Fuel Price Data

A. Introduction

The Fuel price dataset examines the peculiarities of gallon-based gasoline sales. The data collection has 8 variables and 708 observations. The data is from August 2010 to July 2011, it consists of everyday change in the price. Other variables such as weekdays from Sunday to Saturday and additional columns such as site A and site B are added to the data set. Volume column in the dataset is considered as a target variable. All the statistical analysis and visualization operations are performed in Microsoft Excel sheets. 10 regression models are analyzed using 'Data Analysis' tool in excel to get a better model.

B. Descriptive Statistics

Summary	Vol	PO	AvgCompPrice	MinCompPrice	MaxCompPrice
Mean	3343.15	3.54	3.506	3.45	3.55
Median	2840	3.4	3.35	3.3	3.4
Minimum	1472	2.96	2.92	2.88	1.34
Maximum	6147	4.28	4.255	4.22	2.96
Standard Deviation	1158.45	0.43274	0.43066	0.4278	0.4296
Range	4675	1.32	1.335	1.34	1.34

Table 1. Shows Summary Statistics of the variables

From Table 1. It can be observed that Volume variable has the highest standard deviation of 1158.45 and other variable has standard deviation around 0.4. Standard deviation is a statistic measures that evaluates the distribution of a dataset in respect to its mean. The standard deviation is higher when the data points are more spread out than the mean. As the standard deviation is higher for volume the range is also higher. Range is calculated by subtracting the lowest from the highest value. An extremely wide range indicates a high level of variability, whereas an extremely narrow range indicates low levels of variability. Observing the minimum and maximum volume, it can be noted that the difference is almost of 4 to 5 times.

