Predicting the Market Penetration of

Alternative Fuel Stations and Vehicles - United States

Priya Sathish

Abstract

Transportation sector is dependent on Petroleum based fuel that results in higher emission level. The United States consumes approximately 20 million barrels of petroleum per day, about three-fourths of which is used for transportation. Transportation also has a significant economic impact on American businesses and families, accounting for nearly one-sixth of the average household's expenses. Improving efficiency and reducing costs in this sector can thereby make a notable impact on our economy.

Increased economic and energy security aren't the only benefits. Widespread use of alternative fuels and advanced vehicles could reduce the emissions that impact our air quality and public health. Production and distribution of these alternative fuels that can be economically available is a constraint for the increase in the number of alternative fuel driven vehicles.

One known fact is "attaining environmentally beneficial transportation fuel at an affordable price" is the way of the future. This project focuses on the number of alternative fuel stations and the number alternative fuel powered vehicles across the United States, its growth trend and developing a predictive model.

Load Data

Station data and vehicle inventory data has been collected from https://afdc.energy.gov/data_download/ (https://afdc.energy.gov/data_download/). Data wrangling performed and saved in file Alt_Fuel_DW.ipynb from which data from the following cleaned up csv files will be loaded for Exploratory Data Analysis.

Load Shape files

We load the United States shapefile for mapping.
'shape_us_state/tl_2017_us_state.shp'

https://medium.com/@erikgreenj/mapping-us-states-with-geopandas-made-simple-d7b6e66fa20d

We load the geo-json file to map CA.
'ca_california_zip_codes_geo.min.json'

Data Wrangling Explained

Acquired data from https://afdc.energy.gov and loaded csv files as pandas dataframes removing rows with invalid data by setting the parameter error_bad_lines as False. Switching error bad lines turns off the stdout error messages from showing.

Station Data:

Strip leading and trailing white spaces from Zipcode column in the dataframe and selected first 5 characters to be the Zipcode to avoid long codes and converted to numeric datatype. Look for invalid zipcodes and drop them from the dataframe. Clean column names by removing white spaces.

Set proper indices. Drop duplicate rows if any exist.

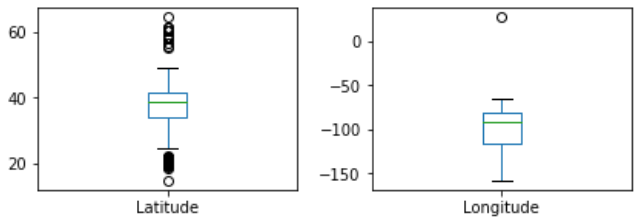
Create a new dataframe by performing a deep copy from the raw data filtering only rows with a valid open date.

We are going to do an anaylsis of the growth of alternative fuel stations across United States, so it is necessary that we have valid open date for the existing stations. Convert the date fields to pandas datetime format.

Create a new column for the Year the station was opened by retrieving the year from the open date. Convert the OpenYear column to contain numeric values.

Select only required fields from the raw data and save as new dataframe containing the station ID, FuelTypeCode, StationName, StreetAddress, IntersectionDirections, City, State, ZIP, GeocodeStatus, Latitude, Longitude, Country, OpenDate and OpenYear.

Check for outliers for Latitude and Longitude. From the boxplots, we can see that there are few outliers with Latitude and Logitude of 0 also. We can drop those rows which will enable accurate mapping plots.



Save the station data frame to new csv files for future use to perform Exploratory Data Analysis and ML.

Vehicle Data:

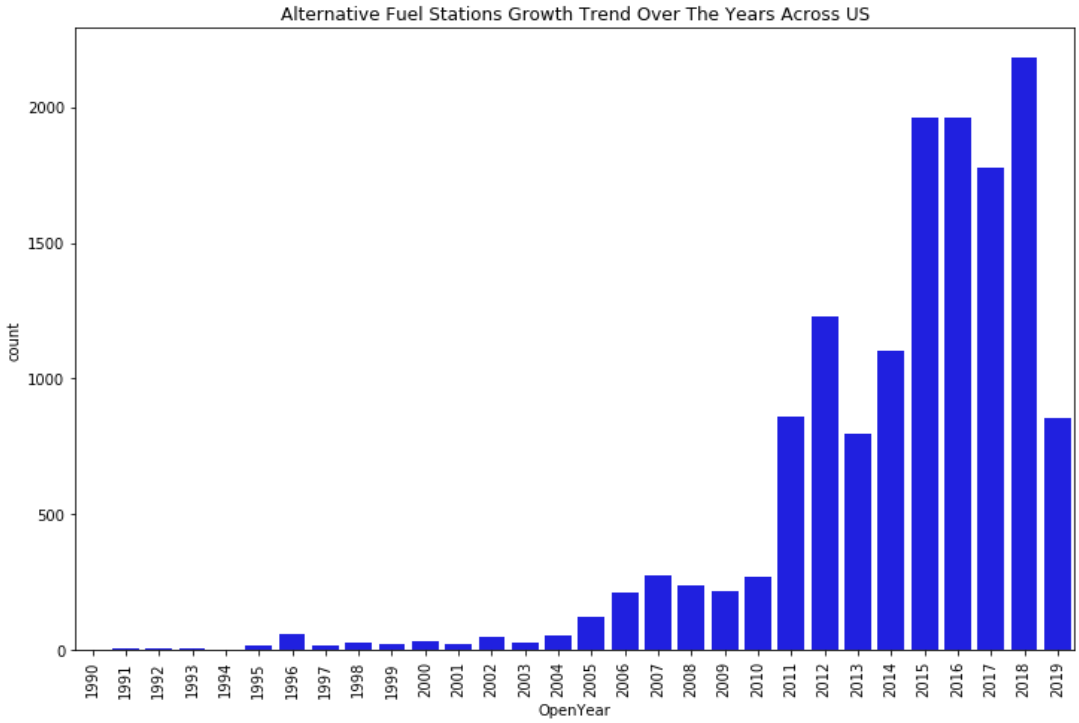
The data frame contains Year as the column header. We want Fuel Type to be the column header.
We transpose the data frame and make first row as the column header to contain Fuel Types as the column names.

PEVs are electric vehicles. Let us keep all electric vehicles under one Fuel Type ELEC.
Add the number of PEVs to the number of ELEC vehicles and group them as one Fuel Type and drop the PEVs column.

Set the index of the data frame to Year.

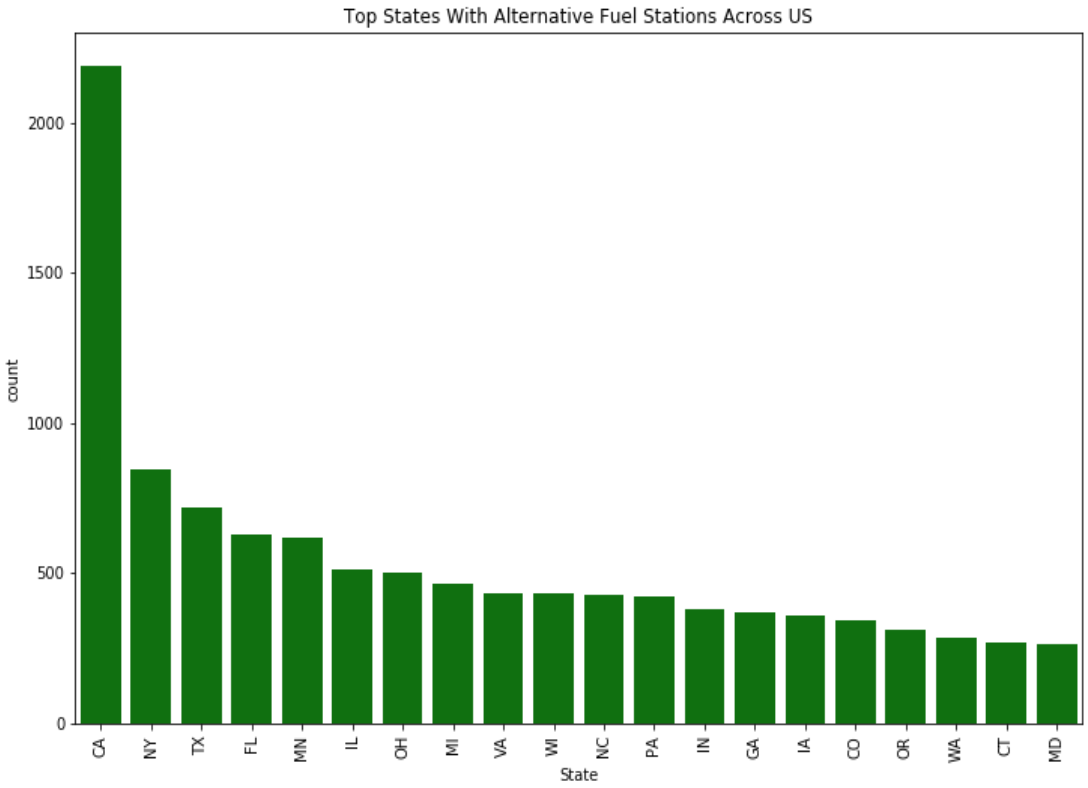
Save the vehicle data frame to new csv file for future use to perform Exploratory Data Analysis and ML.

Perform Exploratory Data Analysis



The trend shows overall growth in the alternative fuel stations year on year, with remarkable growth since 2011. This might be due to increased awareness created by Clean Cities Coalitions in the community. The mission of Clean Cities coalitions is to foster the economic, environmental, and energy security of the United States by working locally to advance affordable, domestic transportation fuels, energy efficient mobility systems, and other fuel-saving technologies and practices. Clean Cities coalitions are comprised of businesses, fuel providers, vehicle fleets, state and local government agencies, and community organizations.

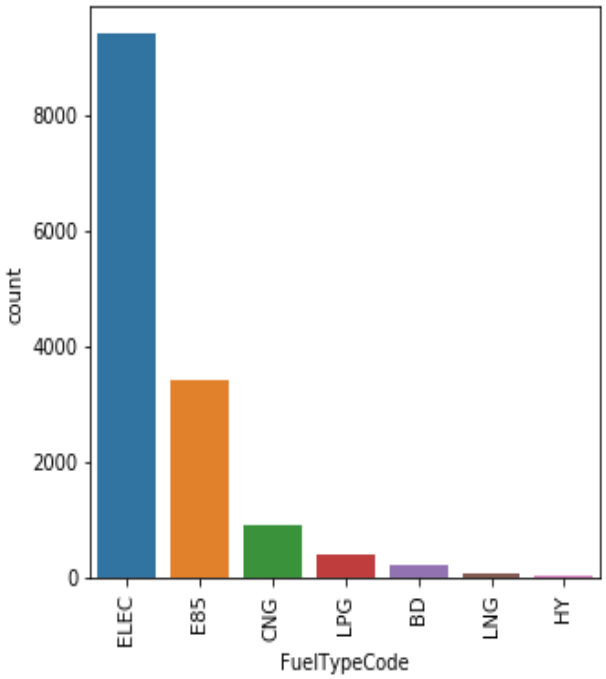
Note: Number of alternative fuel stations in 2019 is accumulated sum until the month of April. At this rate it is expected to surpass the total number of stations above 2500 for the year ending.



Synopsis of the above plot shows highest concentration of alternative fuel stations in California followed by New York and Texas states. Upon research State of California has launched a campaign to reduce consumption of petroleum and diesel fuels and shifting towards cleaner fuels. This campaign is known as Low Carbon Fuel Standard (LCFS) http://www.cadelivers.org/lowcarbon-fuel-standard/ (http://www.cadelivers.org/low-carbon-fuel-standard/).

From 2011 to 2018 the LCFS has avoided 13.7 billion gallons of petroleum, increased 74% use of clean fuels, invested 2.8 billion in clean fuel production.

