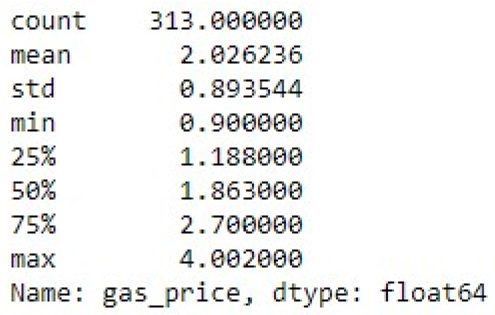
Data Evaluation

Upon loading the forecasting_take_home_data.xlsx data provided I noticed that the data was made up of a two column, 313 observations in total with dates ranging from 01-01-1992 to 01-01-2018. Its two columns comprising of a date, at the monthly level with one record per month, and a price which represented the cost of a gallon of regular gasoline after tax in the United Sates. This data was sourced from the U.S. Energy Information Administration.

To start, I investigated some basic summary statistics of the data and noticed that the data was slightly left skewed due to the relationship of the mean and median illustrated below in Figure 1.



```
count    313.000000
mean       2.026236
std        0.893544
min        0.900000
25%        1.188000
50%        1.863000
75%        2.700000
max        4.002000
Name: gas_price, dtype: float64
```

Figure 1: Summary Statistics

In order to illustrate this skew further, I showed a histogram which validated my original findings that the gasoline prices observations mostly adhered to lower prices ($1.00 - $1.50) as seen in Figure 2.

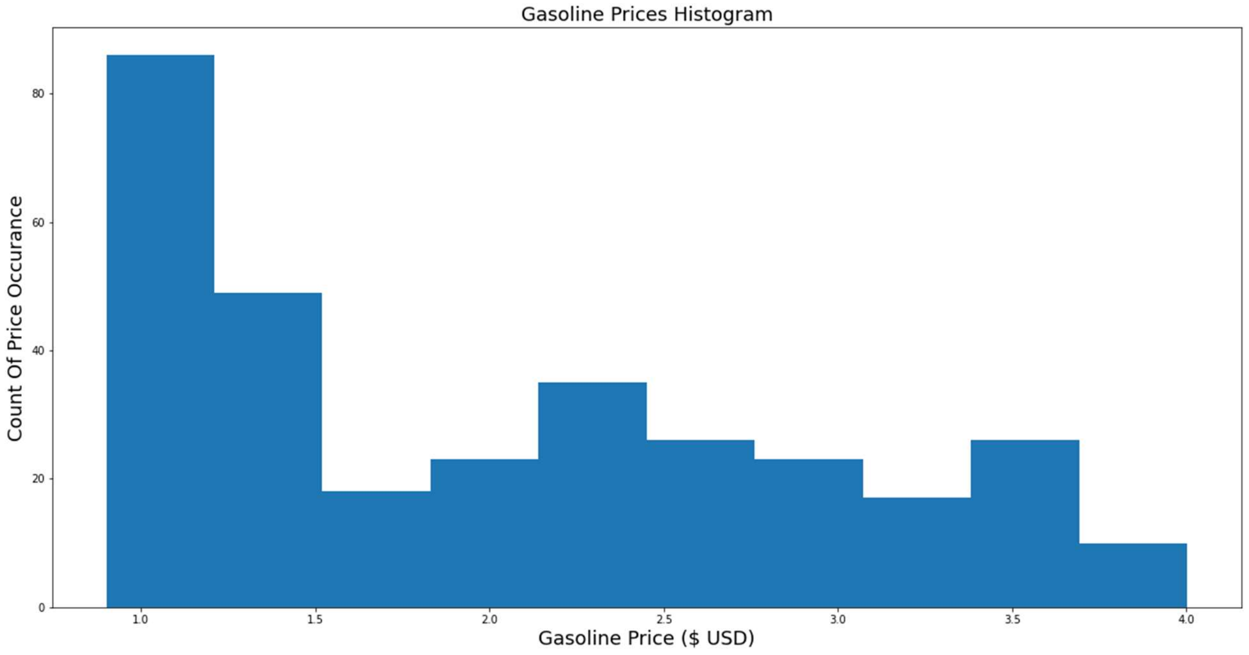


Figure 2: Histogram of Monthly Gasoline Prices (U.S.)

Lastly in my data evaluation I decided to plot the data. Immediately, I noticed the gradual increase of prices since 2000, peaking at roughly $4.00 per gallon, with two large declines in price after 2008. I began to hypothesize that maybe these trends have to do with economic declines in the U.S. economy, potentially in relationship with recessions.

