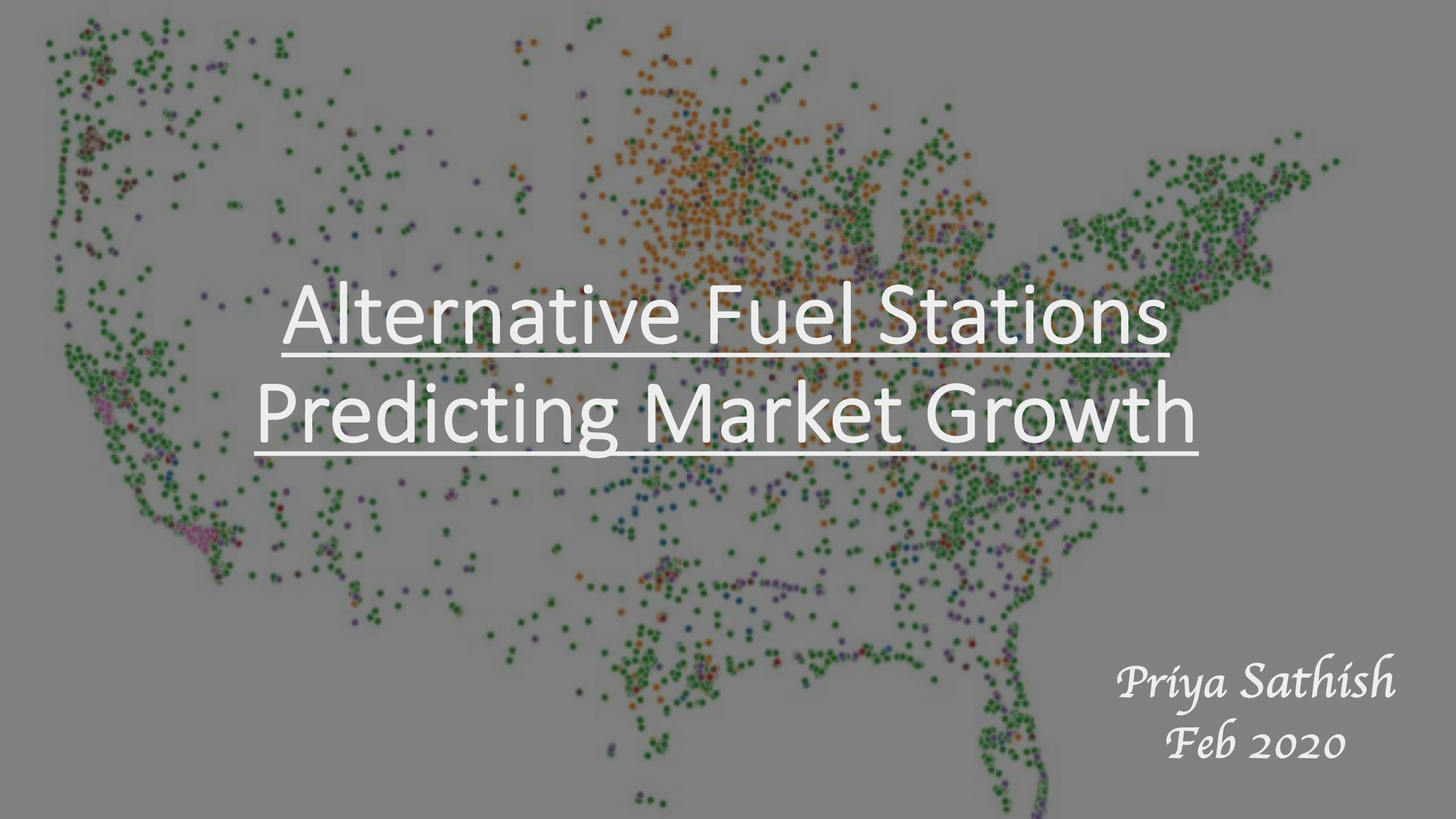


Alternative Fuel Stations Predicting Market Growth

Priya Sathish

Feb 2020

Problem Synopsis

Transport sector uses Petroleum based fuel conventionally

- Economic impact
- Environmental quality
- Fuel cost
- Dependency
- Energy security

……. driving the necessity for Alternative Fuel sources, its availability and access.

Business Case

Goal:

Predicting growth of Alternative Fuel Stations and Alternative Fuel Powered Vehicles across the United States

Stakeholders:

Industries that produce Alternative Fuels

Automobile industry

Vehicle owners with sustainability mindset

Government agencies

Fuel stations with multiple fuel types

Dataset Acquisition

https://afdc.energy.gov/data_download/

Data Wrangling

Challenges

- Missing values
- Outliers
- Duplicates
- Format issues

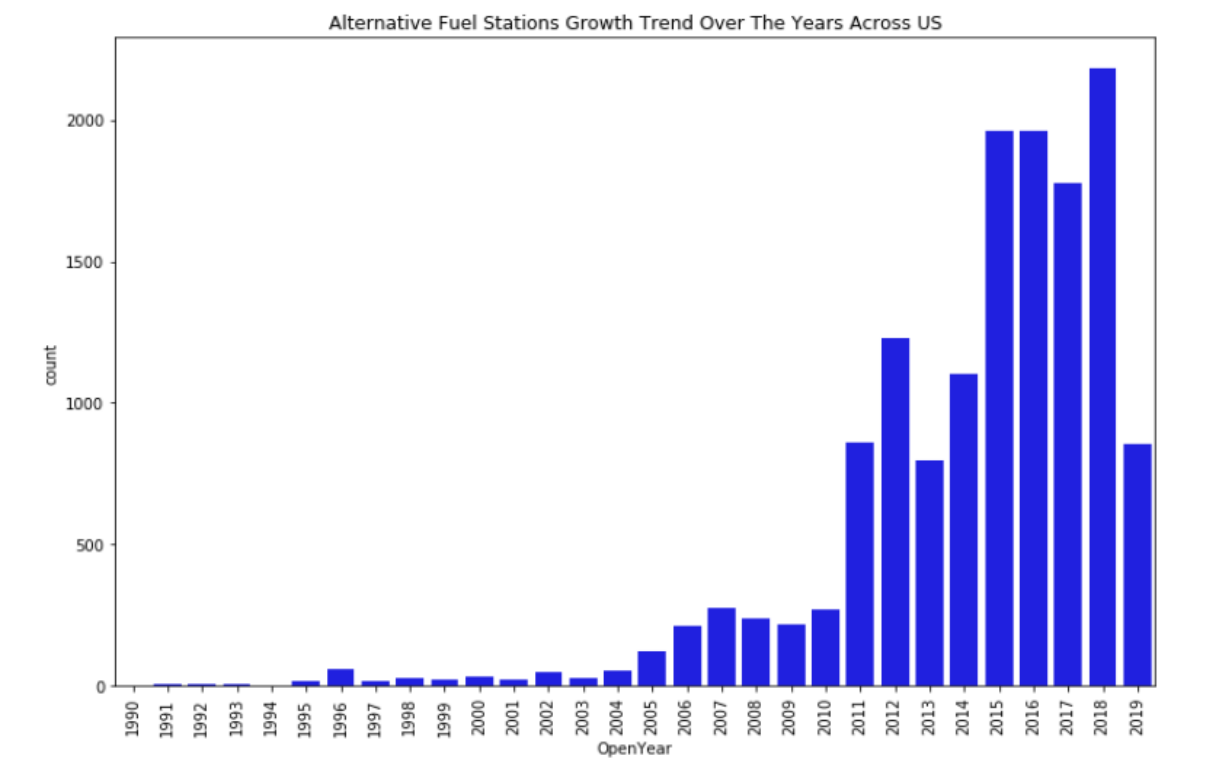
Station Features

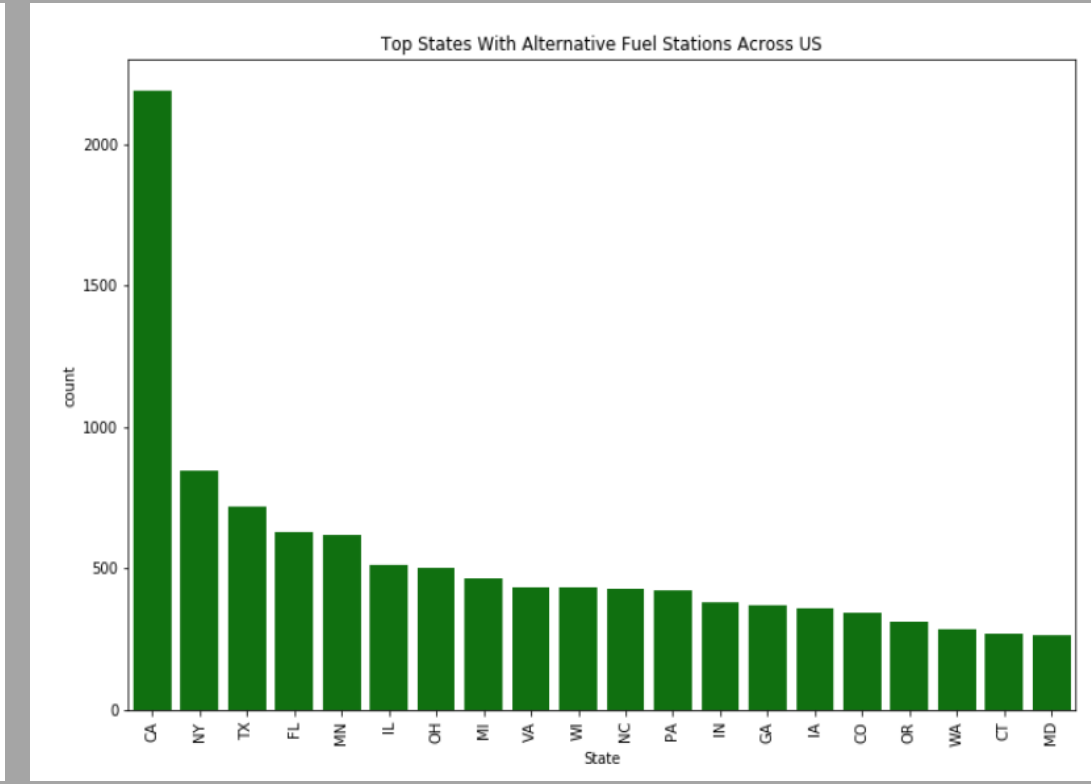
- Station ID
- Fuel type code
- State
- Geocode status
- Latitude
- Longitude
- Open date