The LCFS is working to transform the fuels market from one that relies almost entirely on petroleum-based fuels to a diversified one that uses a variety of alternative fuels. This market-based transition to clean, low carbon fuels is leading to technology innovation that's helping California meet its long-term climate, clean air, and public health goals. Just as the state requirements are helping shift our electricity mix to more renewables like wind and solar, the LCFS is also delivering the clean fuels we need today and going forward.

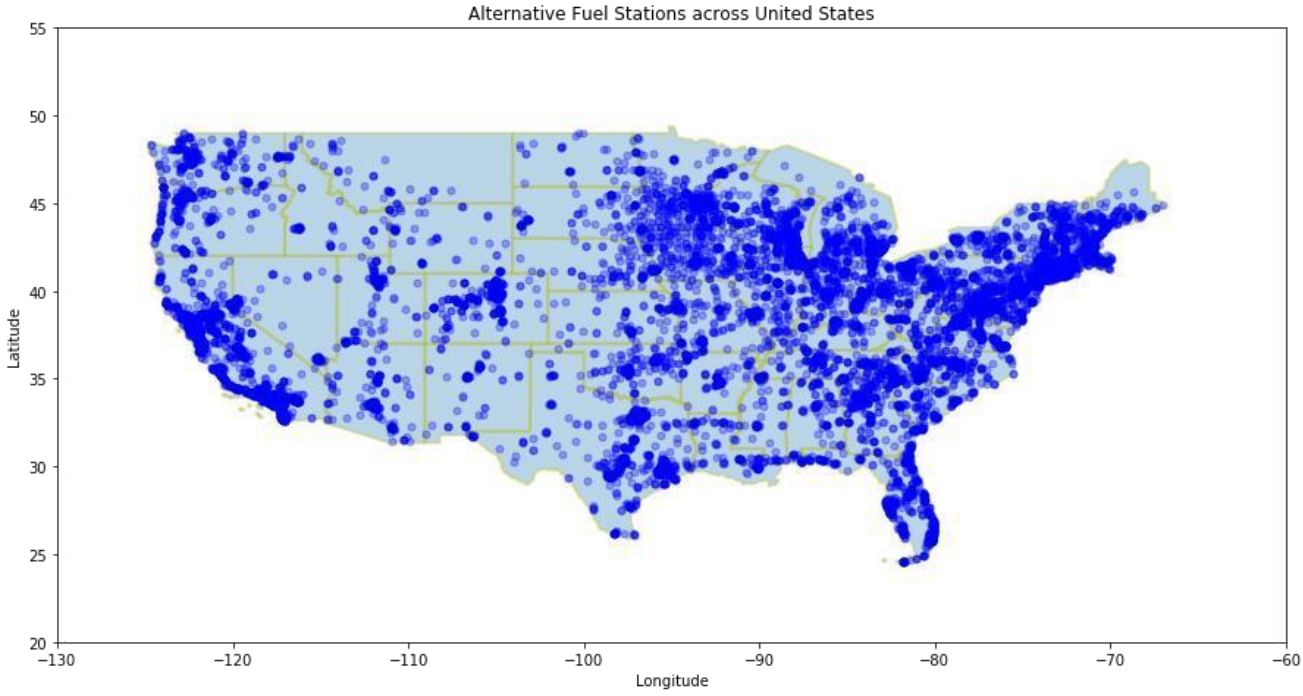


About 90% of the alternative fuel used in transportation sector is comprised of electricity and ethanol85. Among them 65% of the fuel is electric based. The reason for the high number of electric charging stations is attributed to No Tail Pipe Emissions and environmentally friendly.

Plug-in electric vehicles (PEVs) are capable of drawing electricity from off-board electrical power sources (generally the electricity grid) and storing it in batteries. Electricity for charging vehicles is especially cost effective if drivers are able to take advantage of offpeak residential rates offered by many utilities. In many cities, PEV drivers also have access to public charging stations at libraries, shopping centers, hospitals, and businesses. Charging infrastructure is rapidly expanding, providing drivers with the convenience, range, and confidence to meet more of their transportation needs with PEVs.

The next popular alternative fuel is E85(ethanol) that is predominantly produced by fermenting and distilling starch crops mostly corn. This is 100% renewable fuel and roughly 1 acre of corn can be processed into 330 gallons of combustible ethanol.

Alternative Fuel Stations Distribution in United States

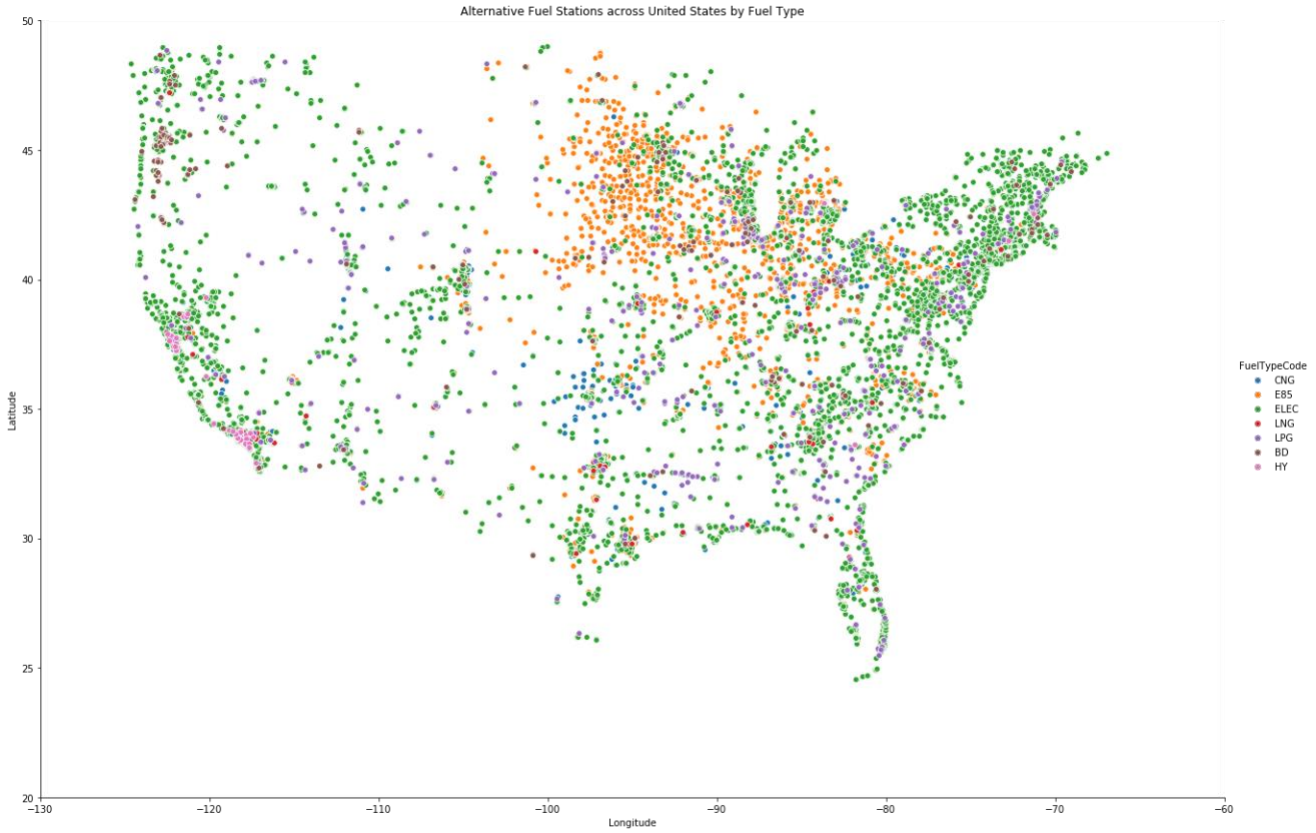


The above geographic plot shows the concentration of alternative fuel stations across United States. It clearly indicates that alternative fuel stations are popular along the coastal line and as well where the population density is high.

Many alternative fuels exist, but few are as bountiful, easily produced and cost effective as traditional fossil fuels. The emissions impact and energy output provided by alternative fuels vary, depending on the fuel source.

Few commonly consumed alternative fuels include Ethanol, Natural Gas, Biodiesel, electricity (batteries), hydrogen (fuel cells), non-fossil methane, propane and other biomass sources.

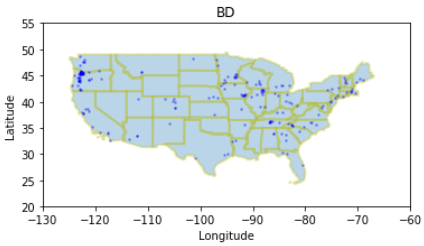
Alternative Fuel Stations Distribution by Fuel Type

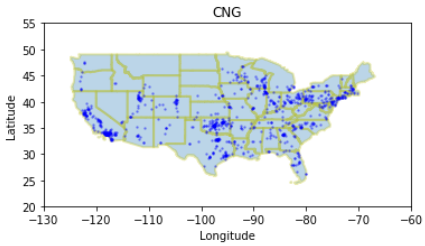


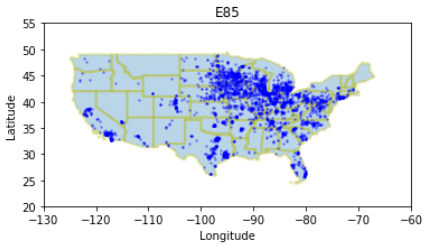
The above geo plot shows various types of alternative fuel stations across United States. Key points are:

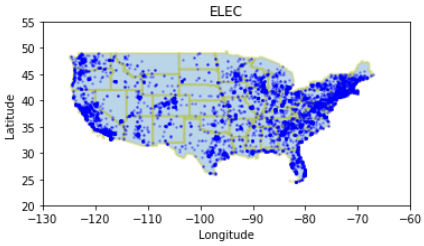
- Electric fuel stations are prominent along the coastline.
- E85 fuel stations are concentrated in North Central states.
- HY stations are concentrated mainly in CA along the coastline.

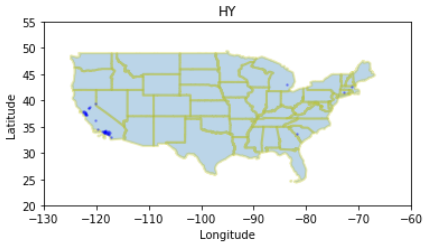
Individual plots to show the concentration of stations with respective fuel types across the United States

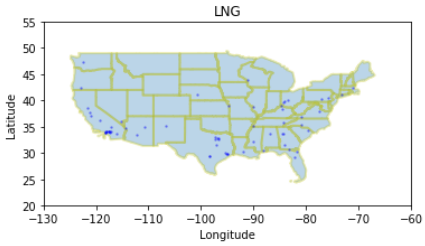


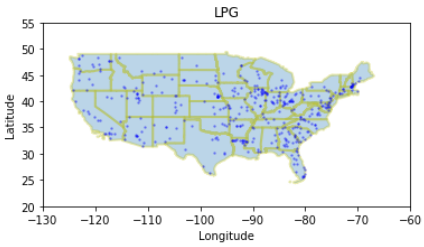












Biodiesel:

Produced from vegetable oil or animal fats, the projected production of biodiesel in the US is nearly 12 billion gallons. North Carolina seems to be leading the way in widespread availability of biodiesel.

Compressed Natural Gas:

If you've seen a municipal bus that's powered by natural gas, it's using compressed methane. With the highest popularity in oil rich nations outside of the US, it's seeing limited use domestically.

E85 (85% Ethanol gas):

An ethanol fuel blend of 85% ethanol, 15% gasoline. Flex fuel vehicles can use E85. Popular in corn growing states where there are local and Federal subsidies.

Electric vehicle charging stations:

Heavy subsidies and the success of Tesla have increased the market viability and variety of all electric vehicles. An increase in pay-as-you-go charging stations as shown in this map make it feasible to travel for longer distances than previously possible.