C. Visualization

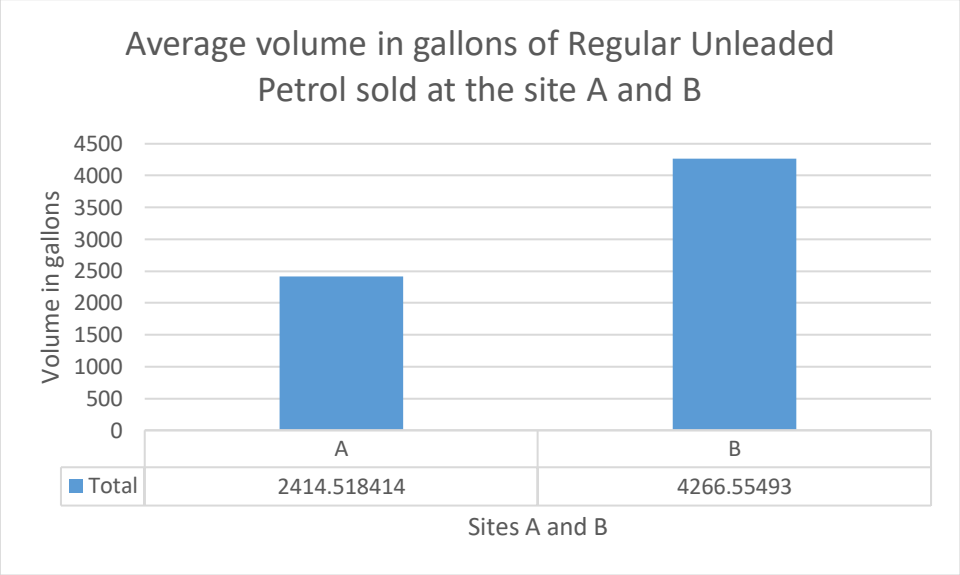

Fig 1. Shows bar graph of Sites Vs Volume of petrol

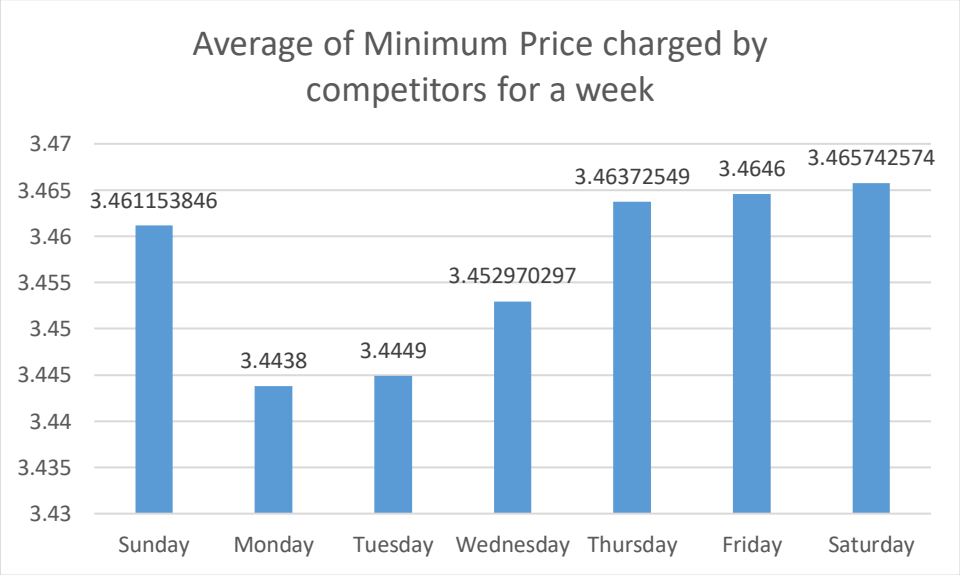

Fig 2. Shows Bar graph of week Vs Minimum Competitive Price

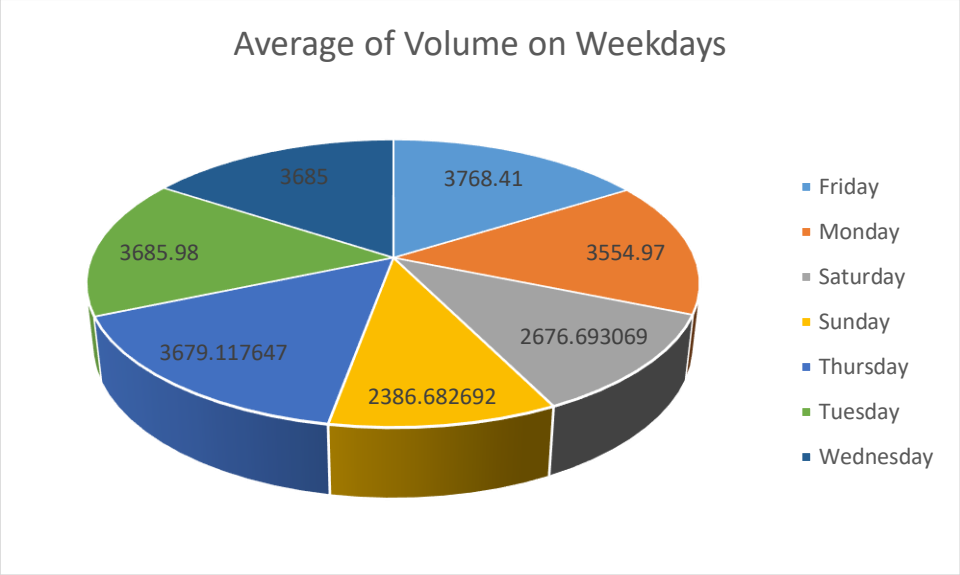

Fig 3. Shows Average of volume on Weekdays

D. Measures of Association

CORRELATION	Vol	Log(Vol)	P0	AvgCompPrice	MinCompPrice	MaxCompPrice
Vol	1	0.991297345	-0.03325316	-0.025931788	-0.045346888	-0.020886219
Log(Vol)	0.991297345	1	-0.03879098	-0.030954856	-0.049799333	-0.025986212
P0	-0.033253156	-0.038790979	1	0.998368124	0.993905809	0.998453745
AvgCompPrice	-0.025931788	-0.030954856	0.998368124	1	0.997272709	0.998757161
MinCompPrice	-0.045346888	-0.049799333	0.993905809	0.997272709	1	0.993296034
MaxCompPrice	-0.020886219	-0.025986212	0.998453745	0.998757161	0.993296034	1

Table 2. Shows Correlation between different variables

E. Regression Analysis

It is observed that there are 4 numeric variables to perform Regression analysis excluding the target variables. Regression Analysis is been performed by filtering Site A and Site B. The first model is been created using the 2 variables which are P0 and AvgCompPrice the R-square value which triggered was 0.0173 for Site A which is very low. Similarly all the 4 variables were used and the R-square value was increased to 0.08261 for Site A and 0.144686 for Site B.

With the help of the column 'Weekday', 7 dummy variables Sunday, Monday, Tuesday, Wednesday, Thursday, Friday and Saturday were created. Model 3 and Model 4 were created which consisted of Weekdays.

Few new columns using the existing variables were also created using functions like Log to the base 10, previous day values like price, new variable if the price went up or down compared to the previous day. If logarithm is used creating new variables, then Log(Volume) was used as a Target variable. The highest R-Square value obtained was 0.5775 for Site A and 0.3817 for Site B as shown in the below table.