Vehicle Features

- Year
- Fuel type
- Sales

Exploratory Data Analysis



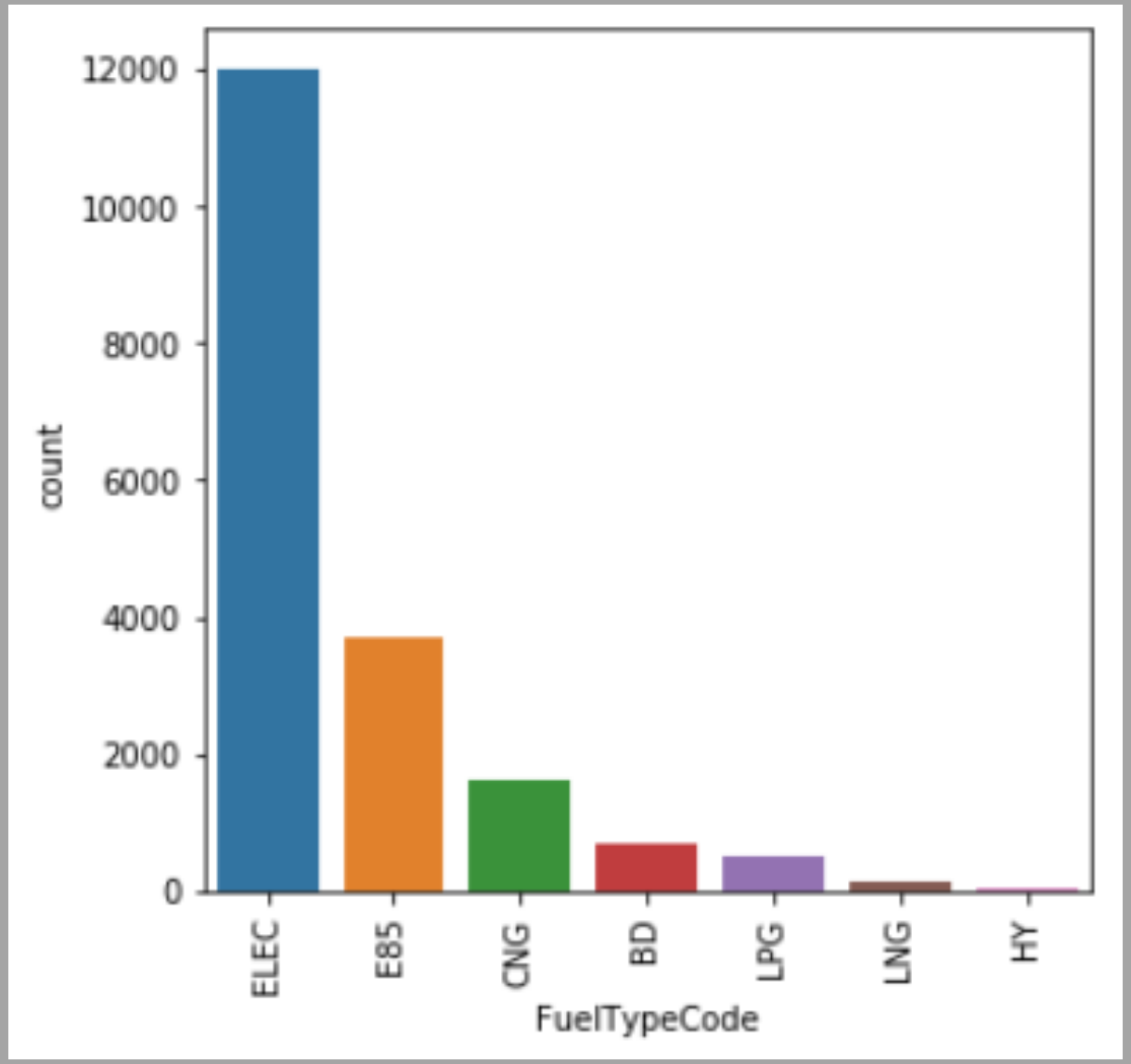


Growth trend of alternative fuel stations across US
Exponential growth since 2011 – awareness created by Clean Cities Coalitions. https://cleancities.energy.gov/about/

Higher concentration of alternative fuel stations in California is suspected due to Low Carbon Fuel Standard campaign. http://www.cadelivers.org/lowcarbon-fuelstandard/ (http://www.cadelivers.org/low-carbon-fuel-standard/)

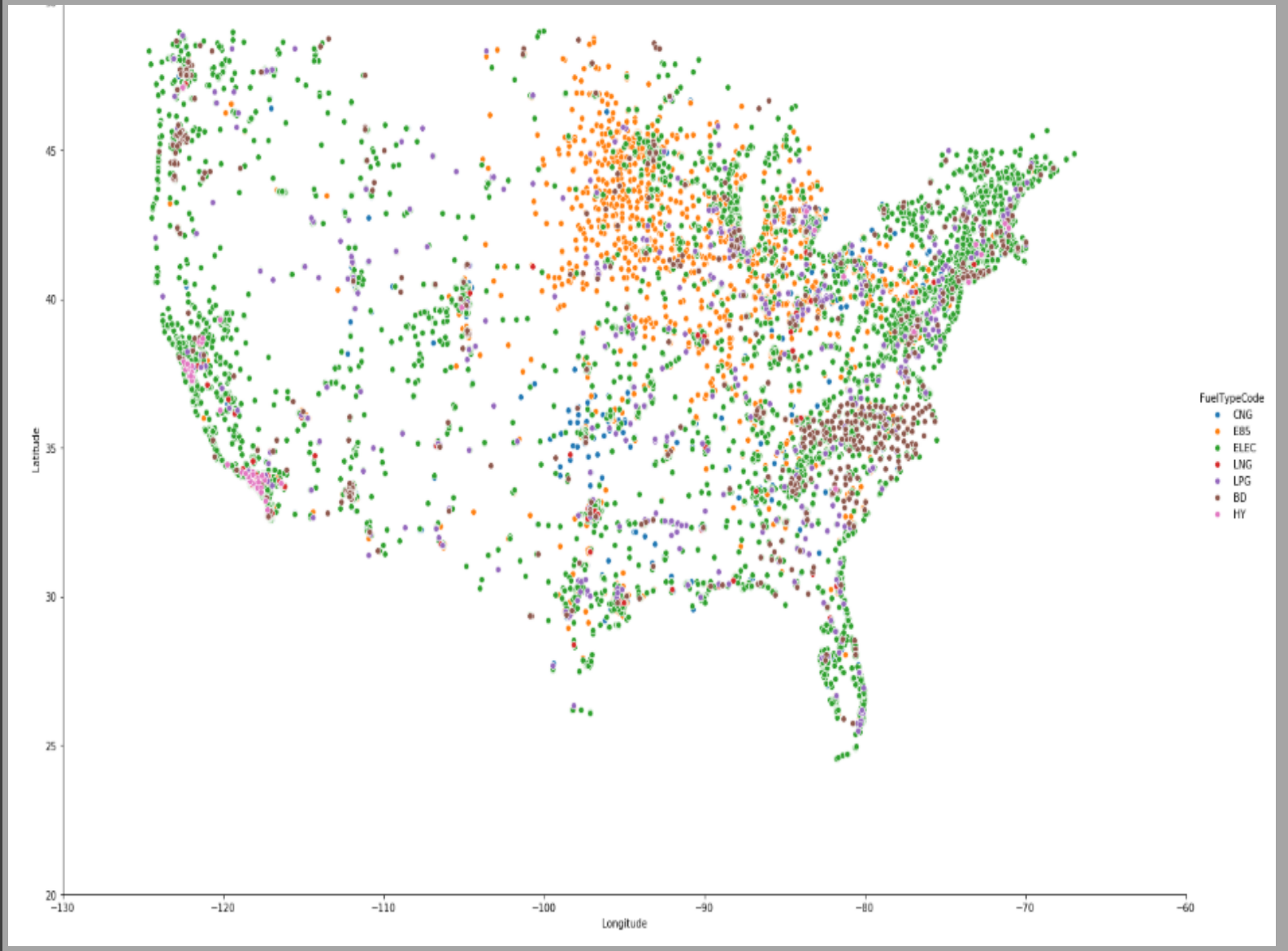
Alternative Fuel Types

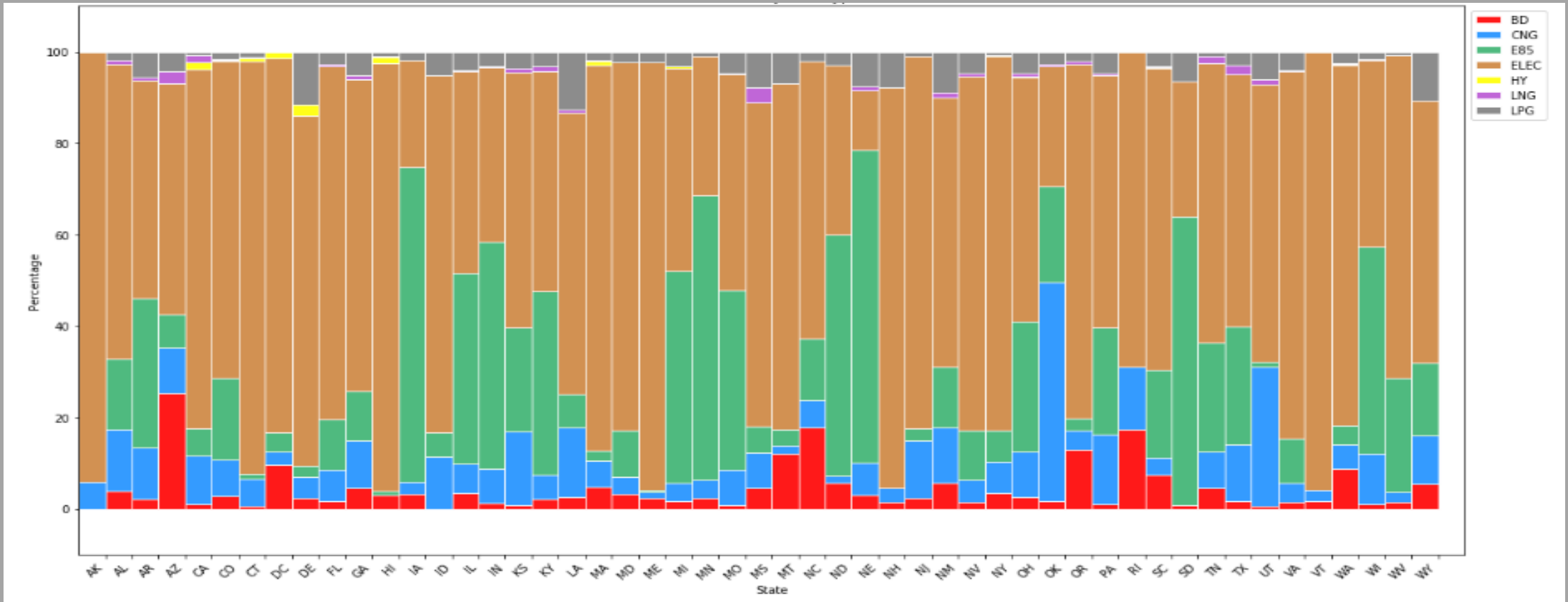
- Over 90% of alternative fuels are comprised of electricity and ethanol85
- Higher number of electric vehicles is attributed to no tail pipe emissions, environmental friendly
- Convenient access to public charging stations and no cost
- Ethanol85 is 100% renewable fuel manufactured from corn.