Hydrogen:

A *hydrogen highway* is a chain of hydrogen-equipped filling stations and other infrastructure along a road or highway. Italy and Germany are collaborating to build a hydrogen highway between Mantua in northern Italy and Munich in southern Germany.

Liquified Natural Gas:

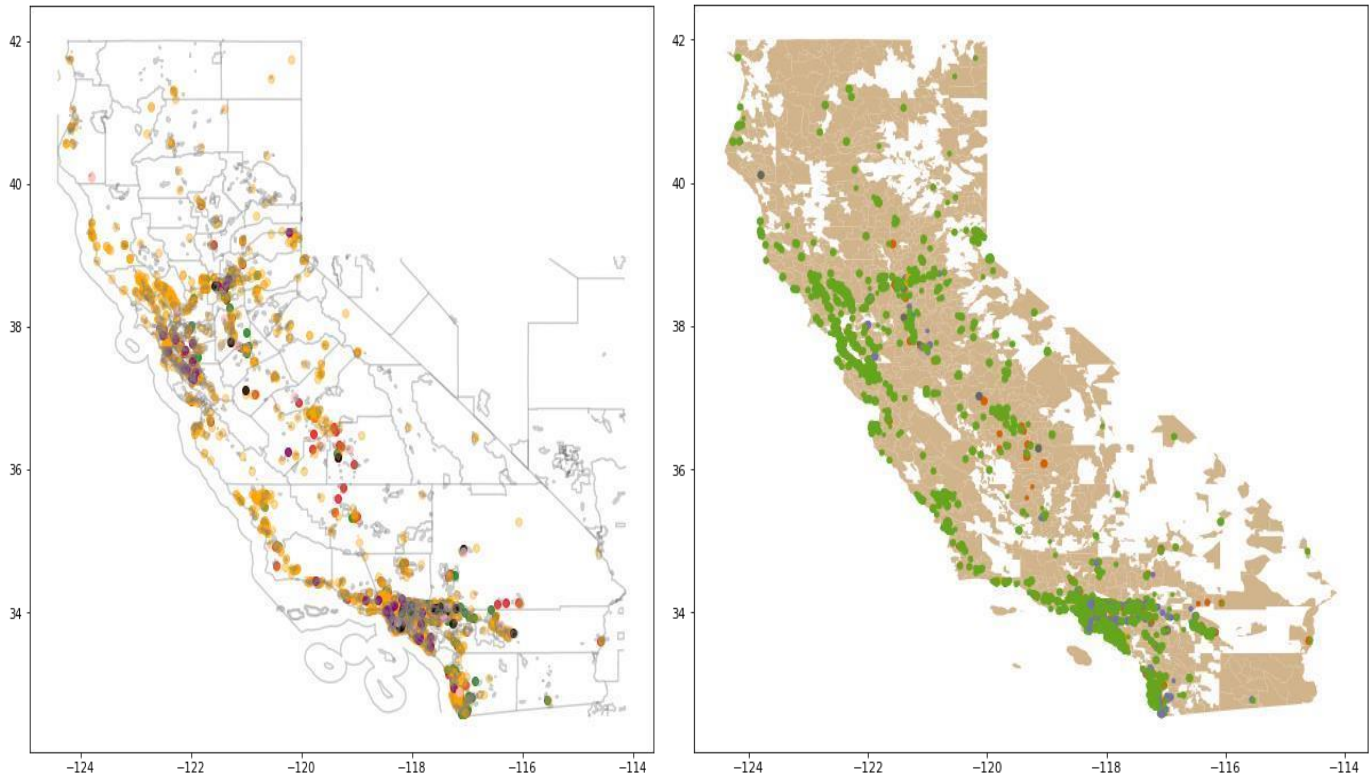
Liquified natural gas is primarily methane that has been converted to liquid for ease of storage or transport.

Propane:

With propane's wide availability and lower maintenance costs, it's an attractive option for light duty industrial vehicles.

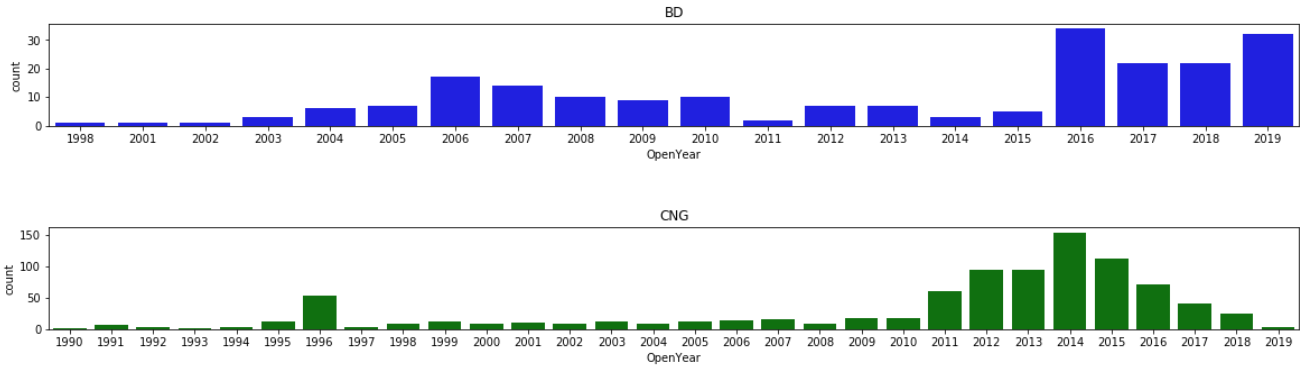
Example

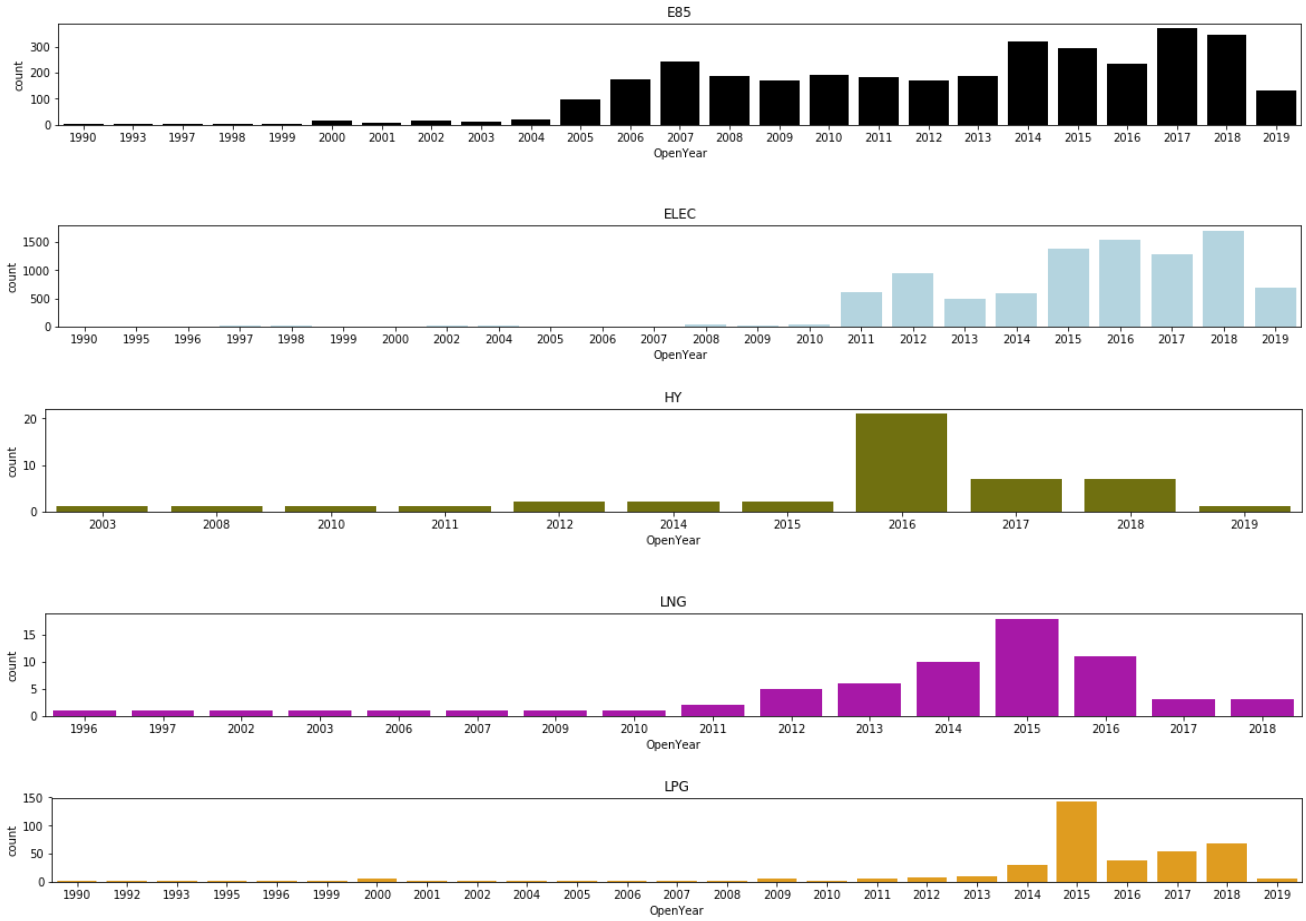
California has the highest number of alternative fuel stations as per the data and as we can see tremendous growth in the electric fuel stations over the past years.



Closer look at the distribution of alternative fuel stations by different fuel types in the state of California

Trend of Fuel Stations by Fuel Type – 1990's onward

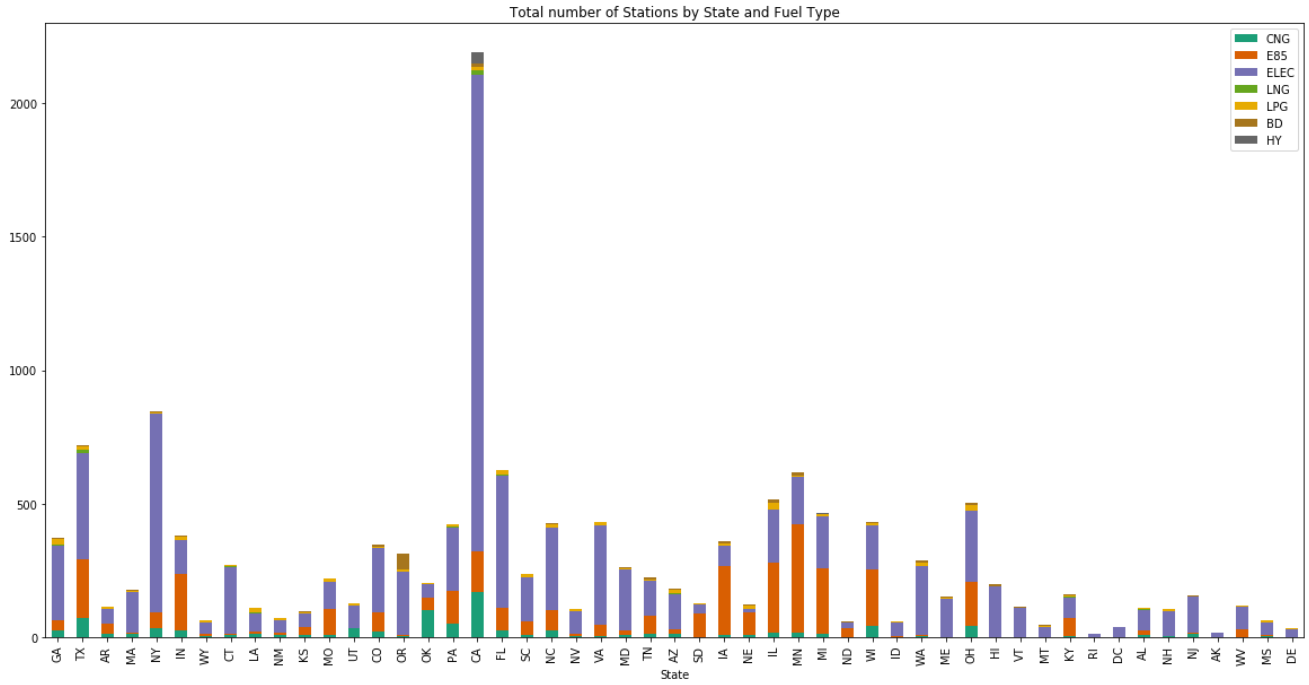




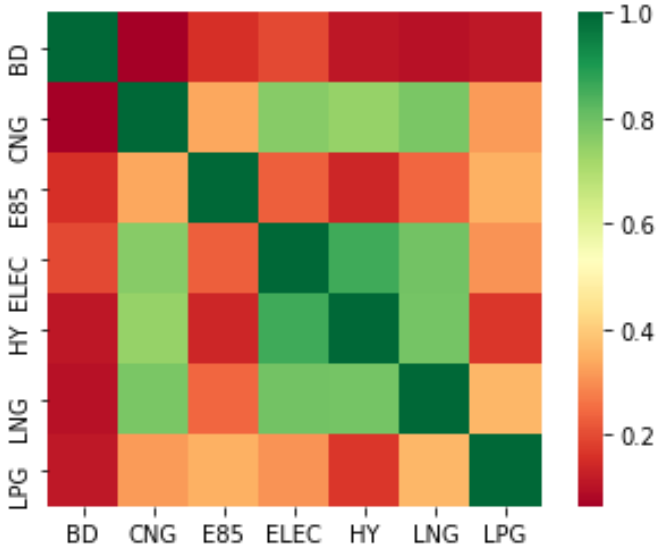
Why there is decline in CNG and LNG in transportation industry?