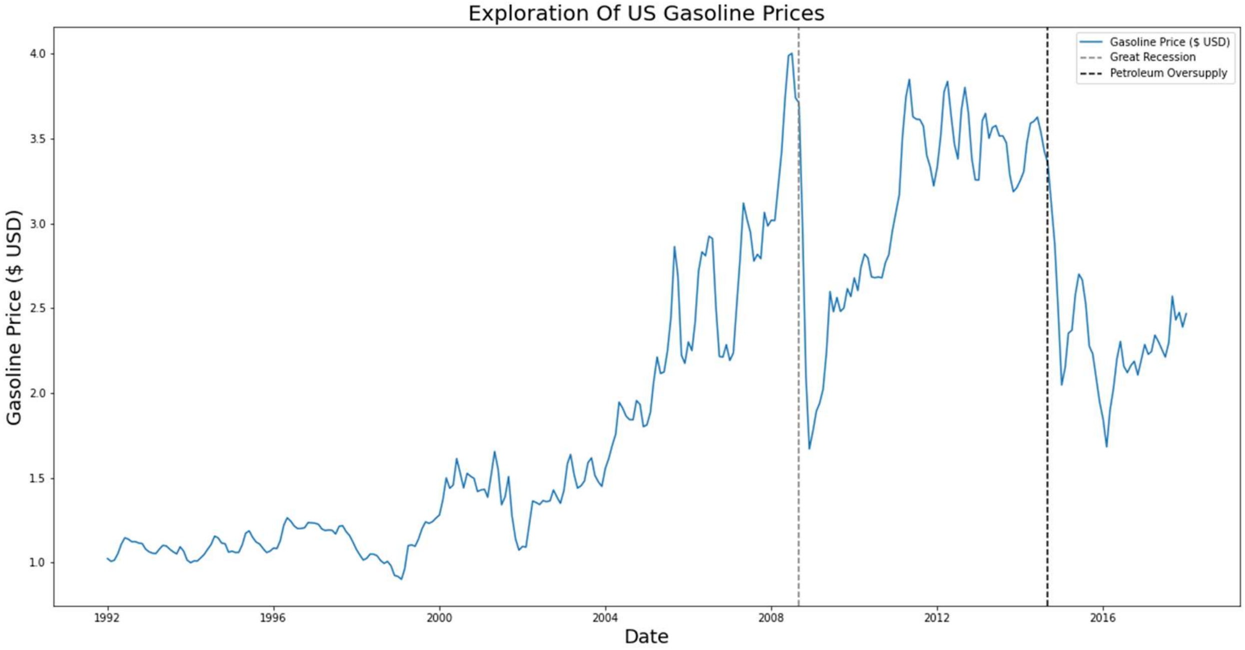


Figure 3: Plot of US Gasoline Prices Over TIme

After further investigation I found that yes, the first drop in gasoline prices was while the Great Recession in the U.S., but the second decline did not seem to be associated with a notable recession. I actually found that the second large decline (2014) was not due to a recession but instead due to an over supply of petroleum on the world market (https://www.bls.gov/opub/btn/volume-4/pdf/the-2014-plunge-in-import-petroleum-prices-what-happened.pdf).

FBProphet Data Modelling

As requested, per the instructions of the homework, I began to model using the FBProphet package, which I had not used before. To highlight, the first thing I learned fitting the model was that by default it modelled and predicted at the daily level. In turn, I was required to modify the instantiation of the Prophet class to model only considering monthly datapoints (removing weekly and daily seasonality). I also learned of its built in functions to utilize cross validation within a time series, specifying different horizons or training and testing periods.

Once I had completed the cross validation, saving the performance metrics, I highlighted three common evaluation metrics I was interested in, Mean Squared Error, Root Mean Squared Error, and Mean Absolute Percentage Error. I chose not to use any evaluation metrics that focused on the median as they are preferable used when one wishes to let outliers or large variations in data not affect the overall metrics. In our case with this data, we do not see in noticeable data errors or outliers that we wish to discard.

I found upon reviewing the FBProhpet white pages that the main way to enhance the model was via hyperparametere tuning instead of a selection of different timeseries models. Therefore I then modified an example of code from their diagnostics white pages (https://facebook.github.io/prophet/docs/diagnostics.html) to capture the evaluation metrics I highlighted earlier. I found the combination of two settings shown in the example, changepoint_prior_scale and seasonality_prior_scale, that minimized the error in all three of the evaluation metrics was 0.05 and 1.0, respectively.

Finally testing the model on the test data set, which was devised using a 70%/30% split of train/test, I found that the results were very poor as seen in Figure 4 and showing in the metrics in Figure 5.

