FEI Propane Case Study

Introduction

In this Case Study we will evaluate the forecast accuracy on the Natural Gas Index on FEI Propane and make meaningful inferences using a variety of widely accepted Machine Learning tools.As most of the models used are quite advanced and may be unfamiliar to the reader, a brief introduction will be explained to establish full comprehension for everyone. The data used here contains a variety of Natural Gas & Crude Index closing prices from January 2010 to April 2020.

Far East Index (FEI) Propane Forecast

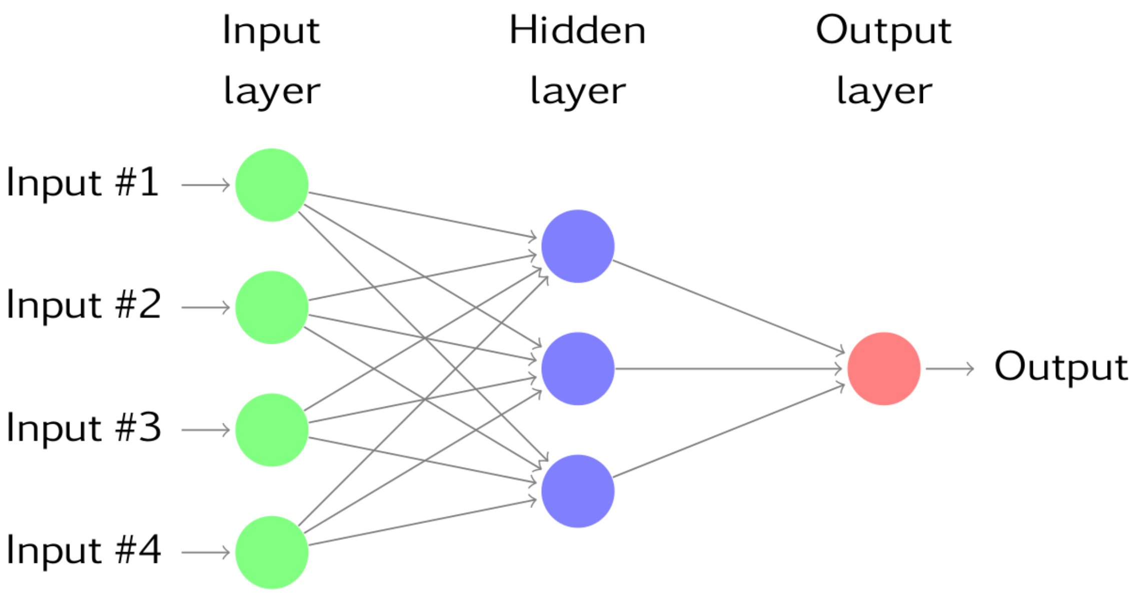
In this section we will focus on forecasting the price of FEI Propane for the next 12 months (as of April of 2020) and assess the performance of the models.To accpomplish this goal we will only be using its time-ordered values with no other variables included.

It should be noted that Time-Series forecasting is a unique form of prediction process because the order of values matters unlike other possibly familiar prediction cases. Consequently, the futher out we forecast the more uncertain and inaccurate we are.

Autoregressive Neural Network Model {NNAR(p,k)}

The first model we will examine is the effectiveness of Autoregressive Neural Network models.Neural Networks are Artificial Intelligence models inspired by biological neural networks in the brain. Consisting of artifical neurons and synapse like connections Neural Networks are able to classify, predict, cluster, and associate patterns within a given dataset. Familiar applications are Natural Language Processing, Image Classification, and Predictive Text. It should be noted that Neural Networks provide high prediction accuracy, they are widely considered a ‘Black-Box’ because we are unsure how to inpret how the computer came to its conclusion.

Although neural networks come in many different types the one used for this case study is a ‘Single Layer Feed-Forward Network’ where our Time-Series values are fed into the input layer, optimized through a hidden layer, and then fed out to provide us our Forecasts. The outputs of one layer is the input of the next layer. A visual representation of this can be seen in the image below.



figName,out.width="30%"

NNAR Forecast Accuracy

As previously mentioned, the model used is an Autoregressive Neural Network NNAR(p,k) which uses lagged values from the Propane Time-Series to model and forecast into the future. Using lagged values for modelling is the "Autoregressive" portion of the Neural Network.

Although we can define exactly how many lagged values (p) and hidden nodes (k) we choose not to as different time-series will require different (p) & (k) values for optimal results. Henceforth, this is exactly the case when cross-validating our model in a time-series and we let the computer automate this portion. Cross-Validation can be described here as testing a statistical model using different time-ordered windows, predicting onto the next set of data values, and then evaluating the accuracy.

For this model we use a rolling window size of 84 months to model and forecast the next 12 months in the full FEI Propane dataset. The output below showcases the NNAR model accuracy for each forecast length using indicators: Mean Error (ME), Root Mean Square Error (RMSE), and Mean Absolute Error (MAE).

