Introduction to Data Analytics: Capstone Project
Predicting Fuel Economy Using EPA FE Trends Report

Group 3:
Noelle Baker
Brian Link
Edward Katynski
Rong (Tim) Situ

December 17, 2021

HYUNDAI

Outline

1. Project Question and Data Set	Noelle Baker
2. Initial Data Processing	
- Targeted Data Set	Noelle Baker
- Initial Heat Map	Rong (Tim) Situ
- Initial Factor Reduction	
- Factor Manipulation	
3. Exploratory Data Analysis	Edward Katynski
- Data Head and Data Types	
- Check Data Characteristics	
- Reduced Factor Heat Map	
- Check Factor Distribution and Outliers	
- Check FE Relationship with Factors	
4. Data Analysis	Rong (Tim) Situ
- Linear Regression – With All Factors, Forwards, and Backwards	
- Check Effect of Each Factor on FE	
5. Check Model Fit	Brian Link
- Model Evaluation and Validation – Q-Q and Residual Plots	
6. Final Results, Conclusions, and Lessons Learned	Brian Link

Project Question and Data Set

Can you accurately predict future vehicle fuel economy using past vehicle fuel economy and other vehicle characteristics?

Using a data set gathered by the EPA, we will be exploring fuel economy trends of all light-duty passenger vehicles certified by the EPA.
We will be using these trends to develop a model to predict the fuel economy based on given characteristics.

- Database location: https://www.epa.gov/automotive-trends/explore-automotive-trends-data#DetailedData

- Database contains:
  - Data and characteristics of all vehicles produced and certified since 1975
  - Size: 1782 KB
  - 5170 Data points, 50+ variables

- Data aggregation:
  - Data set aggregated up to the Manufacturer – Model Year – Regulatory Class (Car or Truck) level
    - Data for individual Models not available in this data set
  - Factors primarily take one of two forms:
    1. A fleet average: e.g., Vehicle Weight or Horsepower
    2. A percentage of total fleet with that characteristic: e.g., Drivetrain – Front or Powertrain – Gasoline

Initial Data Processing

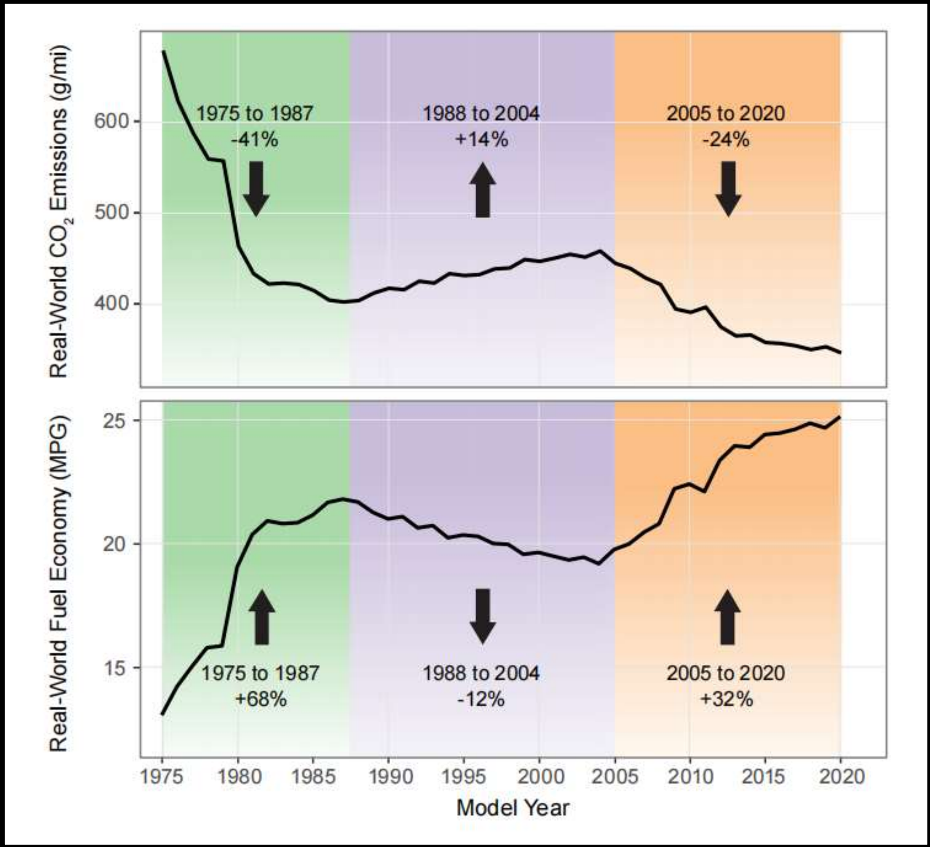
Targeted Data Set

- Fuel economy has changed in three distinct groups since 1975.
  1. 1975 – 1987: fuel shortage in the U.S. led to demand for more fuel efficient vehicles
  2. 1988 – 2004: fuel economy improvements were stagnant, with no pressure from customers or regulatory agencies
  3. 2005 – 2020: fuel economy improvements forced by new regulations

- 2021 Preliminary Data: we're excluding this data which is preliminary and incomplete.