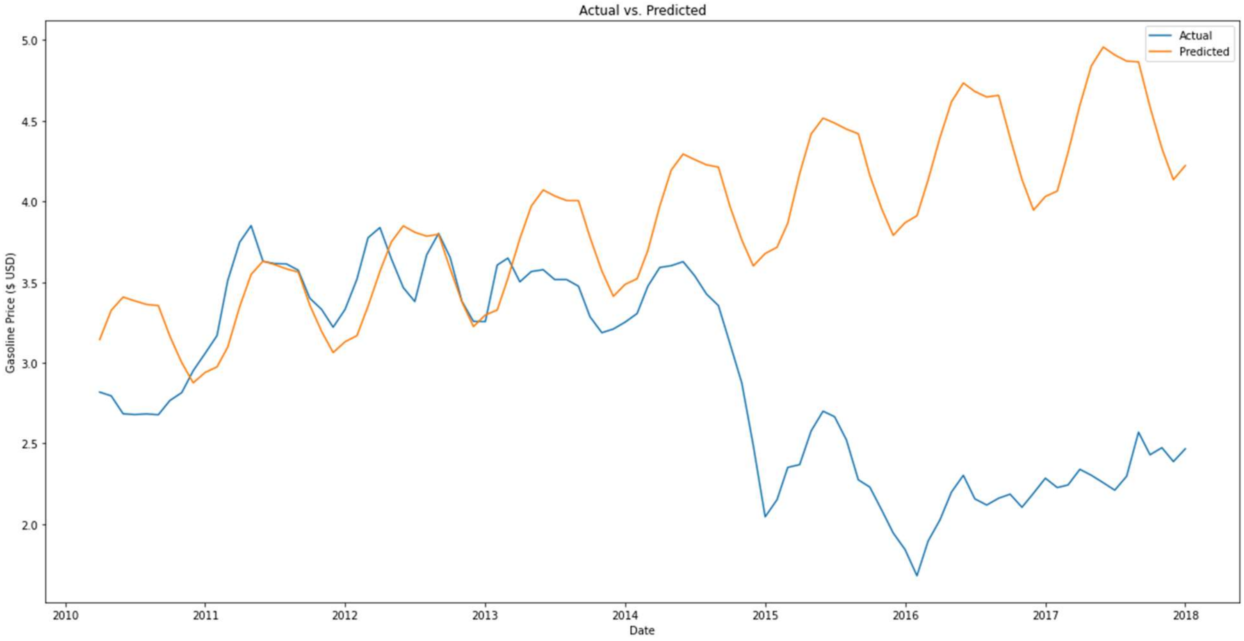
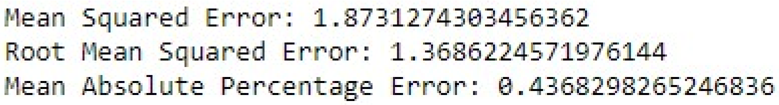


Figure 4: Final Model Actual vs. Predicted



```
Mean Squared Error: 1.8731274303456362
Root Mean Squared Error: 1.3686224571976144
Mean Absolute Percentage Error: 0.4368298265246836
```

Figure 5: Final Evaluation Metrics

If I were to deduce why we saw these poor results, I would suspect that the trend component of the model played too strong of a role in the models' predictions. It assumed that the price would increase on the upward trend that was seen from years 1992 to 2010, but instead did not consider that these prices are affected by many outside factors unpredictable by previous months price.

If I were to try and enhance the model in the future, I would attempt to find other correlating features from outside sources, such as U.S. GDP, the price of the S&P 500 index, price per barrel of oil, or even trade relationships amongst the U.S. oil trade partners.

In summary this activity was a great introduction to FBProphet and docker. I found it intriguing how intuitive and easy it is to fit timeseries models which little to no experience. The largest decisions I had to face during this modelling exercise was around the evaluation metrics instead of choosing which type of model adheres best to the data or which features to use.

Code Instructions

The code that I had wrote should be straightforward and work upon clicking the "run all" command at the top of Jupyter Notebook's. Even though my docker container did not work perfectly on my own computer, it works when not installing pystan or fbprohpet, I still included them in the requirements.txt file to be installed on creation.

If you find that the notebooks/files/data are not showing up in the local run of Jupyter Notebooks then you may have to add back in the volume attachment command to the driver.sh bash script. '--volume $(pwd)/notebooks:/workspace/notebooks \'

It is also worthy to note that I compiled a document (Issues Log – Working Notes.pdf) of all the issues that I ran into while completing this modelling exercise along with solutions that I found to get around them.

Added Files:

- CRC – Data Analysis.pdf
- Issues Log – Working Notes.pdf
- FBProphet Error.txt
- crc_gas_price_forecasting.ipynb

Modified Files:

- Driver.sh
- Requirements.txt