Section1:Introduction

The following is a detailed report outlining the causes and predictions towards the changing fuel prices; We have imported U.S. Gasoline and Diesel Retail Prices 1995-2021 dataset(www.eia.gov/dnav/pet/pet_pri_gnd_dcus_nus_w.htm) or you can find it (your google drive link here)

The change and in prediction price can be found by using regression models particularly Linear Regression which we have applied, But before that we have to pre-process, analyze data and its trends. The following is our dataset consisting of 14 columns and 1361 rows

For Data Dictionary refer to section 2

	Date	A1	A2	A3	R1	R2	R3	M1	M2	M3	P1	P2	P3	D1
0	01/02/1995	1.127	1.104	1.231	1.079	1.063	1.167	1.170	1.159	1.298	1.272	1.250	1.386	1.104
1	01/09/1995	1.134	1.111	1.232	1.086	1.070	1.169	1.177	1.164	1.300	1.279	1.256	1.387	1.102
2	01/16/1995	1.126	1.102	1.231	1.078	1.062	1.169	1.168	1.155	1.299	1.271	1.249	1.385	1.100
3	01/23/1995	1.132	1.110	1.226	1.083	1.068	1.165	1.177	1.165	1.296	1.277	1.256	1.378	1.095
4	01/30/1995	1.131	1.109	1.221	1.083	1.068	1.162	1.176	1.163	1.291	1.275	1.255	1.370	1.090
5	02/06/1995	1.124	1.103	1.218	1.076	1.062	1.159	1.169	1.157	1.288	1.270	1.250	1.368	1.086
6	02/13/1995	1.121	1.099	1.218	1.074	1.058	1.158	1.166	1.153	1.285	1.265	1.243	1.367	1.088
7	02/20/1995	1.115	1.093	1.213	1.067	1.052	1.153	1.160	1.148	1.280	1.259	1.239	1.363	1.088
8	02/27/1995	1.121	1.101	1.211	1.073	1.060	1.152	1.164	1.153	1.276	1.265	1.246	1.362	1.089
9	03/06/1995	1.123	1.103	1.209	1.076	1.063	1.149	1.167	1.157	1.275	1.263	1.244	1.358	1.089
10	03/13/1995	1.116	1.096	1.202	1.069	1.056	1.141	1.158	1.150	1.268	1.256	1.238	1.353	1.088
11	03/20/1995	1.114	1.095	1.201	1.068	1.055	1.140	1.158	1.149	1.267	1.254	1.236	1.351	1.085
12	03/27/1995	1.121	1.102	1.198	1.075	1.063	1.138	1.162	1.153	1.265	1.259	1.241	1.349	1.088
13	04/03/1995	1.133	1.116	1.198	1.087	1.077	1.140	1.174	1.167	1.266	1.270	1.255	1.350	1.094
14	04/10/1995	1.149	1.134	1.207	1.103	1.094	1.149	1.190	1.186	1.273	1.286	1.273	1.357	1.101

Data wrangling is perform on Date column to add more feature such as year and month

```
df['month'] = pd.DatetimeIndex(df['Date']).month
```

```
df['year'] = pd.DatetimeIndex(df['Date']).year
```

```
df.info()
```