Fuel Stations across US by Fuel Type

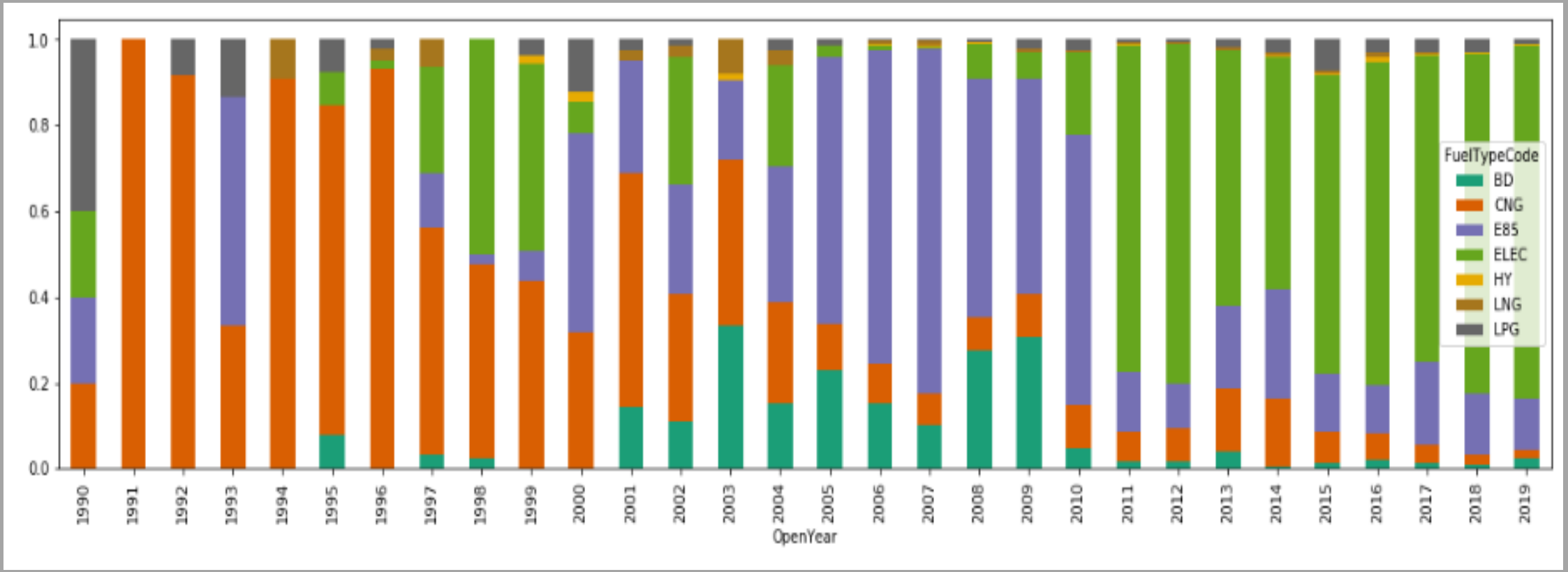
- Electric fuel stations are prominent along Coast line
- Ethanol85 stations are concentrated in Central Plain states due to high corn harvest
- Hydrogen stations are densely located in California
- Bio diesel stations mostly located in South Eastern states





% Distribution by Fuel Type for each State

- Electric stations are higher in states except for those states where E85 is available.
- E85 is higher in states where corn is harvested.

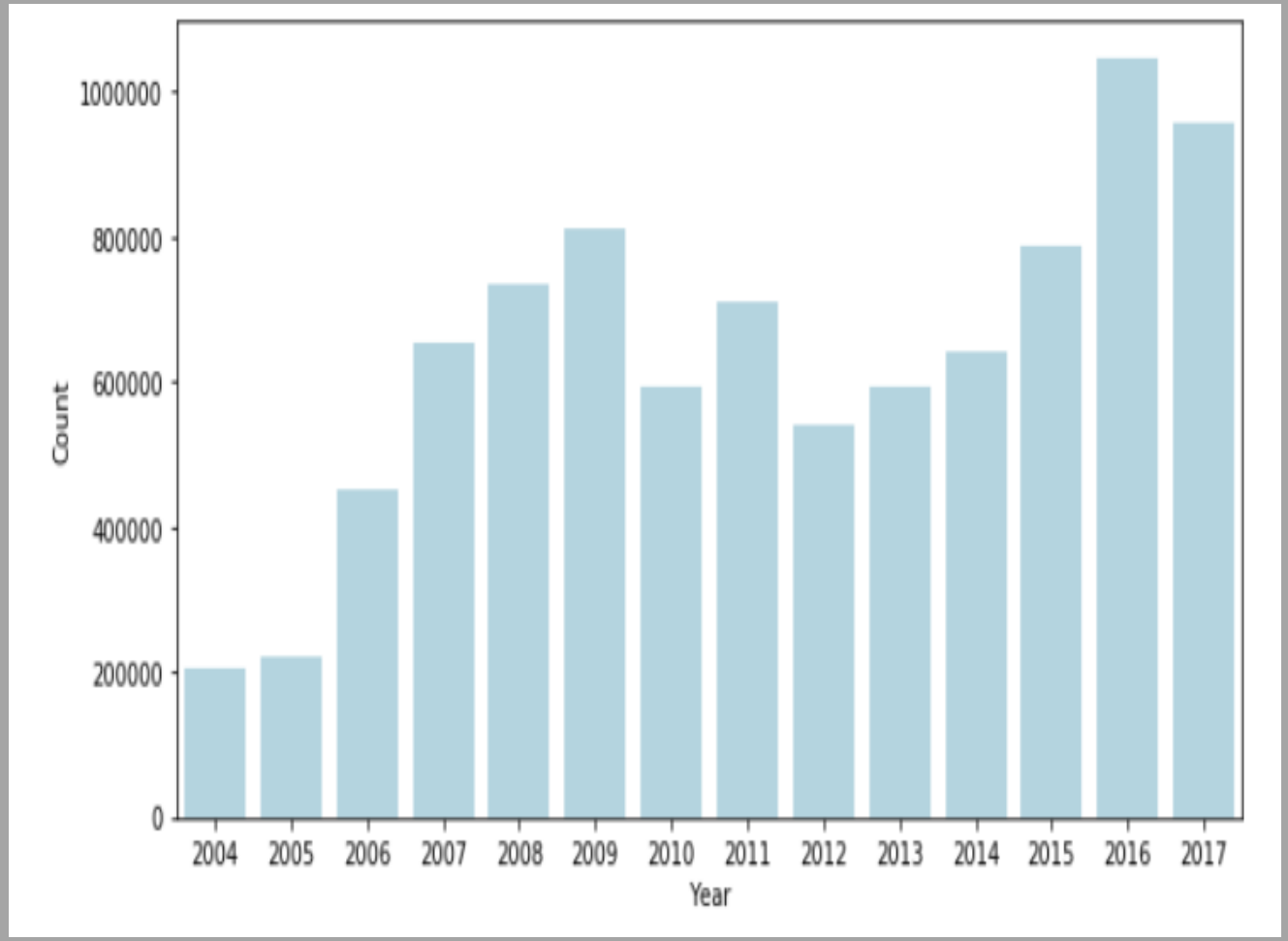


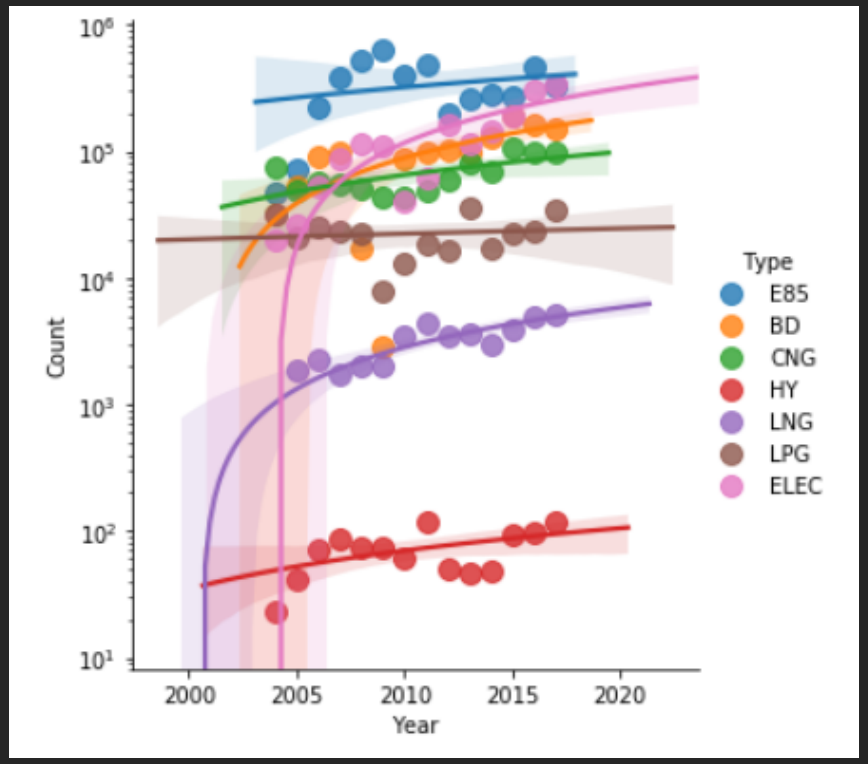
Shift in Fuel Type over years

- In 90's most fuel stations were CNG
- In 2000, there was a shift towards BD and E85
- Since 2010, there is exponential growth in Electric fuel stations

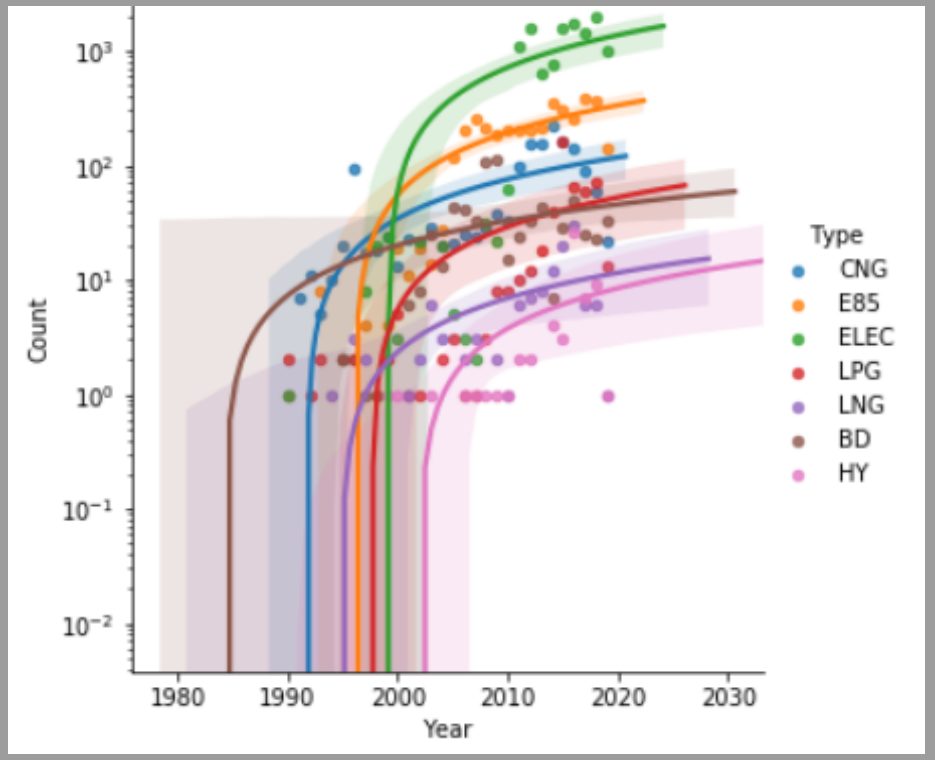
Vehicle Sales by Year

- Alternative fuel equipped vehicles
- Dip in sales during recession
- With strong economy in last few years the growth rate for these vehicles is exponential
- This is also driven by environmental factors and User awareness





