Predicting the Fuel Efficiency of Vintage Cars

A MACHINE LEARNING AND DATA VISUALIZATION PROJECT

May 5, 2021

Authored by: Alex Jones, Jessica Pardo, and Dana Woodruff

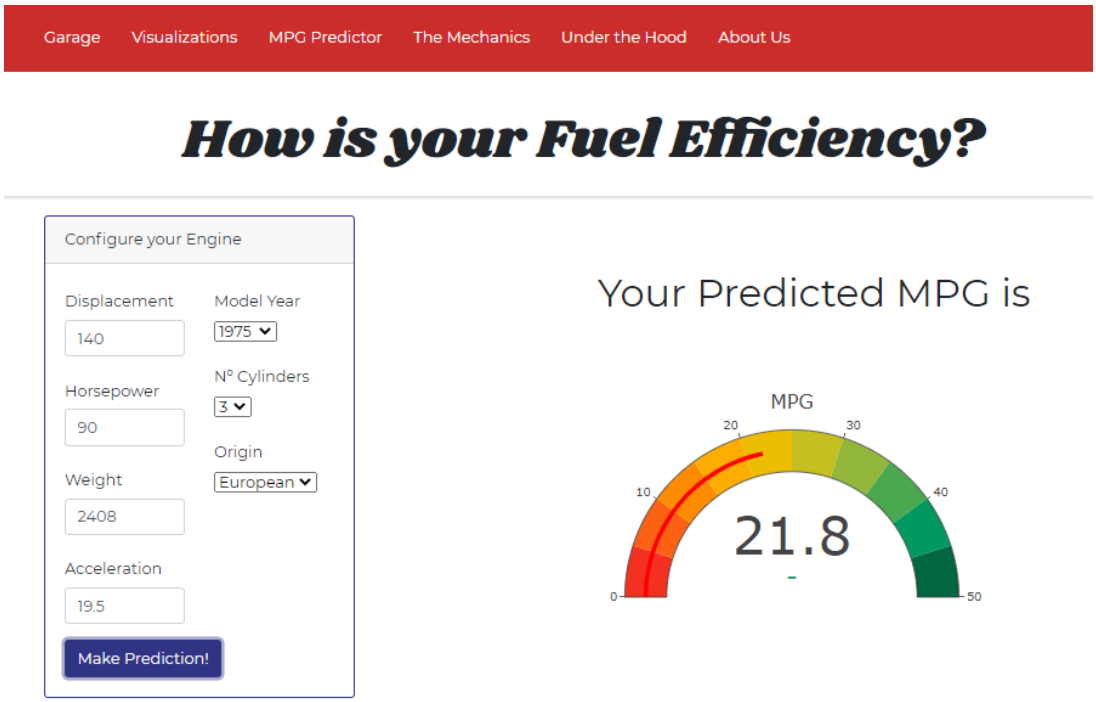
1. INTRODUCTION

As 1970 dawned the world's cars averaged 149 horsepower and 17 miles per gallon… gas was cheap and the roar of the engines drowned out Elvis Presley and Creedence Clearwater Revival on the radio.

October 1973 brought the Yom-Kippur War. Early in the war, the U.S supplied Israel with arms, angering the Arab delegation of OPEC, which responded with an embargo of oil sales to the U.S. and other industrial centers. And so began the decade's first oil crisis that sent oil prices skyrocketing upwards and auto manufacturers scrambled to offer more fuel efficient cars.

Hop in and take a journey throughout the 1970s with us as we first visualize oil prices and fuel efficiency throughout the decade. Then we'll pop the hood and see what changed to bring about an 88% improvement in fuel efficiency by 1982.

Our next stop is with Machine Learning models. As you're flipping through the latest digital copy of Hemmings, debating between a 1970 Pontiac GTO "The Judge" or an iconic 1975 Rolls Royce Silver Shadow, the model will predict the gas mileage you'll experience with your "new" vintage beauty.



2. DATA

Data is gratiously sourced from Kaggle and the University of California, Irvine.

The original data .csv file is relatively clean. It is a small dataset, approximately 400 records, and Excel was used for the minimal cleaning required. Six null values in "horsepower" field were replaced with the manufacturers' specified values for Tableau visualizations. The six values were replaced with median values for the machine learning models.

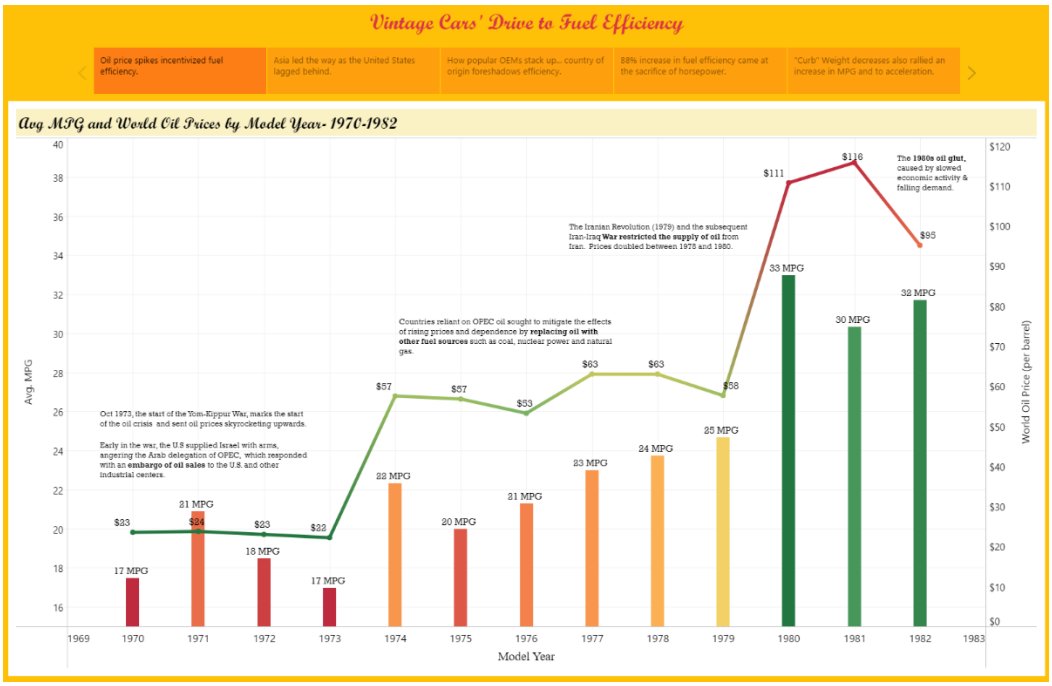
Make and model values were separated into independent fields utilizing Excel's native "text to columns" functionality, for better Tableau visualization prospects. "Make" was listed as unique values to spot misspellings which were then corrected and was capitalized for better tableau visualization. the clean .csv was read into Tableau.

