Statistical Analysis Examples

by Zarko Cosovic

Table of Contents

Basic Statistics 2

Linear Models 10

ANOVA and ANCOVA 15

Basic Statistics

The following two plots represent randomly generated data points from exponential distribution and normal distribution.

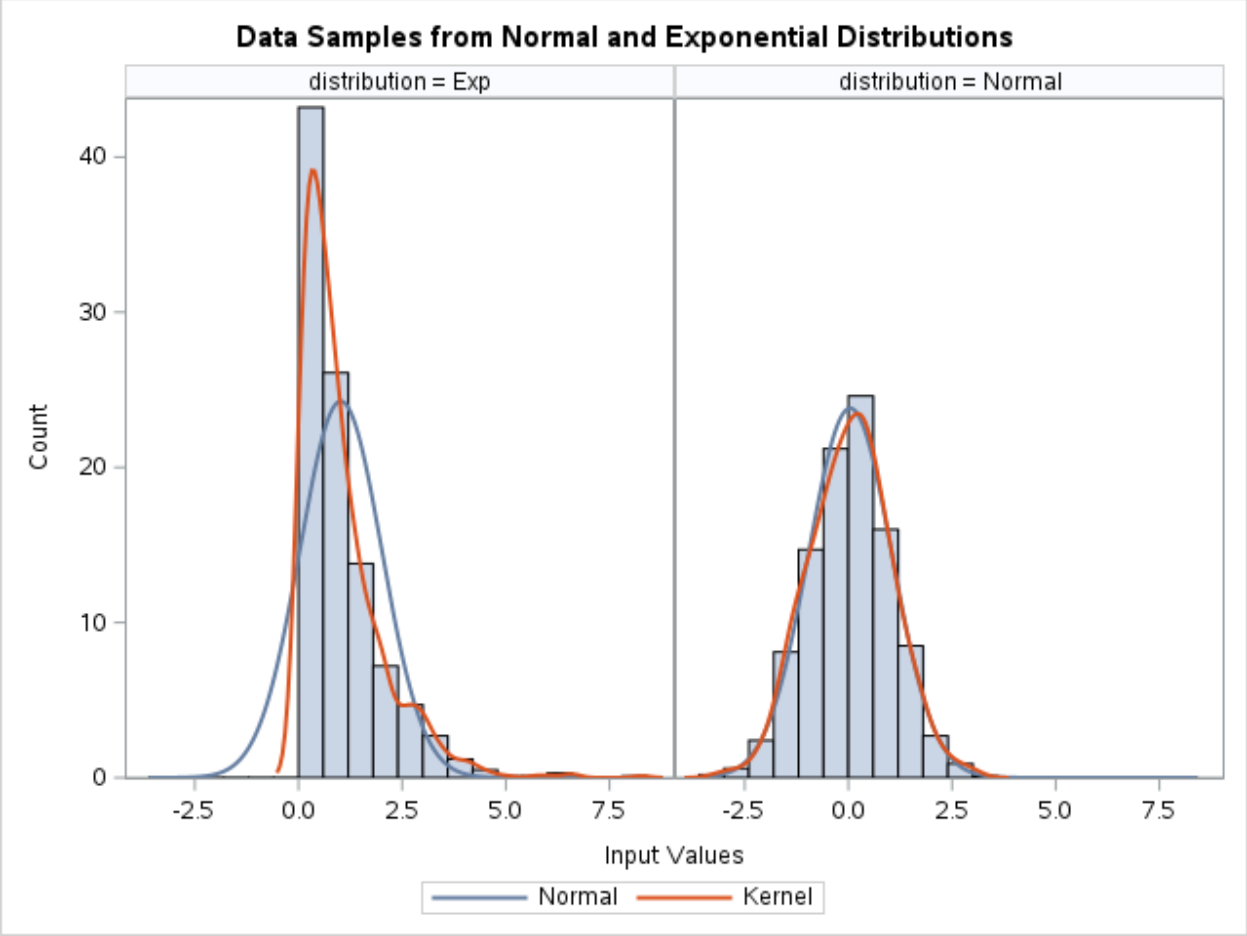

Figure 1. Histograms for data sampled from Exponential and Normal Distributions.

The histogram on the left shows the distribution of the data which was sampled from exponential distribution. We can see that data is skewed to the right and that the normal density curve is not a good approximation of the exponential sample. The histogram on the right shows distribution of the data which was sampled from normal distribution. We can see that the data is explained well using the normal density line, which makes sense since the data was sampled from the normal distribution.