Going forward we will evaluate the MAE as the primary example of accuracy because it provides us with a better interpretation of how our predictions differ from the true values in either the positive or negative direction.

The MAE output can be interpreted as "On average, when forecasting 1 month at any point of the FEI Time-Series the NNAR model has an average error of $.072 from the true Propane price"

```
##                                ME       RMSE       MAE
## Forecast Horizon  1  0.01071716 0.08811405 0.0717233
## Forecast Horizon  2  0.02226396 0.15662280 0.1250173
## Forecast Horizon  3  0.03105469 0.20230985 0.1630588
## Forecast Horizon  4  0.03449088 0.23278719 0.1931504
## Forecast Horizon  5  0.04003131 0.24298911 0.1961124
## Forecast Horizon  6  0.04933966 0.24396519 0.1953665
## Forecast Horizon  7  0.06122874 0.23836090 0.2005022
## Forecast Horizon  8  0.07641788 0.24900366 0.2039970
## Forecast Horizon  9  0.08388497 0.25972490 0.2101026
## Forecast Horizon  10 0.07694094 0.27853687 0.2298226
## Forecast Horizon  11 0.06089730 0.31376202 0.2548913
## Forecast Horizon  12 0.04281565 0.33070579 0.2892758
```

Autoregressive Integrated Moving Average Model {ARIMA (p,d,q)}

The next model evaluated is the Autoregressive Integrated Moving Average (ARIMA) which is a classical approach to time-series forecasting that aims to describe the autocorrelations in the data. ARIMA models are subject to three main parameters: the number of lagged observations (p), the number of times that the observations are differenced (d), and the size of the moving average window (q).

As was with the previous NNAR model, we will cross-validate the ARIMA model with a rolling 84-month modelling window to forecast the next 12 months in the time-series and let the computer optimize the model parameters.

Below we can see the Cross-Validation output for the automated ARIMA model and we can initially conclude that the model performs quite well and slightly better than the Neural Network model with a 1 Month Forecast MAE of .065.

```
##                                   ME       RMSE        MAE
## Forecast Horizon  1   6.764334e-05 0.08077439 0.06479378
## Forecast Horizon  2  -7.948284e-03 0.13044973 0.10342312
## Forecast Horizon  3  -1.721332e-02 0.16453332 0.13361963
## Forecast Horizon  4  -2.622672e-02 0.18841629 0.15705515
## Forecast Horizon  5  -2.987113e-02 0.19934024 0.17030795
## Forecast Horizon  6  -2.736865e-02 0.19845246 0.17330239
## Forecast Horizon  7  -2.149638e-02 0.19429321 0.17099874
## Forecast Horizon  8  -1.093255e-02 0.19535243 0.16446635
## Forecast Horizon  9  -8.182292e-03 0.20310441 0.17125314
## Forecast Horizon  10 -1.856736e-02 0.21796195 0.18246545
## Forecast Horizon  11 -3.822135e-02 0.24315619 0.20014995
## Forecast Horizon  12 -5.849420e-02 0.25850995 0.21872968
```

Hybrid Model

In this next model we evaluate the accuracy of using a hybrid model using both Neural Networks and ARIMA models. Combining these two models will be shown to be beneficial later in the case study, but for now we can see that the model performs better than the NNAR model but slightly worse than the ARIMA with a 1 Month Forecast MAE of .067.

```
##                               ME       RMSE        MAE
## Forecast Horizon  1  0.007483698 0.07928821 0.06561002
## Forecast Horizon  2  0.022513579 0.14164870 0.11648617
## Forecast Horizon  3  0.040780689 0.18527679 0.15770138
## Forecast Horizon  4  0.055788827 0.21158768 0.18215686
## Forecast Horizon  5  0.069959263 0.22470484 0.18718760
## Forecast Horizon  6  0.080842565 0.22602563 0.18991406
## Forecast Horizon  7  0.085283358 0.22100518 0.18165237
## Forecast Horizon  8  0.089616104 0.22651372 0.18388708
## Forecast Horizon  9  0.093859185 0.23979418 0.19244911
## Forecast Horizon  10 0.093538875 0.24907522 0.20499816
## Forecast Horizon  11 0.088741487 0.27927797 0.22923029
## Forecast Horizon  12 0.086010081 0.30066896 0.25307931
```

K-Nearest Neighbors Model (KNN)

Lastly, we will introduce the K-Nearest Neighbors (KNN) model and its accurcy when forecasting propane prices. KNN are widely used models for classification, prediction, and spatial analysis. This algorithm finds K examples that are most similar (called Nearest Neighbors) and uses the Euclidean distance metric to predict the target value.

In our case, it was shown that K=7 nearest neighbors with using lags 1-8 will produce the optimal forecasting model. Below you will see the cross-validation output using a rolling origin. Because KNN models are categorically different than the previously introduced models, the cross-validation output looks different but the interpretation is the same. Here all terms mean the same with a new term “MAPE” being introduced which stands for “Mean Absolute Percent Error”.