Data fields include make, model, model year, horsepower, engine displacement, engine cylinders, acceleration, fuel efficiency, and vehicle weight.

A second .csv was imported that provides inflation adjusted world oil prices for each of the twelve years.

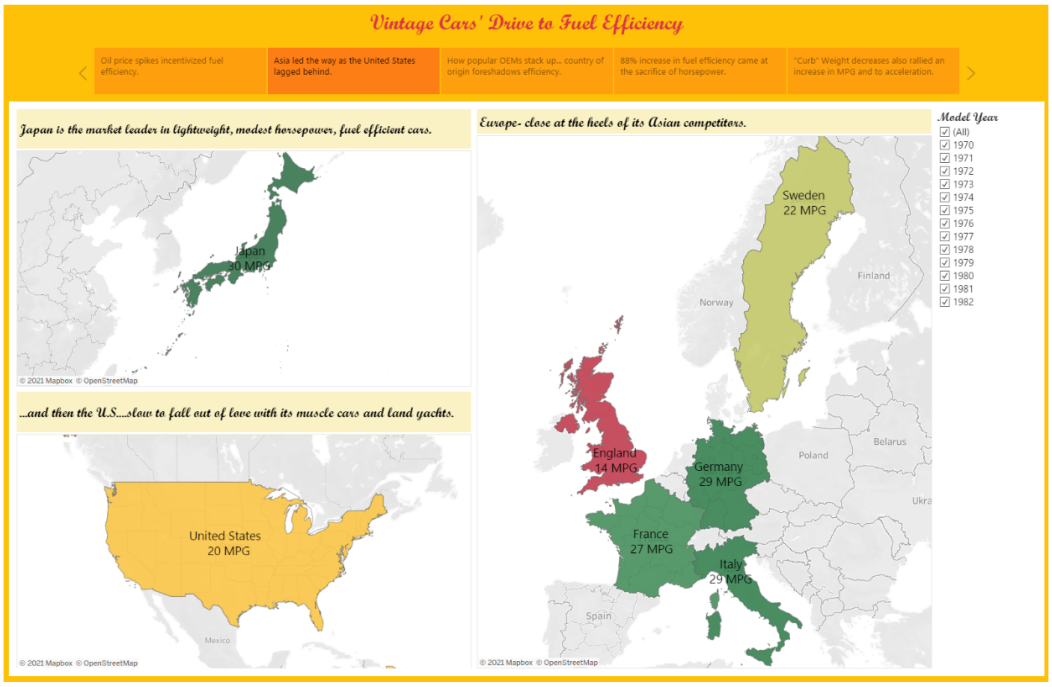
3. TABLEAU VISUALIZATIONS

Ten worksheets each have a visualization. The visualizations are brought together on six dashboards which are then presented as a story. The main filter serves to retrieve data for each year unless the data is presented as a time series. The story captions summarize each dashboard and guide the user through the dashboards.