Top of the Document

Next plot represents randomly generated data points from two linear regression models.

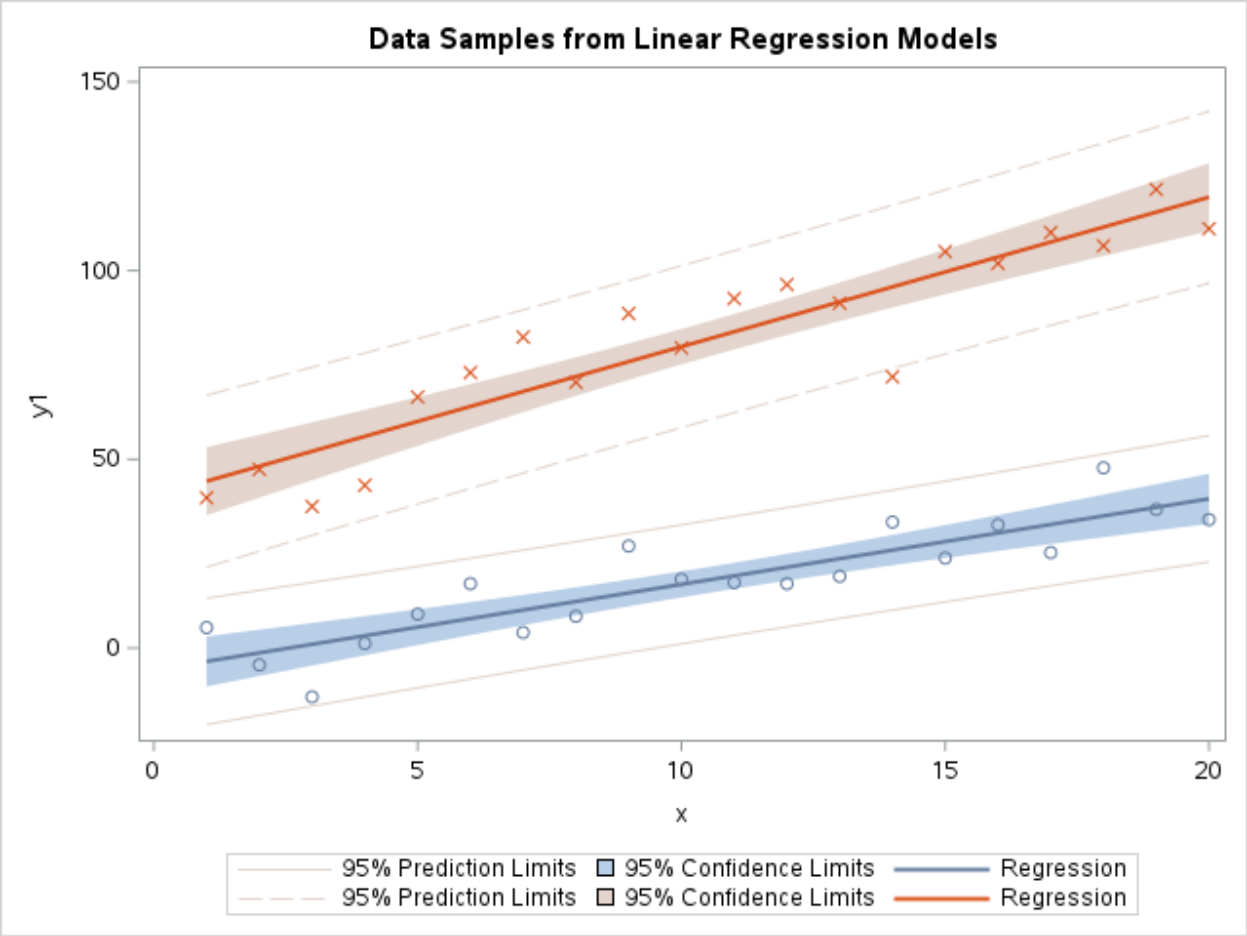

Figure 2. Scatterplot with linear regression lines as well as prediction and confidence limits.

The Scatterplot above is used to analyze the data points which were generated using linear regression models. Lines of best fit laid over the data points outline the trends of the data. The 95% confidence limits highlight the area around the line of best fit. Also outlined in the plot are the 95% prediction limits. Alpha was set to 0.05 for the calculations of the limits.