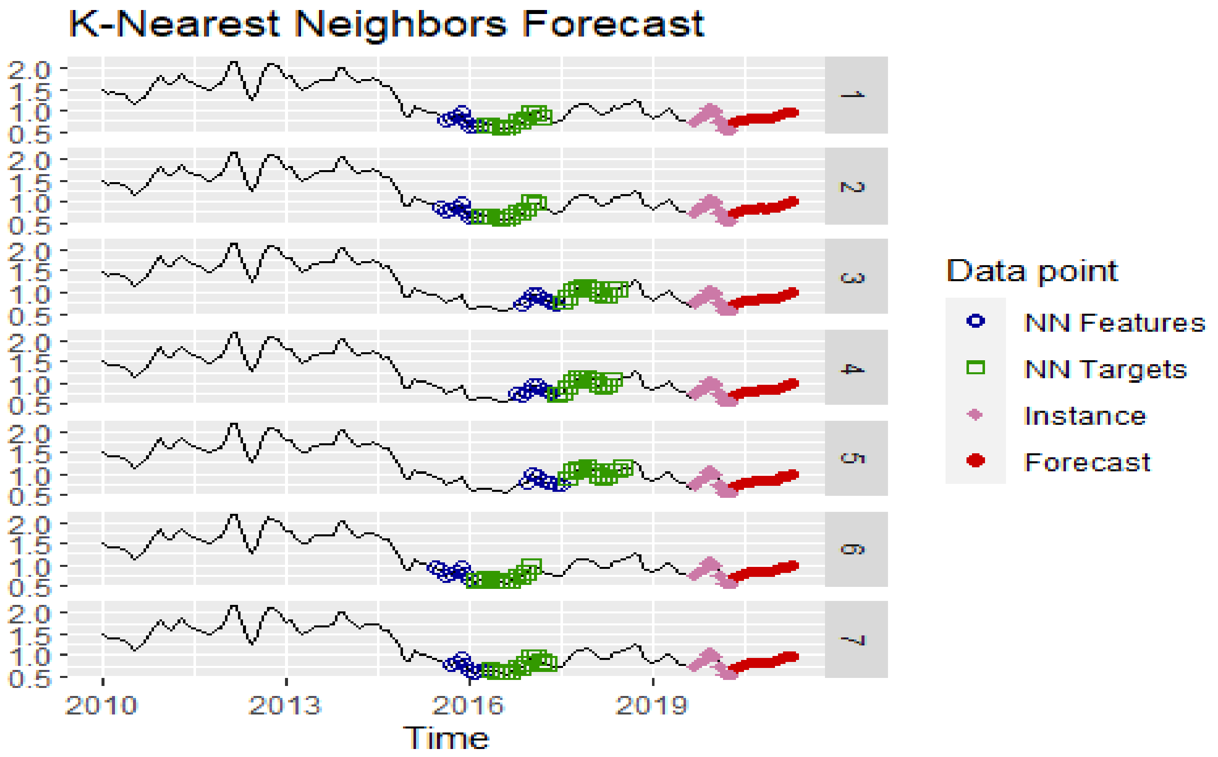
The below output shows us that the 1 Month Forecast has a mean absoute error that is greater than all the rest of the models with a 1 Month Forecast MAE of .081, however in the next section we will see that the KNN model is still highly signficant for our goal.

The graphs below the output provides a visualization on how this process works.

```
##            Jan        Feb        Mar        Apr        May        Jun        Jul
## 2020                                         0.7090617 0.7568312 0.7963286
## 2021 0.8682419 0.9168754 0.9571371 0.9831157                              
##            Aug        Sep        Oct        Nov        Dec
## 2020 0.8168637 0.8269720 0.8391105 0.8346231 0.8401492
## 2021                                                  
```

```
## [1] 7
## [1] 93
```

```
##              h=1        h=2        h=3        h=4        h=5        h=6
## RMSE  0.10663174  0.1500553  0.1842157  0.1850181  0.1927028  0.2242406
## MAE   0.08157971  0.1181825  0.1389356  0.1558504  0.1720041  0.2006191
## MAPE 11.49396661 17.2218919 21.0614233 22.9797119 24.1899054 27.3750637
##             h=7       h=8        h=9       h=10       h=11       h=12
## RMSE  0.2544692  0.2655005  0.2155291  0.1573804  0.1984349  0.2124305
## MAE   0.2322503  0.2401311  0.1791898  0.1447562  0.1949703  0.2124305
## MAPE 31.4108962 32.5042927 25.8853909 25.5446088 35.9794781 39.5468735
```



Model Comparison & Summary

In this portion we will compare all the introduced models and evaluate how they perform on data from this last year. It was shown that the ARIMA model had the lowest cross validated errors while the KNN model had the highest. However, the cross-validated metrics were performed across the entire time series from January 2010 to April 2020. Here we will forecast the last 12 months available using the previous models and compare their accuracy.