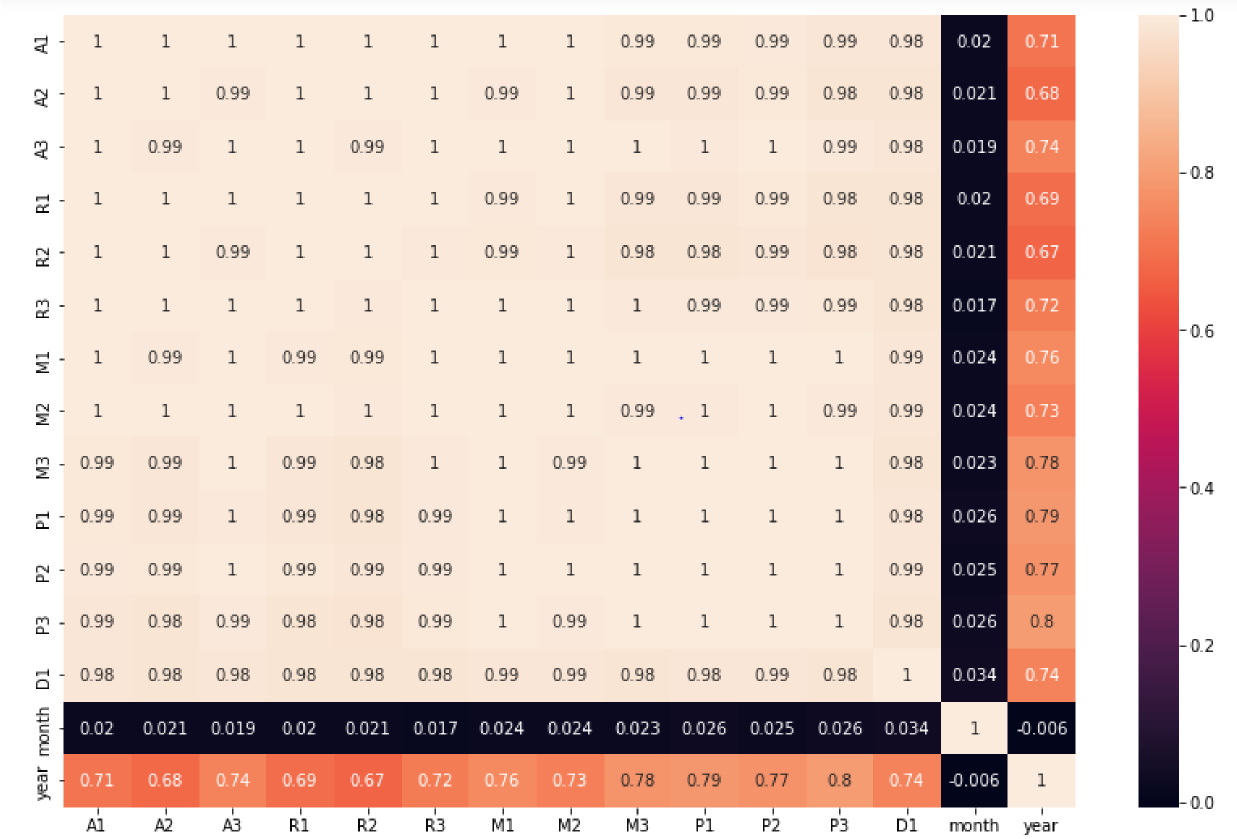
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1361 entries, 0 to 1360
Data columns (total 16 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   Date    1361 non-null   object
 1   A1      1361 non-null   float64
 2   A2      1361 non-null   float64
 3   A3      1361 non-null   float64
 4   R1      1361 non-null   float64
 5   R2      1361 non-null   float64
 6   R3      1361 non-null   float64
 7   M1      1361 non-null   float64
 8   M2      1361 non-null   float64
 9   M3      1361 non-null   float64
 10  P1      1361 non-null   float64
 11  P2      1361 non-null   float64
 12  P3      1361 non-null   float64
 13  D1      1361 non-null   float64
 14  month   1361 non-null   int64
 15  year    1361 non-null   int64
dtypes: float64(13), int64(2), object(1)
memory usage: 170.2+ KB
```

The stats of data can be seen below:

A2	A3	R1	R2	R3	M1	M2	M3	P1	P2	P3	D1
61.000000	1361.000000	1361.000000	1361.000000	1361.000000	1361.000000	1361.000000	1361.000000	1361.000000	1361.000000	1361.000000	1361.000000
2.234511	2.396873	2.225170	2.178511	2.329126	2.382822	2.320970	2.508877	2.519840	2.472096	2.609244	2.404699
0.843815	0.883311	0.850143	0.835549	0.876739	0.882107	0.858521	0.908861	0.911055	0.894472	0.925587	0.998646
0.926000	1.039000	0.907000	0.885000	0.974000	1.008000	0.979000	1.112000	1.100000	1.074000	1.191000	0.953000
1.433000	1.550000	1.421000	1.393000	1.489000	1.517000	1.482000	1.616000	1.607000	1.573000	1.695000	1.418000
2.251000	2.458000	2.237000	2.175000	2.367000	2.481000	2.404000	2.627000	2.693000	2.640000	2.769000	2.479000
2.825000	3.060000	2.828000	2.765000	2.976000	3.033000	2.930000	3.206000	3.209000	3.127000	3.318000	3.070000
4.102000	4.301000	4.114000	4.054000	4.247000	4.229000	4.153000	4.387000	4.344000	4.283000	4.459000	4.764000

There is no need of normalization because there is negligible variance in dataset

Before applying the model we need to analyze the correlation between different features as shown below



Month seems to be highly no correlated with other features. However, the other oil shows a strong correlation between them, It can be assume that the effect on oil price happens globally and every petroleum product get effected.