Regression Models	Variables Used	R-Square		dF
		Site - B	Site - A	
1	P0, AvgCompPrice	0.076649	0.0173	2
2	P0, AvgCompPrice, MinCompPrice and MaxCompPrice	0.144686	0.08261	4
3	P0, AvgCompPrice, MinCompPrice, MaxCompPrice, Monday, Tuesday, Wednesday, Thursday and Friday	0.354389	0.536033	9
4	P0, AvgCompPrice, MinCompPrice, MaxCompPrice, Monday, Tuesday, Wednesday, Thursday, Friday, Saturday and Sunday	0.359261	0.548745	11
5	P0, AvgCompPrice, MinCompPrice, MaxCompPrice, Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, Sunday, Log(P0) and Log(AvgPrice)	0.376854	0.577273	13
6	P0, AvgCompPrice, MinCompPrice, MaxCompPrice, Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, Sunday Log(P0), Log(AvgPrice) Log(MinCPrice) Log(MaxCPrice)	0.381766	0.577538	15
7	P0, P0(t-1), AvgCompPrice, P0 increment, P0 Decrement, Log(AvgPrice), MinCompPrice, Log(P0(t-1)), Monday, Tuesday, Wednesday, Thursday, Friday, Saturday and Sunday	0.380107	0.572997	15
8	P0, P0(t-1), P0 increment, P0 Decrement, AvgCompPrice, Log(AvgCompPrice), MinCompPrice, Monday, Tuesday, Wednesday, Friday, Saturday, Sunday, A, B, Log(P0(t-1)), r5(t), and log(r5(t))	0.298678	0.362702	16

Table 3. Shows R-square values of different models for Site A and Site B.