In the below tables we can see that KNN performs extremely well when forecasting on the most recent year with with more than 90% accuracy when forecasting 6 Months ahead. Meanwhile the ARIMA model forecasts a static values which usually means there is no trend within the data. The graph below the tables provides a visualization of the forecasts to the testing data.

Therefore, we can conclude that although the ARIMA model performs better on average we cannont solely rely on this to give us the best predictions. Instead we should continue to test all models especially when one model gives us a flat forecast to predict into the future. In other words, although the KNN model was superior when forecasting 12 months ago this may not be the case in the current environment.

```
## Fitting the auto.arima model

## Fitting the nnetar model
```

```
#May 2019-April 2020 Model Predictions

Model_Predictions
```

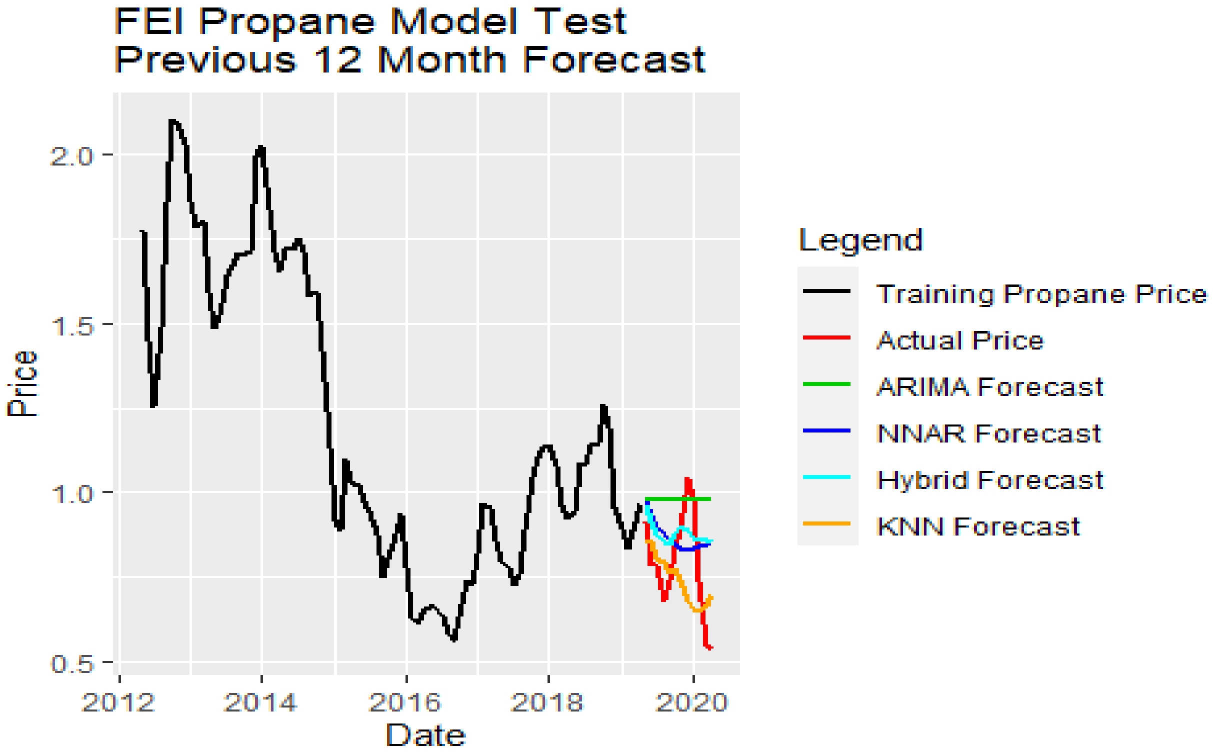
```
##          FEI Price NNAR Forecast
## May 2019 0.9120053     0.8593155
## Jun 2019 0.7873523     0.8527742
## Jul 2019 0.7854460     0.8067154
## Aug 2019 0.6790307     0.7951632
## Sep 2019 0.7390321     0.7624366
## Oct 2019 0.8398403     0.7742385
## Nov 2019 0.8866420     0.7419170
## Dec 2019 1.0445572     0.6822588
## Jan 2020 0.9844621     0.6575360
## Feb 2020 0.7398752     0.6511142
## Mar 2020 0.5489879     0.6708179
## Apr 2020 0.5371614     0.6966113
```

```
#May 2019-April 2020 Model Accuracy

Model_Accuracy
```

```
##          NNAR Accuracy ARIMA Accuracy Hybrid Accuracy KNN Accuracy
## May 2019     0.9341979      0.9269132       0.9485618    0.9422265
## Jun 2019     0.8142767      0.7570229       0.8392406    0.9169090
## Jul 2019      0.8561465      0.7540063       0.8834677     0.9729206
## Aug 2019      0.7091885      0.5587385       0.7403275     0.8289731
## Sep 2019      0.8335024      0.6757533       0.8532879     0.9683309
## Oct 2019      0.9896390      0.8347061       0.9565401     0.9218877
## Nov 2019      0.9428788      0.8962165       0.9840184     0.8367718
## Dec 2019      0.7962670      0.9369146       0.8485351     0.6531560
## Jan 2020      0.8477921      0.9941071       0.8808966     0.6679139
## Feb 2020      0.8605489      0.6772624       0.8346549     0.8800324
## Mar 2020      0.4557791      0.2173363       0.4328153     0.7780825
## Apr 2020      0.4279025      0.1780880       0.4113968     0.7031620

Model_Train_Plot
```