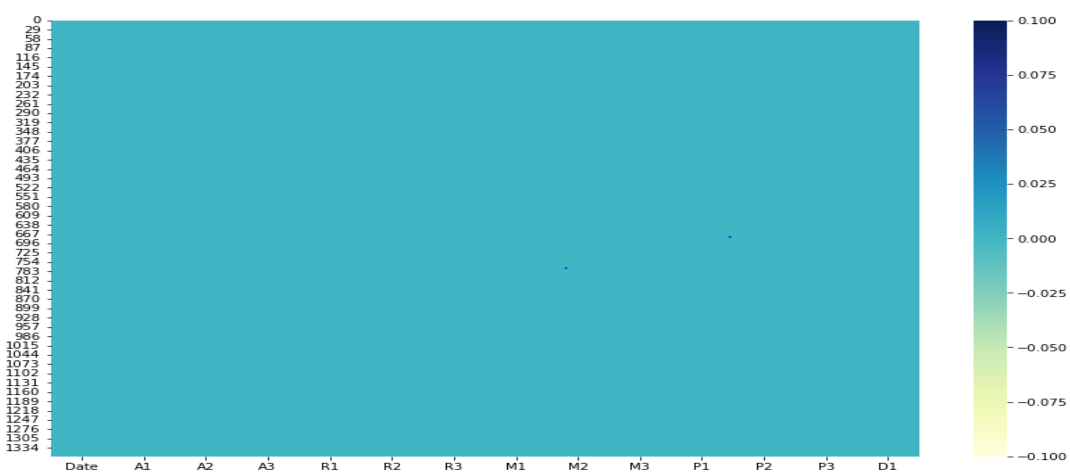
Section2:Data Dictionary:

- A1 = Weekly U.S. All Grades All Formulations Retail Gasoline Prices (Dollars per Gallon)
- A2 = Weekly U.S. All Grades Conventional Retail Gasoline Prices (Dollars per Gallon)

- A3 = Weekly U.S. All Grades Reformulated Retail Gasoline Prices (Dollars per Gallon)
- R1 = Weekly U.S. Regular All Formulations Retail Gasoline Prices (Dollars per Gallon)
- R2 = Weekly U.S. Regular Conventional Retail Gasoline Prices (Dollars per Gallon)
- R3 = Weekly U.S. Regular Reformulated Retail Gasoline Prices (Dollars per Gallon)
- M1 = Weekly U.S. Midgrade All Formulations Retail Gasoline Prices (Dollars per Gallon)