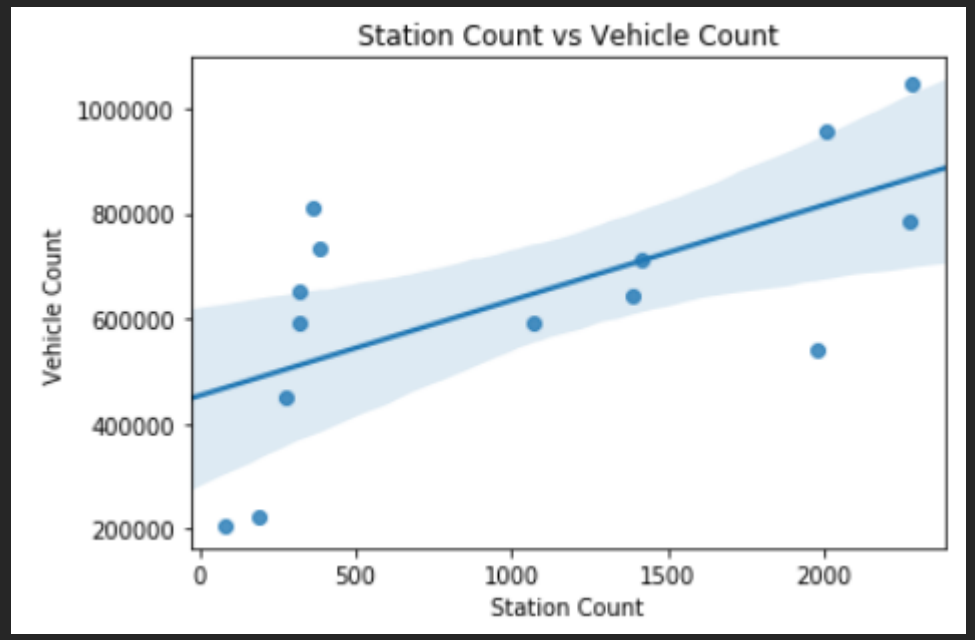
Regression plot – Vehicle Sales by Fuel Type

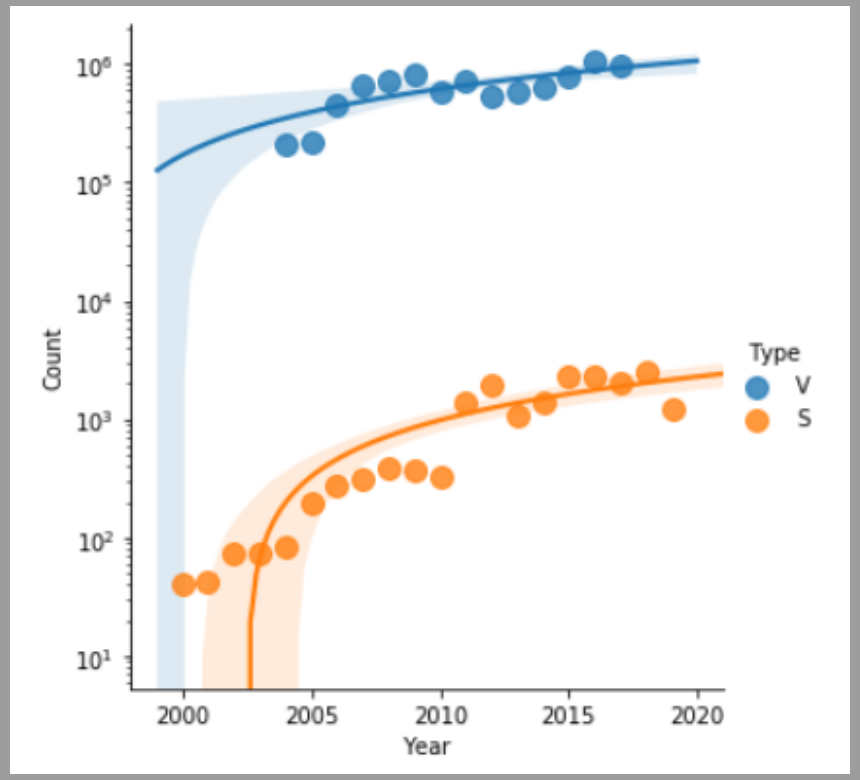


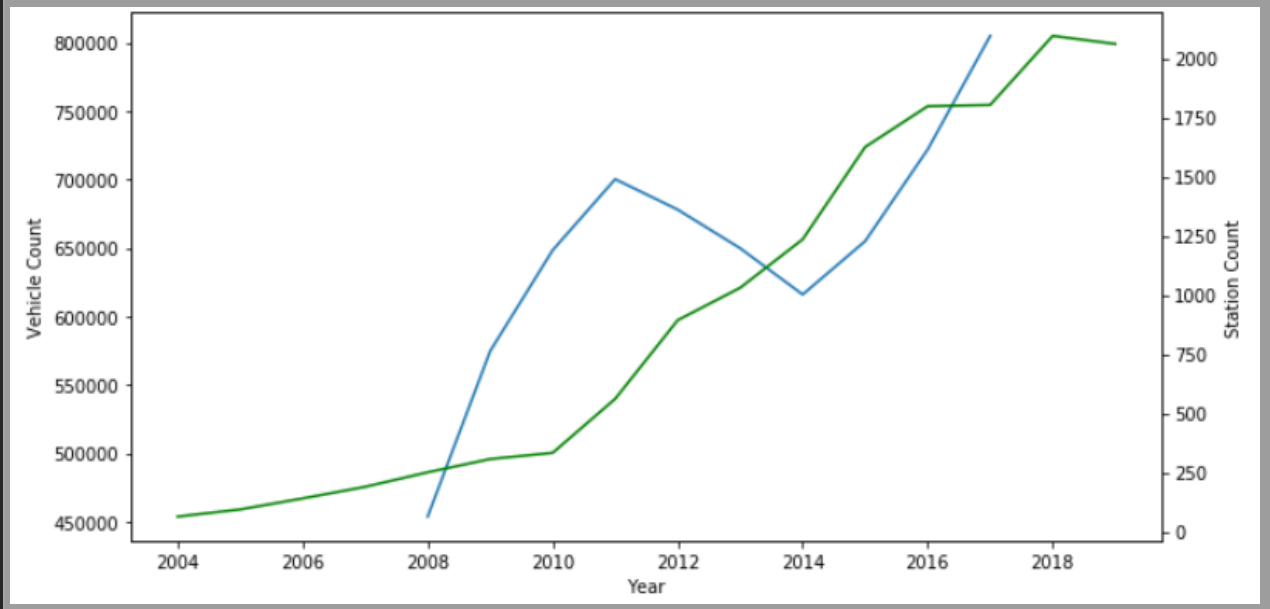
Regression plot – Stations by Fuel Type

Correlation between Stations and Vehicles

- Strong correlation from the year 2005
- Pearson co-efficient 64%







EDA Summary

Models and Algorithms that can be implemented with the dataset which could provide value to the stakeholders:

- Predict the number of Alternative Fuel Stations
- Predict the number of Alternative Fuel Vehicles
- Predict best location for the next Alternative Fuel Station
- Assess individual fuel cost trend's effect on the rate of change in the number of Alternative Fuel Stations and vehicles equipped with that fuel type

Model Development

Models Under Consideration to Predict the growth of Alternative Fuel stations

- ARIMA - Autoregressive Integrated Moving Average Model (Traditional Forecasting)

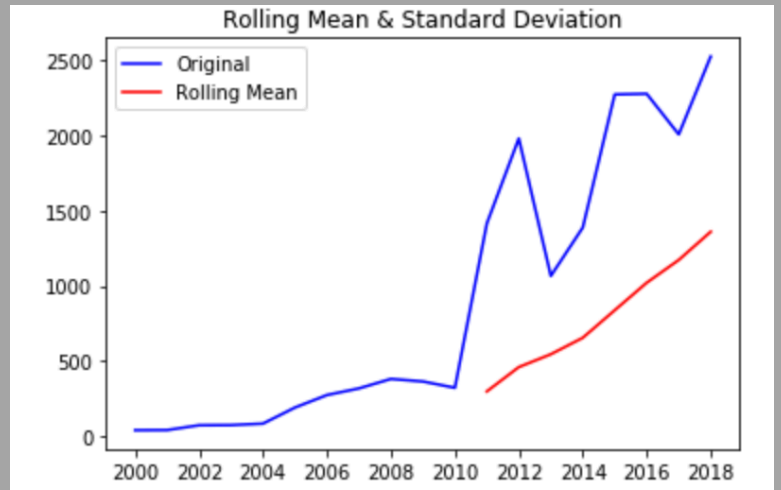
– chosen due to the nonstationary property of the data

- LSTM - Long Short-Term Memory Network (Deep Learning based Forecasting)

– chosen for its learning capacity by preserving and training the features of given data for a longer period

- Goal is to compare the ARIMA and LSTM models with respect to their performance in reducing error rates. (66%Split)
- Assessment Metric: Root Mean Squared Error (RMSE) – Measures the differences or residuals between actual and predicted. The main benefit is that it penalizes large errors.
- Perform Hyperparameter tuning on the best model.

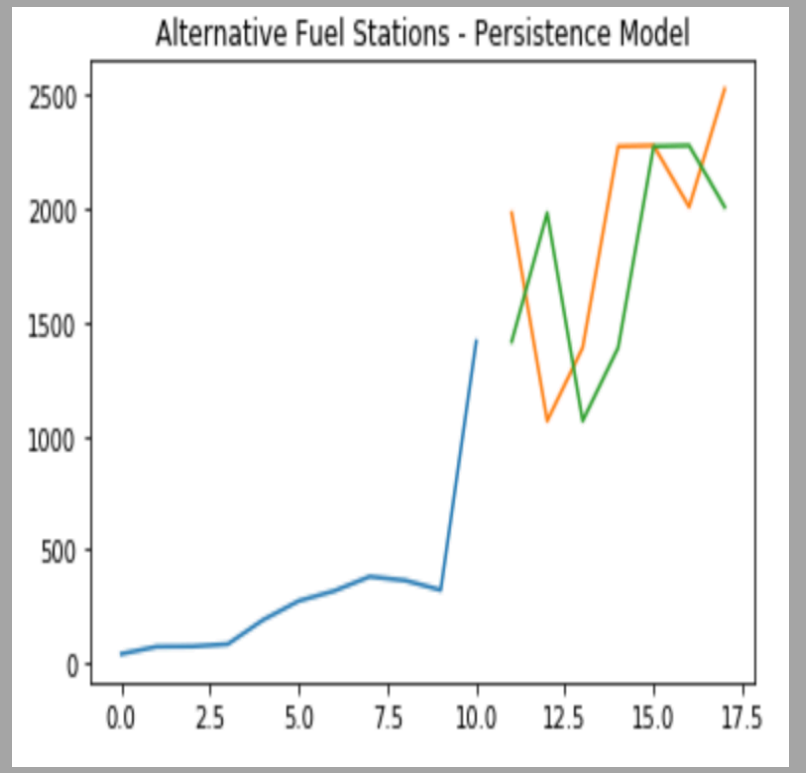
Dickey Fuller Test and Naïve Forecast




```
Results of Dickey-Fuller Test:
Test Statistic                   0.000000
p-value                          0.958532
#Lags Used                       8.000000
Number of Observations Used     10.000000
Critical Value (1%)             -4.331573
Critical Value (5%)             -3.232950
Critical Value (10%)            -2.748700
```

Null Hypothesis (H0): If failed to be rejected, it suggests the time series has a unit root, meaning it is non-stationary. It has some time dependent structure.

Alternate Hypothesis (H1): The null hypothesis is rejected; it suggests the time series does not have a unit root, meaning it is stationary. It does not have time-dependent structure.