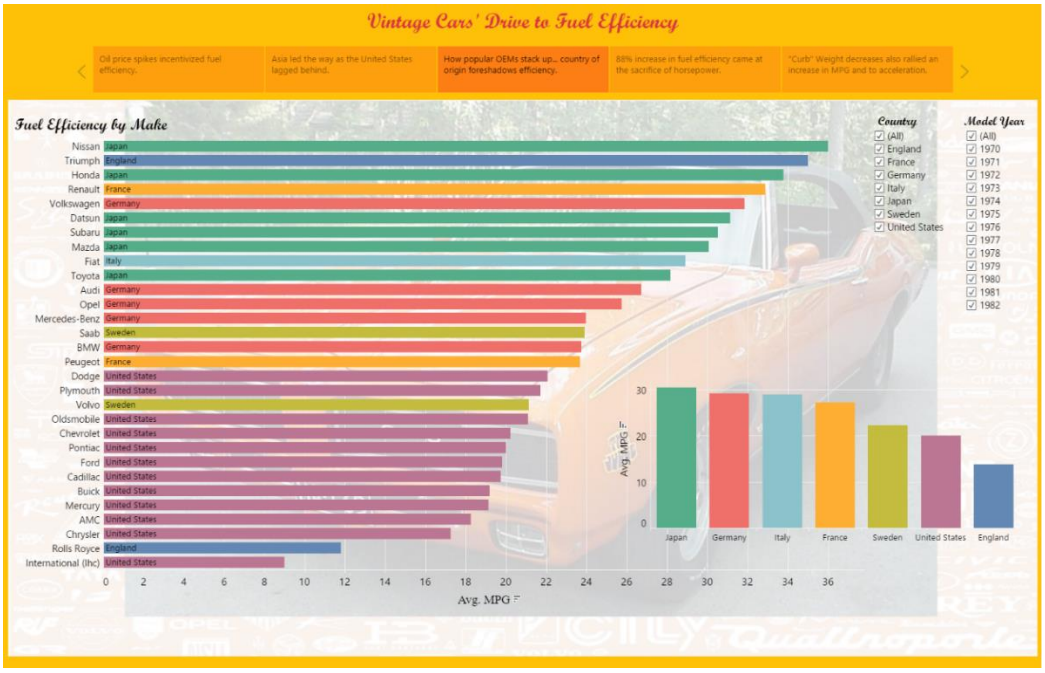


Next the user explores how country of origin influences fuel efficiency. Asia is the frontrunner for the time period with the United States and England trailing the pack.

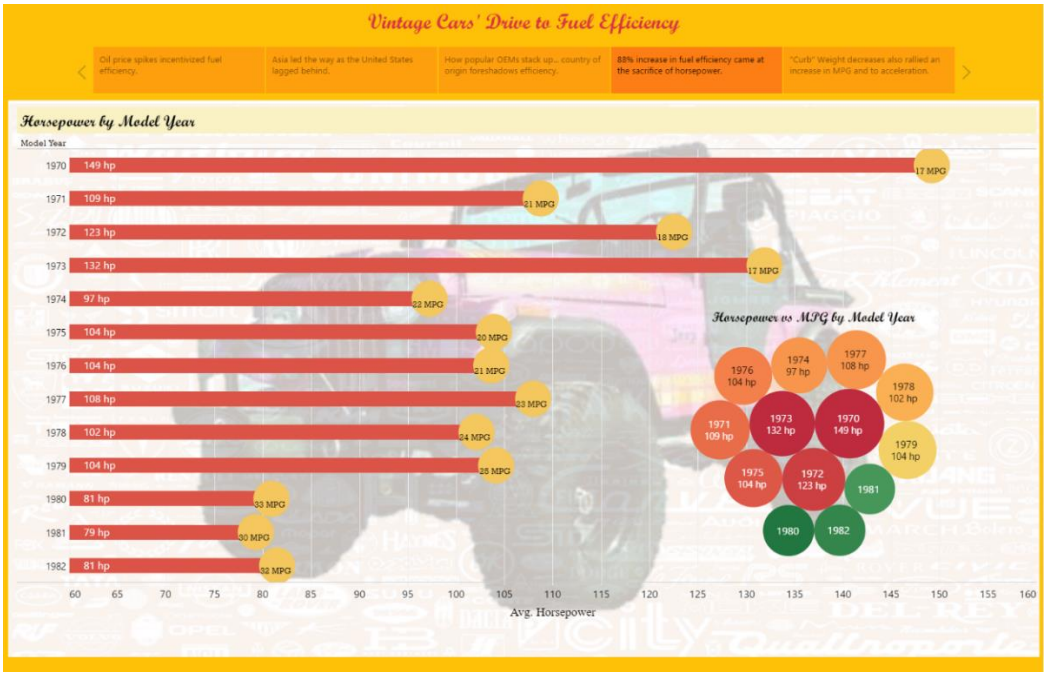
The user can select which year(s) to view and tooltips provides the average metrics for each field by country.



Individual make and fuel efficiency are examined in the third dashboard. The dashboard is generous with labels to provide an easy view of data and, again, tooltips are utilized to provide a wider data view. A summary by country also functions as a legend in the lower right corner.

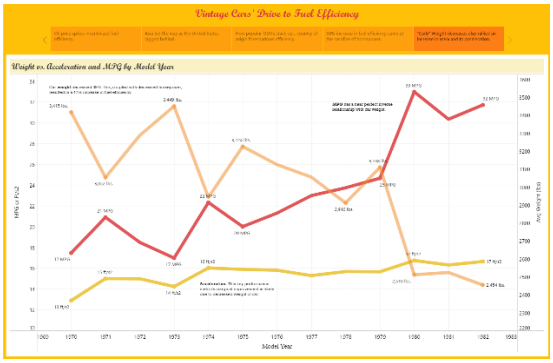


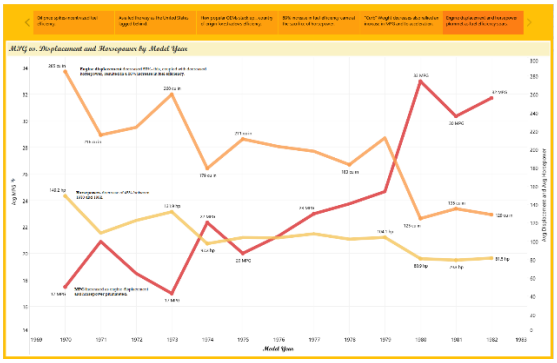
The user then explicity views the 46% reduction in horsepower and an 88% increase in fuel efficiency between 1970 and 1982.



Engine metrics roar to life in the final two dashboards. A 28% decrease in weight, a 45% decrease in horsepower, and a 55% decrease in engine displacement contributed to the 88% MPG improvement and a 31% improvement in acceleration.

Blended and dual axis scales allowed the three independent metrics to show with a shared x-axis.





4. MACHINE LEARNING

DATA

The dataset is imported into Jupyter Notebook and read into a pandas dataframe. Data is examined for null values, and pandas "Describe" is used to understand the data prior to machine learning model implementation.