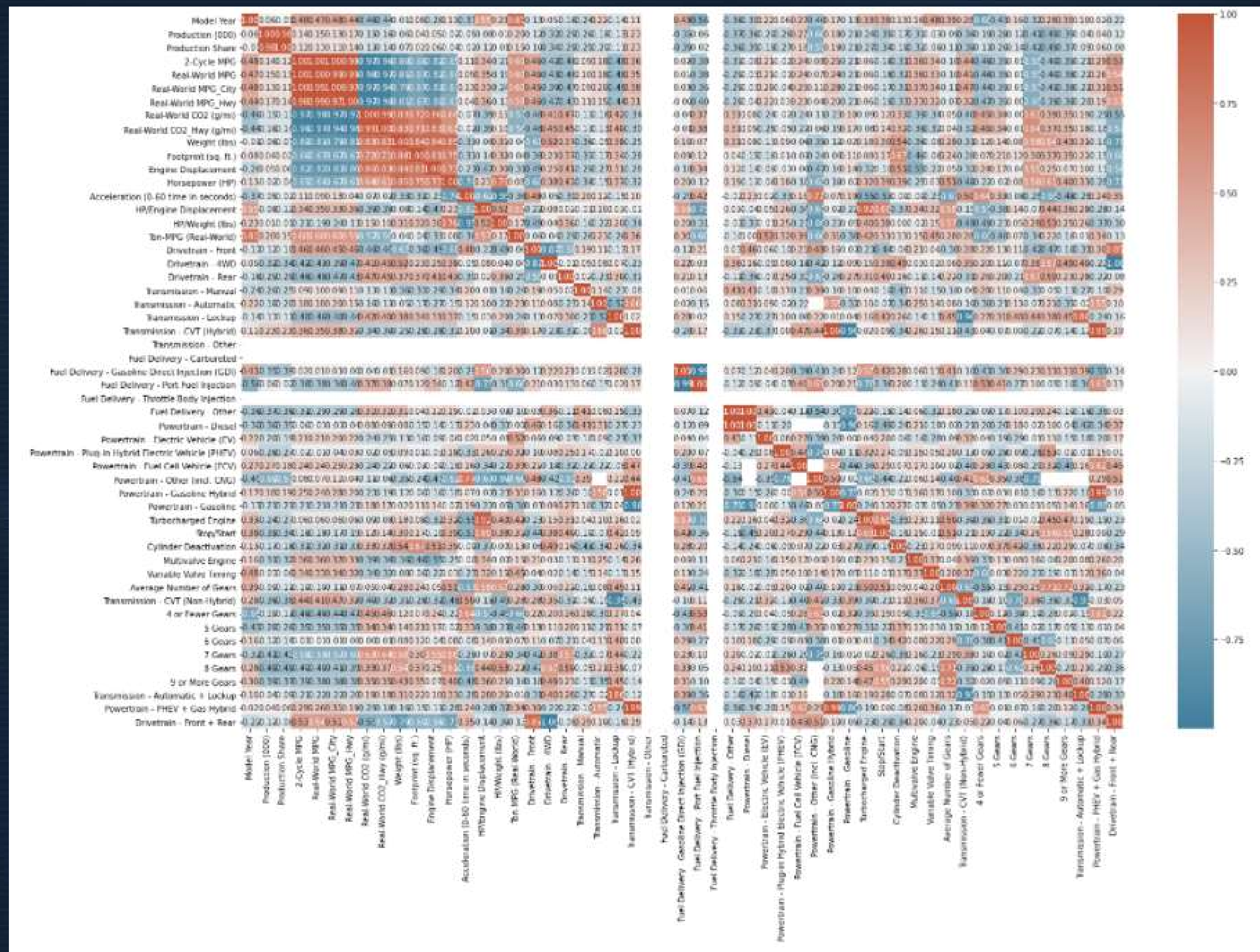
- Final Data Set = 2005 – 2020 MY
  - Using data from only the most recent timeframe will provide the most accurate model for future fuel economy performance.

Trends in Fuel Economy and CO_2 emissions since 1975 MY



Initial Heat Map

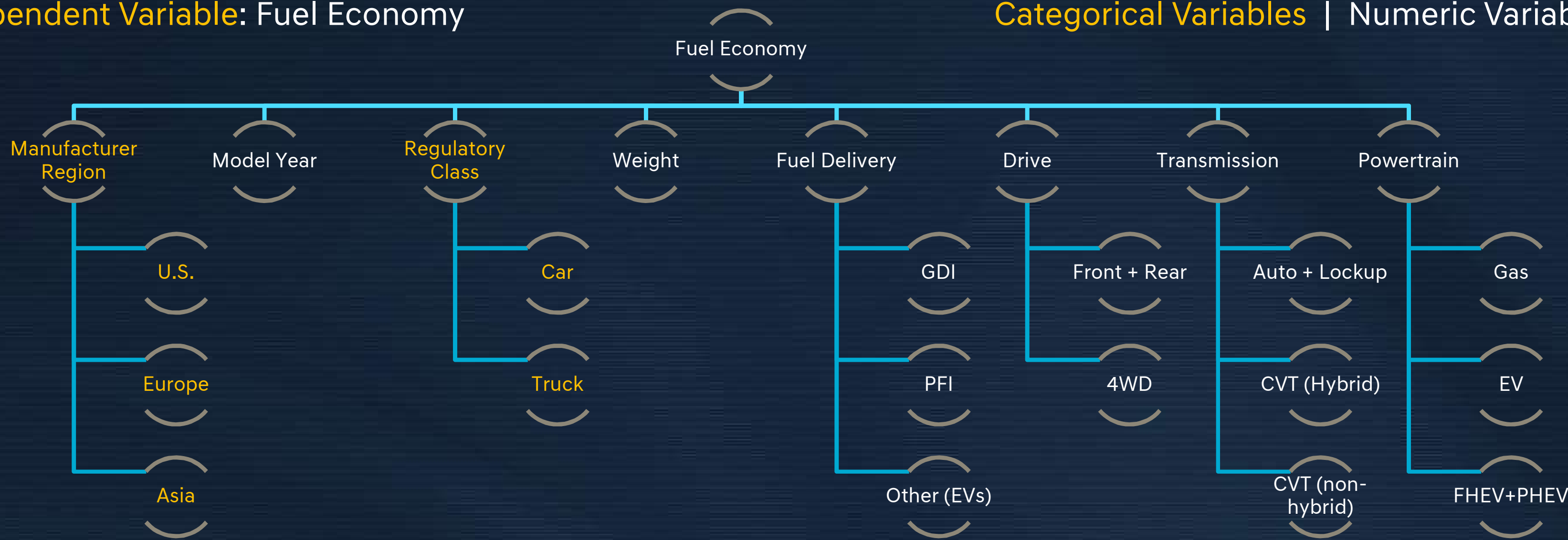
- Difficult to discern interactions with this many factors.
- Factor reduction based on engineering judgement and knowledge of system would be beneficial.

