Sales-Demand Forecasting for Shoe Sales

Introduction

Determining inventory levels are critical for any retail and online operation. If you are out of stock of a unit that a customer wants, it is a missed opportunity that is most likely lost. Conversely, to have items in inventory that no one will purchase is wasted capital. Therefore, it is critical that any business sets its inventory levels correctly and prices its products appropriately. This is where demand forecasting comes in. By using predictive analysis of historical data, it is possible to estimate and predict customers' future demand for products.

Methodology

The methodology of this project is to find sales data for a company over a sufficient amount of time and utilize machine learning to predict sales over the next six months. Several time-series analysis libraries, such as Arima, Sarimax, FBProphet, XGBoost, and TensorFlow, are appropriate for forecasting. We will apply multiple programs to the data, measure the accuracy of each, and determine the most appropriate one to use. We may even find that an ensemble method works the best.

Data

The raw data set for this project was retrieved from towardsdatascience.com and linked to an AWS S3 bucket. It is a collection of shoe sales over three years, representing 18 shops in 4 countries, including The United States, Canada, the United Kingdom, and Germany. The data set contains 14 features and 14,971 customer sales.

```
df_shoes.head()
```

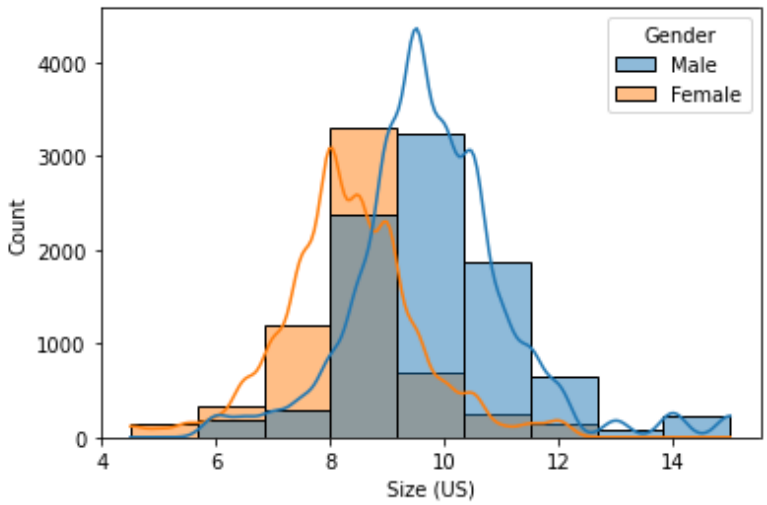
	invoice_no	date	country	product_id	shop	gender	size_us	size_europe	size_uk	unit_price	discount	year	month	sale_price
0	52389.0	1/1/2014	United Kingdom	2152.0	UK2	Male	11.0	44	10.5	159.0	0.0	2014.0	1.0	159.0
1	52390.0	1/1/2014	United States	2230.0	US15	Male	11.5	44-45	11.0	199.0	0.2	2014.0	1.0	159.2
2	52391.0	1/1/2014	Canada	2160.0	CAN7	Male	9.5	42-43	9.0	149.0	0.2	2014.0	1.0	119.2
3	52392.0	1/1/2014	United States	2234.0	US6	Female	9.5	40	7.5	159.0	0.0	2014.0	1.0	159.0
4	52393.0	1/1/2014	United Kingdom	2222.0	UK4	Female	9.0	39-40	7.0	159.0	0.0	2014.0	1.0	159.0

Exploratory Data Analysis

When looking at the data, some of the relationships which can be determined are as follows:

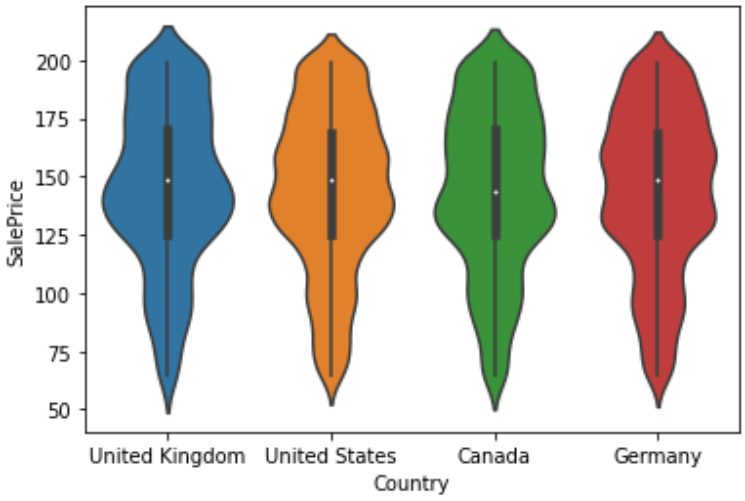
Shoe size follows a normal distribution:

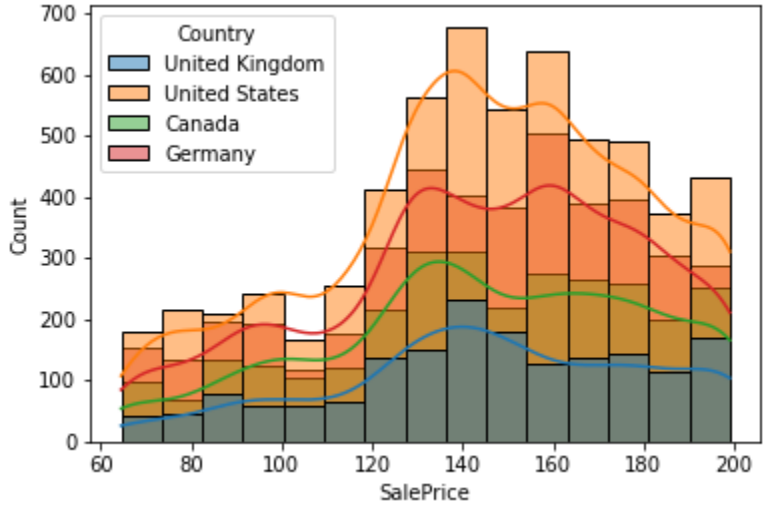
```
sns.histplot(x='Size (US)', data=df, bins=9, hue='Gender', kde=True);
```

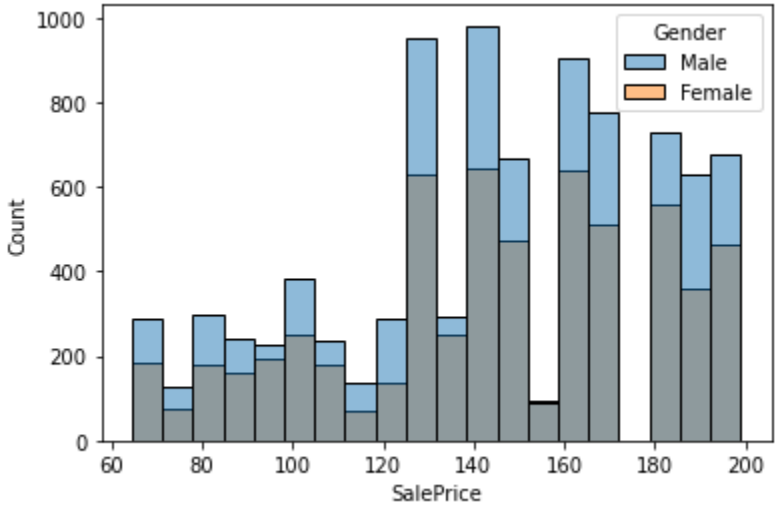


This means when ordering inventory, it should follow this ratio. In Men's, for every three pairs of size 10(US) shoes, two pairs of size 9 and 11 should be ordered and 1 pair of size 12 and size 8. Subsequently in Women's pairs the same ratio should exist with size 9 being the median.

Sale price seems to be similar across all countries:



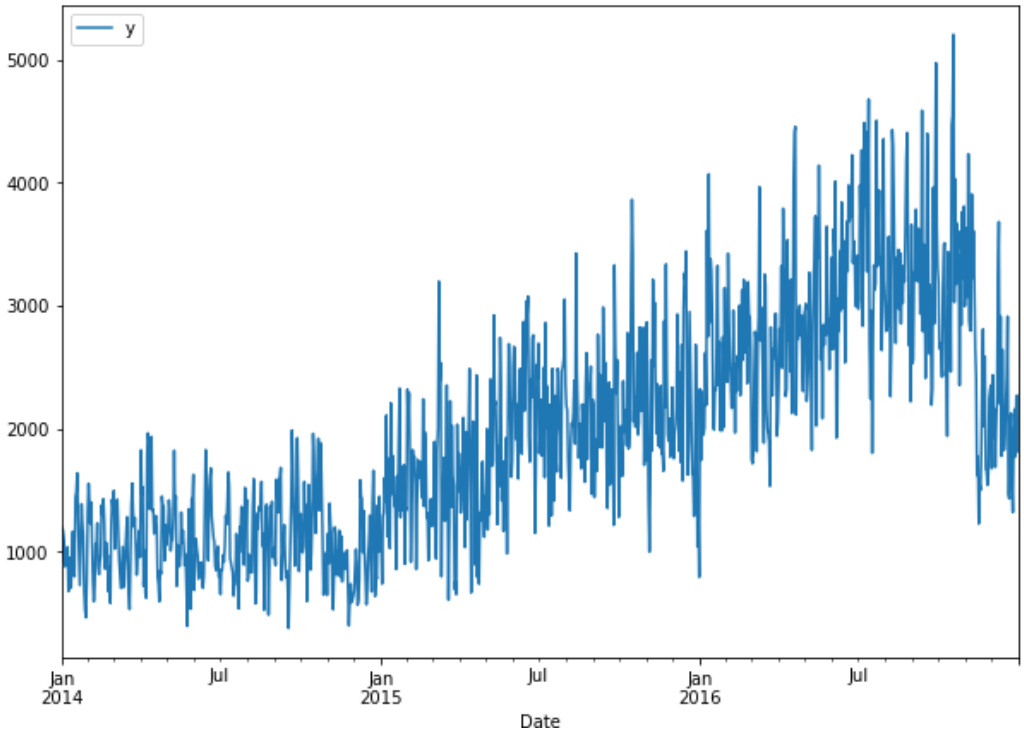




Sales price seems to follow a normal distribution in each country with the exception of Germany which show a definitive bimodal distribution and possibly a slight bimodal variance in the United States. It also appears from the data that men purchase more high price shoes than women.

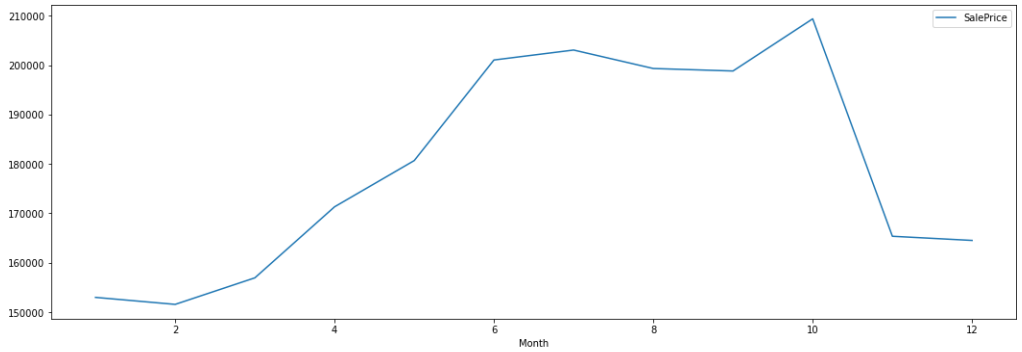
Preprocessing

The first step in time series analysis is to do a series decomposition looking at trend, seasonality, and noise. Plotting the current sales data over time makes it difficult to tell at a glance any of these beyond a general trend.

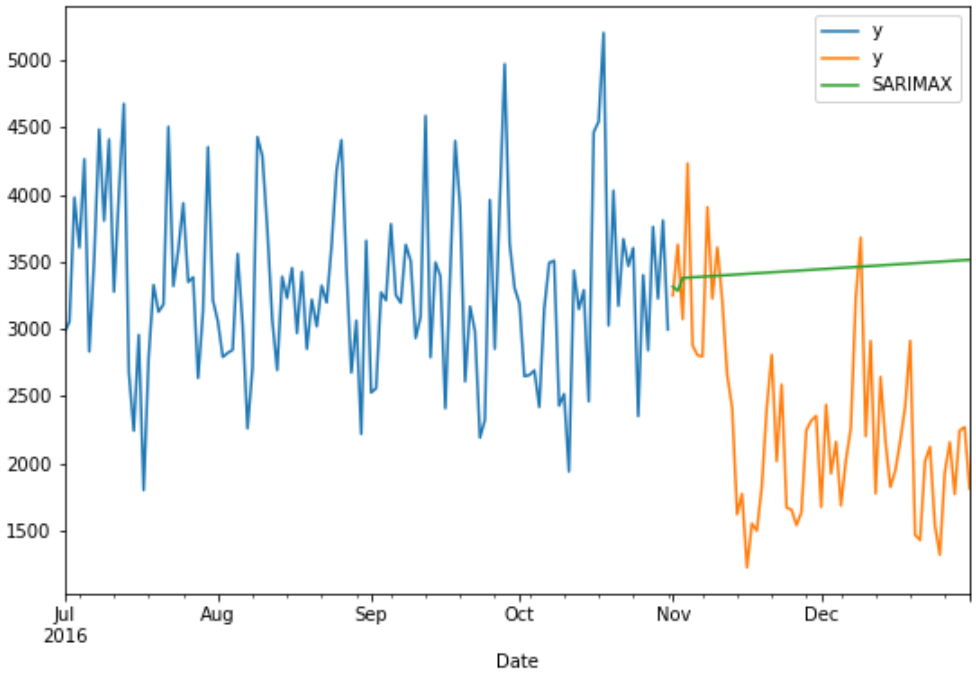


I utilized the Python module statsmodel to decompose the plot and smooth out the data.