- M2 = Weekly U.S. Midgrade Conventional Retail Gasoline Prices (Dollars per Gallon)
- M3 = Weekly U.S. Midgrade Reformulated Retail Gasoline Prices (Dollars per Gallon)
- P1 = Weekly U.S. Premium All Formulations Retail Gasoline Prices (Dollars per Gallon)
- P2 = Weekly U.S. Premium Conventional Retail Gasoline Prices (Dollars per Gallon)
- P3 = Weekly U.S. Premium Reformulated Retail Gasoline Prices (Dollars per Gallon)
- D1 = Weekly U.S. No 2 Diesel Retail Prices (Dollars per Gallon)

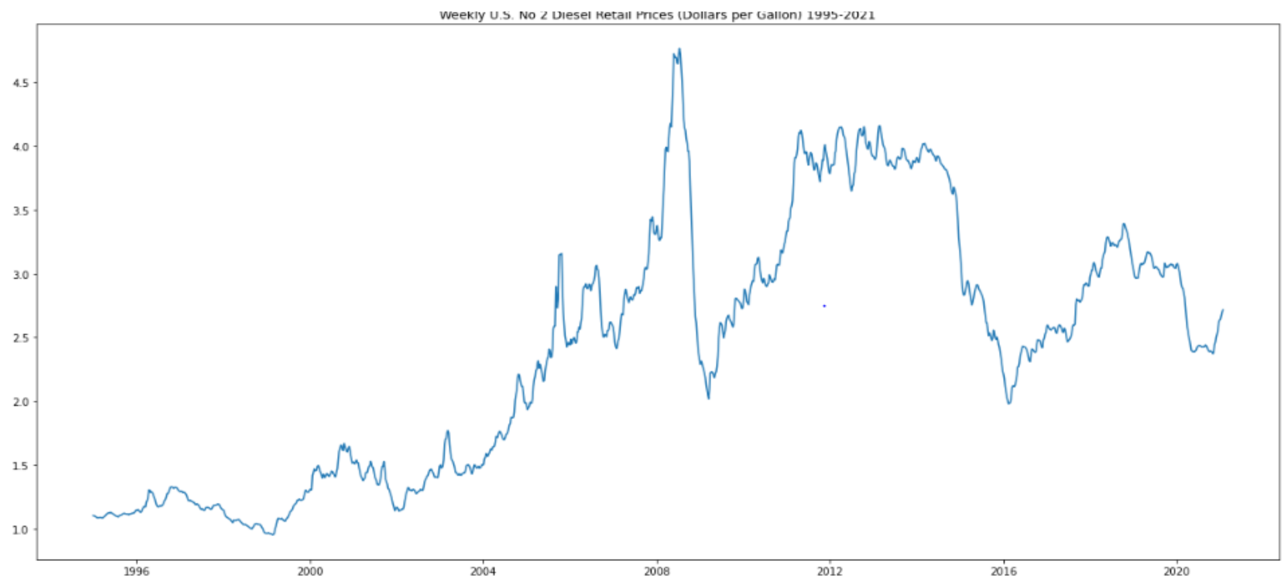
Training and Testing data:

Before applying regression model we have to split our dataset into test and train data, beside that our data must not have null values to check that we have drawn heat map.