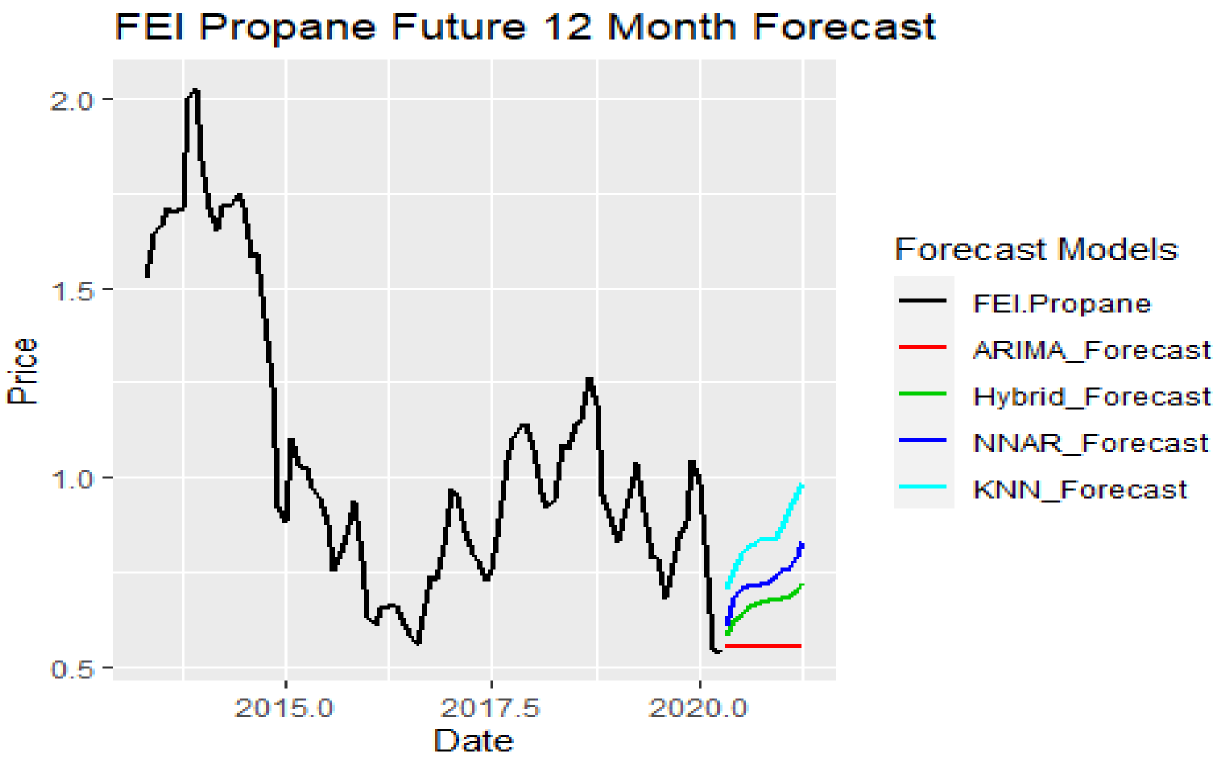


Future Forecasts

Below is a table of the future 12 month forecasts as of April 2020 using the previous 84 months in the time-series. We can see again that the ARIMA Model is again producing static forecasts while the rest of the models are projecting the FEI Propane price to increase again. We can also see that the KNN model is forecasting the greatest price increase while the Neural Network model has a much lower relative increase from the March close price of $.537.

```
## Fitting the auto.arima model

## Fitting the nnetar model
```



```
#12 Month Forecast from May 2020 to May 2021
forecasts.dataframe

##          KNN      NNAR     ARIMA    Hybrid
## 1  0.7090617 0.6110862 0.5540395 0.5838709
## 2  0.7568312 0.6802343 0.5540395 0.6200316
## 3  0.7963286 0.7084299 0.5540395 0.6405805
## 4  0.8168637 0.7142205 0.5540395 0.6545987
## 5  0.8269720 0.7155818 0.5540395 0.6656837
## 6  0.8391105 0.7186828 0.5540395 0.6729001
## 7  0.8346231 0.7246892 0.5540395 0.6769609
## 8  0.8401492 0.7407113 0.5540395 0.6782669
## 9  0.8682419 0.7544191 0.5540395 0.6788400
## 10 0.9168754 0.7627572 0.5540395 0.6852461
## 11 0.9571371 0.7914052 0.5540395 0.7037560
## 12 0.9831157 0.8334284 0.5540395 0.7264347
```

Inference on the Propane Far East Index

In this section we will evaluate the relationships between the FEI Propane and the other available variables from January 2010 through April 2020. Although naturally we would like to know which variables have the biggest effect on FEI Propane we will not dive into this topic very much. This is because causality is an extremely complex subject studied by Psychologists and Economists for years and is often quite difficult to prove. Hence the old trope "Correlation does not mean Causation" because while it is easy to prove correlation between factors is present, but to prove one causes an effect on the other instead of an unknown variable is difficult. This is especially true in statistical modelling because we use correlation to help predict outcomes but our results are just mathematical formulations of the data and may not explain the true reasons of the behaviors of variables.