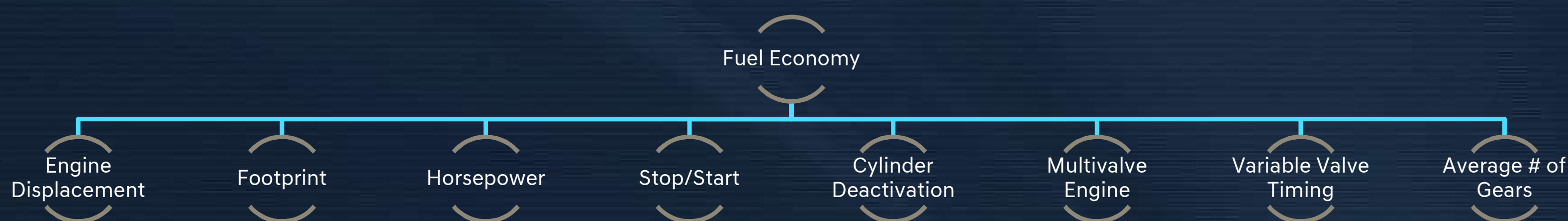


Initial Factor Reduction

- Dependent Variable: Fuel Economy

Categorical Variables | Numeric Variables

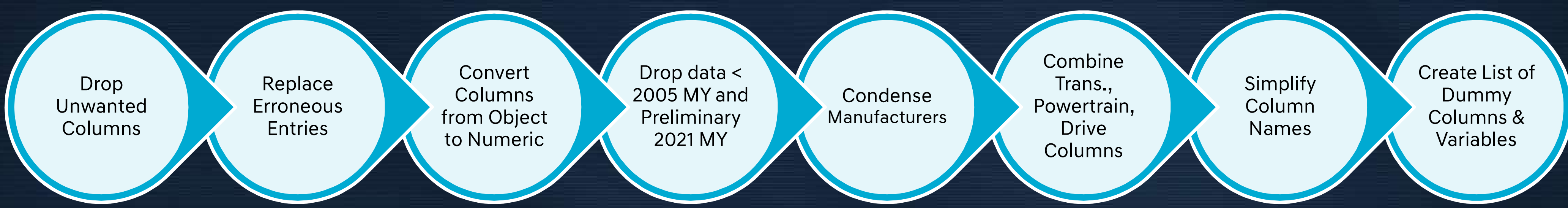




- Excluded variables: Vehicle Type, Production, Production Share, Real World MPG (Composite, City, Hwy), Real World CO2 (Composite, City, Hwy), HP/Engine Displacement, HP/Weight, Ton-MPG, Transmission-Other, Transmission – Manual, Fuel Delivery (Carbureted, Throttle Body Injection), Powertrain (FCEV, CNG, Other), Differentiate Gears (≤4, 5, 6, 7, 8, ≥9)

Factor Manipulation

- Initial factor manipulation needed to prepare data set for analysis



Exploratory Data Analysis

Data Head

- Print data head with remaining columns

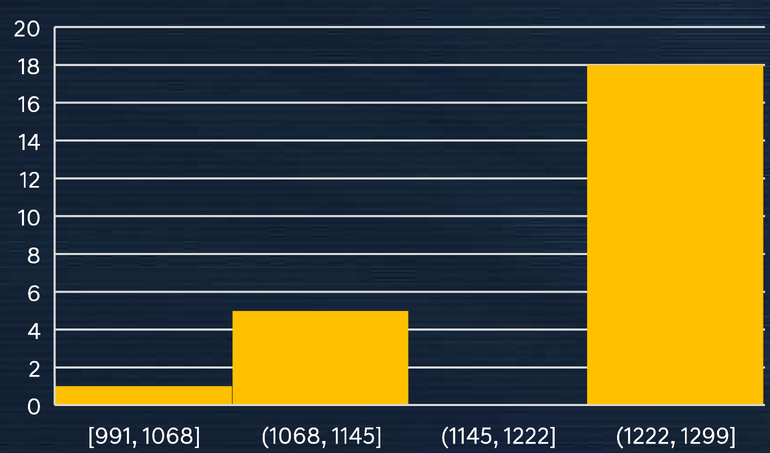
- Observations:
  - Data has been reduced to 24 columns
  - Manufacturer was categorized and then transformed into two dummy variables: Domestic and European. Asian is not listed (Asian indicated by Domestic = 0 and European = 0)
  - Similarly for Reg Class – a single dummy variable was created to represent both Car and Truck. A Car is indicated by Truck = 0

	MY	MPG	Weight	Footprint	EngDisp	HP	AWD	CVT_Hybrid	PortFuelInj	FuelOther	...	MultiVlv	VVT	Gears	CVT	AT	PT_PHEV	NotAWD	MFG_Domestic	MFG_European	RegClass_Truck
0	2013	NaN	NaN	NaN	NaN	NaN	0.00	0.00	0.00	0.00	...	0.00	0.00	NaN	0.00	0.00	0.00	0.00	0.00	1.00	0.00
1	2014	NaN	NaN	NaN	NaN	NaN	0.00	0.00	0.00	0.00	...	0.00	0.00	NaN	0.00	0.00	0.00	0.00	0.00	1.00	0.00
2	2015	35.27	4000.00	48.55	122.05	240.00	0.00	0.00	0.00	0.00	...	1.00	1.00	8.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
3	2016	34.79	4000.00	48.56	122.05	240.00	0.00	0.00	0.00	0.00	...	1.00	1.00	8.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
4	2017	31.25	4000.00	48.56	122.05	240.00	0.00	0.00	0.00	0.00	...	1.00	1.00	8.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00

Check Data Types

- Observations:
  - Data types appropriately defined as either numeric (float64) or categorical (uint8)
  - Non-null count is high