With growing popularity of EV's combined with steady improvement in battery technology, the pervasiveness of the grid, and services to enable things like high-speed charging makes the EV's market inevitable.

Secondly, the high cost of filling stations has also kept a lid on CNG / LNG. The CNG or LNG driven vehicles get worse mileage than regular gas or EV's.
Nevertheless, CNG is increasingly touted as the fuel of the future for the U.S., in part because of declining prices and large reserves and growing demand.



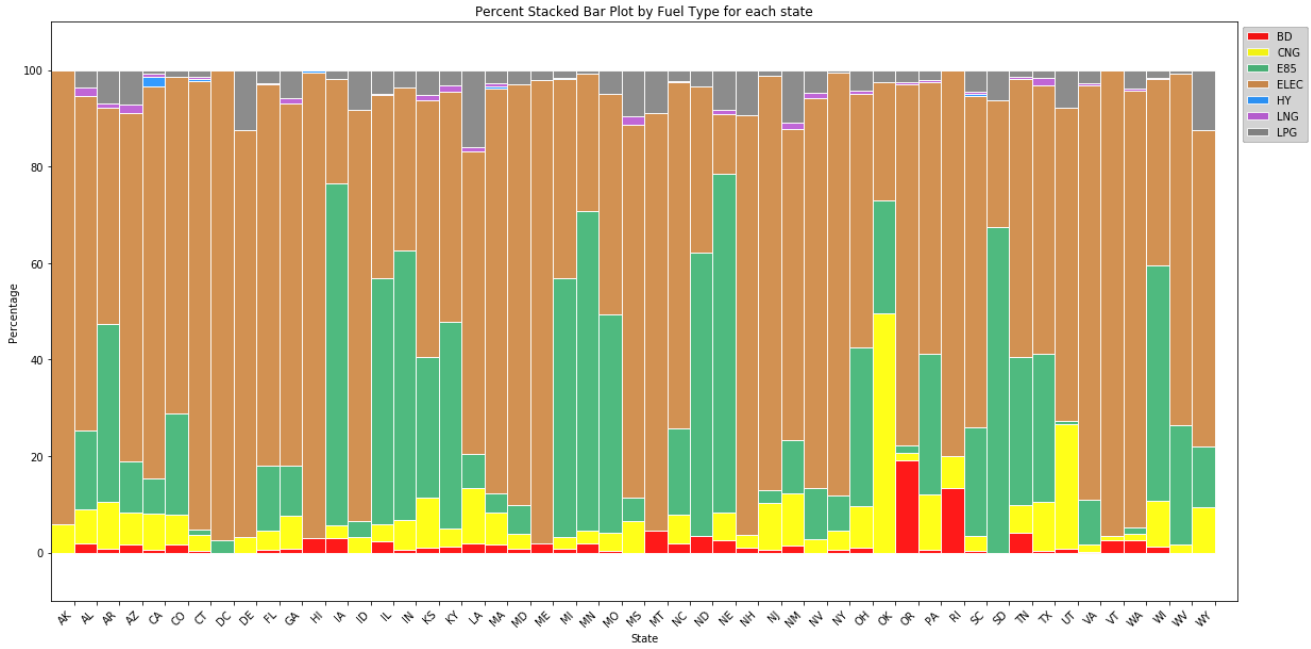
The stacked bar plot shows the distribution of the different alternative fuel stations in each state. CA has the most Electric, CNG, LNG and Hydrogen fuel stations. MN has the most E85 fuel stations. IL has the most LPG fuel stations. OR has the most BD fuel stations.



Correlation Heat Map

The heat map shows correlation among various alternative fuel stations by type. It shows high correlation amongst CNG, HY and ELEC. There is no correlation amongst BD, E85 and LPG.

Percentage Distribution of Alternative Fuels in Each State

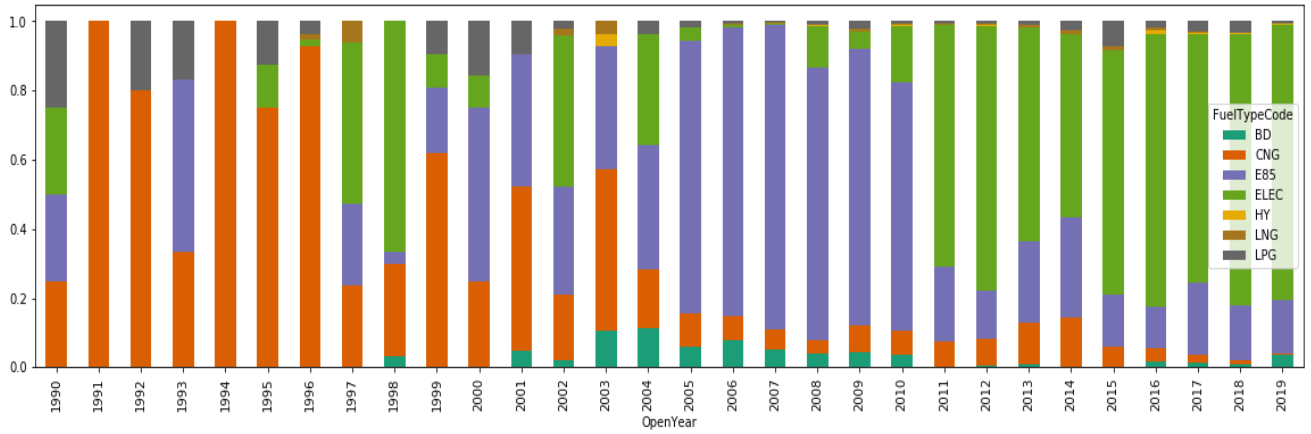


In general, electric fuel stations are predominantly higher in proportion in most of the states except for those where ethanol type fuel is produced more which is supported by the corn fields. This can be seen in states like NE, IA, MN, SD and WI.

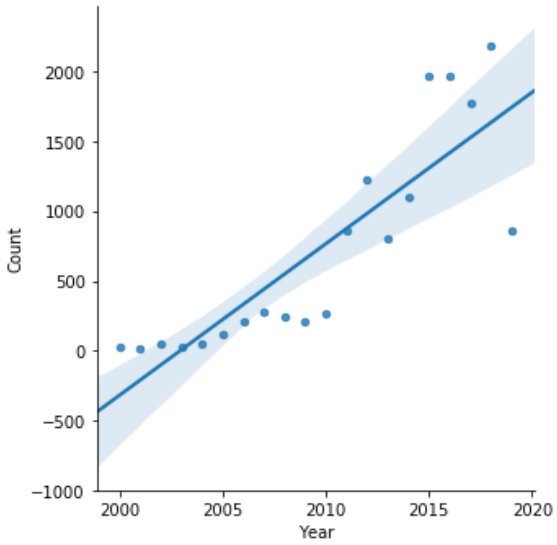
State regulations and incentives are spurring CNG station investment and vehicle adoption. It is no accident that states like Oklahoma and Utah have the highest concentrations of CNG stations and vehicles. Oklahoma provides waivers of state tax liability to in-state suppliers of CNG station components. Utah requires that CNG be supplied by regulated utilities (effectively capping prices).

CA has adopted HY based fuel to meet its long term climate, clean air and public health goals. HY fuel stations are clustered in major metropolitan cities in CA that are also tech savvy. CA has about 40 HY fuel stations for about 6500 fuel cell cars as of May 2019. There is now 1 retail station per 163 vehicles. Though it is in growing trend it is expected to not reach the states set goal of 100 HY fueling stations by 2020.

Percentage Distribution of Alternative Fuels by Year

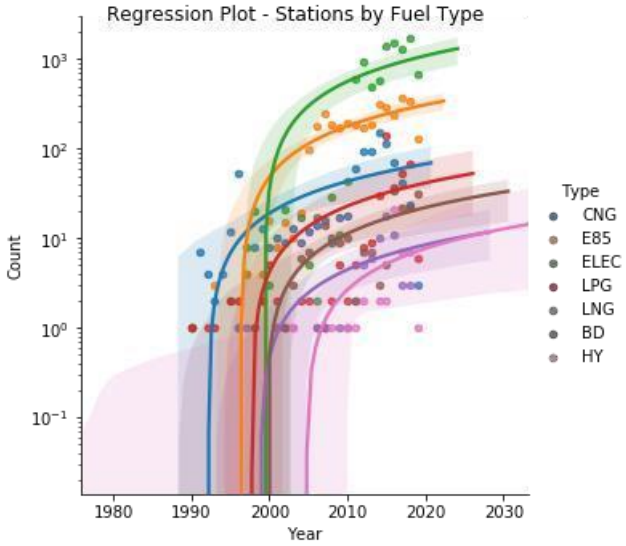


In 90's the majority of alternative fuels stations opened were Biodiesel based which has shifted towards electric and E85 type alternative fuels mainly due to advancement in technology and affordable cost. In the past decade it is distictly focused on growth towards electric based stations. This combined with higher number of electric vehicles justify the growth of number of stations.

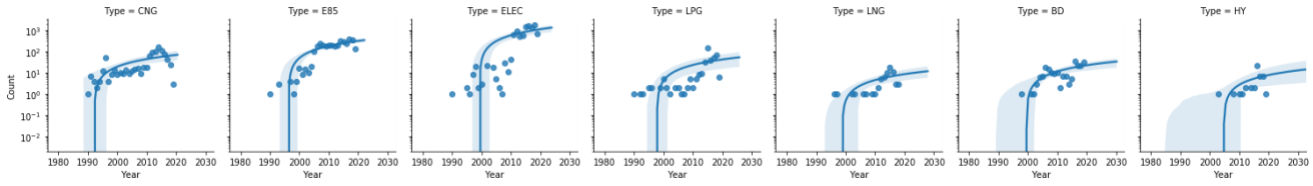


Over the last 15 years the total number of alternative fuel stations is in a steeper growth that is at a rate of 100 to 125 per month fuel stations year over year.

Regression plot with log scale on y axis enables the separation between various fuel type stations data points for better visualization.



