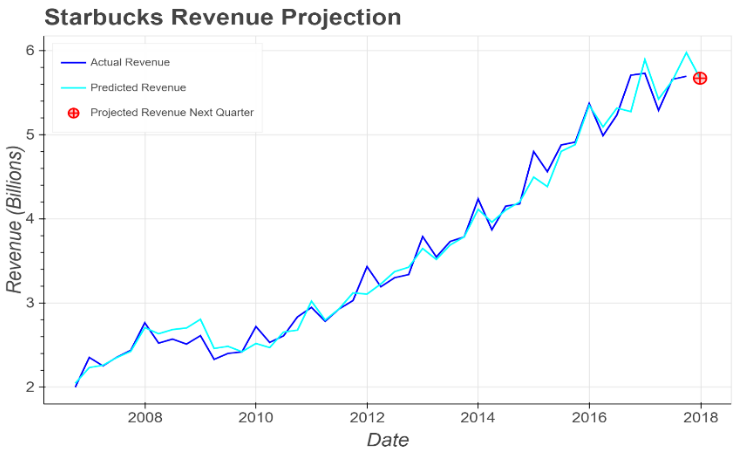


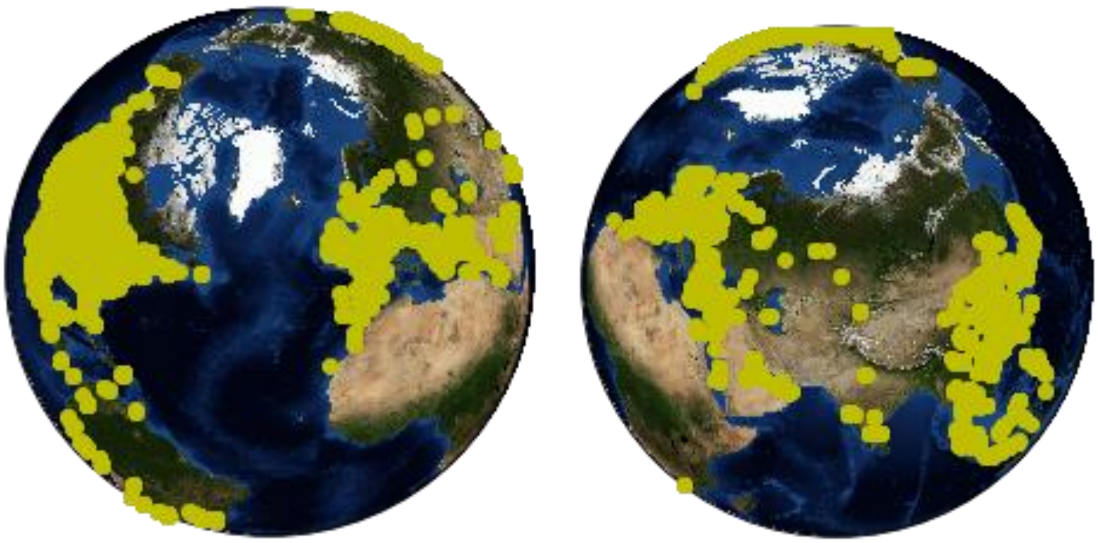
Figure 1: Starbucks quarterly revenue projection

Appendix

1. World Wide Store analysis:

a) Visual of Starbucks worldwide store concentration:

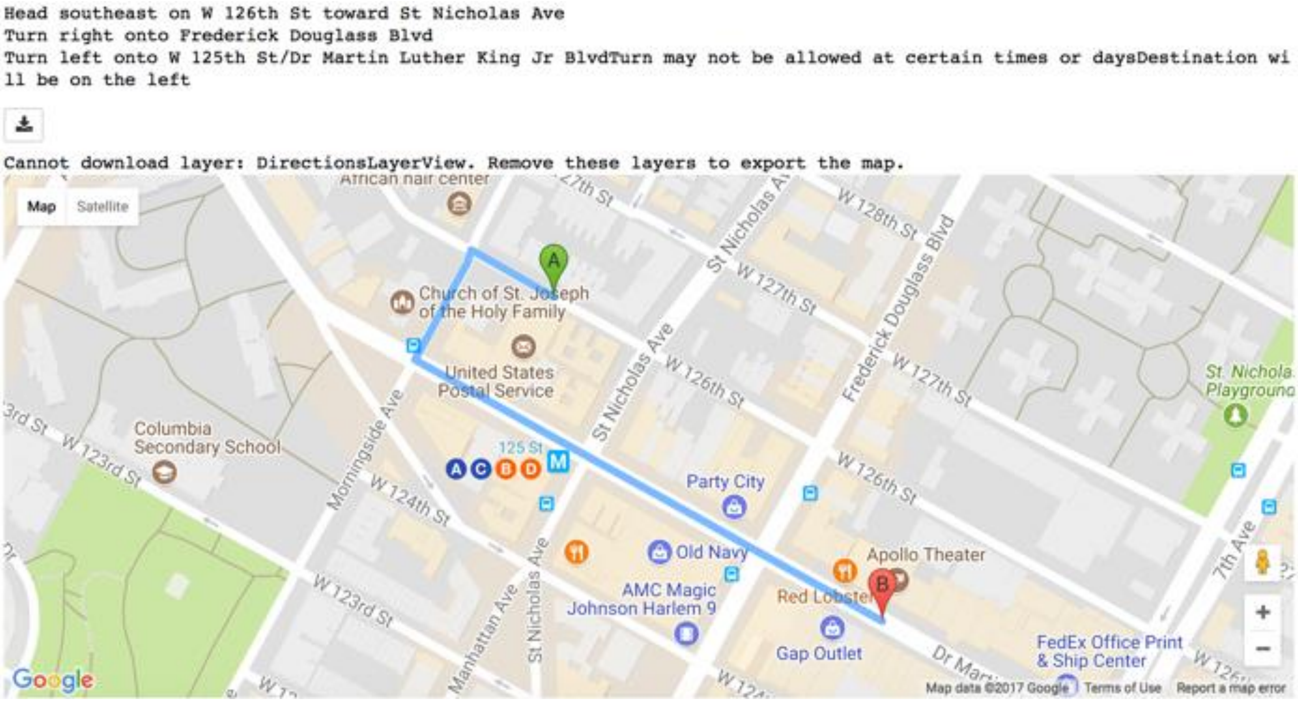
Using matplotlib, we created the plots below where a green dot represents a Starbucks store. The maps show a concentrated number of Starbucks stores in America and East Asia.



b) Sample output of function that tells you where the farthest Starbucks is from a given location:

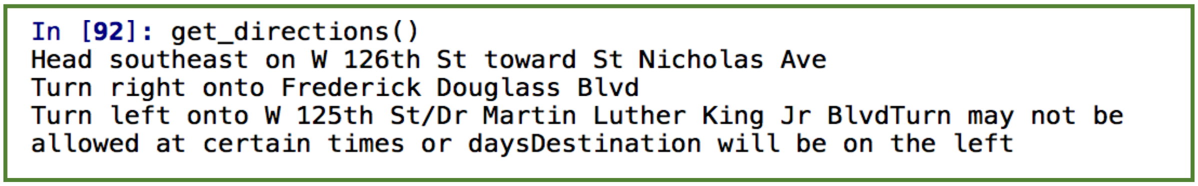
```
Brand                                Starbucks
Store Name                              Napier
Ownership Type                        Licensed
Street Address             Emerson St, Napier
City                                    Napier
State/Province                               N
Country                                     NZ
Postcode                                   NaN
Phone Number                      06 834 2447
Timezone            GMT+12:00 Pacific/Auckland
Longitude                               176.92
Latitude                                -39.49
Name: 33842-97882, dtype: object
```

c) Sample output showing where the nearest Starbucks is, and how to get there:



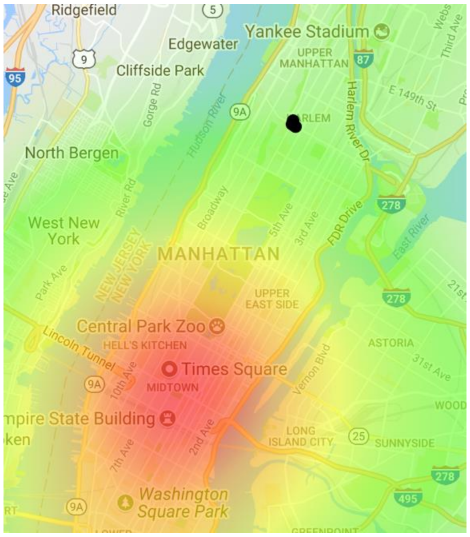
```
Head southeast on W 126th St toward St Nicholas Ave
Turn right onto Frederick Douglass Blvd
Turn left onto W 125th St/Dr Martin Luther King Jr BlvdTurn may not be allowed at certain times or daysDestination wi
ll be on the left
```

Cannot download layer: DirectionsLayerView. Remove these layers to export the map.

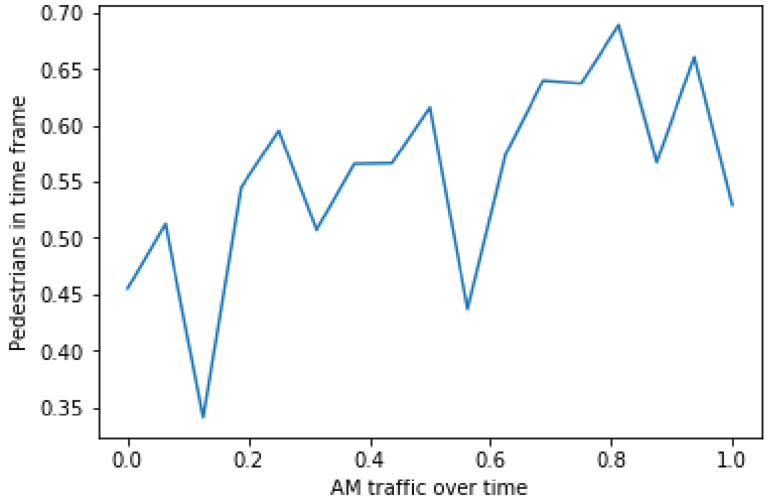


```
In [92]: get_directions()
Head southeast on W 126th St toward St Nicholas Ave
Turn right onto Frederick Douglass Blvd
Turn left onto W 125th St/Dr Martin Luther King Jr BlvdTurn may not be
allowed at certain times or daysDestination will be on the left
```

For locations in the city of New York it overlays the map with pedestrian traffic density estimates for that timeframe (AM or PM), as follows:

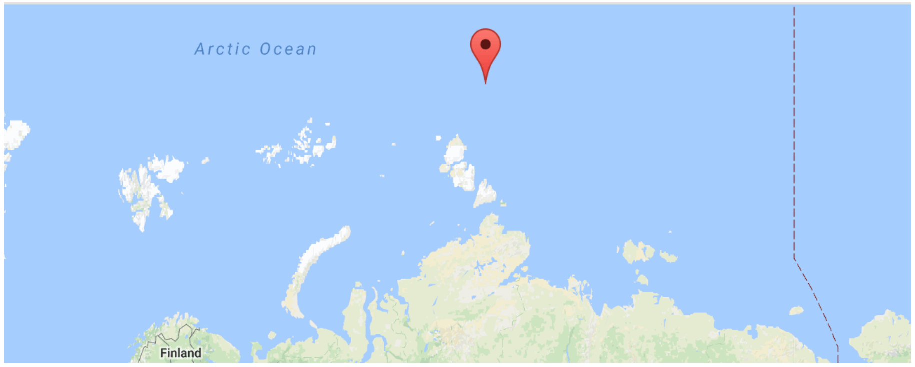


d) Sample output of the relative traffic at the nearest Starbucks over time (again for locations in New York City). It turns out that ours is relatively quiet.