Non-null Count (Max = 1245)




```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1245 entries, 0 to 5140
Data columns (total 24 columns):
 #   Column          Non-Null Count   Dtype
---  ------          --------------   -----
 0   MY              1245 non-null    int64
 1   MPG             1143 non-null    float64
 2   Weight          1143 non-null    float64
 3   Footprint        991 non-null    float64
 4   EngDisp         1143 non-null    float64
 5   HP              1143 non-null    float64
 6   AWD             1245 non-null    float64
 7   CVT_Hybrid      1245 non-null    float64
 8   PortFuelInj     1245 non-null    float64
 9   FuelOther       1245 non-null    float64
 10  PT_EV           1245 non-null    float64
 11  PT_ICE          1245 non-null    float64
 12  ISG             1245 non-null    float64
 13  CylDeact        1245 non-null    float64
 14  MultiVlv        1245 non-null    float64
 15  VVT             1245 non-null    float64
 16  Gears           1133 non-null    float64
 17  CVT             1245 non-null    float64
 18  AT              1245 non-null    float64
 19  PT_PHEV         1245 non-null    float64
 20  NotAWD          1245 non-null    float64
 21  MFG_Domestic    1245 non-null    uint8
 22  MFG_European    1245 non-null    uint8
 23  RegClass_Truck  1245 non-null    uint8
dtypes: float64(20), int64(1), uint8(3)
```

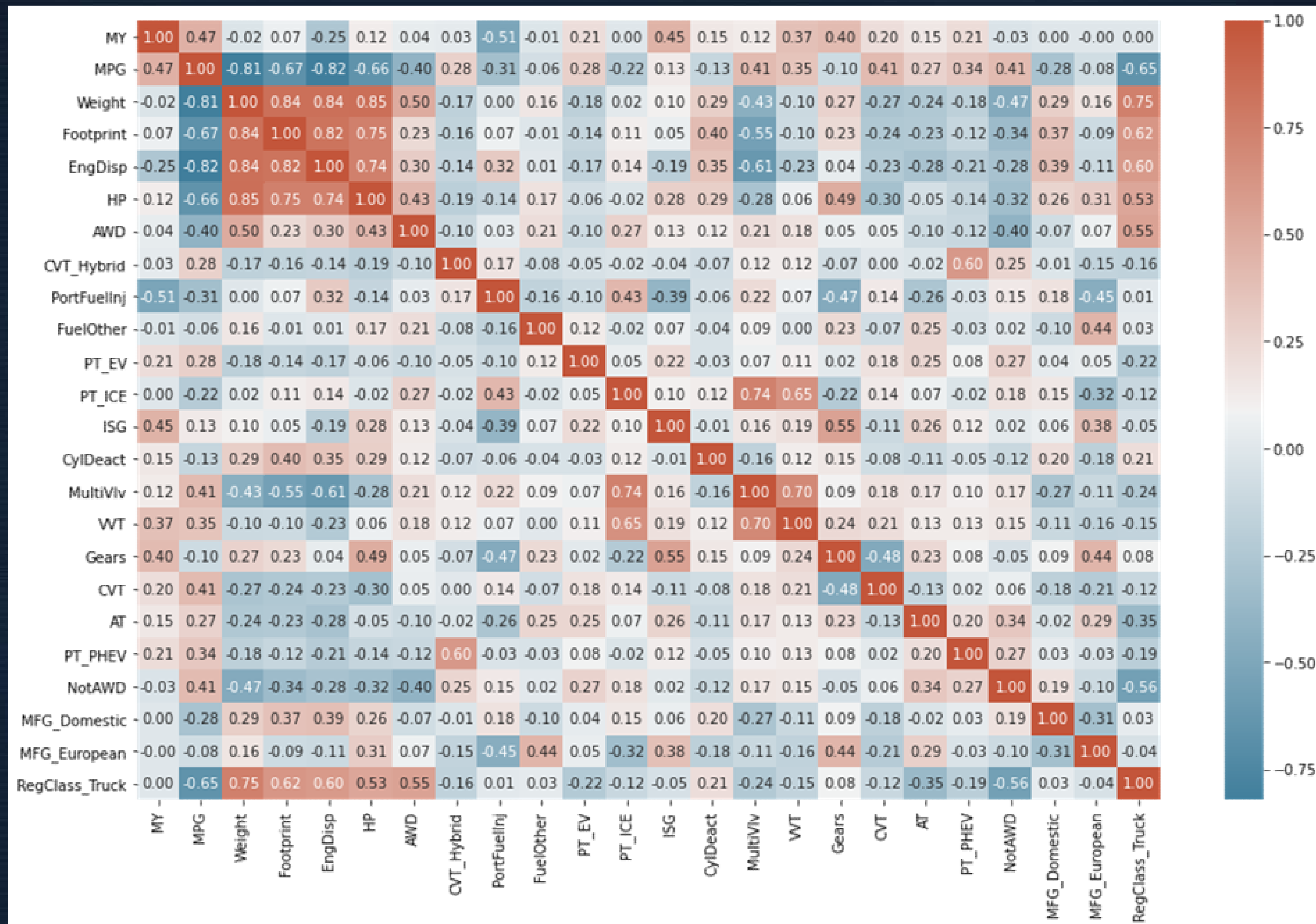
Check Data Characteristics

- Observations:
  - Differences between the means and medians shows some skewness, which will be more apparent in the histograms.
  - Data values are in the expected range for these factors.

	MY	MPG	Weight	Footprint	EngDisp	HP	AWD	CVT_Hybrid	PortFuelInj	FuelOther	...	MultiVlv	VVT	Gears	CVT	AT	PT_PHEV	NotAWD	MFG_Domestic	MFG_European	RegClass_Truck
count	1245	1143	1143	991	1143	1143	1245	1245	1245	1245	...	1245	1245	1133	1245	1245	1245	1245	1245	1245	1245
mean	2013	29.47	4138.19	49.60	182.02	229.54	0.36	0.01	0.60	0.01	...	0.84	0.79	5.72	0.11	0.13	0.01	0.29	0.25	0.22	0.54
std	4.32	5.70	633.19	5.50	47.14	49.23	0.38	0.03	0.44	0.03	...	0.33	0.36	1.41	0.26	0.32	0.03	0.40	0.43	0.41	0.50
min	2006	15.41	2997.37	42.56	94.12	131.00	0.00	0.00	0.00	0.00	...	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
25%	2009	25.17	3626.34	45.84	144.25	187.24	0.00	0.00	0.03	0.00	...	0.95	0.77	5.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
50%	2013	28.71	4000.00	47.53	171.30	223.73	0.20	0.00	0.90	0.00	...	1.00	1.00	6.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00
75%	2017	33.45	4578.64	51.66	213.26	266.79	0.71	0.00	1.00	0.00	...	1.00	1.00	6.30	0.00	0.00	0.00	0.72	1.00	0.00	1.00
max	2020	45.55	6668.90	68.43	366.14	379.28	1.00	0.27	1.00	0.21	...	1.00	1.00	10.00	1.00	1.00	0.28	1.00	1.00	1.00	1.00