From the plot of the persistence model predictions it is clear that the model is 1-step behind reality. There is a rising trend and year-to-year noise in the stations count, which highlights the limitations of the persistence technique.

ARIMA

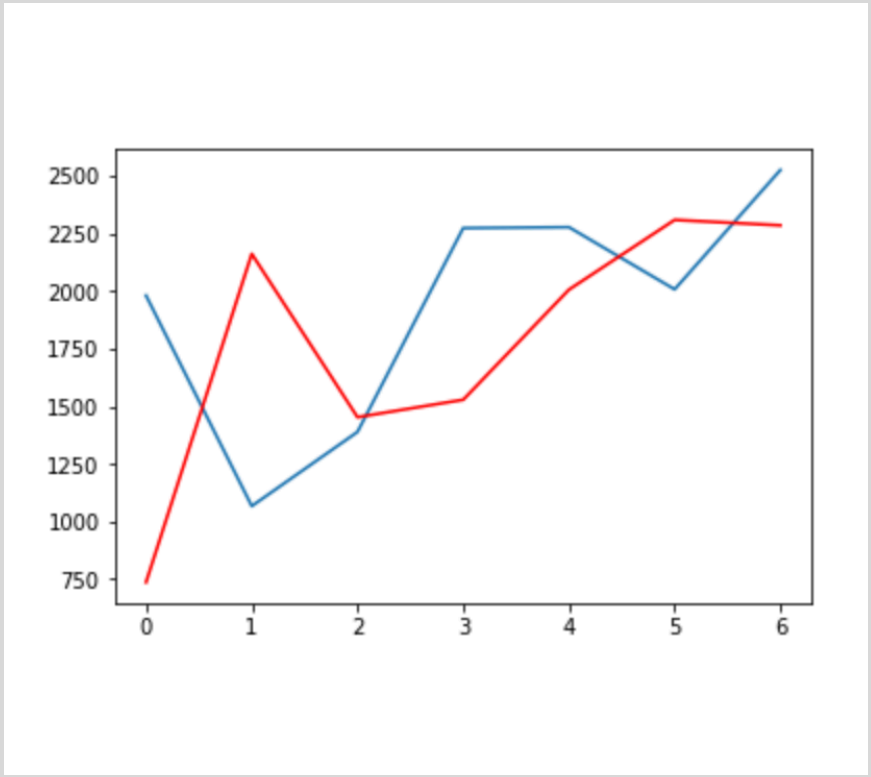
From the plots we determine the parameters p, q, d and fit an ARIMA(0,1,1) model. This sets the lag value to 0 for autoregression, uses a difference order of 1 to make the time series stationary, and uses a moving average model of 1.

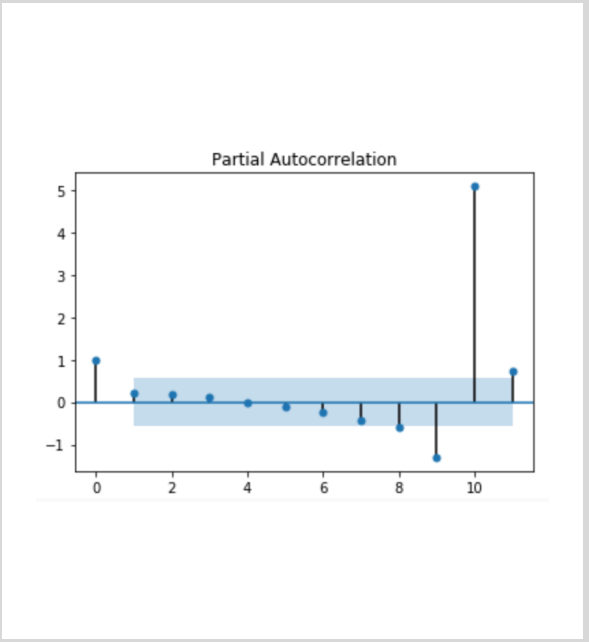
A line plot is created showing the expected values (blue) compared to the rolling forecast predictions (red).

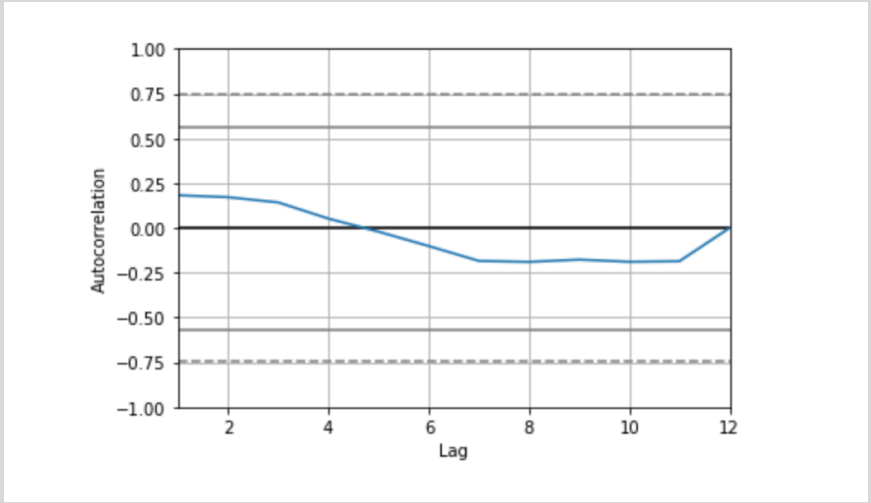
Test RMSE: 710.1796

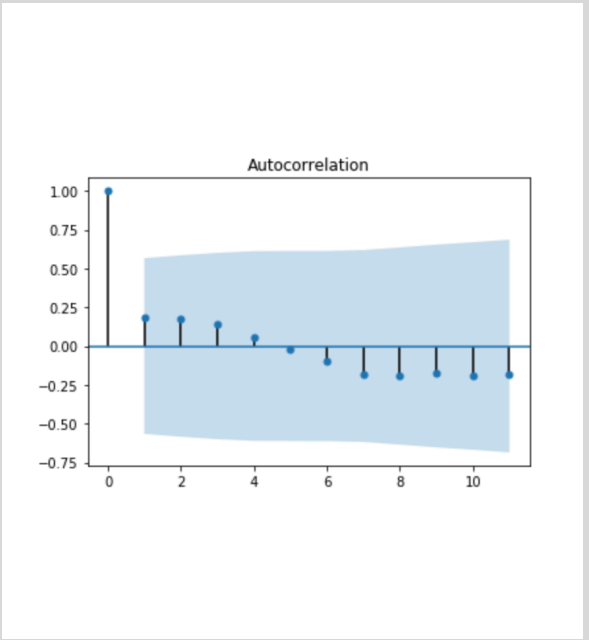
```
# one-step out-of sample forecast
forecast = model_fit.forecast(steps=5)[0]
forecast
```

```
array([2285.93941588, 2419.5423177 , 2553.14521953, 2686.74812136,
       2820.35102319])
```



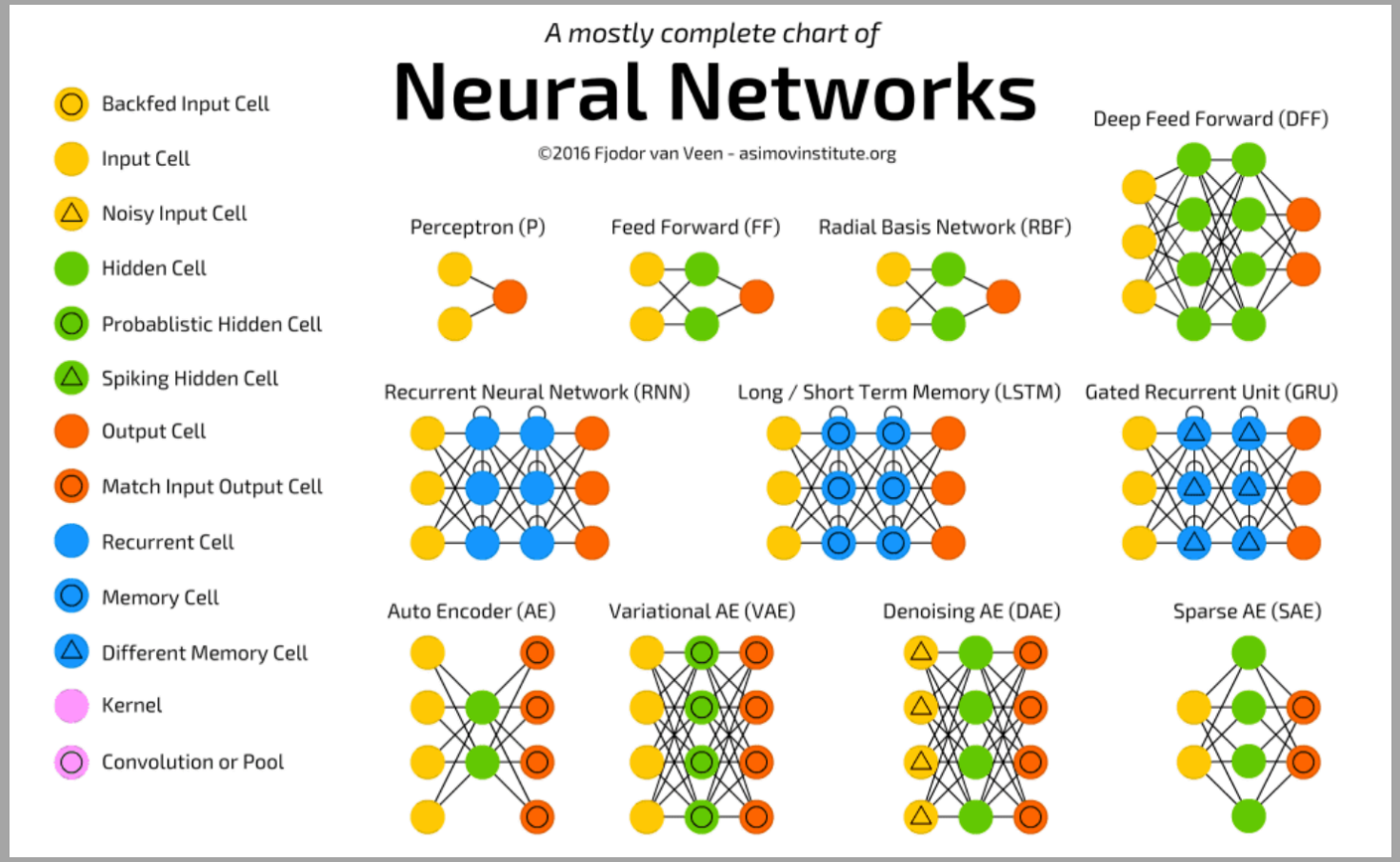






Glimpse of Neural Network

Why LSTM?