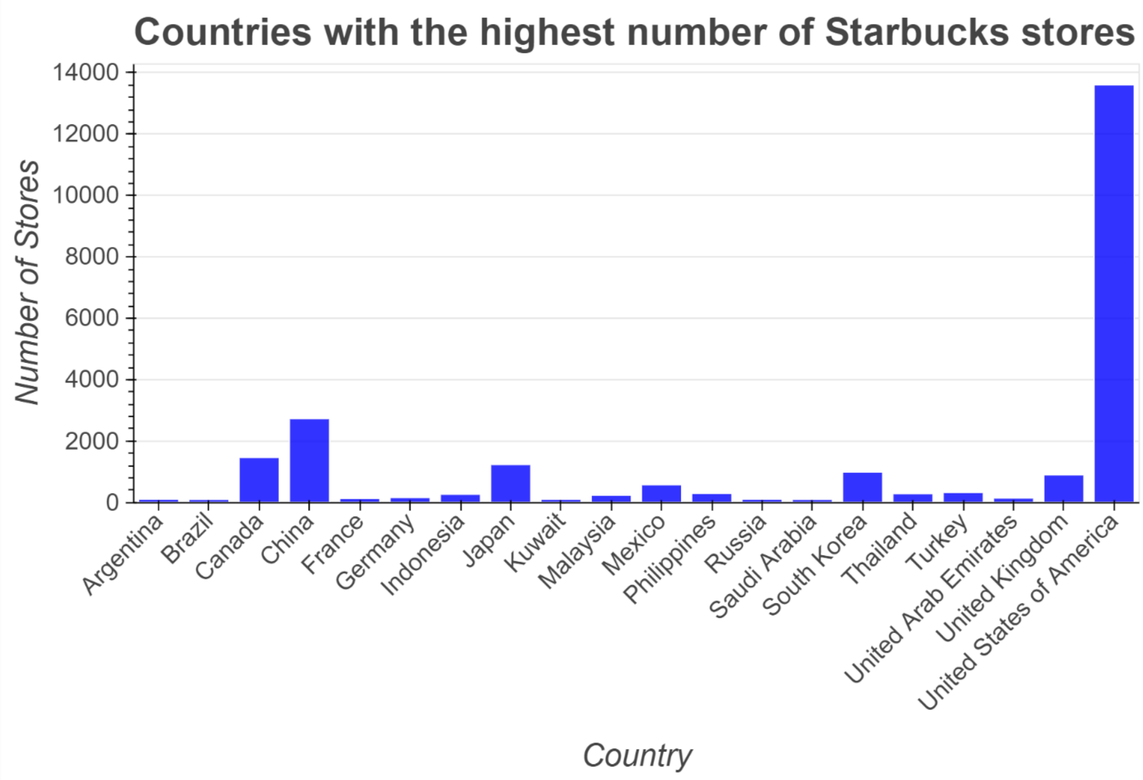


e) The farthest you can get from Starbucks on the globe :