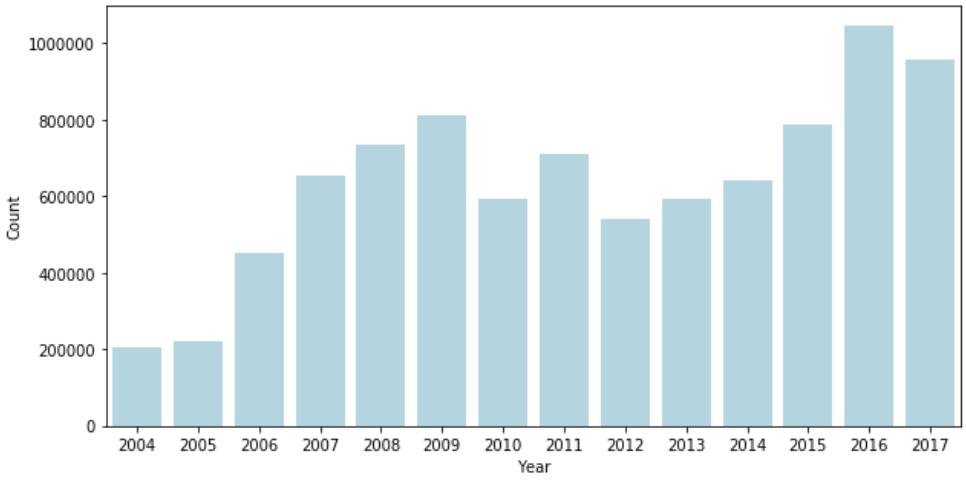
sns lm plot to show the regression by fuel type

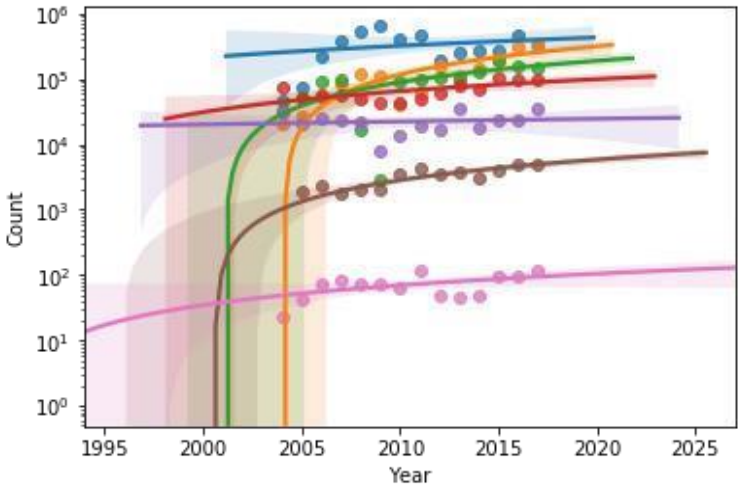


Individual regression plots of alternative fuel stations for each fuel type.

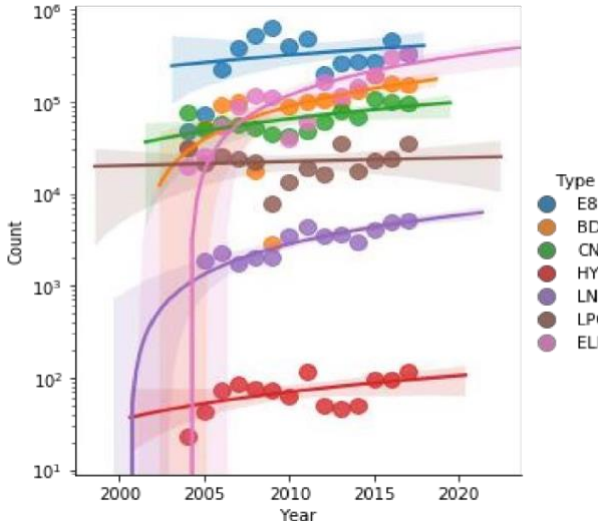
Vehicle Data

Alternative fuel equipped vehicles trend over the years

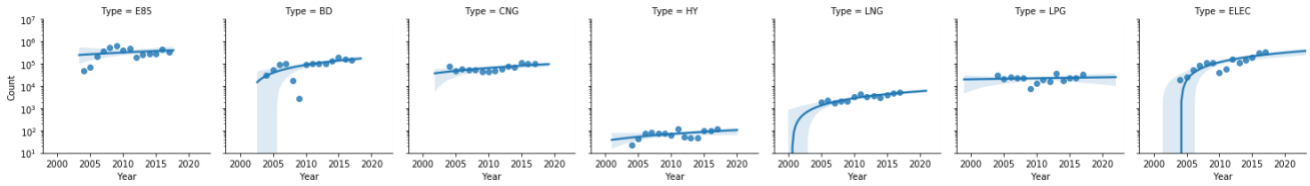




Regression plot with log scale for vehicles equipped with alternative fuels shows an increase in growth trend which coincides with the growth trend of number of alternative fuel stations.



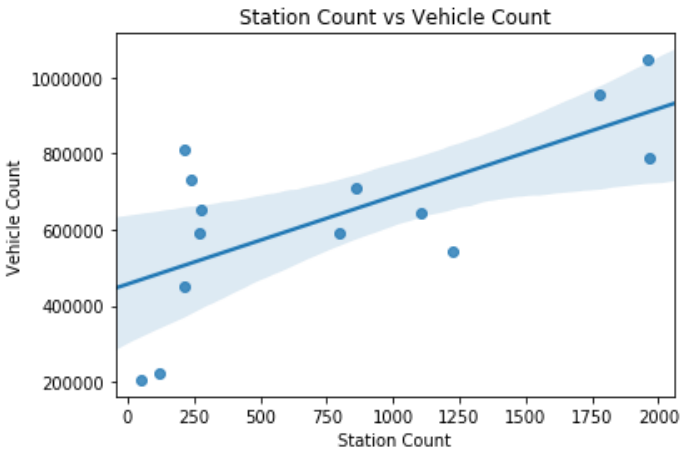
sns lmplot for the same vehicle data.

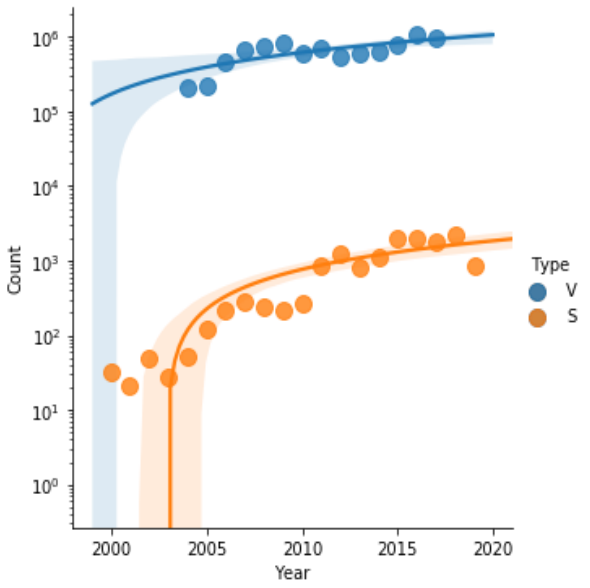


Individual regression plots for vehicles equipped with alternative fuels shows growth trend except for CNG and LNG that flattens which is again supported by stagnation in stations trend.

Let's look at the correlation between number of alternative fuel stations and number of alternative fuel vehicles across US using the scipy stats pearsonr function that returns the pearsonr coefficient and the pvalue.
Selecting a signifance level of 0.05, and so if the pvalue is less than 0.05, we can find out if the correlation coefficienct returned is significant.

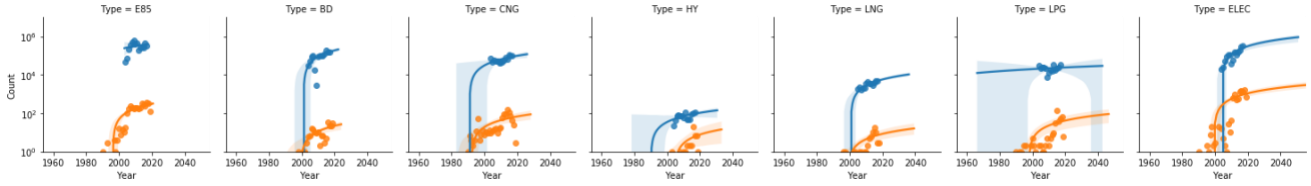
Calculate the pearsonr coefficient between the stations and vehicles using scipy stats:
scipy.stats.pearsonr(result_corr_df.Count_x, result_corr_df.Count_y) = (0.6822392170959737, 0.007186427141207331)
With pearsonr coefficient of about 68% and p value of 0.007 which is less than the significance level of 0.05 we can conclude that the correlation is significant.

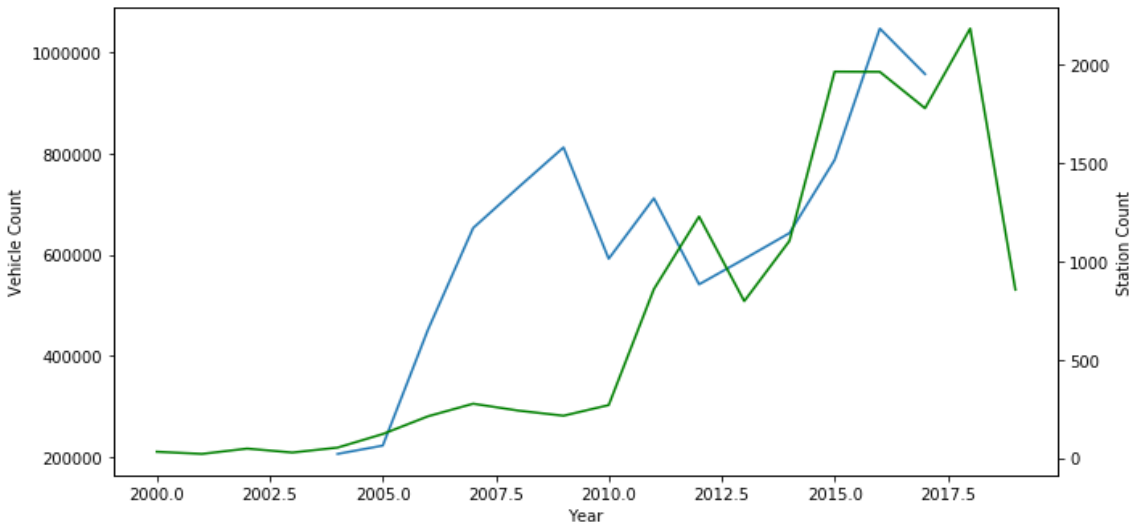




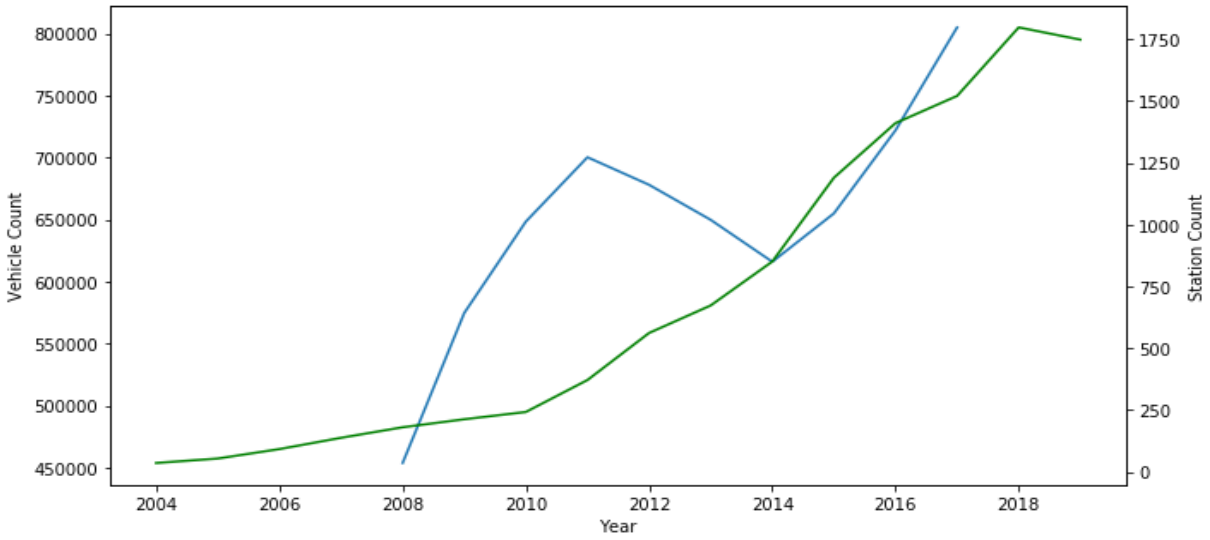
We can see the trend between the number of alternative fuel stations and vehicles equipped with alternative fuels being parallel to each other from the year 2005 onwards showing strong correlation.