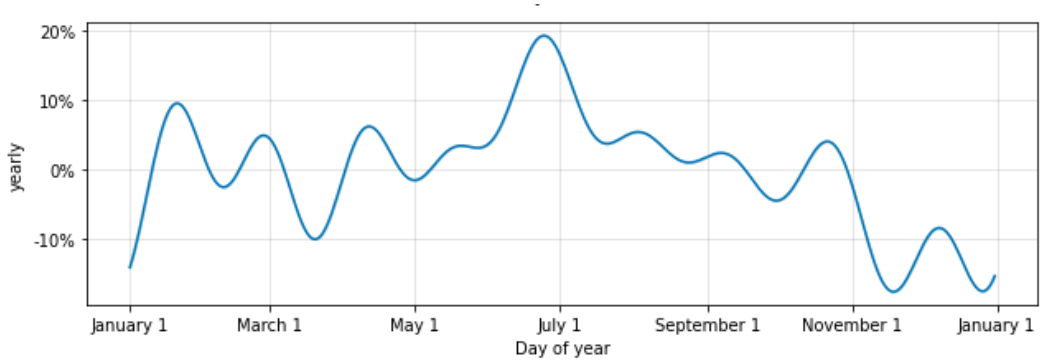
After smoothing out the data we can definitely see a seasonal trend.



This means that I would not utilize and Arima model, but instead start with a Sarimax model since it has a seasonal component to it. I started with a Sarimax model; however, it didn't seem to capture an unexpected drop which occurred in the training set.



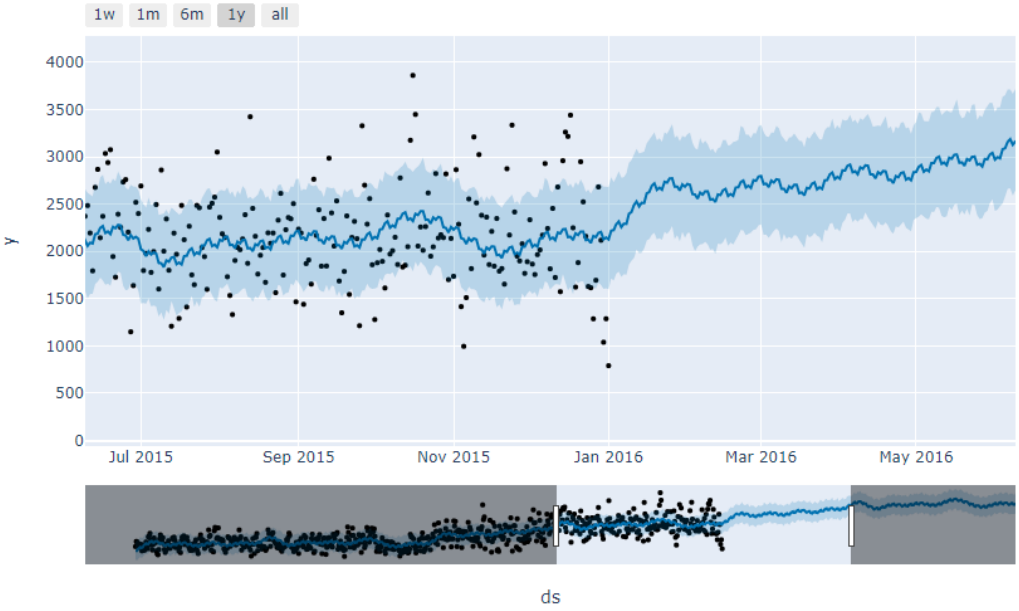
I then went to an FBProphet model to see if it could handle the unexpected dip in sales better. The Prophet model seemed to anticipate the yearly trend much better than the Sarimax model.



Upon cross validating the models, I found that the FBProphet model had a much lower Mean Absolute Percentage Error than the Sarimax model and that is the model I chose to utilize for the prediction.

Analysis

Utilizing the FBProphet model we could see that there is a general trend upwards over the coming two quarters.



This would indicate that inventory levels should be slightly elevated from last years levels.

Future improvements

- In the Future I would like to change this to a multivariate time series analysis by adding things like holidays into the model
- Also since shoe model was part of the data set, I would like to see how sales coincide with the release of a new model.