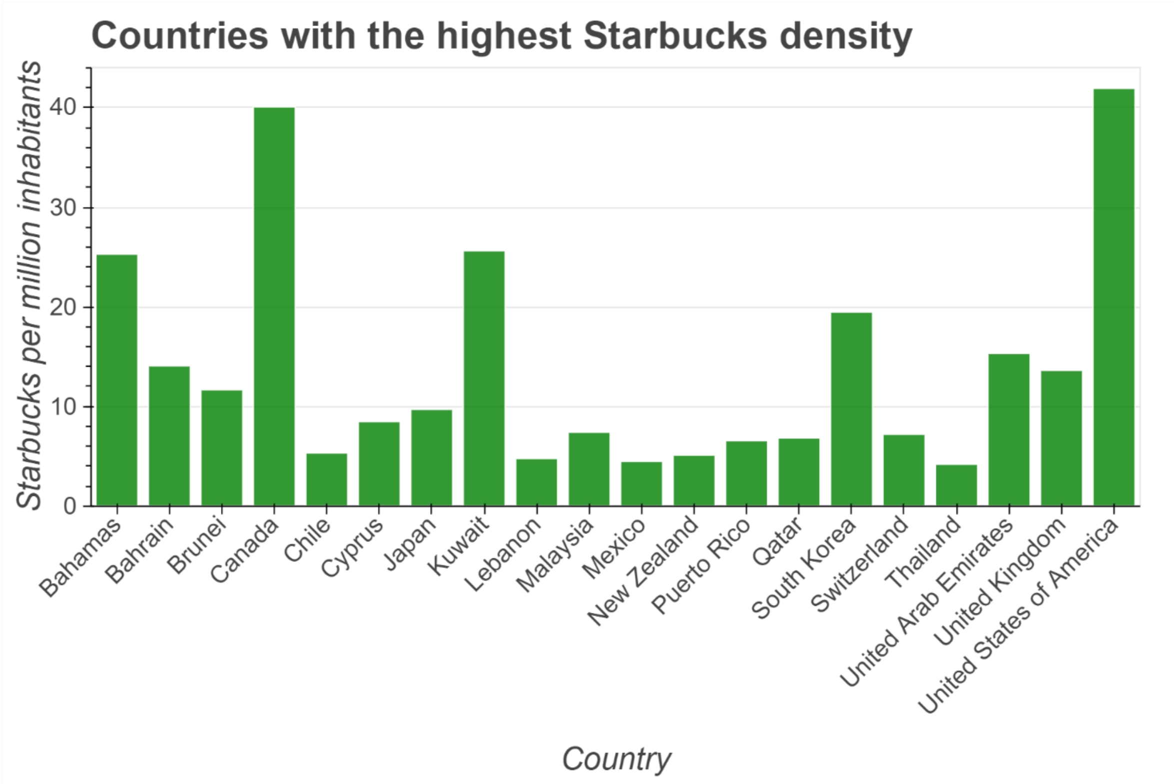
Apparently, the arctic circle north of Russia is the farthest you can get, over 5300 km away from the nearest Starbucks location. The coordinates of this point using latitude and longitude are displayed in the image below. We want to emphasize that this may not be an absolute optimal solution, because it was based on algorithms we developed as this was a highly complex combinatorial problem.