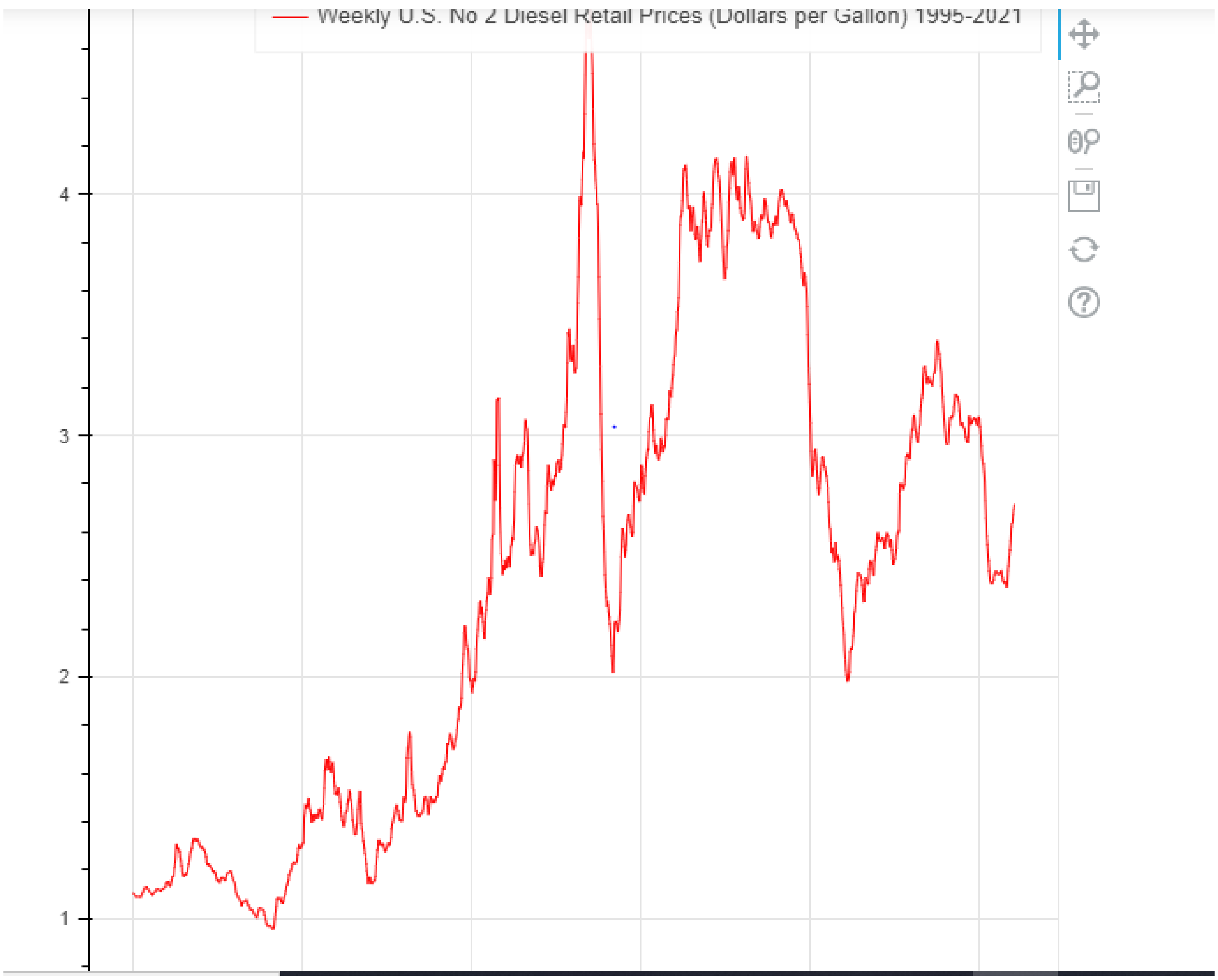


It can be clearly seen that there exist no null values indicating that we can apply our regression model.

We have to predict Diesel Retail Prices (Dollars per Gallon) “D1”. Let’s analyze the trend first



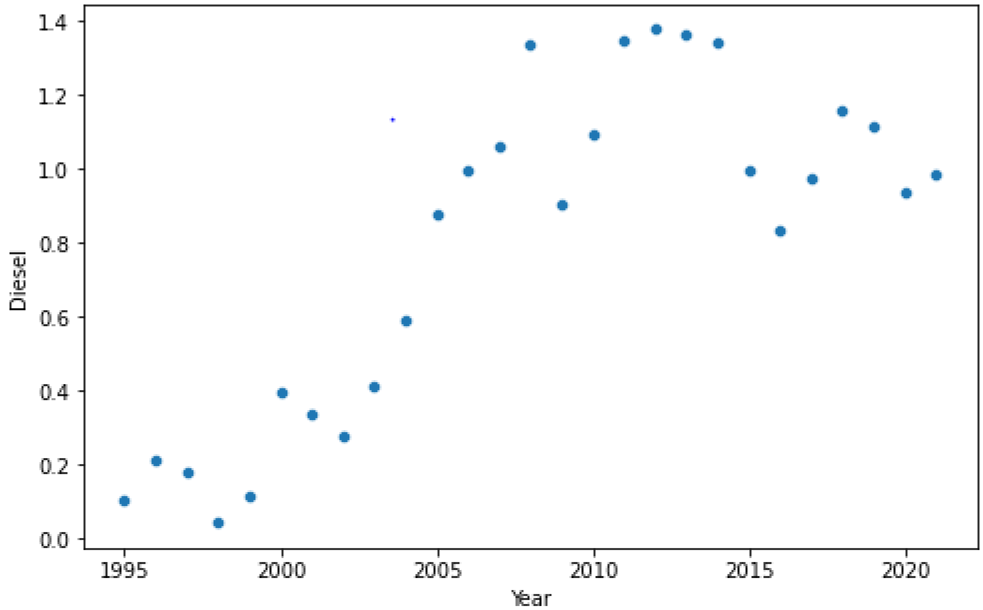
This graph illustrates the fluctuation in diesel prices; It can be clearly seen a rising trend since year(2000) that dominated in year 2018 which is then followed by a significant incline. However it can be assume that in 2020 the prizes will be again increasing which is what we have to predict.