Individual regression plots to see the trend between the number of stations and number of vehicles for each fuel type. Strong correlation seems to be the case with alternative fuel types like BD, CNG, LNG and ELEC whereas with HY, the number of stations seem to progress at a higher rate than the number of HY fuel cell based vehicles.

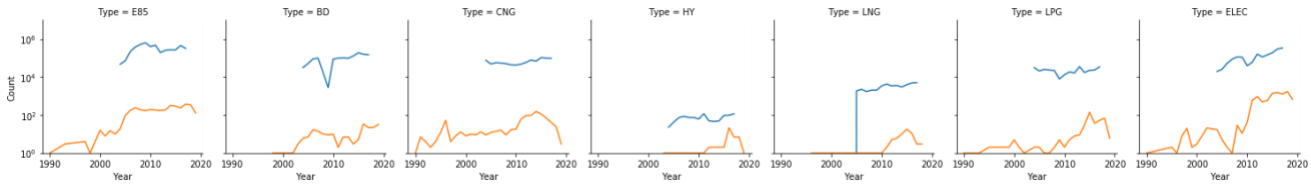




We can clearly see that from 2005 onwards the alternative fuels station count and alternative fuels vehicle count trends similar and follow each other showing that some correlation exists between them.



This plot shows the rolling mean of the previous plot with lines being smoothed out. Overall, the number of alternative fuel stations and alternative fuel vehicles seems to be on the rise in the past 15 years.



These individual plots clearly show increase in E85, BD and Electric vehicles and stations while CNG, LNG, LPG and HY show decline in the number of stations.

Deep Learning

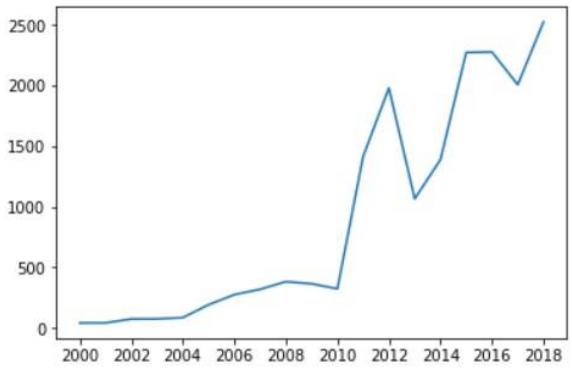
ARIMA and LSTM Models were implemented to predict number of alternative fuel stations facilities will be opening in a given year or month.

https://github.com/glazenda/Capstone1_Alt_Fuels/blob/master/Alt_Fuel_ML.ipynb

Establishing a baseline is essential on any time series forecasting problem. A baseline in performance gives an idea of how well all other models will actually perform on the problem. A baseline in forecast performance provides a point of comparison.

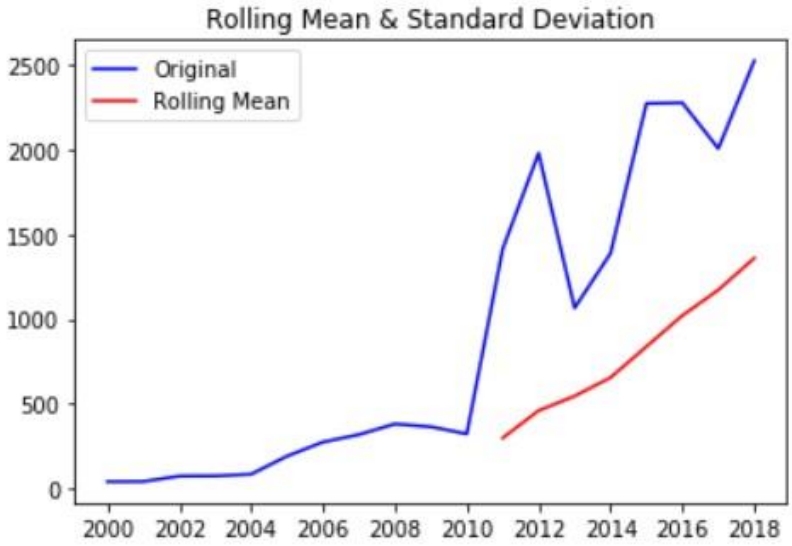
Data for Baseline Prediction

	Year	Count
22	2000	41
23	2001	42
24	2002	74
25	2003	75
26	2004	85



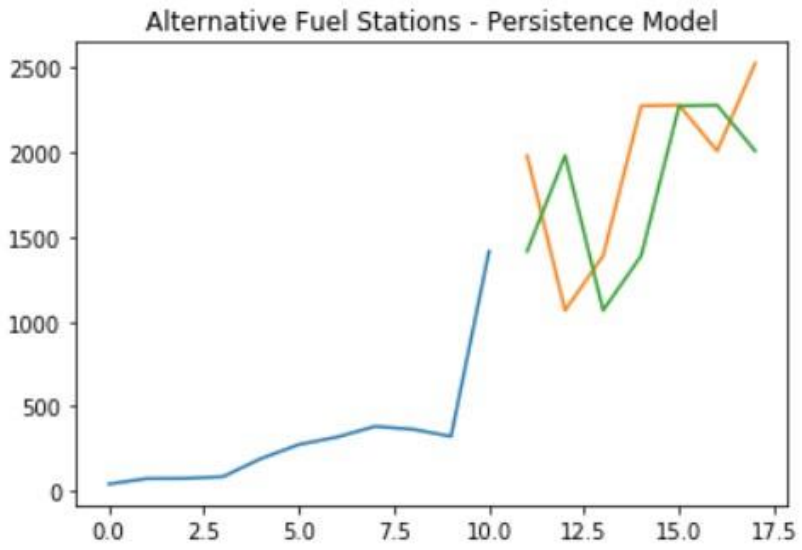
The mean is clearly increasing with time and this is not a stationary series.

Trend - The number of stations has grown over time and the trend also is influenced by various economical factors (production cost, availability etc.) and sustainability interest of the users. These factors are not reflected in our prediction data.




```
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
```

We can define our persistence model as a function that returns the value provided as input.
When we evaluate this model on the test set using walk forward validation method, we get test MSE: 340395.714 In this case, the error is more than 340000 over the test dataset.



Plot to show the training dataset and the diverging predictions from the expected values from the test dataset.

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 yearto-year noise in the stations count, which highlights the limitations of the persistence technique.

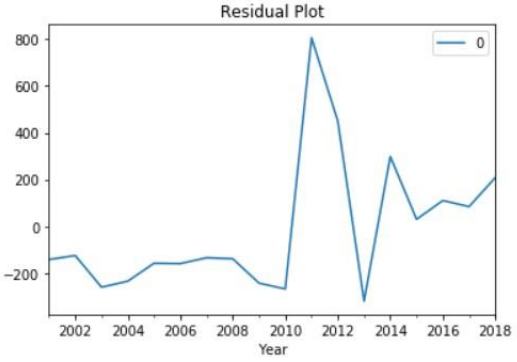
ARIMA

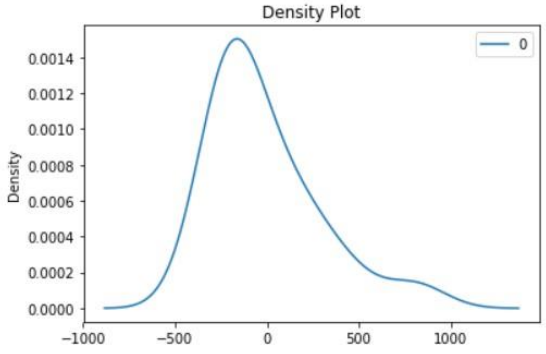
Autoregressive Integrated Moving Average Model

Le al o ake a q ick look a an a oco ela ion plo of he ime e ie .

We fit an ARIMA (3,1,1) model. This sets the lag value to 3 for autoregression, uses a difference order of 1 to make the time series stationary, and uses a moving average model of 1.

```
                             ARIMA Model Results
==============================================================================
Dep. Variable:                D.Count   No. Observations:                   18
Model:                 ARIMA(3, 1, 1)   Log Likelihood                -128.008
Method:                       css-mle   S.D. of innovations            285.628
Date:                Fri, 09 Aug 2019   AIC                            268.017
Time:                        21:24:52   BIC                            273.359
Sample:                  01-01-2001   HQIC                           268.753
                       - 01-01-2018
==============================================================================
                    coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          143.1406     37.898      3.777      0.002      68.863     217.418
ar.L1.D.Count    0.8332      0.197      4.232      0.001       0.447       1.219
ar.L2.D.Count   -0.5349      0.236     -2.268      0.041      -0.997      -0.073
ar.L3.D.Count    0.7017      0.179      3.924      0.002       0.351       1.052
ma.L1.D.Count   -1.0000      0.002   -507.008      0.000      -1.004      -0.996
                                    Roots
=============================================================================
                  Real          Imaginary           Modulus         Frequency
-----------------------------------------------------------------------------
AR.1            1.0000           -0.0000j            1.0000           -0.0000
AR.2           -0.1188           -1.1878j            1.1938           -0.2659
AR.3           -0.1188           +1.1878j            1.1938            0.2659
MA.1            1.0000           +0.0000j            1.0000            0.0000
-----------------------------------------------------------------------------
```

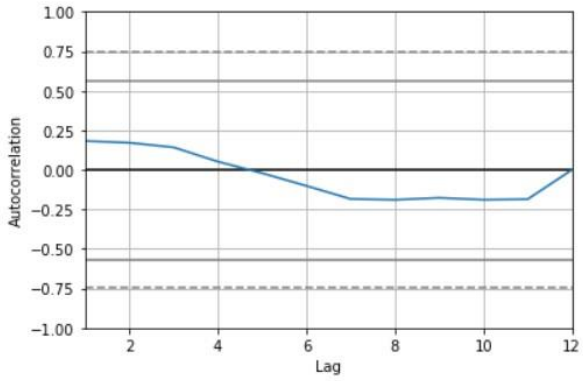


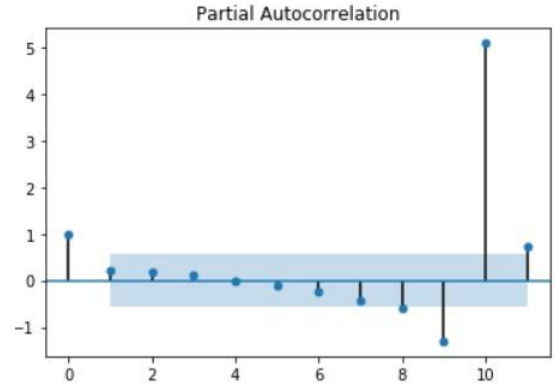


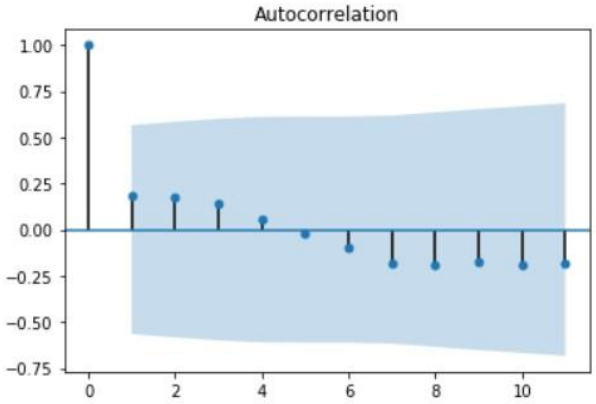

```
count   18.000000
mean   -10.944758
std    295.466134
min   -320.256652
25%   -216.029793
50%   -136.712713
75%    103.628189
max    806.593338
```