Data Gathering and Preprocessing

```
# Read cvs file using pandas
mpg_df = pd.read_csv('../Data/data.csv')
mpg_df
```

t[2]:

	mpg	cylinders	displacement	horsepower	weight	acceleration	year	origin	make	model	country
0	9.0	8	304.0	193.0	4732	18.5	70	1	international (IHC)	1200d	United States
1	10.0	8	307.0	200.0	4376	15.0	70	1	chevrolet	c20	United States
2	10.0	8	360.0	215.0	4615	14.0	70	1	ford	f250	United States
3	11.0	8	380.0	189.0	4648	11.0	70	2	rolls royce	silver shadow	England
4	11.0	8	380.0	189.0	4648	11.0	71	2	rolls royce	silver shadow	England
...	...	...	...	...	...	...	...	...	...	...	...
404	43.4	4	90.0	48.0	2335	23.7	80	2	volkswagen	dasher (diesel)	Germany
405	44.0	4	97.0	52.0	2130	24.6	82	2	volkswagen	pickup	Germany
406	44.3	4	90.0	48.0	2085	21.7	80	2	volkswagen	rabbit c (diesel)	Germany
407	44.6	4	91.0	67.0	1850	13.8	80	3	honda	civic 1500 gl	Japan
408	46.6	4	86.0	65.0	2110	17.9	80	3	mazda	glc	Japan

409 rows × 11 columns

Examination for correlations is made both as a dataframe and a visualization.

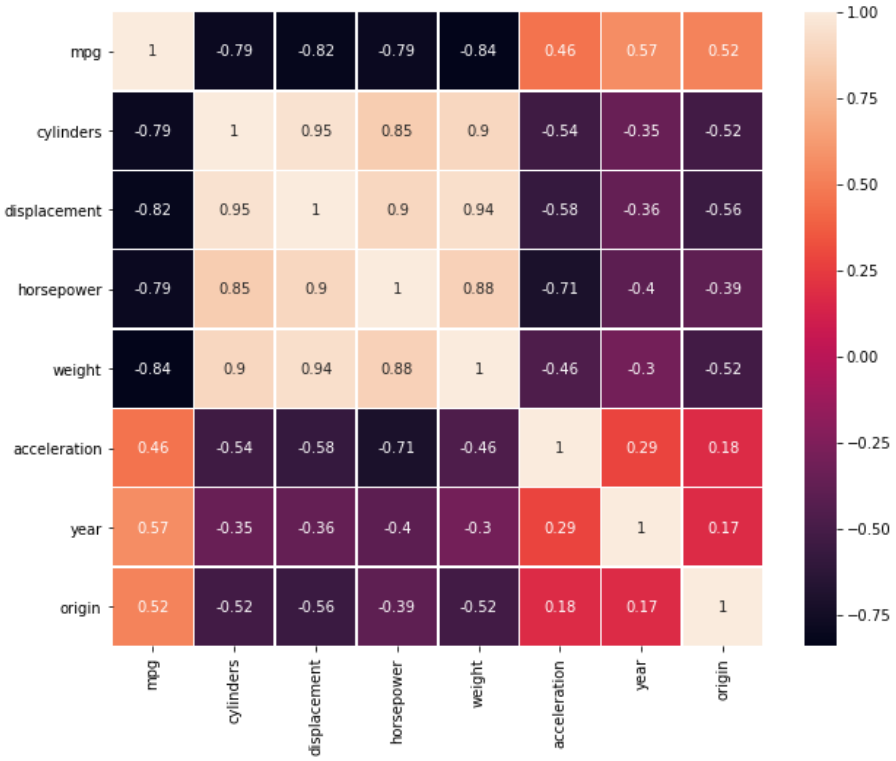
Correlations

```
linear_feat.corr()
```

11]:

	mpg	cylinders	displacement	horsepower	weight	acceleration	year	origin
mpg	1.000000	-0.788159	-0.816540	-0.787206	-0.841886	0.456455	0.567691	0.524289
cylinders	-0.788159	1.000000	0.953080	0.849093	0.901268	-0.536239	-0.345548	-0.523433
displacement	-0.816540	0.953080	1.000000	0.903833	0.938297	-0.577335	-0.363767	-0.557914
horsepower	-0.787206	0.849093	0.903833	1.000000	0.876306	-0.712086	-0.399472	-0.393686
weight	-0.841886	0.901268	0.938297	0.876306	1.000000	-0.464254	-0.302058	-0.523480
acceleration	0.456455	-0.536239	-0.577335	-0.712086	-0.464254	1.000000	0.286633	0.175036
year	0.567691	-0.345548	-0.363767	-0.399472	-0.302058	0.286633	1.000000	0.174139
origin	0.524289	-0.523433	-0.557914	-0.393686	-0.523480	0.175036	0.174139	1.000000