We have divide our splitting parameter as X(YEAR) & y(Price) we have train our data till 2019 which is then test on year 2020.

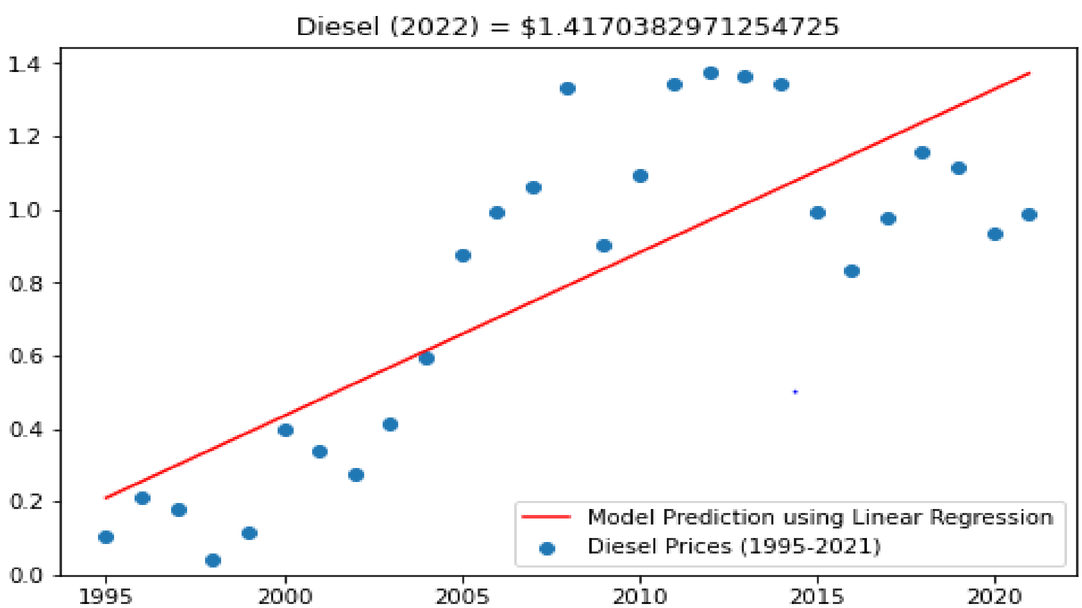
Analysis

WE have drawn a scatterplot between year and price as shown below



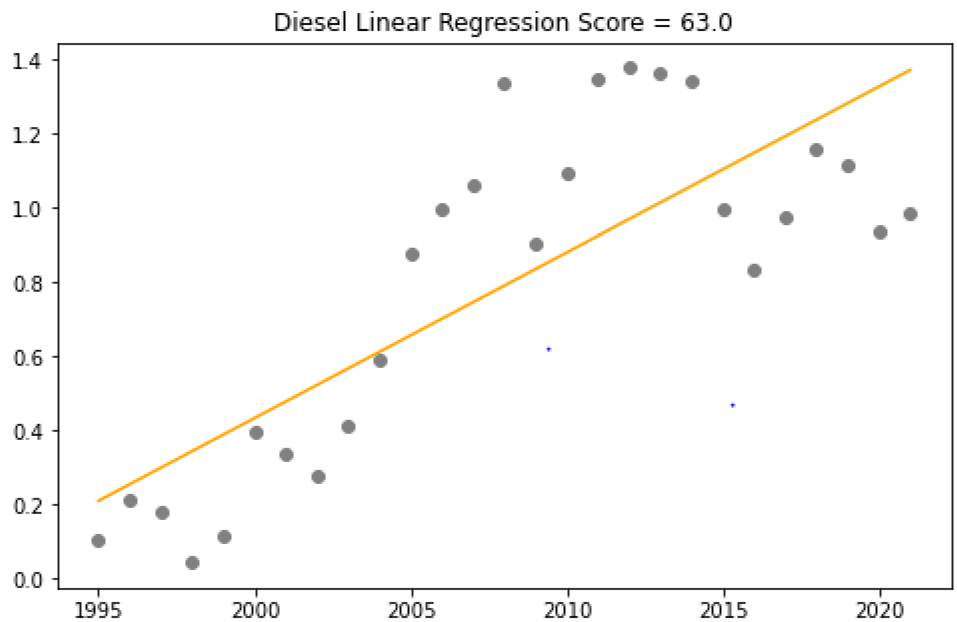
After training model we predict and the year 2020 and it can be clearly seen a rise in fuel price which