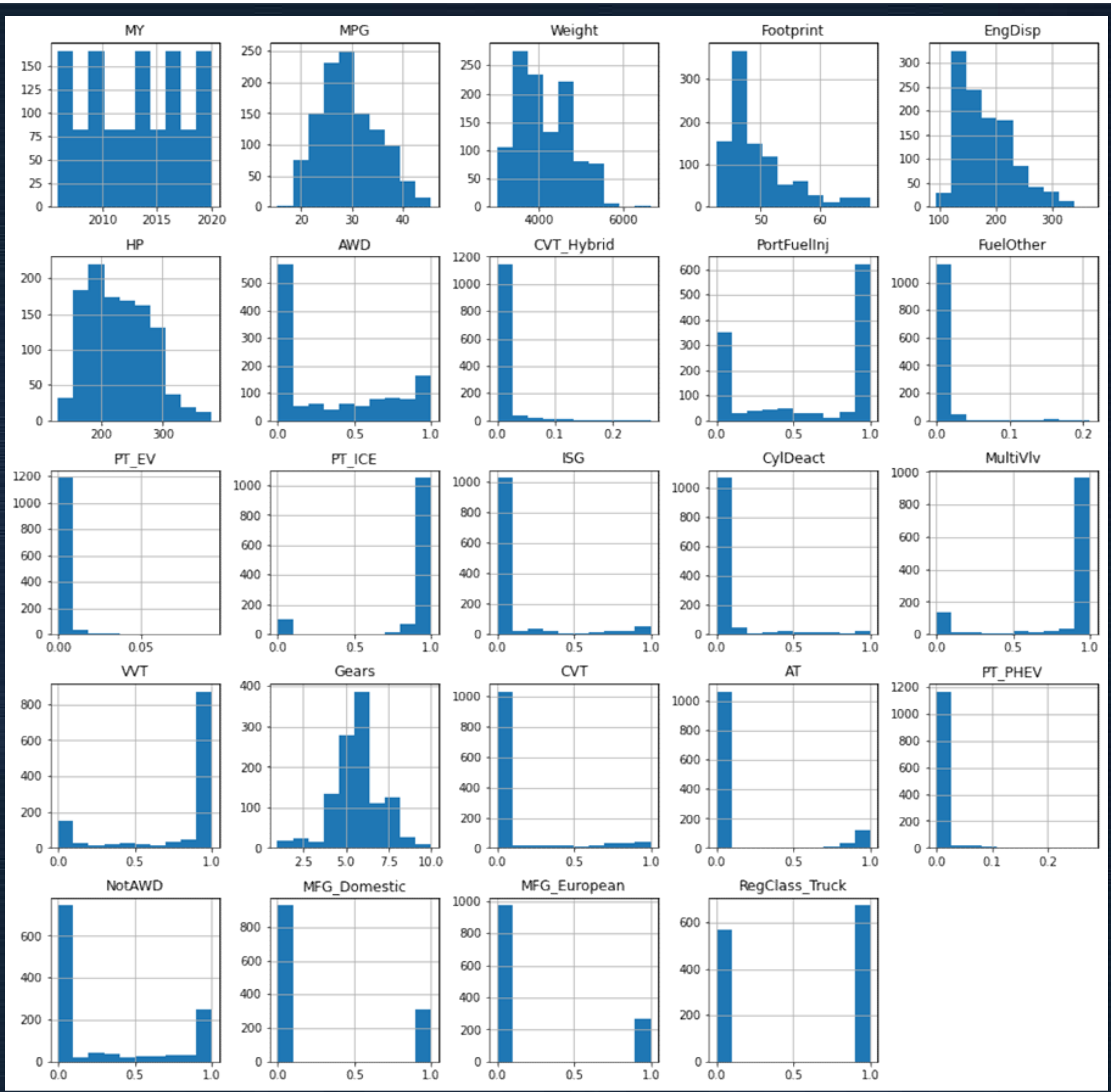
Heat Map

- Observations:
- Simplified heat map with reduced number of factors.
- Identify factors which are poor predictors of FE
  - Values near 0 for ISG (stop-start), Fuel Other, CycleDeact, Gears, and MFG_European – expected to drop out of final regression
- Strong predictors are likely to be present in a final regression
  - Values closer to 1 for Weight, Footprint, Engine Displacement



Check Factor Distribution and Outliers

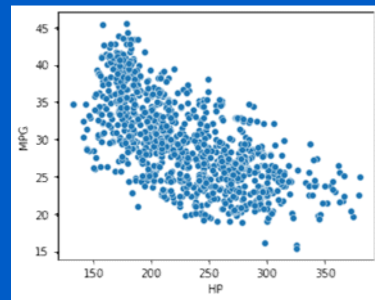
- Histograms of numerical factors
- Observations:
  - Expected left-skewing can be seen in FE (MPG), Weight, Footprint, Engine Displacement, and HP.
  - Other factors with the scale 0 – 1 (or <1) show percentages of fleet.
    - Factors dominated by 1 show prevalent features like PT_ICE (vehicles with gasoline engines)
    - Factors dominated by 0 show rare factors like PT_PHEV (plug-in hybrid electric vehicles)

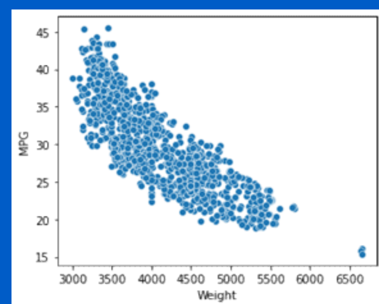


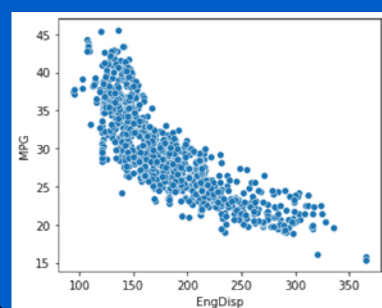
Check Fuel Economy Relationship with Factors

- Scatter plots vs. Fuel Economy
- Observations:
  - Examples of 24 remaining factors.
  - Trends can be observed in factors with stronger correlation, both positive and negative.
  - Some factors, with weak correlation, do not have a visible trend.
  - Model Year (below), shows discrete values and a positive correlation.