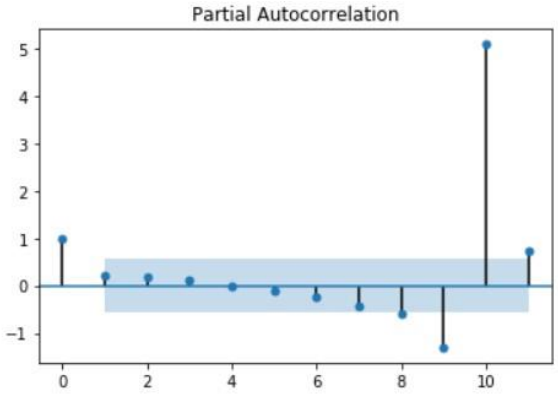
Running the example prints a summary of the fit model. This summarizes the coefficient values used as well as the skill of the fit on the on the in-sample observations. First, we get a line plot of the residual errors, suggesting that there may still be some trend information not captured by the model. Next, we get a density plot of the residual error values, suggesting the errors are Gaussian, but may not be centered on zero. The distribution of the residual errors is displayed.
The results show that indeed there is a bias in the prediction (a non-zero mean in the residuals).

Rolling Forecast Model

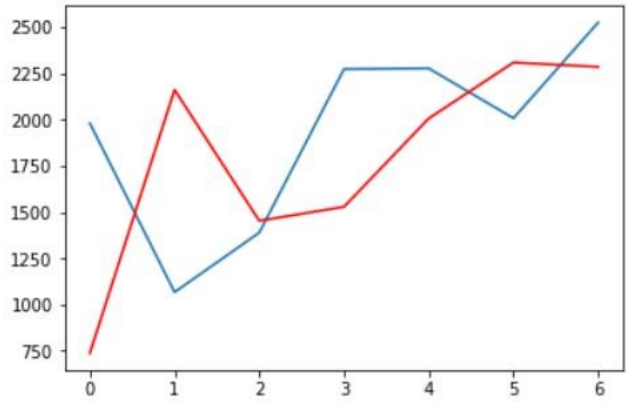








From the above plots for the rolling forecast model, we fit an ARIMA (0, 1, 1) model this sets the lag value to 0 for auto regression, uses a difference order of 1 to make the time series sationary, and uses a moving average model of 1.



```
predicted=736.277166, expected=1981.000000
predicted=2162.083684, expected=1067.000000
predicted=1453.029757, expected=1389.000000
predicted=1528.641705, expected=2274.000000
predicted=2006.533178, expected=2278.000000
predicted=2309.595233, expected=2008.000000
predicted=2285.939416, expected=2526.000000
Test MSE: 504355.001
```

A line plot is created showing the expected values (blue) compared to the rolling forecast predictions (red). We can see the values show some trend and are in the correct scale.

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

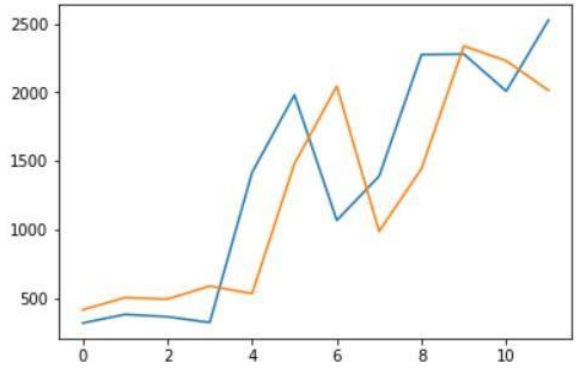
```
array ([2285.93941588, 2419.5423177, 2553.14521953, 2686.74812136,2820.35102319])
```

The model predicted the growth rate of number of fuel stations to be at the rate of 5% YoY.

LSTM - Yearly

Long Short-Term Memory Network

LSTM forecast model for a one-step univariate time series.




```
Year=2007, Predicted=415.462310, Expected=319.000000
Year=2008, Predicted=504.610349, Expected=382.000000
Year=2009, Predicted=492.788682, Expected=365.000000
Year=2010, Predicted=588.580606, Expected=323.000000
Year=2011, Predicted=533.633508, Expected=1416.000000
Year=2012, Predicted=1478.874415, Expected=1981.000000
Year=2013, Predicted=2043.874415, Expected=1067.000000
Year=2014, Predicted=986.150428, Expected=1389.000000
Year=2015, Predicted=1443.251027, Expected=2274.000000
Year=2016, Predicted=2336.874415, Expected=2278.000000
Year=2017, Predicted=2229.048656, Expected=2008.000000
Year=2018, Predicted=2013.250462, Expected=2526.000000
Test RMSE: 521.525
```

From the above plot, the calculated RMSE of 521 and whereas the RMSE for ARIMA model is 710, which confirms in this case LSTM performs better over ARIMA.

LSTM - Monthly

We will split the alternative fuels dataset into two parts: a training and a test set. 66% of data will be taken for the training dataset and the remaining 34% of data will be used for the test set.

Models will be developed using the training dataset and will make predictions on the test dataset. A rolling forecast scenario will be used, also called walk-forward model validation. Each time step of the test dataset will be walked one at a time. A model will be used to make a forecast for the time step, then the actual expected value from the test set will be taken and made available to the model for the forecast on the next time step.

We will convert our loaded alternative fuel data set into a supervised learning problem. We convert the dataset to be stationary so that it is easier to model and very likely result in skilfull forecast. Then we remove the trend by differencing the data, which leaves us with a difference series or the changes to the observation from one timestep to the next. To transform timeseries to scale, we use the default activation function for LSTM's which is the hyperbolic tangent(tanh), for which the output values are between -1 and 1. The data has to be provided in matrix format, so we reshape our NumPy arrays before transforming.

We use LSTM which is the Recurrent Neural Network (RNN) as it can learn and remember over long sequences and does not rely on a pre-specified window lagged observation as input. In Keras, this is referred to as stateful and involves setting the stateful argument to True when defining an LSTM layer.

By default, an LSTM layer in Keras maintains state between data within one batch. A batch of data is a fixed-sized number of rows from the training dataset that defines how many patterns to process before updating the weights of the network. State in the LSTM layer between batches is cleared by default, therefore we must make the LSTM stateful. This gives us fine-grained control over when state of the LSTM layer is cleared, by calling the reset_states() function. The batch_size must be set to 1. This is because it must be a factor of the size of the training and test datasets. The predict() function on the model is also constrained by the batch size; there it must be set to 1 because we are interested in making one-step forecasts on the test data.

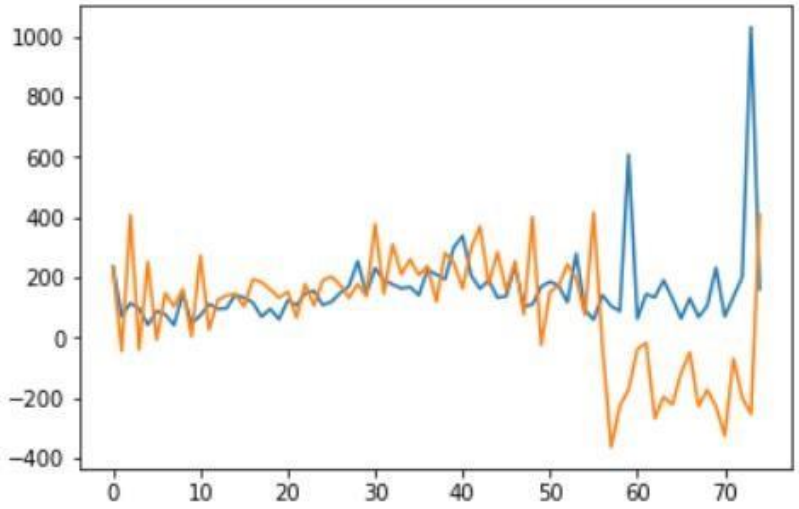
The LSTM layer expects input to be in a matrix with the dimensions: [samples, time steps, features]. Given that the training dataset is defined as X inputs and y outputs, it must be reshaped into the Samples/TimeSteps/Features format.

LSTM Forecast

Once the LSTM model is fit to the training data, it can be used to make forecasts. Again, we have some flexibility. We can decide to fit the model once on all of the training data, then predict each new time step one at a time from the test data. To make a forecast, we can call the predict() function on the model. This requires a 3D NumPy array input as an argument. In this case, it will be an array of one value, the observation at the previous time step.

The predict() function returns an array of predictions, one for each input row provided. Because we are providing a single input, the output will be a 2D NumPy array with one value.

Walk-forward validation on the test data



The chart above shows predicted (orange) versus observed (blue) trend.

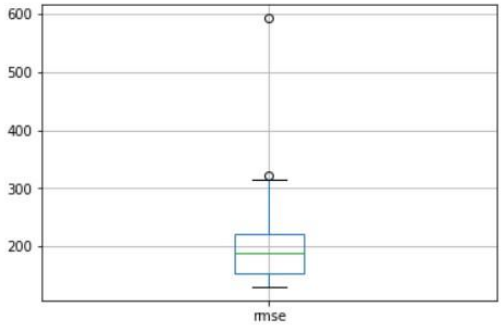
Develop a Robust Result

https://github.com/glazenda/Capstone1_Alt_Fuels/blob/master/Alt_Fuels_ML_Monthly_LSTM_RobustModel.ipynb

A difficulty with neural networks is that they give different results with different starting conditions. One approach might be to fix the random number seed used by Keras to ensure the results are reproducible. Another approach would be to control for the random initial conditions using a different experimental setup. We can repeat the experiment from the previous section 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. We can wrap the model fitting and walk-forward validation in a loop of fixed number of repeats. Each iteration the RMSE of the run can be recorded. We can then summarize the distribution of RMSE scores. To get a good distribution of RMSE scores, we will use 30 repeats and plot them.

LSTM Repeated Experiment Box and Whisker Plot



	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

Tuning LSTM Hyperparameters with Keras for Time Series Forecasting

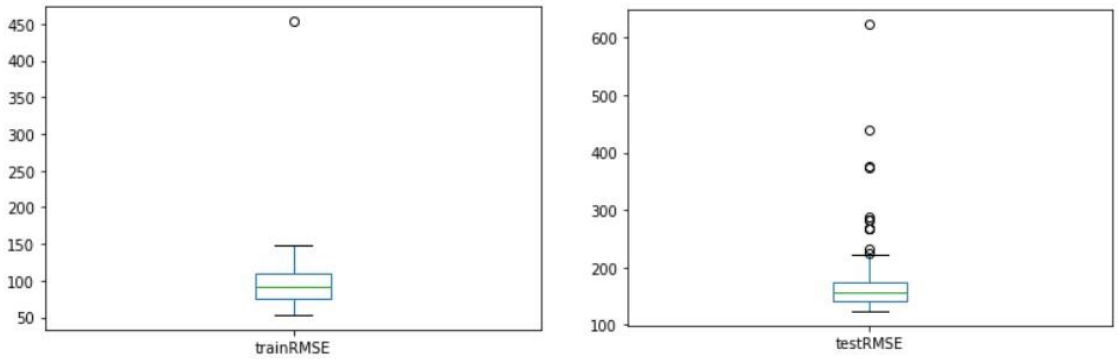
https://github.com/glazenda/Capstone1_Alt_Fuels/blob/master/Alt_Fuels_ML_Monthly_LSTM_HyperparameterTuning.ipynb

Configuring neural network is difficult as there is no good theory on how to do it. So, we will take a systematic approach to explore different configurations for dynamical and objective results point of view for understanding the predictive modeling problem. We will tune and interpret the results of number for training epochs and the number of neurons. We will perform experimental runs, in which each scenario will be run 10 times for a given neuron and for a given epoch. Number of Neurons ranges from 1 to 5 and the number of epoch ranges from 2^0 to 2^{12}. The batch_size must be set to 1. This is because it must be a factor of the size of the training and test datasets. The predict() function on the model is also constrained by the batch size; there it must be set to 1 because we are interested in making one-step forecasts on the test data. The reason of the experimental runs is that, the random initial conditions for an LSTM network can result in very different results each time a given configuration is trained.