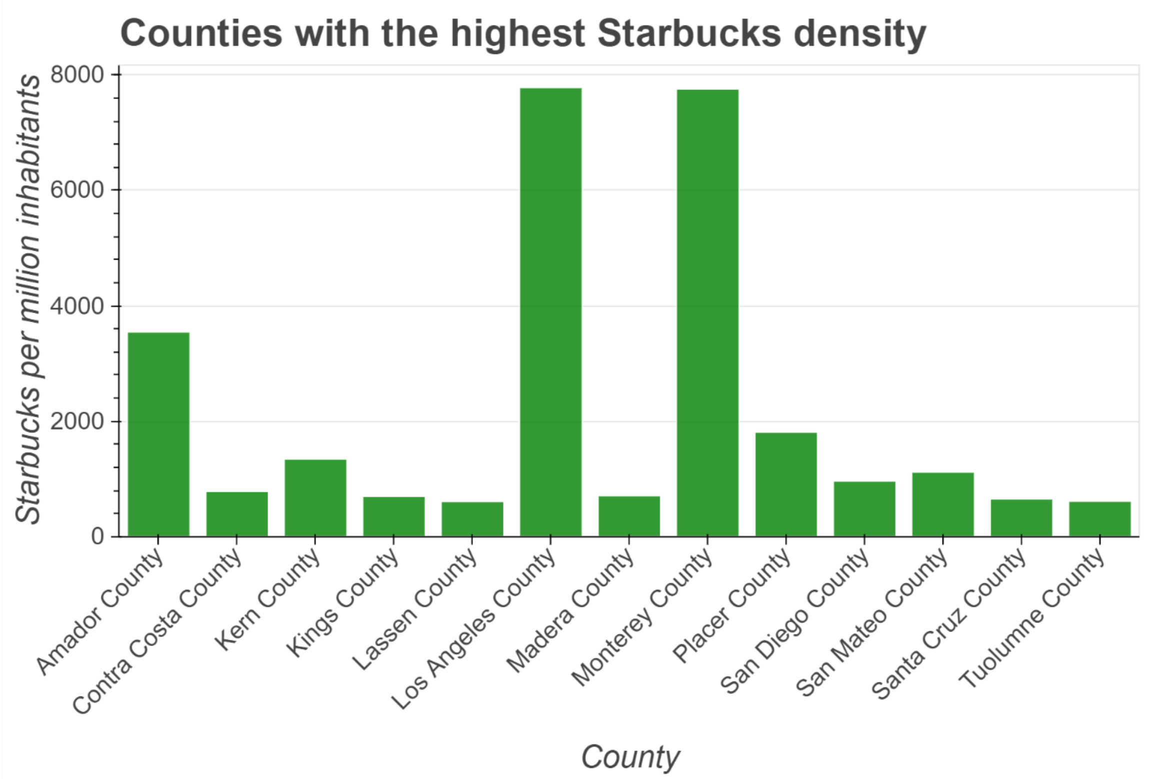
f) The top 20 countries with the highest number of Starbucks stores:



g) The top 20 countries with the highest Starbucks density:

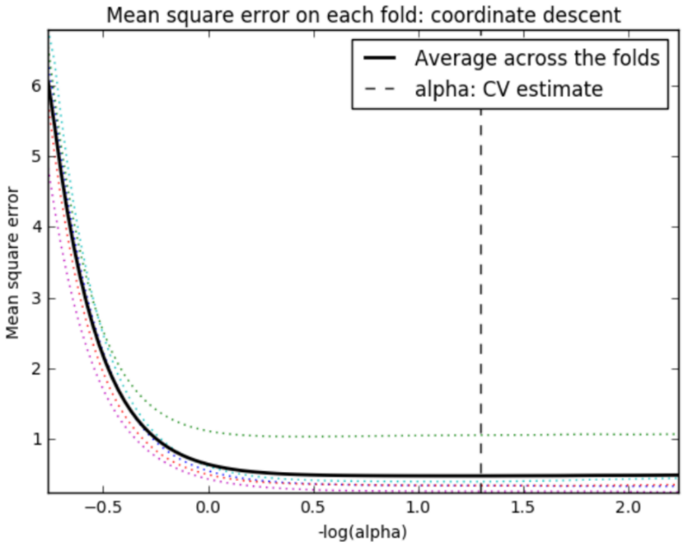


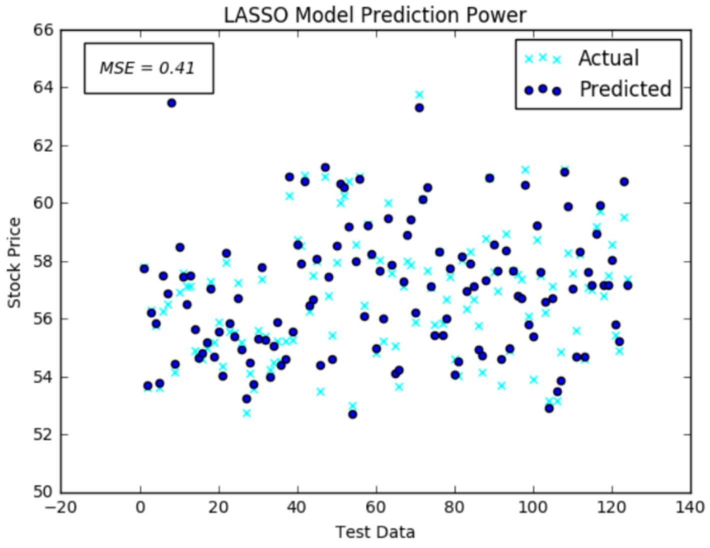
h) The top 20 counties in California with the highest Starbucks density