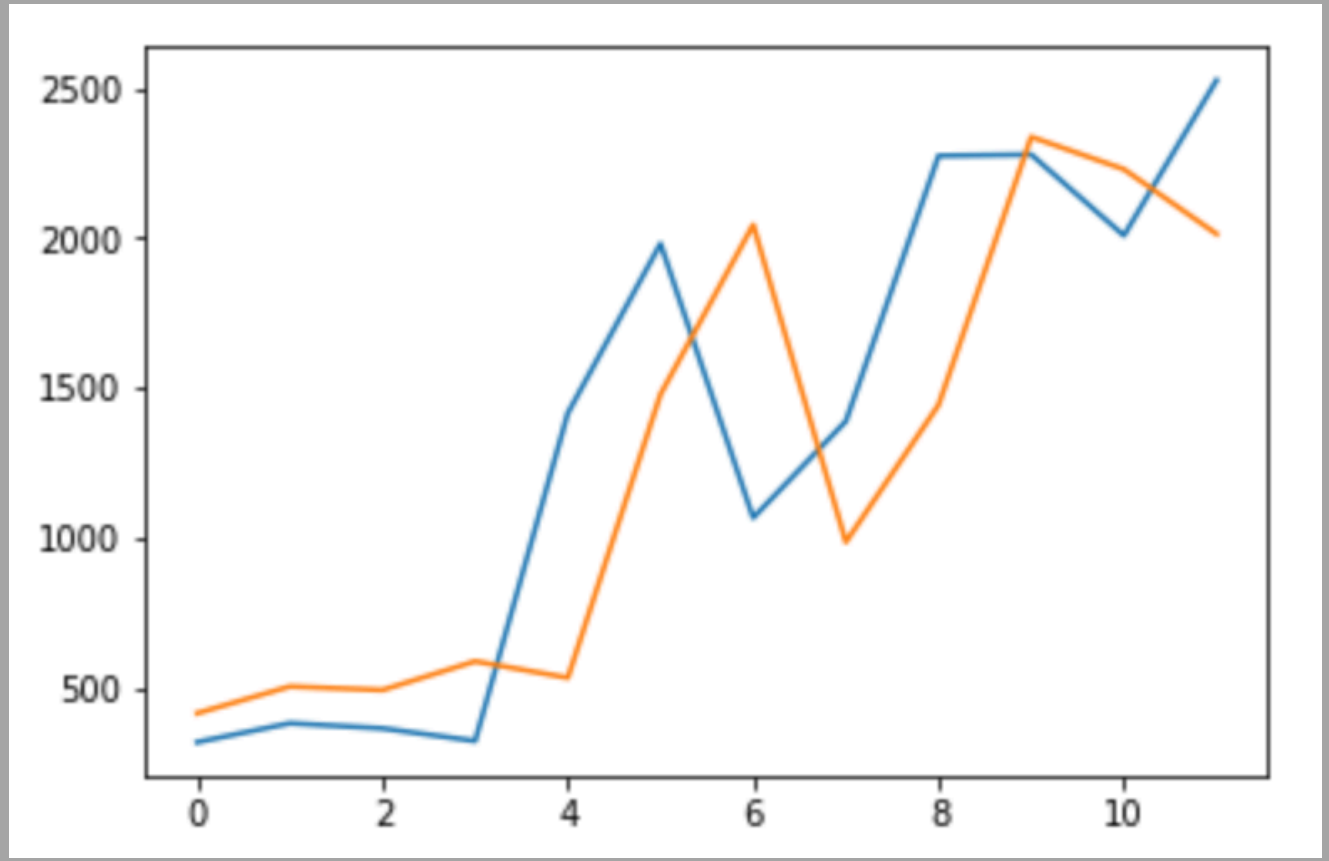
LSTM

- Transform the time series into a supervised learning problem
- Transform the time series data so that it is stationary.
- Transform the observations to have a specific scale.
- Develop LSTM model

A line plot is created showing the expected values (blue) compared to the rolling forecast predictions (orange).

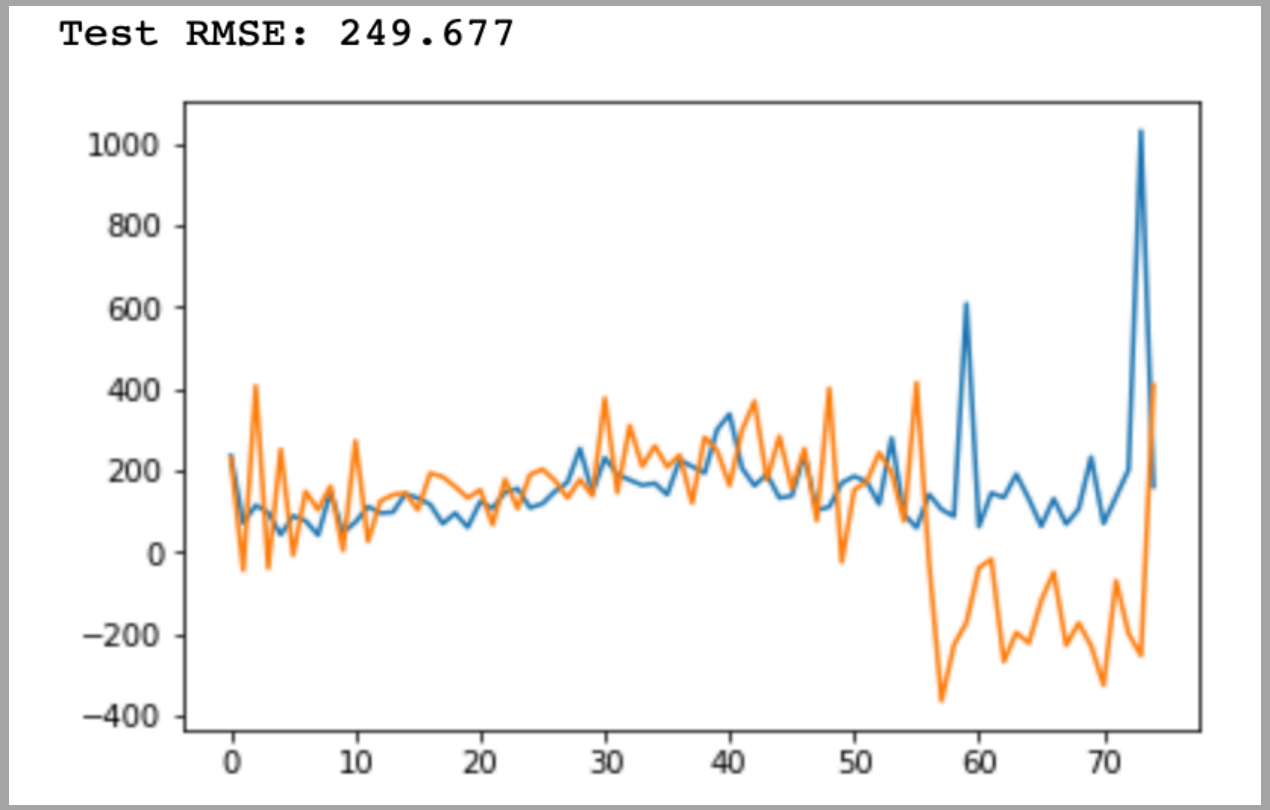
Test RMSE: 521.25

Average reduction in Error Rates obtained by LSTM is 27% compared to ARIMA indicating the superiority of LSTM over ARIMA.



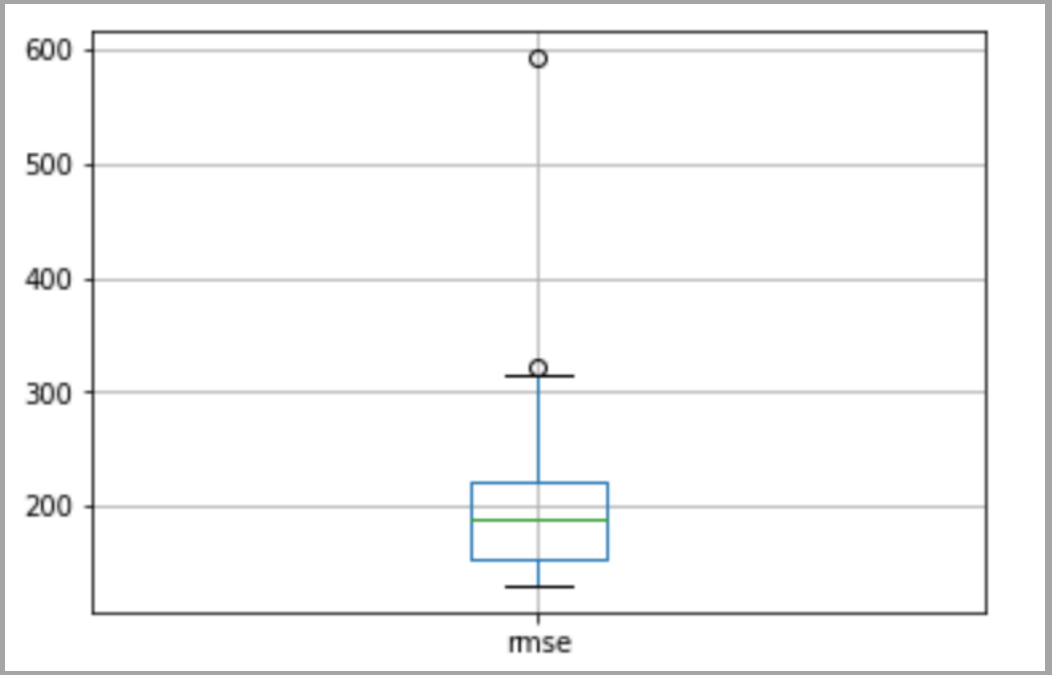
LSTM Model

- Consider Monthly data
- Model Evaluation: A rolling-forecast scenario will be used, also called walk-forward model validation. The root mean squared error (RMSE) will be used as it punishes large errors and results in a score that is in the same units as the forecast data, namely monthly growth in the number of Alternative Fuel Stations.
- Drawbacks: A difficulty with neural networks is that they give different results with different starting conditions.
- Approach: Develop a robust model repeating the experiment multiple times, then take the average RMSE as an indication of how well the configuration would be expected to perform on unseen data on average. This is often called multiple repeats or multiple restarts.



LSTM Robust Model

- We will use 30 repeats as that would be adequate to provide a good distribution of RMSE scores.
- Next Step: Tune the Hyperparameters and explore different configurations.