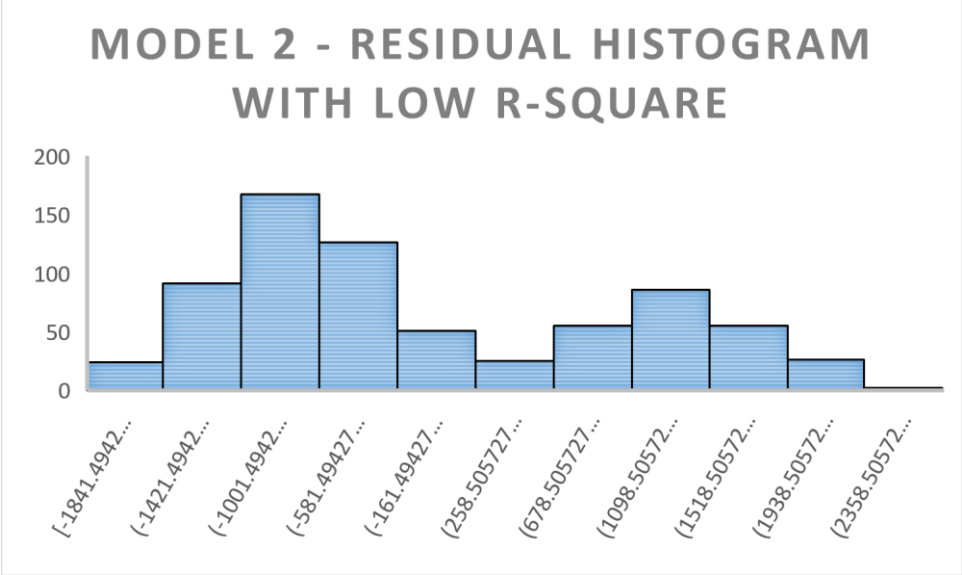

Figure 4. Shows uneven residual Histogram

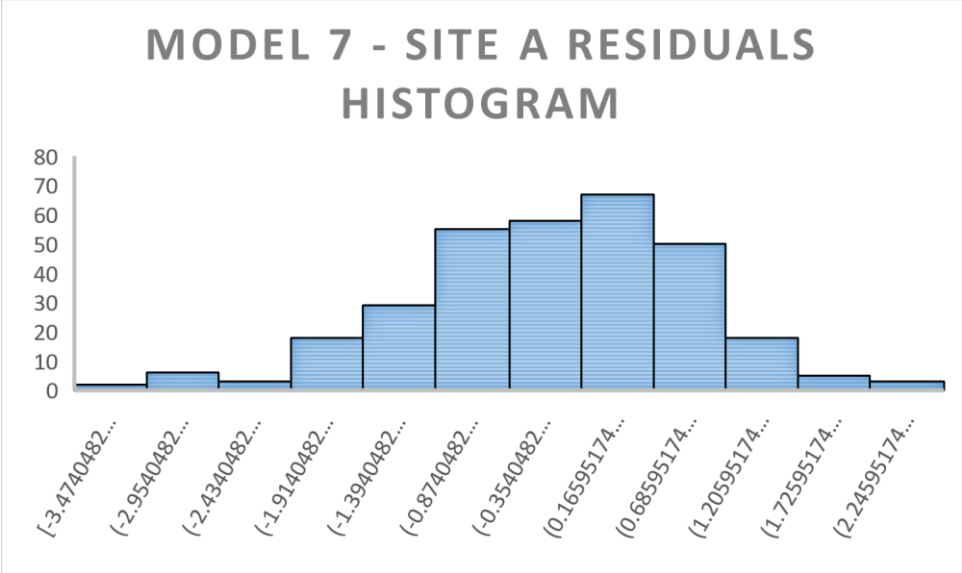

Figure 5. Shows Site-A improvised model Histogram

ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	11	23757456.58	2159768.78	41.58860601	1.88033E-56			
Residual	342	19536721.55	57124.9168					
Total	353	43294178.13						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	3110.293525	130.6260891	23.810661	1.39532E-74	2853.361851	3367.225199	2853.361851	3367.225199
P0	-4002.44604	573.3966839	-6.980239252	1.53798E-11	-5130.274109	-2874.61797	-5130.274109	-2874.61797
AvgCompPrice	-1411.883091	2944.822681	-0.479445876	0.631927902	-7204.127399	4380.361218	-7204.127399	4380.361218
MinCompPrice	2133.644713	1543.589108	1.382262094	0.167793288	-902.4787462	5169.768172	-902.4787462	5169.768172
MaxCompPrice	3158.1462	1425.202426	2.215928168	0.027354328	354.8804246	5961.411976	354.8804246	5961.411976
Monday	-170.0533783	47.88795188	-3.551068101	0.000437427	-264.2453711	-75.86138556	-264.2453711	-75.86138556
Tuesday	23.20411547	48.40697548	0.479354788	0.631992624	-72.00875762	118.4169886	-72.00875762	118.4169886
Wednesday	0	0	65535	#NUM!	0	0	0	0
Thursday	40.44129514	47.91400158	0.844039191	#NUM!	-53.80193544	134.6845257	-53.80193544	134.6845257
Friday	85.39684107	47.96720526	1.780317211	0.075911753	-8.951037122	179.7447193	-8.951037122	179.7447193
Saturday	-424.1913129	48.11337145	-8.816495291	6.10799E-17	-518.8266889	-329.5559368	-518.8266889	-329.5559368
Sunday	-571.1961814	47.64982946	-11.98737095	7.01586E-28	-664.9198054	-477.4725575	-664.9198054	-477.4725575

Figure 6. Shows Site-A Model 4 with Vol as Target variable

ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	15	0.864173047	0.057611536	35.64911799	2.2102E-60			
Residual	339	0.632132724	0.001864698					
Total	354	1.496305772						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0*
Intercept	4.0426341	0.161221193	25.07507862	3.24145E-79	3.725514195	4.359754004	3.725514195	4.35975
P0	0.528053029	0.95387825	0.553585355	0.580227577	-1.348212573	2.404318632	-1.348212573	2.40431
AvgCompPrice	-1.708529627	1.27014624	-1.345144026	0.179477871	-4.206890085	0.789830831	-4.206890085	0.78983
MinCompPrice	0.453587621	0.283122882	1.602087467	0.110067621	-0.103311255	1.010486496	-0.103311255	1.01048
MaxCompPrice	1.220991488	1.269008903	0.962161483	0.336654285	-1.275131842	3.717114818	-1.275131842	3.71711
Monday	-0.031168642	0.008669178	-3.595340116	0.000372072	-0.048220797	-0.014116487	-0.048220797	-0.01411
Tuesday	0.00392335	0.008762843	0.447725652	0.654637176	-0.013313045	0.021159744	-0.013313045	0.02115
Wednesday	0	0	65535	#NUM!	0	0	0	
Thursday	0.005662786	0.008685969	0.651946384	#NUM!	-0.011422397	0.02274797	-0.011422397	0.0227
Friday	0.013086964	0.008679873	1.507736759	0.132553242	-0.003986229	0.030160156	-0.003986229	0.03016
Saturday	-0.080040833	0.008720496	-9.178472172	4.40258E-18	-0.097193931	-0.062887734	-0.097193931	-0.06288
Sunday	-0.109778197	0.008618299	-12.73780277	1.22525E-30	-0.126730275	-0.092826119	-0.126730275	-0.09282
Log(P0)	-9.279724166	7.683458974	-1.207753461	0.227984022	-24.39298394	5.833535608	-24.39298394	5.83353
Log(AvgPrice)	9.384247378	9.071646751	1.034459083	0.301658971	-8.459558959	27.22805372	-8.459558959	27.2280
Log(MinCPrice)	0	0	65535	#NUM!	0	0	0	
Log(MaxCPrice)	-4.447760611	9.658954836	-0.460480527	#NUM!	-23.446794	14.55127278	-23.446794	14.5512

Figure 7. Shows Site-A Model 7 with Log(Vol) as Target variable

F. Summary of Insights

When understanding the multiple linear regression model, the first step is to determine which features in the Fuel price data set should be taken into consideration when calculating the volume of the attribute. The data that was provided was clean enough to allow for the use of multiple regressions.