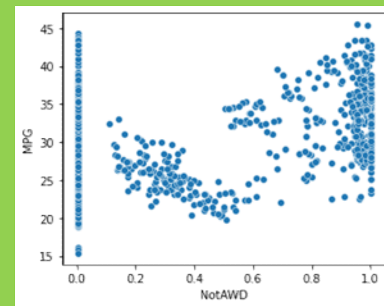
Negative Correlation

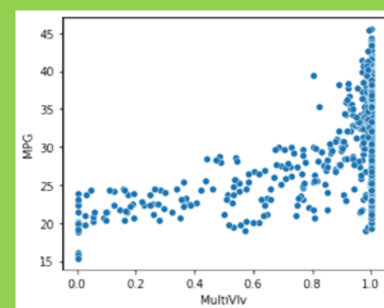


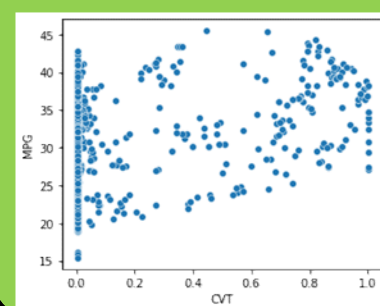




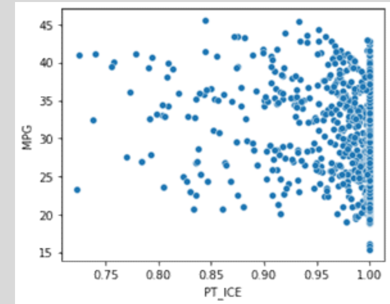
Positive Correlation

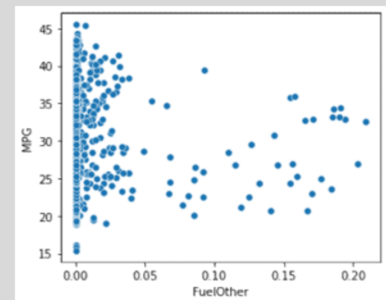


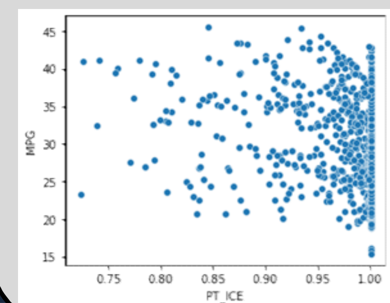




Weak Correlation







Analysis

Linear Regression with All Factors

- First, divide data set into 'test' and 'train' sets in order to verify the final model.
- Next, create a model with all 23 independent variables, and with Fuel Economy as the response variable
- Effect of each factor on FE: (P)
  - 17 factors show significance at 0.05
  - 7 factors fail to show significance at 0.05 level:
    - AT, Fuel Other, HP, ISG, MFG_European, PT_EV, and PT_PHEV
- Model Fit: (R-squared)
  - The model fit is very good at 92.4%
- Next: determine if backwards and forwards regressions show similar results, eliminating the 6 factors not shown as significant here.

OLS Regression Results

Dep. Variable:	MPG	R-squared:	0.924
Model:	OLS	Adj. R-squared:	0.922
Method:	Least Squares	F-statistic:	351.8
Date:	Thu, 16 Dec 2021	Prob (F-statistic):	0.00
Time:	16:44:44	Log-Likelihood:	-1277.1
No. Observations:	686	AIC:	2602.
Df Residuals:	662	BIC:	2711.
Df Model:	23		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-618.4419	57.380	-10.778	0.000	-731.110	-505.774
AT	-0.1995	0.227	-0.879	0.380	-0.645	0.246
AWD	-0.6337	0.278	-2.282	0.023	-1.179	-0.088
CVT	3.0130	0.333	9.045	0.000	2.359	3.667
CVT_Hybrid	14.9988	3.752	3.998	0.000	7.632	22.366
CylDeact	1.4875	0.384	3.873	0.000	0.733	2.242
EngDisp	-0.0325	0.006	-5.753	0.000	-0.044	-0.021
Footprint	0.1110	0.029	3.859	0.000	0.055	0.168
FuelOther	2.9547	4.297	0.688	0.492	-5.482	11.391
Gears	-0.2157	0.075	-2.859	0.004	-0.364	-0.068
HP	-0.0082	0.004	-1.929	0.054	-0.016	0.000
ISG	0.6146	0.378	1.624	0.105	-0.129	1.358
MFG_Domestic	-0.7759	0.224	-3.470	0.001	-1.215	-0.337
MFG_European	-0.4759	0.358	-1.331	0.184	-1.178	0.226
MY	0.3373	0.028	11.934	0.000	0.282	0.393
MultiVlv	-2.0774	0.569	-3.652	0.000	-3.194	-0.961
NotAWD	0.7049	0.212	3.321	0.001	0.288	1.122
PT_EV	21.8344	11.994	1.820	0.069	-1.717	45.386
PT_ICE	-8.9953	3.687	-2.440	0.015	-16.234	-1.756
PT_PHEV	2.3405	2.890	0.810	0.418	-3.334	8.015
PortFuelInj	-1.3194	0.255	-5.178	0.000	-1.820	-0.819
RegClass_Truck	-1.1002	0.259	-4.253	0.000	-1.608	-0.592
VVT	1.4225	0.370	3.839	0.000	0.695	2.150
Weight	-0.0041	0.000	-10.782	0.000	-0.005	-0.003

Omnibus:	5.620	Durbin-Watson:	1.885
Prob(Omnibus):	0.060	Jarque-Bera (JB):	7.397
Skew:	0.018	Prob(JB):	0.0248
Kurtosis:	3.507	Cond. No.	4.41e+06