Top of the Document

This section uses the data from a sample of 1993 model cars from the *1993 Cars Annual Auto Issue* published by *Consumer Reports* and from *Pace New Car and Truck 1993 Buying Guide*.

The following two tables are used to analyze three variables “MPG_City”, “EngineSize”, and “Weight” and the relationships between them.

Pearson Correlation Coefficients, N = 428 Prob > |r| under H0: Rho=0			
	MPG_City	EngineSize	Weight
MPG_City **MPG (City)**	1.00	-0.71 <.0001	-0.74 <.0001
EngineSize **Engine Size (L)**	-0.71 <.0001	1.00	0.81 <.0001
Weight **Weight (LBS)**	-0.74 <.0001	0.81 <.0001	1.00

Table 1. Correlation between Miles Per Gallon in the city, Engine size, and Weight.

From the table above we know that engine size and the weight of a vehicle are positively correlated, while miles per gallon in the city are negatively correlated with both engine size and weight. The very low p-value shows that the findings are significant.

Simple Statistics							
Variable	N	Mean	Std Dev	Median	Minimum	Maximum	Label
MPG_City	428	20.06	5.24	19.00	10.00	60.00	MPG (City)
EngineSize	428	3.20	1.11	3.00	1.30	8.30	Engine Size (L)
Weight	428	3578	758.98	3474.50	1850	7190	Weight (LBS)

Table 2. Summary of statistics for variables MPG City, Engine size, and Weight.

The table above displays basic statistical information about the data, such as number of observations, mean, standard deviation, median as well as minimum and maximum values.

Top of the Document

Next two tables display the frequencies of a vehicle type and the origin of a vehicle, respectively.

Type	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Hybrid	3	0.70	3	0.70
SUV	60	14.02	63	14.72
Sedan	262	61.21	325	75.93
Sports	49	11.45	374	87.38
Truck	24	5.61	398	92.99
Wagon	30	7.01	428	100.00

Table 3. Frequency table for the vehicle types.

The table above shows that sedan is the most common vehicle in the dataset, with over 60% of the vehicles in the dataset being a sedan. Fewest number of cars were hybrid with fewer than 1% of the cars being a hybrid.

Origin	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Asia	158	36.92	158	36.92
Europe	123	28.74	281	65.65
USA	147	34.35	428	100.00

Table 4. Frequency table for the vehicle's country of origin.

The second table shows that most of the vehicles in the dataset come from Asia at nearly 36.92%, followed by USA at 34.35% of the cars, and the fewest vehicles coming from Europe at 28.74%.

Top of the Document

Crossing the previous two tables creates the following contingency table.

Table of Origin by Type							
	Type						
Origin	Hybrid	SUV	Sedan	Sports	Truck	Wagon	Total
Asia	3	25	94	17	8	11	158
	0.70	5.84	21.96	3.97	1.87	2.57	36.92
	1.90	15.82	59.49	10.76	5.06	6.96	
	100.00	41.67	35.88	34.69	33.33	36.67	
Europe	0	10	78	23	0	12	123
	0.00	2.34	18.22	5.37	0.00	2.80	28.74
	0.00	8.13	63.41	18.70	0.00	9.76	
	0.00	16.67	29.77	46.94	0.00	40.00	
USA	0	25	90	9	16	7	147
	0.00	5.84	21.03	2.10	3.74	1.64	34.35
	0.00	17.01	61.22	6.12	10.88	4.76	
	0.00	41.67	34.35	18.37	66.67	23.33	
Total	3	60	262	49	24	30	428
	0.70	14.02	61.21	11.45	5.61	7.01	100.00

Table 5. Contingency table for origin and type of vehicle.

The table above allows us to see the frequency and percentage of the origin of each type of vehicle in the dataset. For example, we can see that all the hybrid vehicles came from Asia, half the trucks came from the USA, and most sports cars came from Europe. The bottom row and the leftmost column display the frequency and percentage information from the previous two tables.

Top of the Document

The following section uses the data from a random sample of three hundred houses sold in Ames, IA during the 2006-2010 period.

First, we analyze the sale price using tables and plots.

Basic Statistical Measures			
Location		**Variability**	
Mean	137524.9	**Std Deviation**	37623
Median	135000.0	**Variance**	1415463276
Mode	110000.0	**Range**	255000
		Interquartile Range	45475

Table 6. Summary of statistics for Sale Price.

Mean price of a house in this dataset is $137,524.90 and the standard deviation is $37,623. The price difference between the cheapest and most expensive house in the sample is $255000.