2. LASSO Regression:

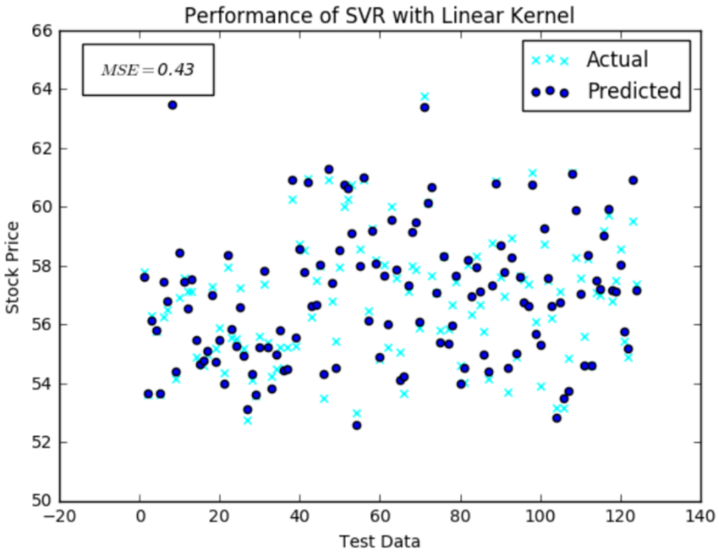
The graphs provided below are for the LASSO Regression Model. The first graph on the left is the output of the cross validation done to determine the optimal Lagrangean penalty for the L_1-norm of the coefficients. The second graph on the right shows the prediction power of the LASSO regression model as well the Mean Squared Error (MSE) of the model on the test data.

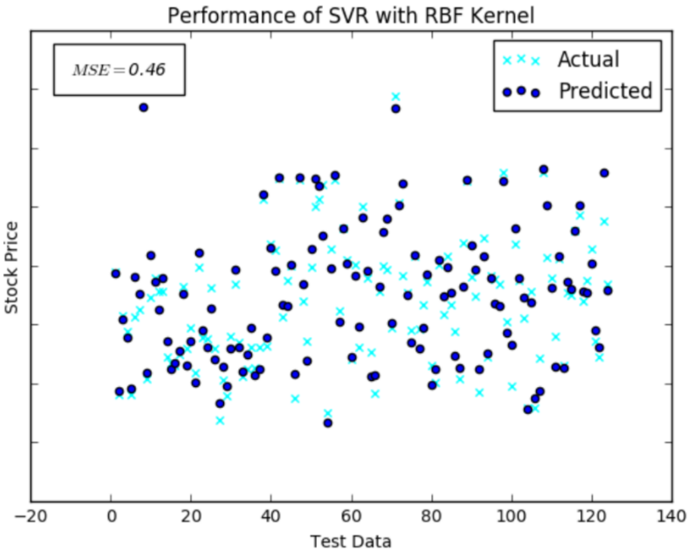




3. Support vector Regression:

For the Support Vector Regression (SVR) Model, we considered linear, radial, sigmoid and polynomial kernels. The sigmoid and polynomial kernels consistently performed very badly, so we excluded them from the model. We determined the optimal parameters for the SVR model for both the linear and radial kernels, namely the cost parameter, epsilon error and gamma (for radial kernel) using a grid search and cross validation. The graphs below provide the prediction accuracy for both kernels. We can actually see that in this instance of time, the LASSO performs better than the SVR model, but we want to emphasize that this is not always the case.


Performance of SVR with Linear Kernel
$MSE = 0.43$
Actual
Predicted
Stock Price
Test Data


Performance of SVR with RBF Kernel
$MSE = 0.46$
Actual
Predicted
Stock Price
Test Data