Nonetheless, we will attempt to do our best using Granger Causality, Correlation Matrices, Principal Component Analysis, Cointegration tests, and Impulse Response Functions to develop some form of inference of the FEI Propane along with the Oil Industry as a whole. Below is a list of variables in our dataset that will be evaluated for inference.

```
##  [1] "Propane_FEI"                 "Propane_ARA"
##  [3] "Propane_MtBelvieu"           "Propane_Aramco"
##  [5] "Propane_Sonatrach"           "Normal_Butane_ARA"
##  [7] "Normal_Butate_AFEI"          "Normal_Butane_MtBelvieu"
##  [9] "Normal_Butane_Aramco"        "Normal_Butane_Sonatrach"
## [11] "US_C5_Seller_Mont_Belvieu"   "Naptha_Buyer_Japan"
## [13] "Naptha_Buyer_NWE"            "WM_WTI_..bbl"
## [15] "WM_Brent_..bbl"              "WM_Henry_Hub_Gas_US..mmbtu"
## [17] "WM_Naphtha_NWE_CIF_..bbl"    "WTI.Gas_Ratio"
```

Granger Causality

We begin this section by utilizing Granger Causality tests which uses statistical tests to determine whether one variable's time series is useful in forecasting FEI Propane monthly prices.

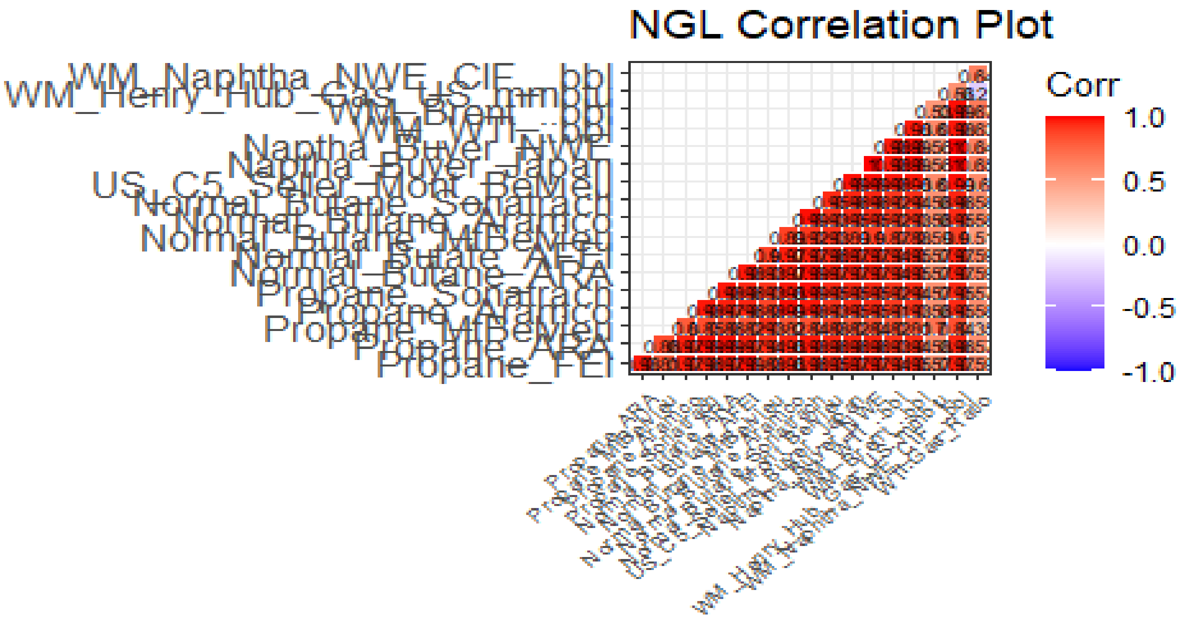
After testing all variables under a Vector Autoregression Model it was shown that the 'Normal Butane Aramco' was the only significant variable for Granger Causality and Instantaneous Causality on FEI Propane. This means that if we were to use another variable to forecasts FEI Propane in conjuction to its own time series it would Normal Butane Aramco.