Forward Linear Regression - Results

- Use Forward Regression strategy on the training data set to add factors, one at a time, until reaching the significance level of 0.05
- Effect of each factor on FE: (P)
  - 17 factors are added until the loop reaches the factor PT_PHEV (with P > 0.05) and stops
  - These factors align with those that showed significance in the full regression model
- Model Fit: (R-squared)
  - The model fit is also very good at 92.3%

OLS Regression Results

Dep. Variable:	MPG	R-squared:	0.923
Model:	OLS	Adj. R-squared:	0.921
Method:	Least Squares	F-statistic:	446.2
Date:	Thu, 16 Dec 2021	Prob (F-statistic):	0.00
Time:	16:44:46	Log-Likelihood:	-1281.9
No. Observations:	686	AIC:	2602.
Df Residuals:	667	BIC:	2688.
Df Model:	18		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-663.9746	51.664	-12.852	0.000	-765.418	-562.531
Weight	-0.0042	0.000	-12.420	0.000	-0.005	-0.004
MY	0.3603	0.026	13.953	0.000	0.310	0.411
PT_ICE	-9.8787	1.748	-5.651	0.000	-13.311	-6.446
CVT	3.1653	0.323	9.799	0.000	2.531	3.800
PortFuelInj	-1.2224	0.244	-5.009	0.000	-1.702	-0.743
CVT_Hybrid	14.7107	2.733	5.383	0.000	9.344	20.077
AWD	-0.6247	0.274	-2.281	0.023	-1.162	-0.087
HP	-0.0080	0.004	-2.259	0.024	-0.015	-0.001
CylDeact	1.5386	0.378	4.074	0.000	0.797	2.280
EngDisp	-0.0335	0.005	-7.401	0.000	-0.042	-0.025

	coef	std err	t	P>|t|	[0.025	0.975]
RegClass_Truck	-1.0226	0.254	-4.020	0.000	-1.522	-0.523
Footprint	0.1217	0.027	4.463	0.000	0.068	0.175
VVT	1.3569	0.361	3.761	0.000	0.648	2.065
MultiVlv	-2.0644	0.522	-3.953	0.000	-3.090	-1.039
NotAWD	0.7194	0.206	3.499	0.000	0.316	1.123
MFG_Domestic	-0.6242	0.188	-3.323	0.001	-0.993	-0.255
Gears	-0.1976	0.074	-2.680	0.008	-0.342	-0.053
PT_PHEV	1.2485	2.623	0.476	0.634	-3.902	6.399

Omnibus:	5.176	Durbin-Watson:	1.933
Prob(Omnibus):	0.075	Jarque-Bera (JB):	6.671
Skew:	-0.005	Prob(JB):	0.0356
Kurtosis:	3.483	Cond. No.	3.96e+06

Backward Linear Regression - Results

- Use Backward Regression strategy on the training data set to remove factors, one at a time, until reaching the significance level of 0.05

- Effect of each factor on FE: (P)
  - Factors are removed until only 16 remain
  - This regression included one fewer factor than the forward regression, excluding HP.

- Model Fit: (R-squared)
  - This model fit is excellent as well, at 92.3%

OLS Regression Results

Dep. Variable:	MPG	R-squared:	0.923
Model:	OLS	Adj. R-squared:	0.921
Method:	Least Squares	F-statistic:	499.1
Date:	Thu, 16 Dec 2021	Prob (F-statistic):	0.00
Time:	16:44:46	Log-Likelihood:	-1284.6
No. Observations:	686	AIC:	2603.
Df Residuals:	669	BIC:	2680.
Df Model:	16		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-663.0111	51.466	-12.883	0.000	-764.065	-561.957
MY	0.3606	0.026	14.025	0.000	0.310	0.411
Weight	-0.0045	0.000	-14.151	0.000	-0.005	-0.004
Footprint	0.1133	0.027	4.239	0.000	0.061	0.166
EngDisp	-0.0382	0.004	-9.457	0.000	-0.046	-0.030
AWD	-0.8185	0.261	-3.136	0.002	-1.331	-0.306
CVT_Hybrid	15.5474	2.476	6.280	0.000	10.687	20.408
PortFuelInj	-1.0951	0.238	-4.599	0.000	-1.563	-0.628
PT_ICE	-10.0901	1.693	-5.961	0.000	-13.414	-6.766
CylDeact	1.5798	0.378	4.180	0.000	0.838	2.322
MultiVlv	-2.4421	0.497	-4.912	0.000	-3.418	-1.466
VVT	1.2734	0.359	3.547	0.000	0.569	1.978
Gears	-0.2678	0.067	-3.987	0.000	-0.400	-0.136
CVT	3.1113	0.323	9.633	0.000	2.477	3.746
NotAWD	0.6859	0.205	3.354	0.001	0.284	1.087
MFG_Domestic	-0.6629	0.187	-3.542	0.000	-1.030	-0.295
RegClass_Truck	-0.8316	0.241	-3.454	0.001	-1.304	-0.359

Omnibus:	6.959	Durbin-Watson:	1.929
Prob(Omnibus):	0.031	Jarque-Bera (JB):	9.334
Skew:	0.071	Prob(JB):	0.00940
Kurtosis:	3.554	Cond. No.	3.93e+06

Check Model Fit

Model Evaluation and Validation

- Both Forward and Backward models have similar fits.
- Backwards model is simpler, with 1 fewer factor, so we choose this one for further validation.