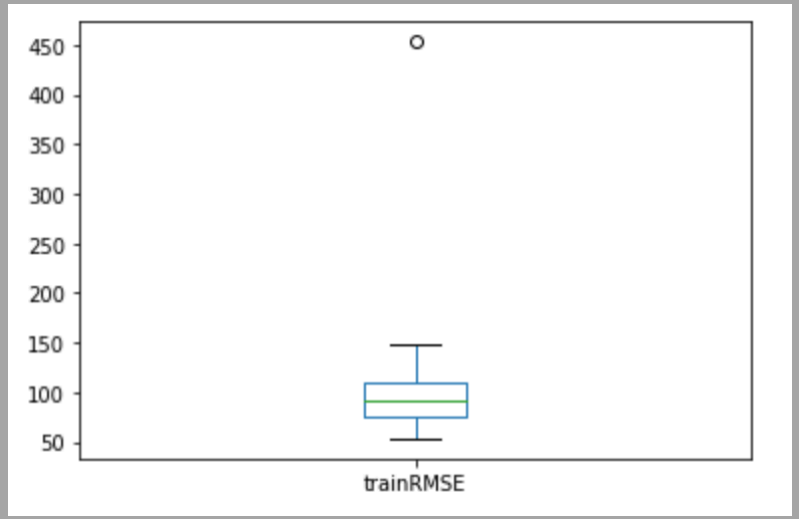


	rmse
count	30.000000
mean	207.626266
std	88.162054
min	130.316560
25%	154.672941
50%	187.908740
75%	220.793513
max	592.663033

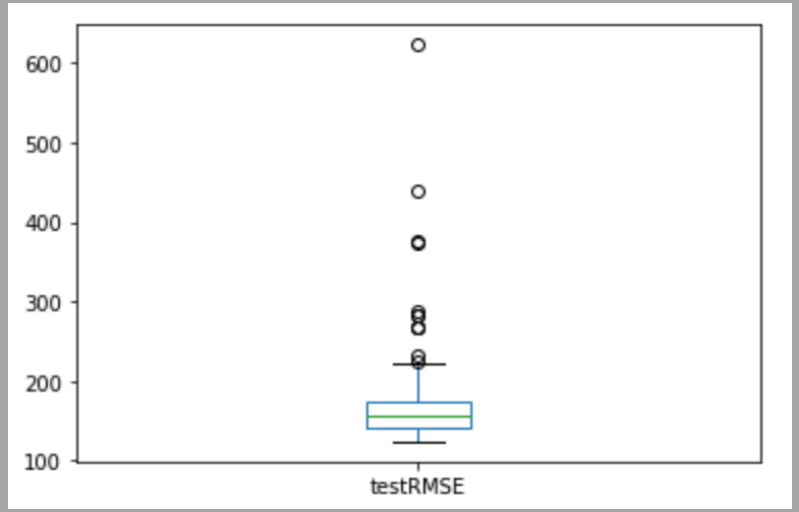
LSTM Hyperparameter Tuning

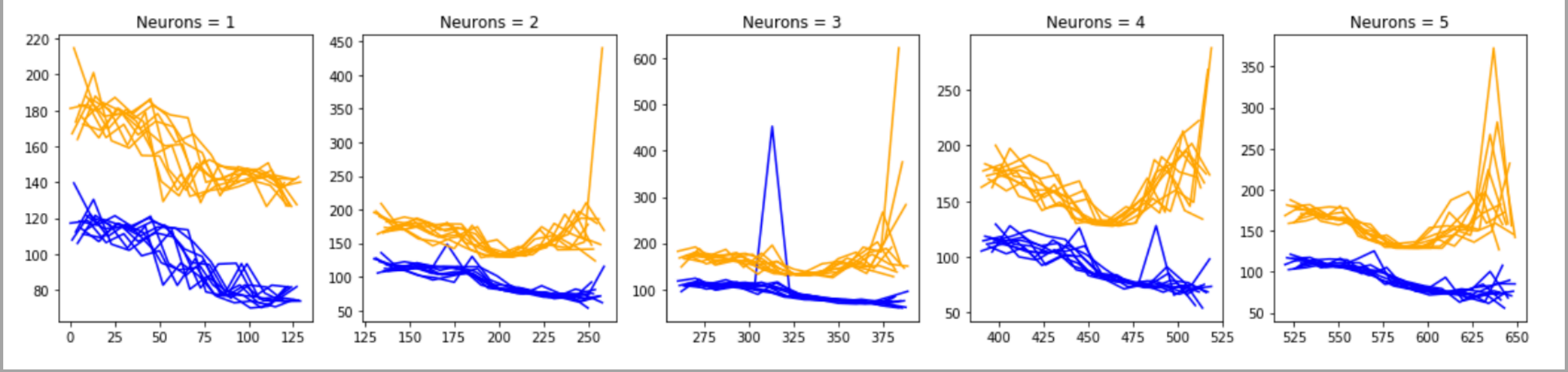
- No good theory on how to configure a neural network.
- Systematic approach to explore different configurations. (Range of Neurons - 1 .. 5, Range of Epochs - 2^0 ..2^12)
- Perform experimental runs for each scenario. (10 Runs)
- Next Step: Tune and interpret the results for number of training epochs and number of neurons.

Overall Range of train RMSE



Overall Range of test RMSE

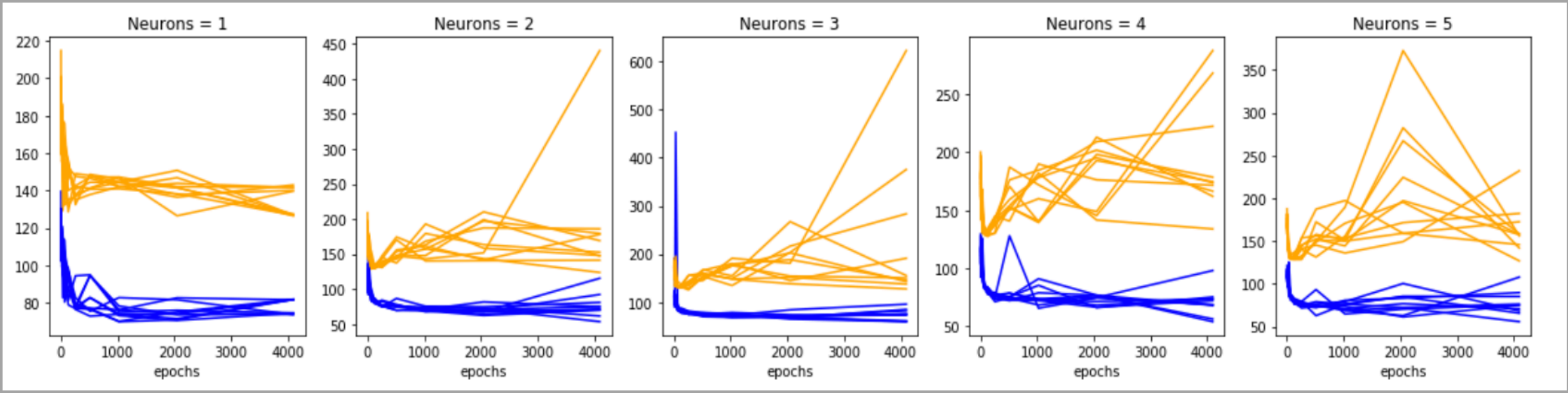




Diagnostic of 1 to 5 neurons

Hyperparameter Diagnostics

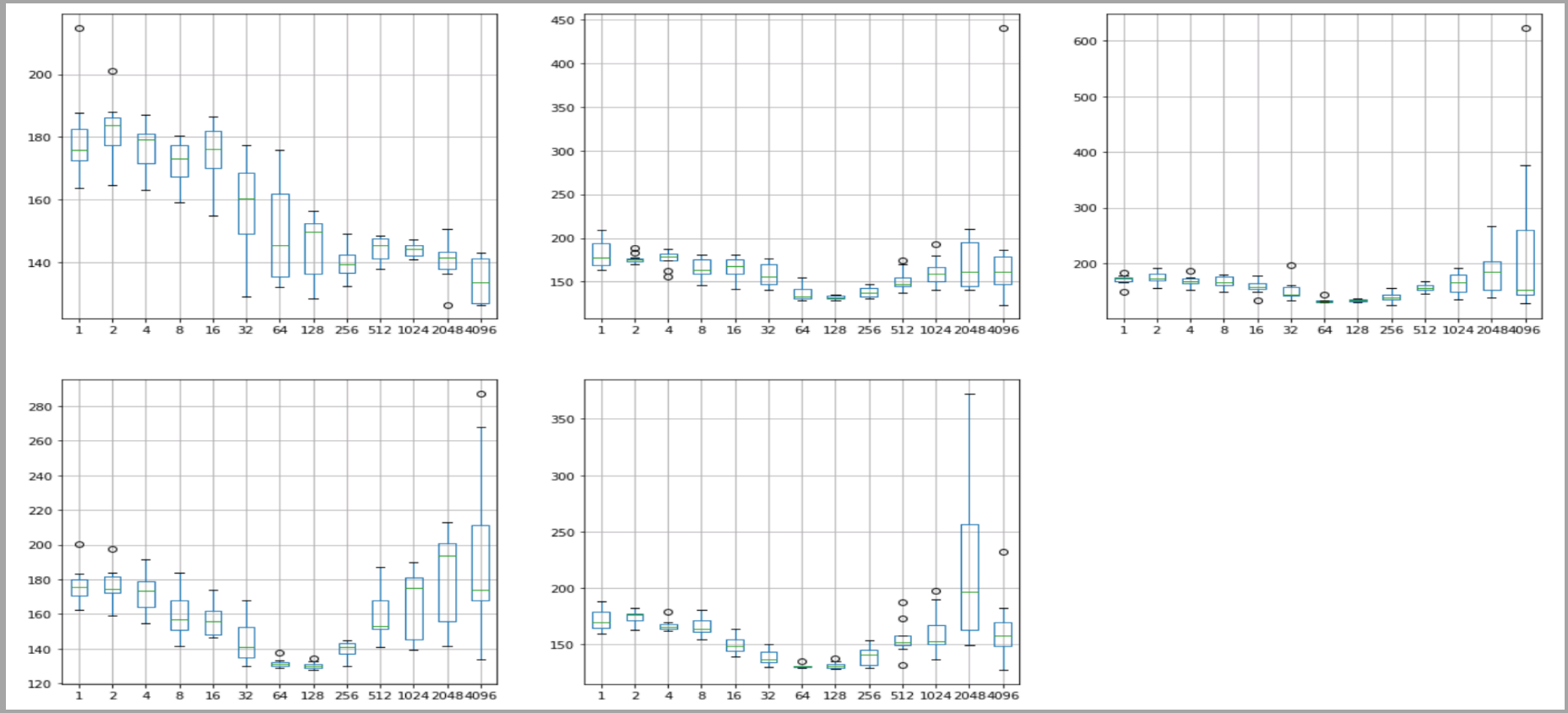
- A line plot of the series of RMSE scores on the train(blue) and test(orange) for each set of neurons is created.
- The number of neurons affects the learning capacity of the network. Generally, more neurons would be able to learn more structure from the problem at the cost of longer training time. More learning capacity also creates the problem of potentially overfitting the training data.