Below we can see the output of our Granger Causality Tests and that we reject our Null Hypothesis (HO) with 95% confidence.

```
## $Granger
##
##  Granger causality H0: Normal_Butane_Aramco do not Granger-cause
##  Propane_FEI
##
## data:  VAR object var1
## F-Test = 4.2187, df1 = 1, df2 = 206, p-value = 0.04124
##
##
## $Instant
##
##  H0: No instantaneous causality between: Normal_Butane_Aramco and
##  Propane_FEI
##
## data:  VAR object var1
## Chi-squared = 15.311, df = 1, p-value = 9.12e-05
```

Industry Correlation Matrix

Next we will investigate the correlations of all variables which can be seen in the graph below. It is clear that almost all of the variables are highly correlated to each other. This most likely speaks to the industry as a whole in which all variables tend to move together in the same direction. The only outliers in this matrix is the Henry Hub Gas and WTI Gas Ratio which only shows only mild positive correlation with the rest of the Industry Indexes.

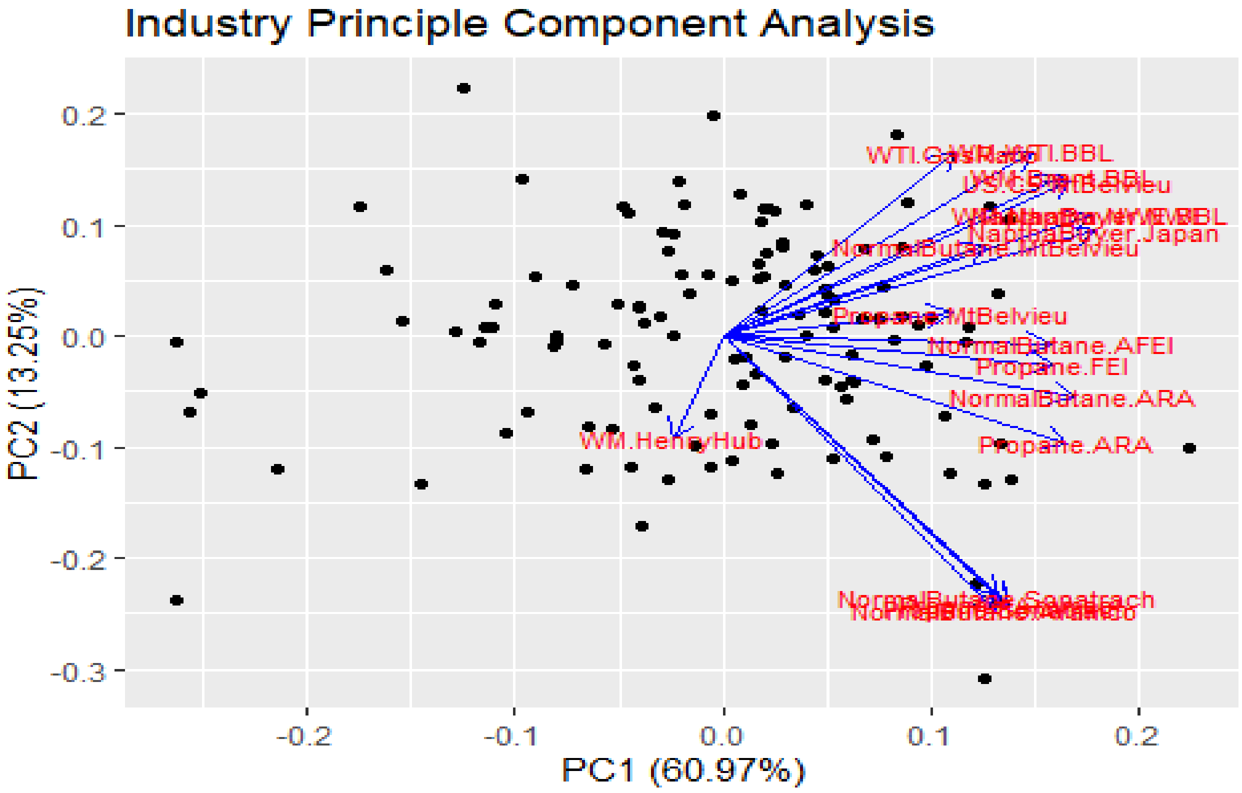


Principal Component Analysis (PCA)

Next we will evaluate a Principal Component Analysis on all industry variables to examine which ones are most closely related to each other. To quickly introduce Principal Component Analysis (PCA); it is an unsupervised exploratory data analysis procedure that highlights the relatedness of variables within the dataset. Each Principal Component describes a certain percentage of variance within the population with all components adding up to 100%. As was with Neural Networks, PCA is also a widely used procedure for almost every industry.

For our analysis we will need to transform all variables into stationary white noise which is done by taking the first difference of each varaible. This is necessary because of the Time-Series properties of this project.

In the Biplot below we can clearly see that Propane FEI is most closely related to Normal Butane AFEI, Normal Butane ARA, and Propane ARA. We can alos see that there are three distinct grops within the first two Principle Components. This is significant because the first two Principle Components explains roughly ~75% of the variance within all the data. Therefore, making strong conclusions about how some variables are related is reasonable (to a certain degree).