It can be observed from table 3. that Introduction of dummy variables Monday, Tuesday, Wednesday, Thursday, Friday, Saturday and Sunday helped the model to improve its R-square value and efficiency. The use of the log to the base 10 function for different attributes, as well as increasing the number of variables, all helped to increase the R-square value. From figure 4 and figure 5 it can be observed that as the model has been improvised the residuals are not uneven. A residual is the difference between actual value and the mean value predicted by the model for that observation. Residual values are particularly valuable in regression methods because they represent the degree to which a model explains for the variance in the observed data.

Figure 1 shows that the average volume of normal unleaded gasoline sold at site A, which is 2414.5 gallons, is smaller than the amount sold at site B, which is 4266.5 gallons. Site B sells roughly twice as much unleaded gasoline as Site A, which is a significant difference. Figure 2 illustrates the lowest priced competitor to the site on week days. Friday and Saturday are the most expensive days of the week, as can be noted when compared to the other days of the week. Monday is the day when the prices are the lowest.

From Table 2. It can be observed that no variable is has any positive correlation between Target variable, Volume. A negative relationship exists between all variables. The Significance F in Figure 6 is quite small, as can be seen. In most cases, we define a statistical significance threshold and use it as the cutoff point for assessing the model's performance and effectiveness. It is identical in meaning to the P-value when it comes to significance F. The most significant distinction is that the

significance F is applied to the whole model as a whole, but the P value is applied only to each individual variable in the model.

According to Figure 6, the coefficient P0 = -4002.44 shows that for every unit rise in the P0 variable, the target variable, volume will decrease by the amount indicated by the coefficient 4002.44. Similarly, for MinComPrice if there is increase in 1 unit of price then volume will increase by 2133.64 units. Figure 6: When dealing with a basic linear equation, it is sometimes refers to as the regression line. If the correlation coefficient of the independent variable is negative, for every unit rise in the predictor variables, the dependent variable will drop by the value of the coefficient, if the correlation coefficient is positive the dependent variable will rise by the value of coefficient for every unit drop in the predictor variables, and vice versa for every unit increase. A p-value is a probability measure that indicates the likelihood of getting the observed findings on the assumption that the null hypothesis is correct. The stronger the statistically significant difference of the observed difference, the lower the p-value is calculated to be. A p-value of 0.05 or less is typically regarded as statistically significant in most situations.

Figure 7 illustrates that when the logarithmic function is taken into account, the significance value has decreased (which is a good sign), and the coefficient P0 has changed from negative to positive. The R-square of the models has also been enhanced. An easy method of turning a highly skewed value into something more normalized is to use logarithmic transformations. When modelling variables having non-linear connections, the likelihood of making mistakes might be biased to the negative.

People who make gasoline use different types of crude oil and different types of technology to make it. People who buy gas also have to pay for the cost of other ingredients that can be mixed in with it, such as fuel ethanol. (EIA, 2022)

G. Conclusion

As evidenced by the wide range of data, estimating the influence of petrol prices on demand is challenging. With most statistical analyses, the process of developing a regression model is iterative, which can be started with a simple model (typically with the independent variable that is most relevant to research question), then add additional variables which are expected to be relevant to the issue under investigation, based on theoretical considerations. Based on Table 3, it can be shown that the R-square value grows in direct proportion to the amount of variables being considered. Model 6 is regarded to be the best model for Site-A, and Model 6 is considered to be the best model for Site-A as well, with R-square values of 57.75 percent and 38.17 percent, respectively.
The location of the sites are not known to get a clear scenario if there consists any transportation cost as well. Different sites must also be having some different dealers. It might also include the costs of distribution, marketing, and profit for the retail dealers who sell and service the gas. Most gasoline is sent from refineries by pipeline to terminals near where people use it. (EIA, 2022)

1. References

EIA, 2022. *Gasoline explained Factors affecting gasoline prices.* [Online]
Available at: https://www.eia.gov/energyexplained/gasoline/factors-affecting-gasoline-prices.php
[Accessed 5 May 2022].