```
Predict Weekly U.S. No 2 Diesel Retail Prices (Dollars per Gallon) 1995-2021
$1.4170382971254725 USD
```

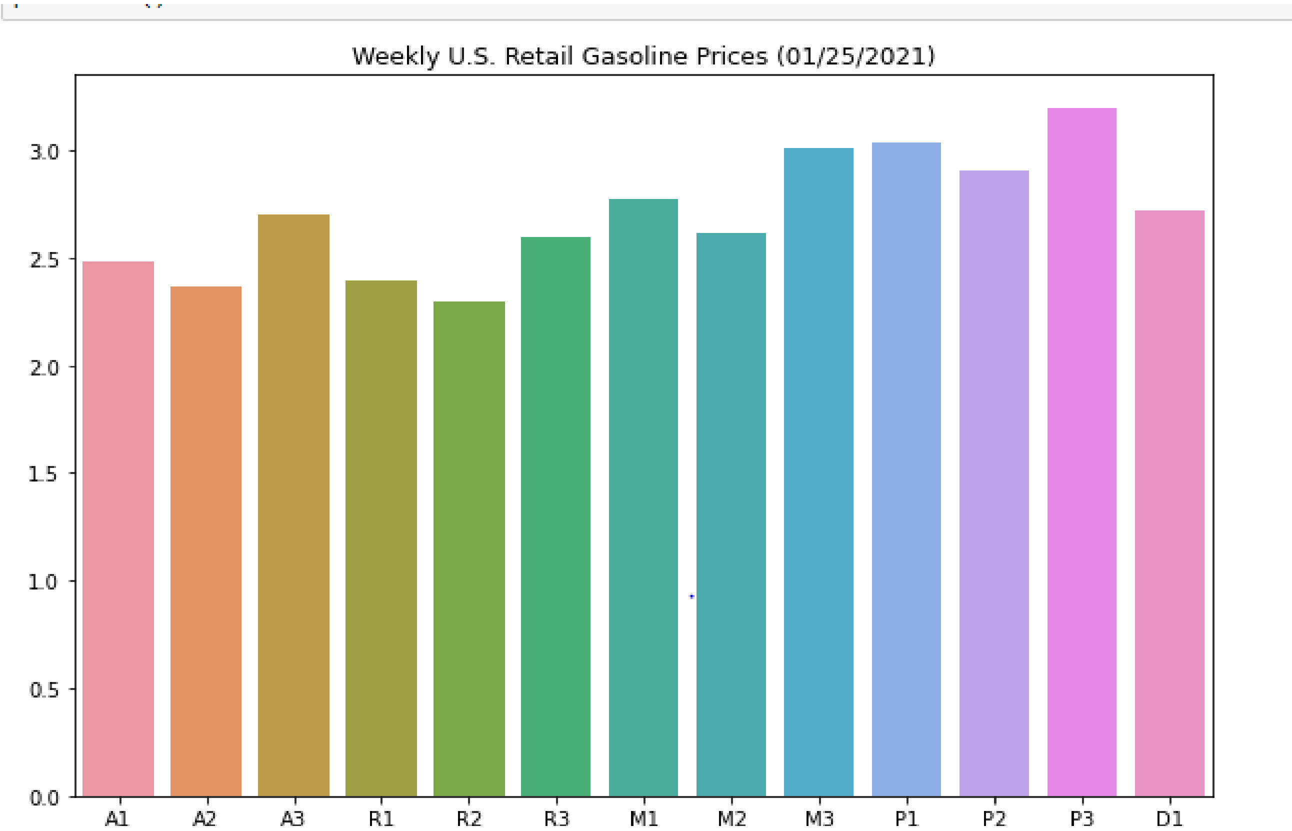


We can interpret that diesel in 2022 will again increase and touch 1.41 $ meanwhile the regression score is found to be 63.0 and coef of 0.04 indicating a periodically rise in price according to coef.