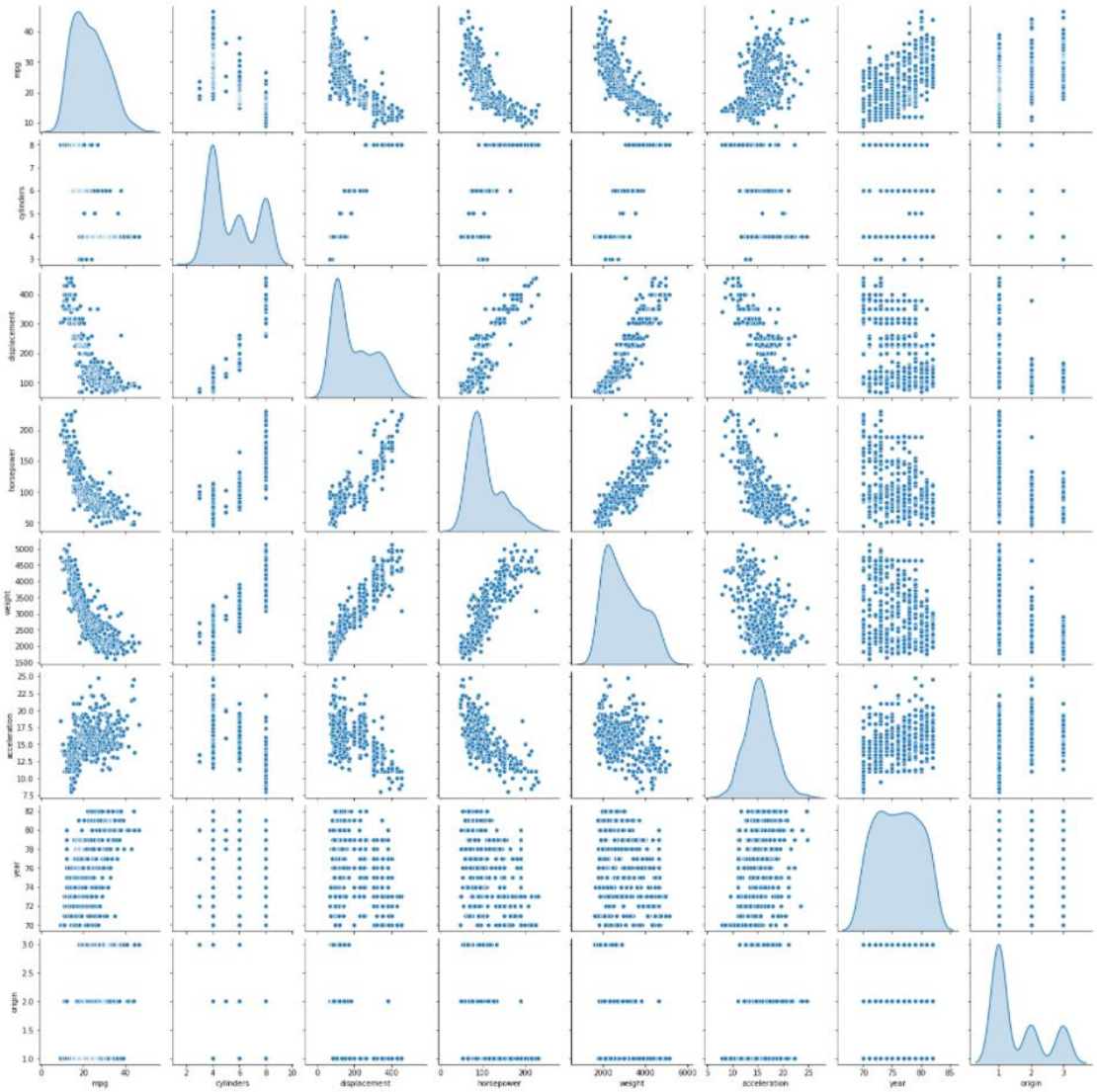
```
plt.figure(figsize=(10,8))
sns.heatmap(linear_feat.corr(), annot=True, linewidths=.5)
plt.show()
```



Pair plots, another correlation tool, clearly demonstrates that "cylinders" and "origin" fields do not show a normal distribution as they represent a specific value and can be considered categorical values.

```
##pairplots to get an intuition of potential correlations

sns.pairplot(mpg_df[["mpg", "cylinders", "displacement", "horsepower", "weight", "acceleration", "year", "origin"]], diag_kin
```

```
<seaborn.axisgrid.PairGrid at 0x297e154cfd0>
```



TRAINING THE MODEL

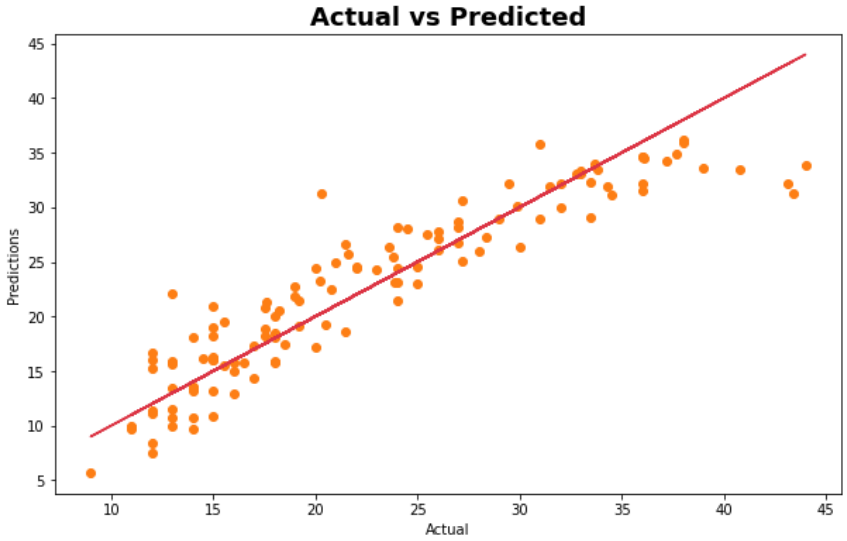
Training of the data begins with dropping columns determined to be categorical in nature. 30% of the data was used as a test set.

MODELS

Linear models explored include Linear, Ridge, Lasso, and ElasticNet.