Cointegration Tests

Lastly, in this section we test for variables that are cointegrated with Propane FEI. Cointegration tests are significant because they test whether correlated time-series are statistically significant and not correlated by chance.

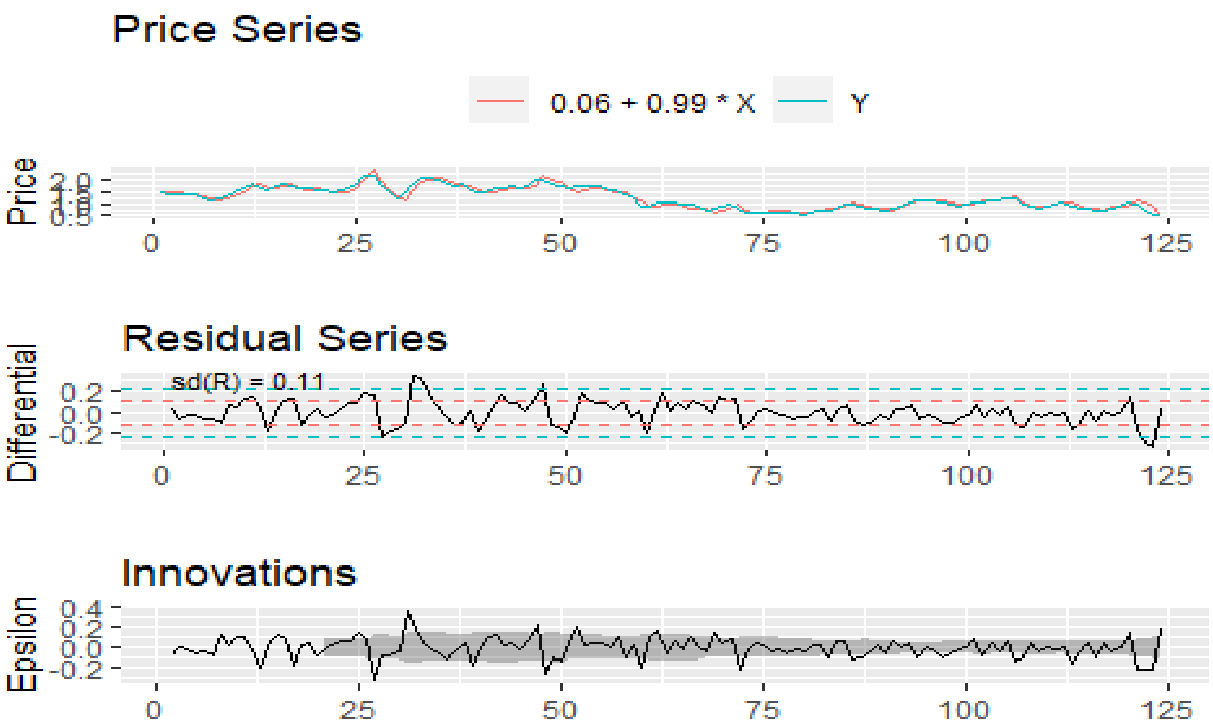
For conciseness we will not show the output for every variable test but we have found that the following variables are highly cointegrated with FEI Propane:

- Propane Aramco
- Normal Butane AFEI
- Normal Butane Aramco
- Normal Butate Sonatrach
- US C5 Seller Mont Belvieu
- Naptha Buyer Japan
- Naptha Buyer NWE
- WM WTI.bbl
- WM Brent.bbl
- WM Naptha NWE CIF.bbl

For demonstration, below is the output for testing cointegration between Propane Aramco and Propane FEI. In the summary table we can see that the tests passes on all levels with 99% confidence. This is also clearly evident in the plots below that show we easily create a function with Propane Aramco to predict Propane FEI.

```
## Y[i] =   0.9908 X[i] +   0.0615 + R[i], R[i] =   0.4346 R[i-1] + eps[i],
eps ~ N(0,  0.1033^2)
##         (0.0236)        (0.0597)                (0.0829)
##
## R[124] = 0.0383 (t = 0.341)
##
## Unit Root Tests of Residuals
##                                                     Statistic    p-value
##   Augmented Dickey Fuller (ADF)                         -5.550    0.00010
##   Phillips-Perron (PP)                                 -63.536    0.00010
##   Pantula, Gonzales-Farias and Fuller (PGFF)             0.399    0.00010
##   Elliott, Rothenberg and Stock DF-GLS (ERSD)           -5.289    0.00010
##   Johansen's Trace Test (JOT)                          -27.881    0.00836
##   Schmidt and Phillips Rho (SPR)                        63.199    0.99990
##
## Variances
##   SD(diff(X))          =   0.145947
##   SD(diff(Y))          =   0.125876
##   SD(diff(residuals))  =   0.123349
##   SD(residuals)        =   0.112332
##   SD(innovations)      =   0.103340
##
```

```
## Half life        =    0.831742
## R[last]          =    0.038253 (t=0.34)
```



Price Series

Residual Series

Innovations