Hyperparameter Diagnostics

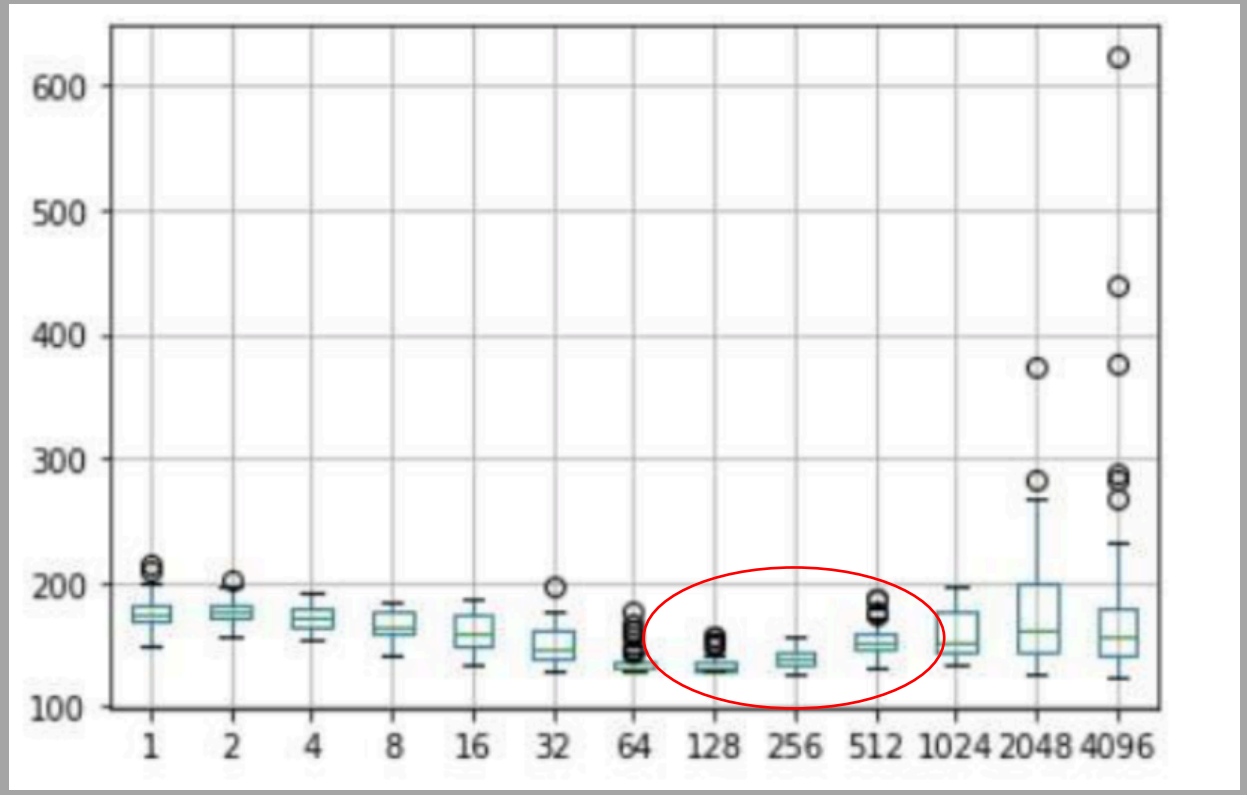
Diagnostic of 2^0 to 2^12 Epochs

- A line plot of the series of RMSE scores on the train(blue) and test(orange) for each set of neurons is created.
- The increasing trend is a sign of overfitting. This is when the model overfits the training dataset at the cost of worse performance on the test dataset. Severe overfitting shows sharp rise in test error.



Hyperparameter Diagnostics

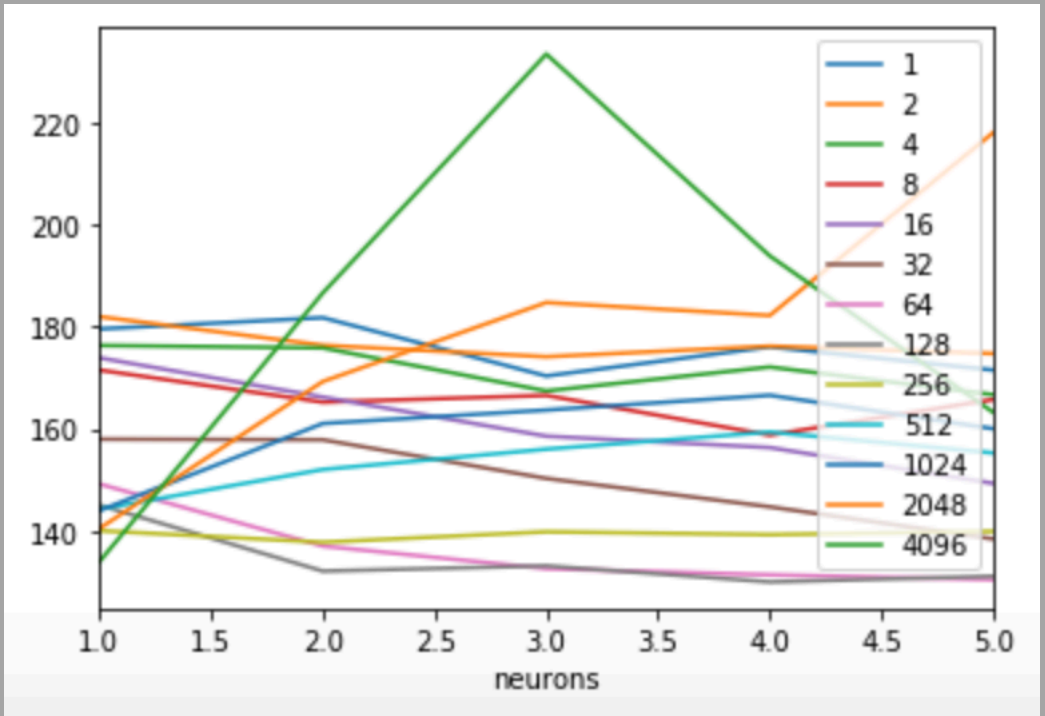
- Box plot showing overall range of test RMSEs for each Epoch by Neurons(1 to 5).



Hyperparameter Diagnostics

Box plot showing overall range of test RMSEs for each Epoch.

- The plot shows that the best possible performance may be achieved with epochs between 128 and 512.
- Tuning a neural network is a tradeoff of average performance and variability of that performance with an ideal result having low mean error with low variability meaning it is generally good and reproducible.



Hyperparameter Diagnostics

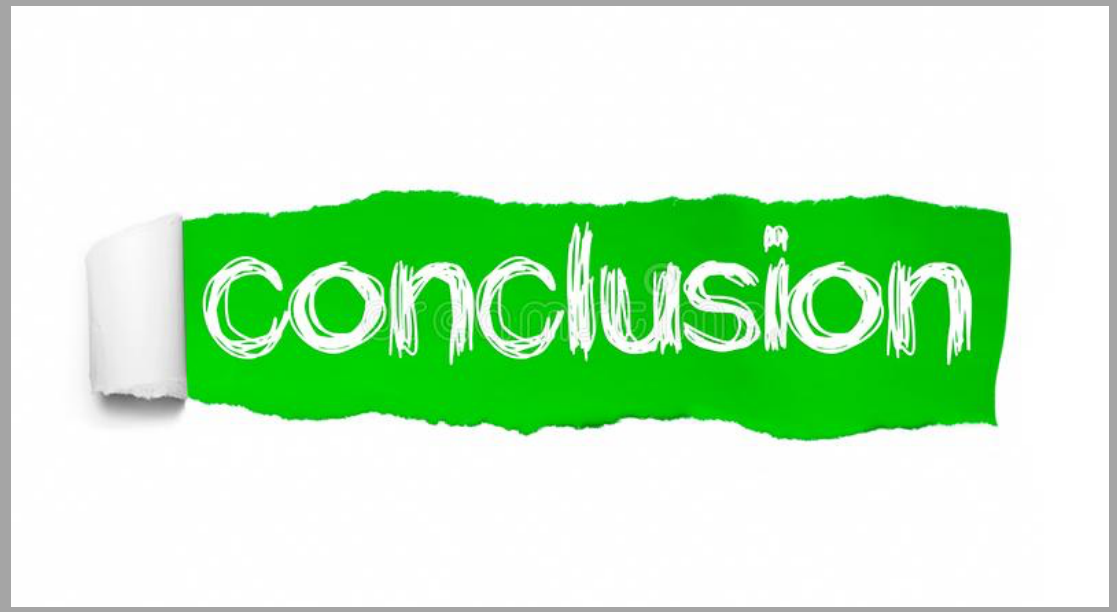
- The plot shows the lowest RMSE range being achieved with 128 epochs.
- The number of neurons affects the learning capacity of the network. Generally, more neurons would be able to learn more structure from the problem at the cost of longer training time. More learning capacity also creates the problem of potentially overfitting the training data.
- 4 Neurons gives the lowest RMSE compared to the others.

	1	2	4	8	16	32	64	128	256	512	1024	2048	4096
neurons													
1	179.543728	181.947533	176.316406	171.502963	173.894587	158.001037	149.242037	145.117036	140.110755	144.318498	144.030747	140.527129	134.035443
2	181.735583	176.350822	175.825733	165.229678	166.226762	157.877889	137.105458	132.211653	137.848912	152.069129	161.058701	169.261081	186.569272
3	170.350301	174.108840	167.417724	166.546551	158.579016	150.324816	132.691824	133.234635	139.896709	156.041915	163.684957	184.700825	233.266278
4	176.025395	176.182347	172.104560	158.786517	156.351783	144.774422	131.481355	130.097134	139.333244	159.397103	166.575353	182.171626	193.817633
5	171.509503	174.729199	166.690929	165.744443	149.359092	138.498329	130.555283	131.192563	139.974753	155.303036	160.037251	217.927600	163.233271

Hyperparameter Diagnostics

Conclusion

- From the mean performance alone, the results suggest a network configuration with 4 neurons as having the best performance over 128 epochs with a batch size of 1.
- This configuration also shows the lowest RMSE and tightest variance.



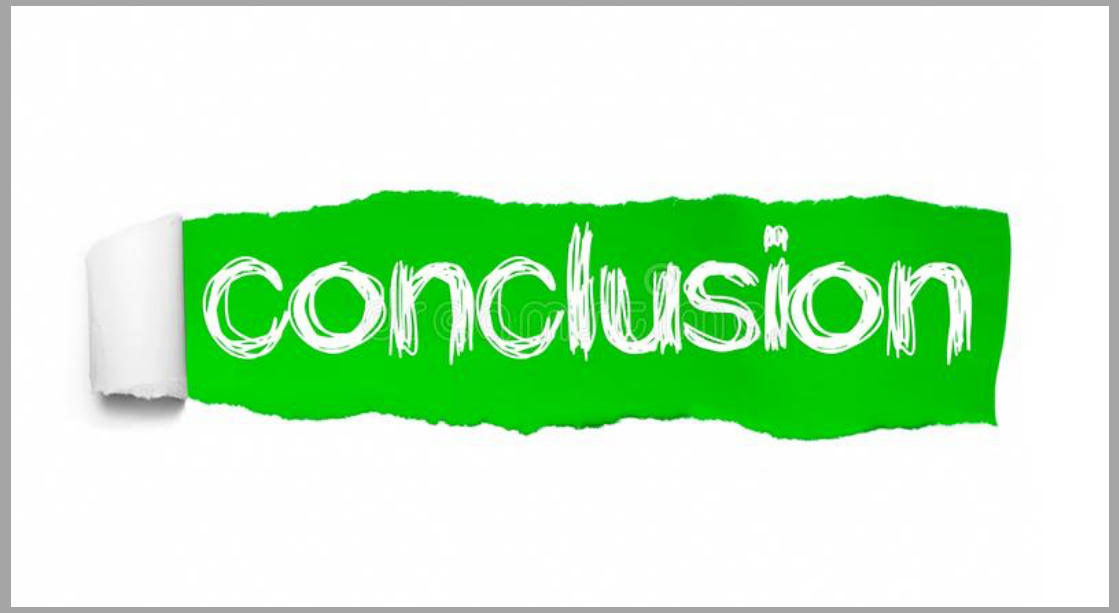
Model Perspective

- Limitation: Did not compare performance with other deep learning models

Approach: Other advanced LSTM versions and Deep Learning models could be implemented and tested for performance

- Limitation: Did not consider other features that might influence the dependent variable.

Approach: Other features like Fuel Cost, number of Vehicles equipped with Alternative Fuels etc. could be included for Multivariate Time Series Analysis.



Business Perspective

- Significant growth in use of Alternative Fuels in the recent past.
- Provides sustainable growth by most importantly reducing Carbon dioxide emissions.
- In the last decade, the average retail alternative fuel prices has been stable or slightly on a decline trend. https://afdc.energy.gov/fuels/prices.html.
- I believe this trend to continue favoring Alternative Fuels due to following factors:
    1. As dependency on petroleum fuels continue to decline.
    2. Increase in Vehicle Owners with sustainability mindset.
    3. Leap in technology to produce Alternative Fuels with lower production cost.