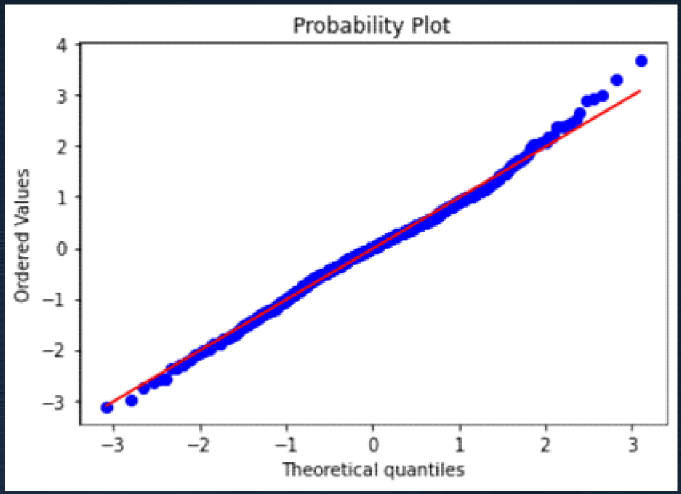
Find MAE, MSE, RMSE and MAPE

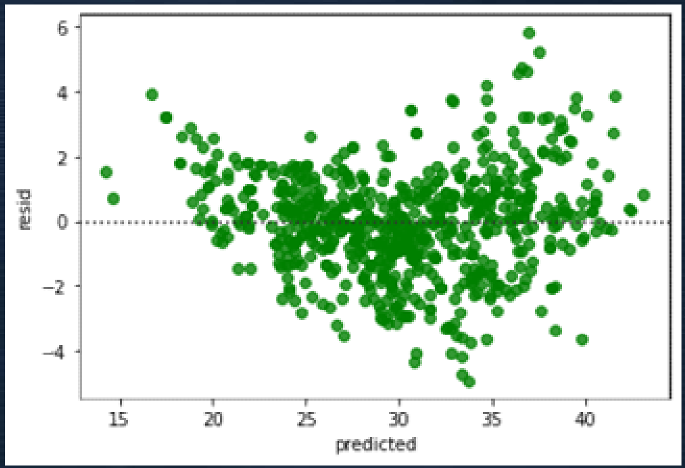
- MAE is 1.21 | MSE is 2.48 | RMSE is 1.57 | MAPE is 306.92

Check model assumptions using Q-Q Plot and Residuals Plot

- Q-Q
    - Compares the actual FE values with the predicted values to determine the probability that they have the same distribution.
    - This plot confirms the two data sets are very well matched, following the 1:1 line, with some deviation at the extremes.

- Residual Plot
    - Checks to see:
        1. if the residuals show a random distribution, indicating a linear regression is appropriate for this data set.
        2. or if there is a pattern in the residuals, indicating a different type of fit may be more appropriate.
    - This residual plot shows a random distribution --> linear regression is appropriate.





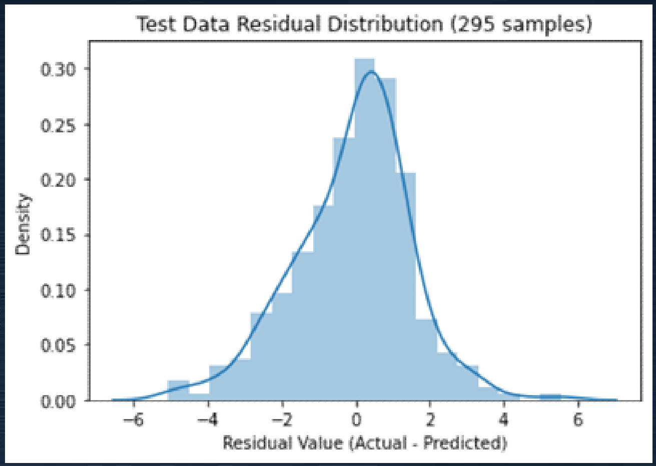
Model Evaluation and Validation

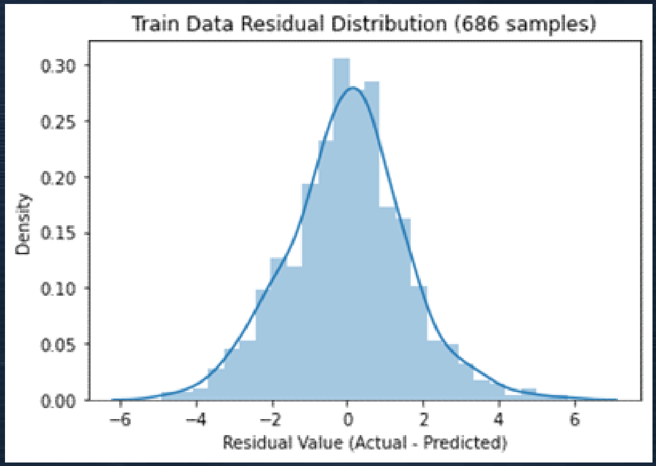
Check for model overfitting

- Run test data set through hard-coded backwards elimination model

Training vs. Test Set Data Results

- Histograms of the 'test' and 'train' sets (to the right) show similar near-normal distributions
- A scatter plot, below, of the paired data (Actual and Predicted FE) are well matched for the 'test' and 'train' datasets





Results / Conclusion

Final Results, Conclusions, and Lessons Learned

- Data set preparation is critical to producing a good model. Preparing a data set that contains missing values and superfluous columns can occupy a large portion of the project time.
- Data at the vehicle model level (e.g. Sonata) would have provided a more accurate tool to predict fuel economy – aggregation at the fleet level loses specificity and dampens the results.
- The full linear regression, backwards regression, and forwards regressions all produced well-fit models. The backwards regression contained the least factors and was therefore the best choice for predicting fuel economy.
- The final model selected had 16 factors that were significant predictors of FE
- The goodness of model fit was confirmed using the Q-Q plot and Residual plot.

Significant Predictors of Fuel Economy

THANK YOU

Appendix A: Python Code Base

```
#!/usr/bin/env python
# coding: utf-8

# ## Group Project: EPA FE Analysis and Prediction
#
# ### Context:
#
#
# ### Objective :
#
# * To identify the different factors that affect fuel economy in surveyed vehicles
# * To make a model to predict if an employee will attrite or not
#
#
# ### Dataset :
# The data contains (**replace** - demographic details), (**replace** - work-related metrics) and (**replace** - attrition flag).
#
# MY: Vehicle Model Year
# MPG: Miles per Gallon ()
# Weight: Vehicle Weight in lbs
# Footprint: Vehicle footprint in Square Feet
# EngDisp: Engine displacement in CC
# HP: Horsepower
# AWD: Percentage of Fleet with All Wheel Drive Capability
# CVT_Hybrid: Percentage of hybrid fleet with CVT
# PortFuelInj: Percentage of fleet with PFI
# FuelOther: Percentage of fleet with alternative fuel
```

** Scroll in presentation mode or by editing TextBox object**