Quantiles	
Level	**Quantile**
Max	290000
90%	187300
75% Q3	159475
50% Median	135000
25% Q1	114000
10%	91150
Min	35000

Table 7. Quantile values for the Sale Price.

Only 10% of the houses are below $91150 and 10% of the houses are above $187300.

Top of the Document

To test whether the data points are normally distributed we use the Q-Q Plot.

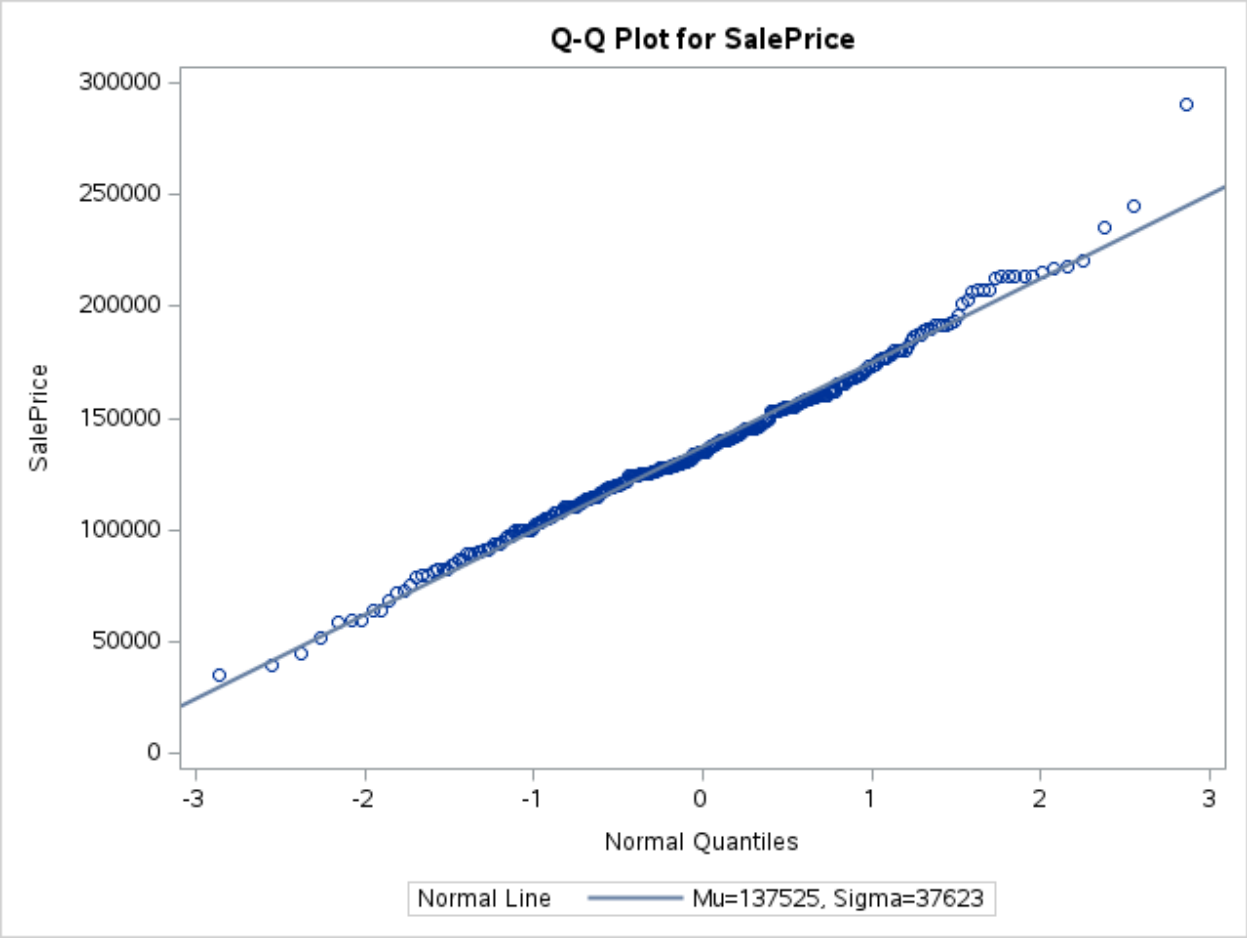

Figure 3. Q-Q Plot test of normality for the Sale Price data.

Using a Q-Q plot we can see that most Sale Price data fall on the Normal line, with minor discrepancies towards the end. In this case we conclude that the data is normally distributed.