A diagnostic approach will be used to investigate the model configurations, where line plots of model skill over time will be created and studied for insight into how a given configuration performs, and how it may be adjusted to show improved performance. The model will be evaluated on both train data set (66%) and test data sets (34%) at the end of each scenario, and the RMSE scores will be saved. T

The results of the experimental runs are saved to a CSV file alt_fuels_hyperparameter_tuning.csv, upon which diagnostics will be performed to investigate model configurations.

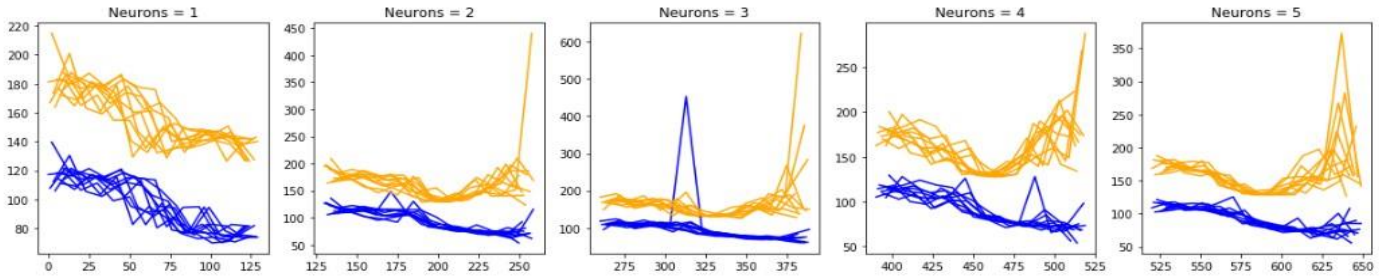
https://github.com/glazenda/Capstone1_Alt_Fuels/blob/master/Alt_Fuels_Hyperparameter_Diagnostics.ipynb



The plots above shows overall train and test RMSE's.

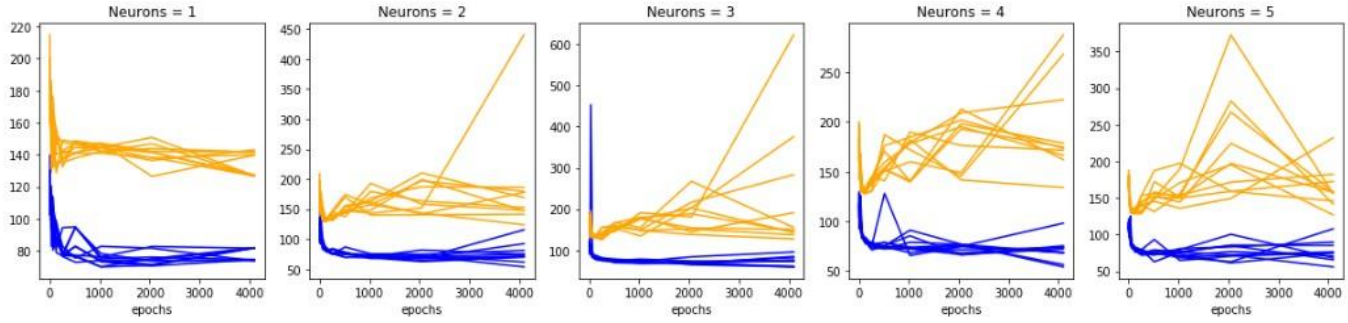
Diagnostic of 1 to 5 neurons

A line plot of the series of RMSE scores on the train and test sets for each set of neurons is created. The train scores are colored blue and test scores are colored orange.

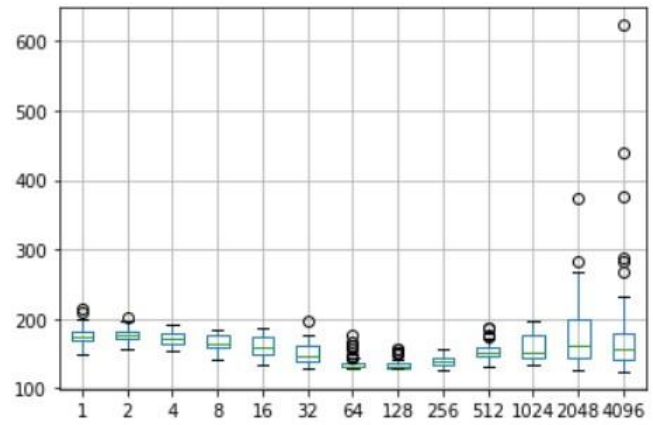


Diagnostic of 2^0 to 2^{12} Epochs

A line plot of the series of RMSE scores on the train and test sets for each set of neurons for each epoch is created. The train scores are colored blue and test scores are colored orange.



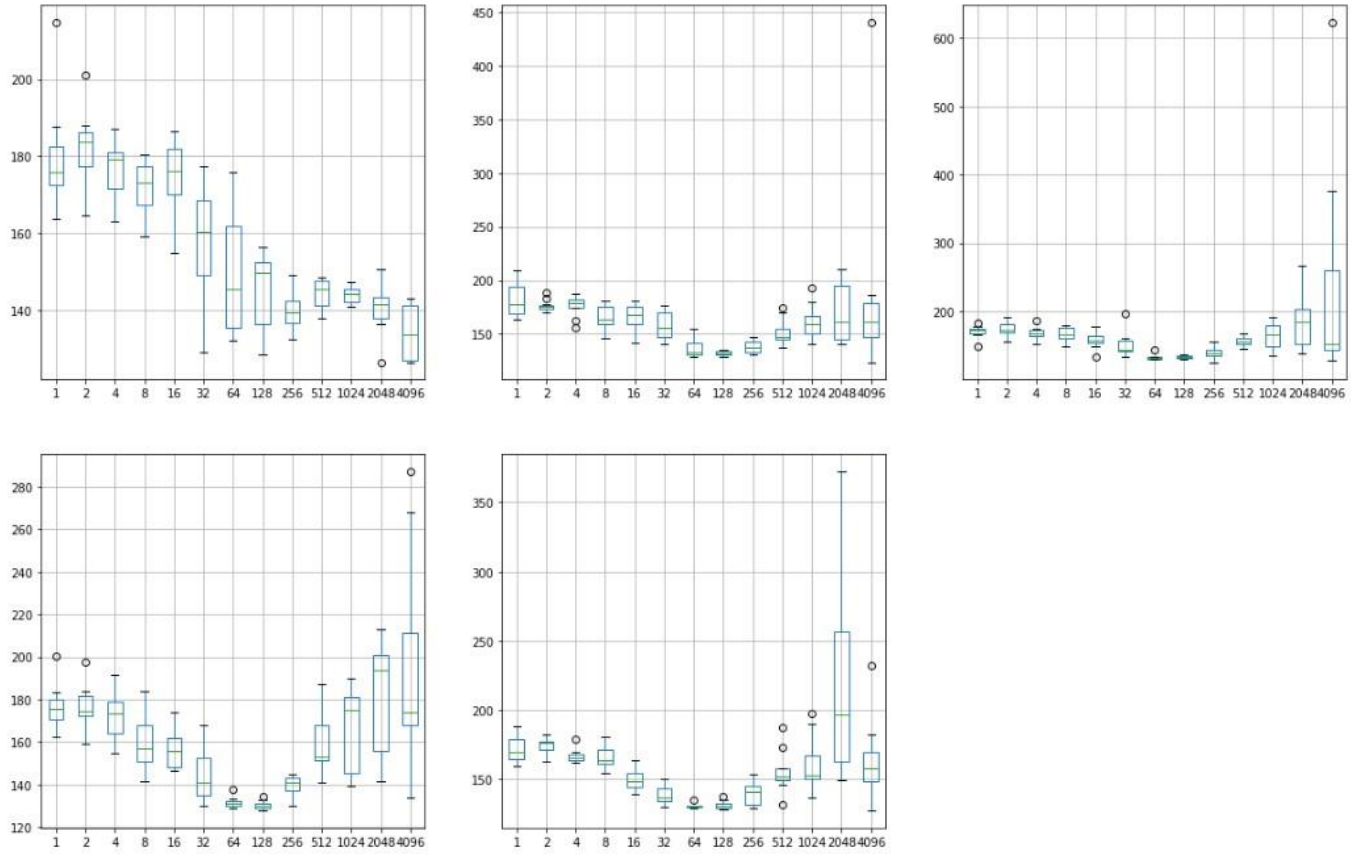
Boxplot to view overall range of test RMSE for each epoch



The distributions are shown on a box and whisker plot to see how they directly compare.

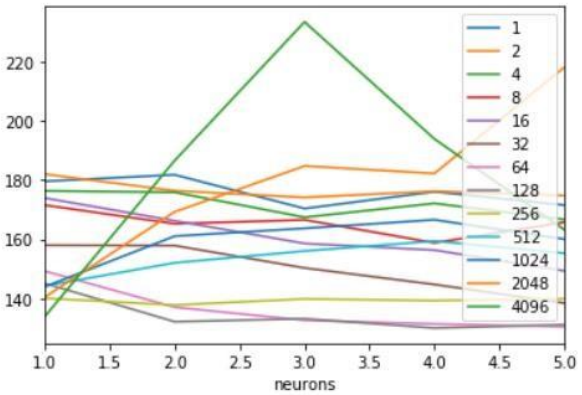
This comparison shows that the choice of setting epochs to 128 is better than the tested alternatives. It also shows that the best possible performance may be achieved with epochs between 128 and 512.

Tuning a neural network is a tradeoff of average performance and variability of that performance, with an ideal result having a low mean error with low variability, meaning that it is generally good and reproducible.



neurons	1	2	4	8	16	32	64	128	256	512	1024	2048	4096
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	174.10884	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	172.10456	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	217.9276	163.233271

The plot below shows the lowest RMSE range being achieved with 128 epochs.



From the above plots we will investigate the effect of varying the number of neurons in the network.

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.

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.

Conclusion:

It is clear from the data and the predictions that there is significant growth in use of alternate fuels in the automobile industry especially in the recent past. Using alternate fuels in automobiles, not only provides sustainable growth and most significantly it reduces the harmful Carbon-dioxide gases emitting to the atmosphere. Several alternate fuels in discussion in this article burns clean and they are environmentally friendly. In the last decade, the average retail fuel prices in the USA has been rather stable and may be slightly on a decline trend (https://afdc.energy.gov/fuels/prices.html). I believe this trend will continue and change in years to come favoring towards alternative fuels due to the following factors:

x as the dependency on petroleum fuels (gasoline and diesel oil) continue to decline, x more environmentally conscious vehicle owners interested in alternate fuel driven vehicles, x increased availability of the gas stations for the commuters / vehicle owners, x leap in technology to produce alternative fuels with lower production cost x fuels that are safer to use

This report favors the industries that are in production of automobile vehicles enabling the use of alternative fuels, industries that produce alternative fuels, vehicle owners who are with sustainability mindset and interested in leaving positive footprints to the environment, government that continues to thrive increasing vehicle fleet with alternative fuels and gas stations providing flexibility with various fuel types.