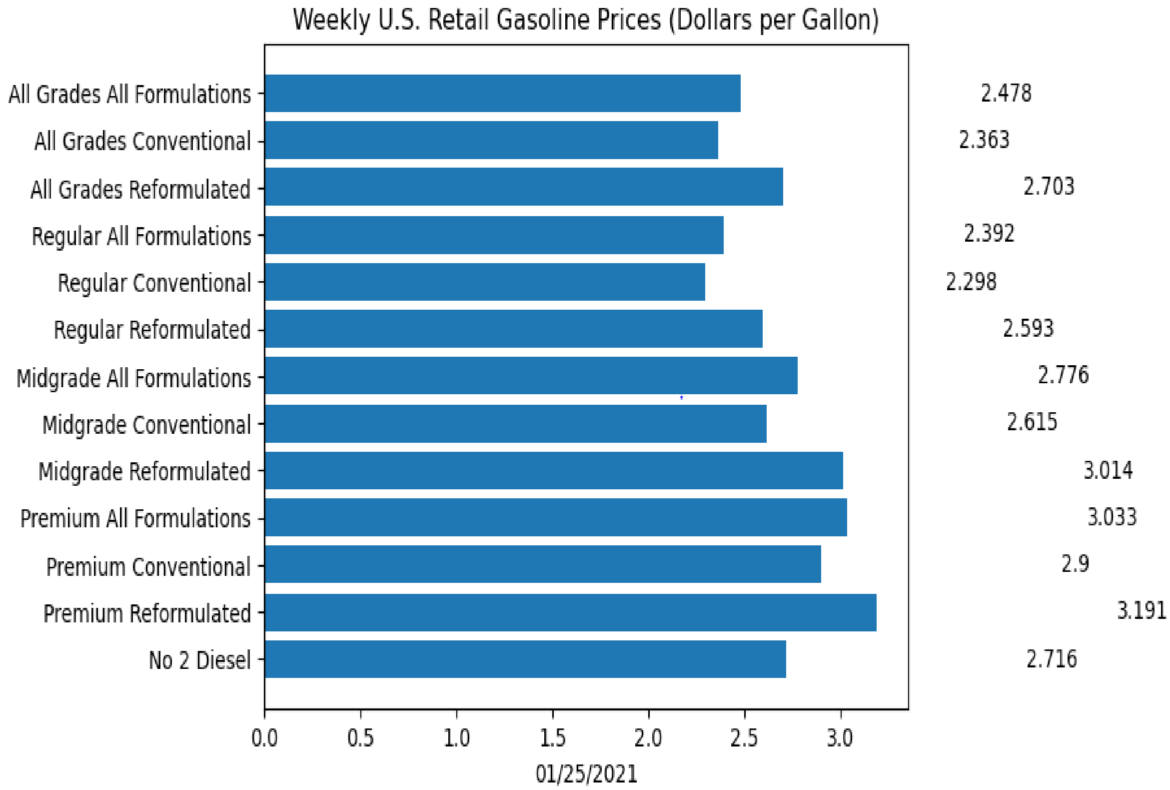
```
Diesel
score = 63.0
coef = [0.04468824]
intercept = -88.9425781818661
SciKit-Learn
```



Apart from D1 prices analysis if we look at weekly price of other fuels so we can find something new



- P3 (Weekly U.S. Premium Reformulated Retail Gasoline Prices) seems to be the most dominating fuel.



We can understand much clearly from above diagram.

Conclusion:

From the above diagram we have find the trends of fuel price which seems to be quite fluctuated. However it has experienced a great decline and its recovering again which is our model predicted too, the probability of touching 1.14$ again. Our model performed quite well on the test data of 2020 with a regression score of 63 and coef of 0.04 which is acceptable