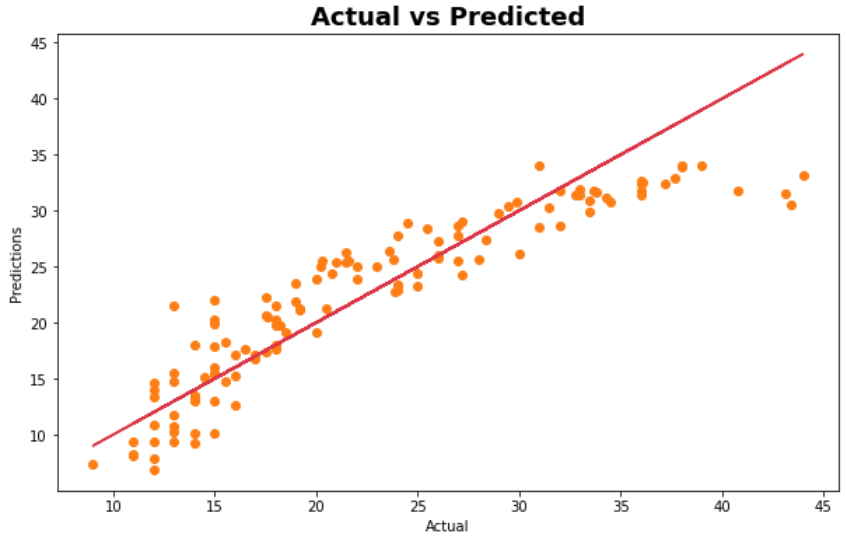
Linear model:

```
Model Evaluation Report
The In Sample R2 Score: 0.8455826329089134
The In Sample RMSE: 2.9674327357677006

The Out Sample R2 Score: 0.8507259859040834
The Out Sample RMSE: 3.3691296516747364
```



Elastic Net model:

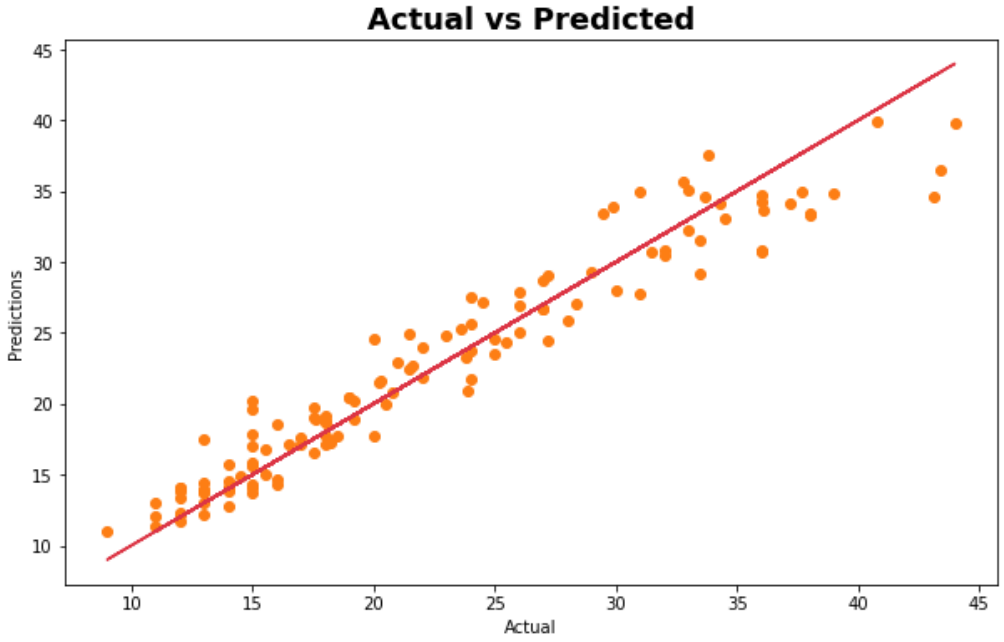
```
Model Evaluation Report
The In Sample R2 Score: 0.803633557413395
The In Sample RMSE: 3.3463121323221747

The Out Sample R2 Score: 0.8346667364178618
The Out Sample RMSE: 3.5457306168830875
```



Random Forest models explored are DecisionTree, Random Forest, AdaBoost, and Gradient Boost.

Random Forest:

```
Model Evaluation Report
The In Sample R2 Score: 0.977772795094775
The In Sample RMSE: 1.1258362765865744

The Out Sample R2 Score: 0.9278229947377916
The Out Sample RMSE: 2.342745149903927
```

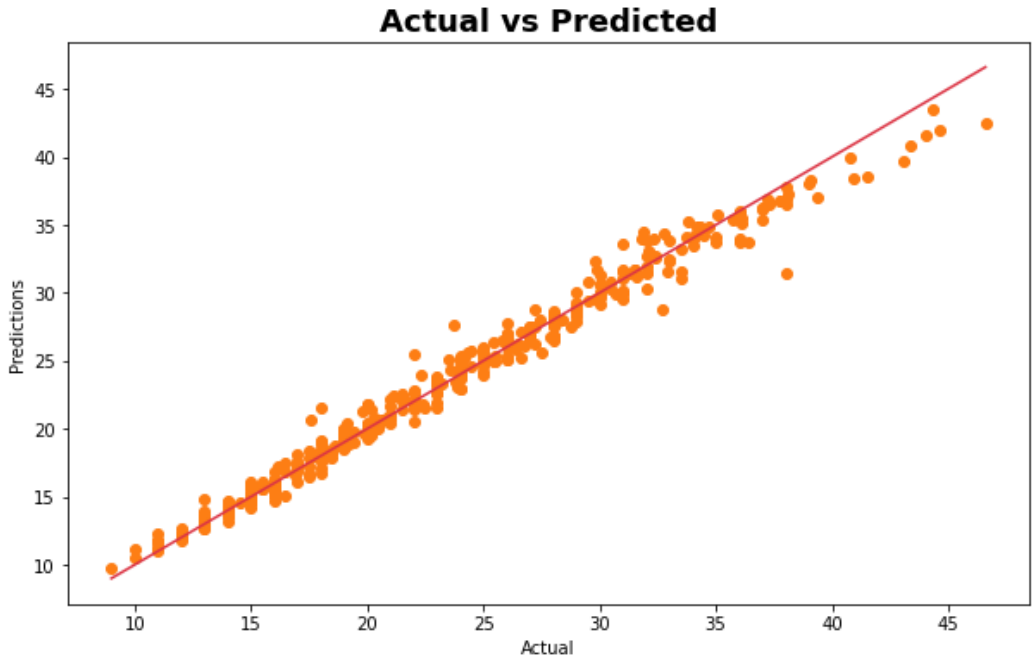


Model results were viewed as a dataframe for easy comparison and selection.

	Model	IN SAMPLE RMSE	OUT SAMPLE RMSE
5	RANDOM FR	1.125836	2.342745
7	GRADIENT BOOSTING	1.356529	2.657023
6	ADA BOOTS	2.447871	3.030495
1	RIDGE	3.005916	3.273203
4	DECISION TREE	0.000000	3.282090
0	LINEAR	2.967433	3.369130
3	ELASTICNET	3.346312	3.545731
2	LASSO	3.350587	3.547672

The Random Forest model was choosen because it has the lowest RMSE and does not overfit the in-sample data.

```
Model Evaluation Report
The In Sample R2 Score: 0.9837653391708724
The In Sample RMSE: 1.010377387191379
```



5. PREDICTING THE PAIN at the PUMP

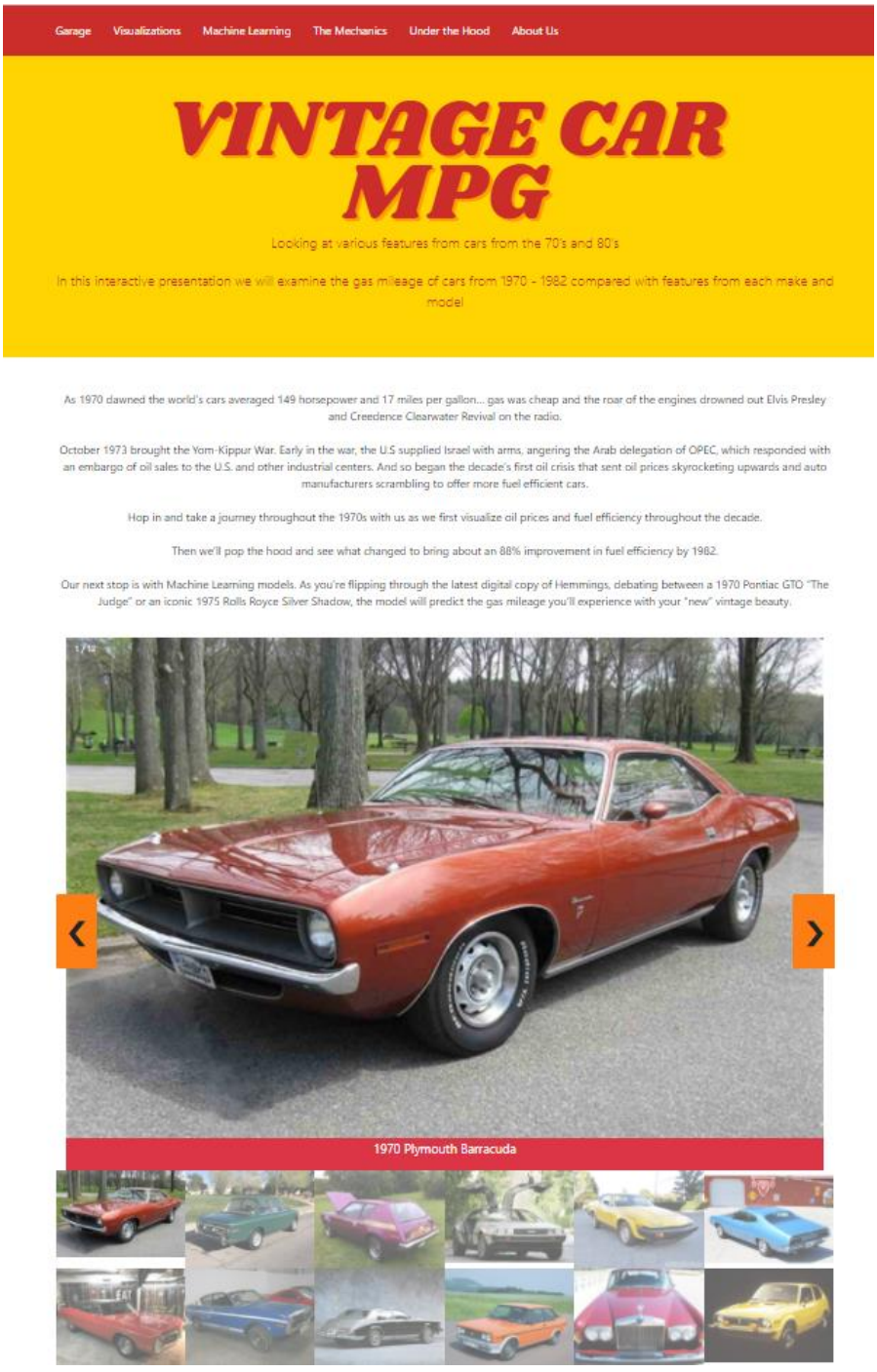
The pickled model is then used to predict a car's fuel efficiency based on characteristics selected by the user that include model year, engine displacement, horsepower, and vehicle weight.

The website displays the predicted result as both text and as a colored gauge.

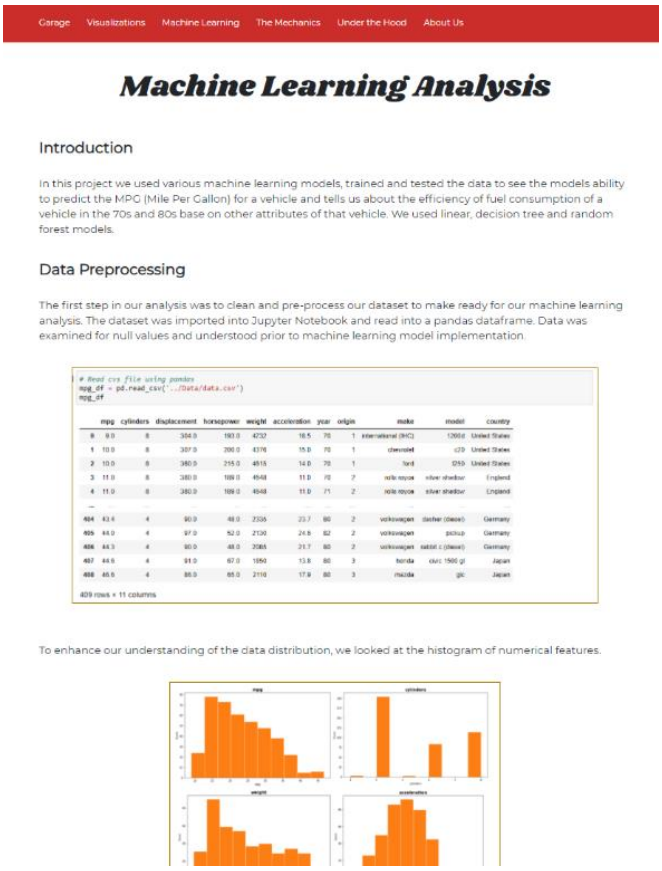
6. DEPLOYMENT

The project is packaged as a full stack web deployment on Heroku.

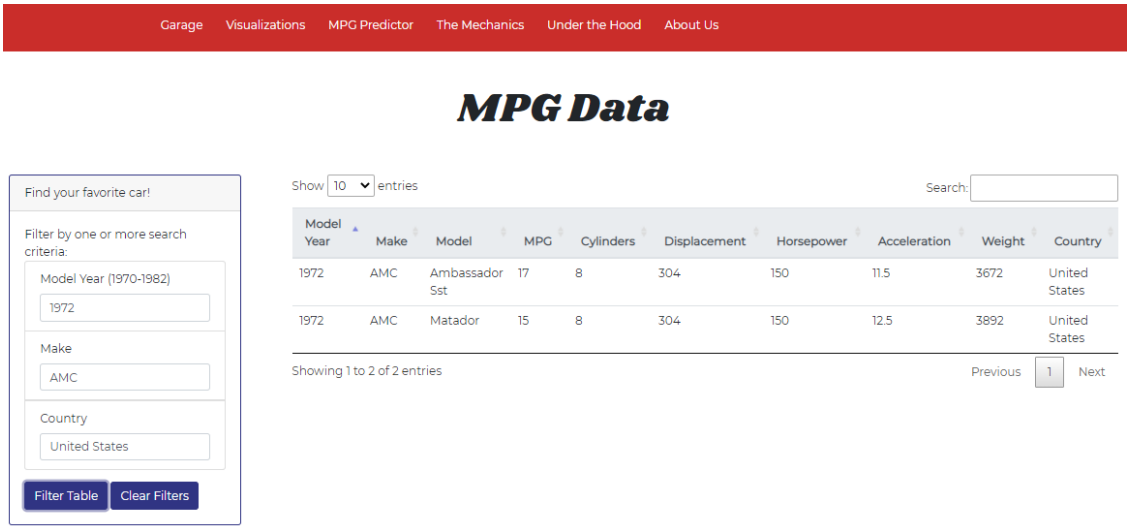
The "Garage" page introduces the user to the project. it includes an interactive slideshow with twelve car images from the time period. A navigation bar allows the user to visit several different pages:



- Visualizations - This is the Tableau storyboard
- MPG Predictor - This is the predictive activity
- The Mechanics - This is the explanation of the Machine Learning behind the prediction.



- Under the Hood - This is the dataset. This can be filtered by one to three metrics, or searched using the search too. Paganation is an added feature.



- About Us - This is the team that crafted the project.

Tableau is deployed in Tableau Public at:
https://public.tableau.com/profile/dana.woodruff#!/vizhome/Vintage-Car-MPG/VintageCars?publish=yes

The website is deployed at: https://vintagecarsmpg.herokuapp.com/

Project materials are held in GitHub at:
https://github.com/jessicapardo/Vintage-Cars-MPG
...with forks at:
https://github.com/danawoodruff/Vintage-Cars-MPG
https://github.com/BaudOptics/Vintage-Cars-MPG

7. REFERENCES

Carnegie Mellon University. *Auto MPG Data Set*. 1983. http://archive.ics.uci.edu/ml/datasets/Auto+MPG. 21 April 2021.

Hallman, Carly. *Gas Prices Through History*. 2016. https://www.titlemax.com/discovery-center/planes-trains-and-automobiles/average-gas-prices-through-history/. 23 April 2021.

History.com Editors. *Energy Crisis (1970s)*. 30 August 2010. https://www.history.com/topics/1970s/energy-crisis. 22 April 2021.

Macrotrends. *Crude Oil Prices - 70 Year Historical Chart*. 2021. https://www.macrotrends.net/1369/crude-oil-price-history-chart. 22 April 2021.

Wikipedia. *1970s energy crisis*. 15 April 2021. https://en.wikipedia.org/wiki/1970s_energy_crisis. 23 April 2021.