Top of the Document

Using a plot to view the distribution of the data.

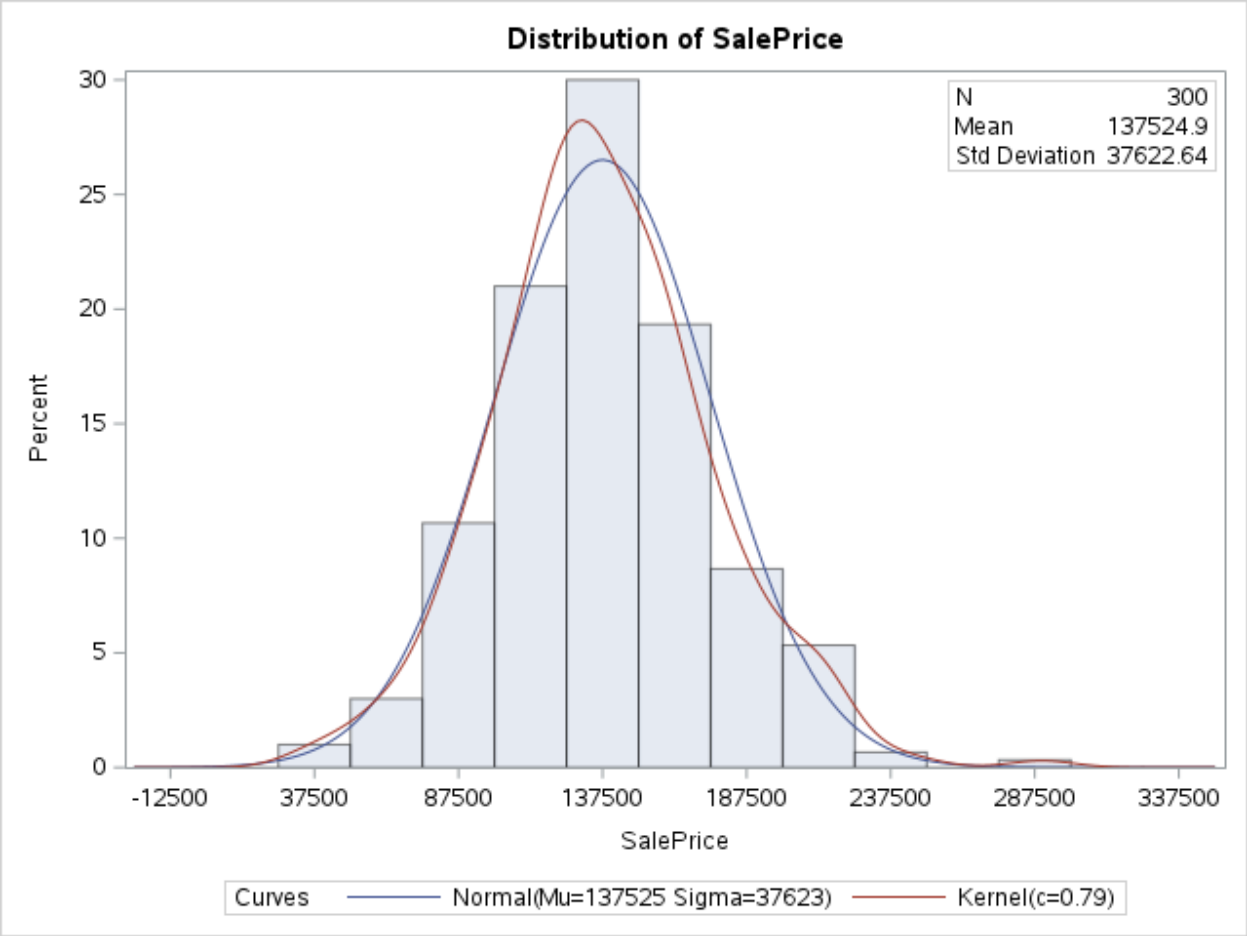

Figure 4. Histogram for Sale Price data outlined with normal and kernel density curves.

Histogram shows the sale price data appears normally distributed as data are fairly symmetric around the mean price of $137524.9. We can also see that the normal density curve approximates the sale price data well and is closely aligned with the kernel density line.

Top of the Document

Linear Models

Next section presents data modeling results for a random sample of three hundred houses sold in Ames, IA during the 2006-2010 period. The goal is finding out which features have the most effect on the sale price of a house.

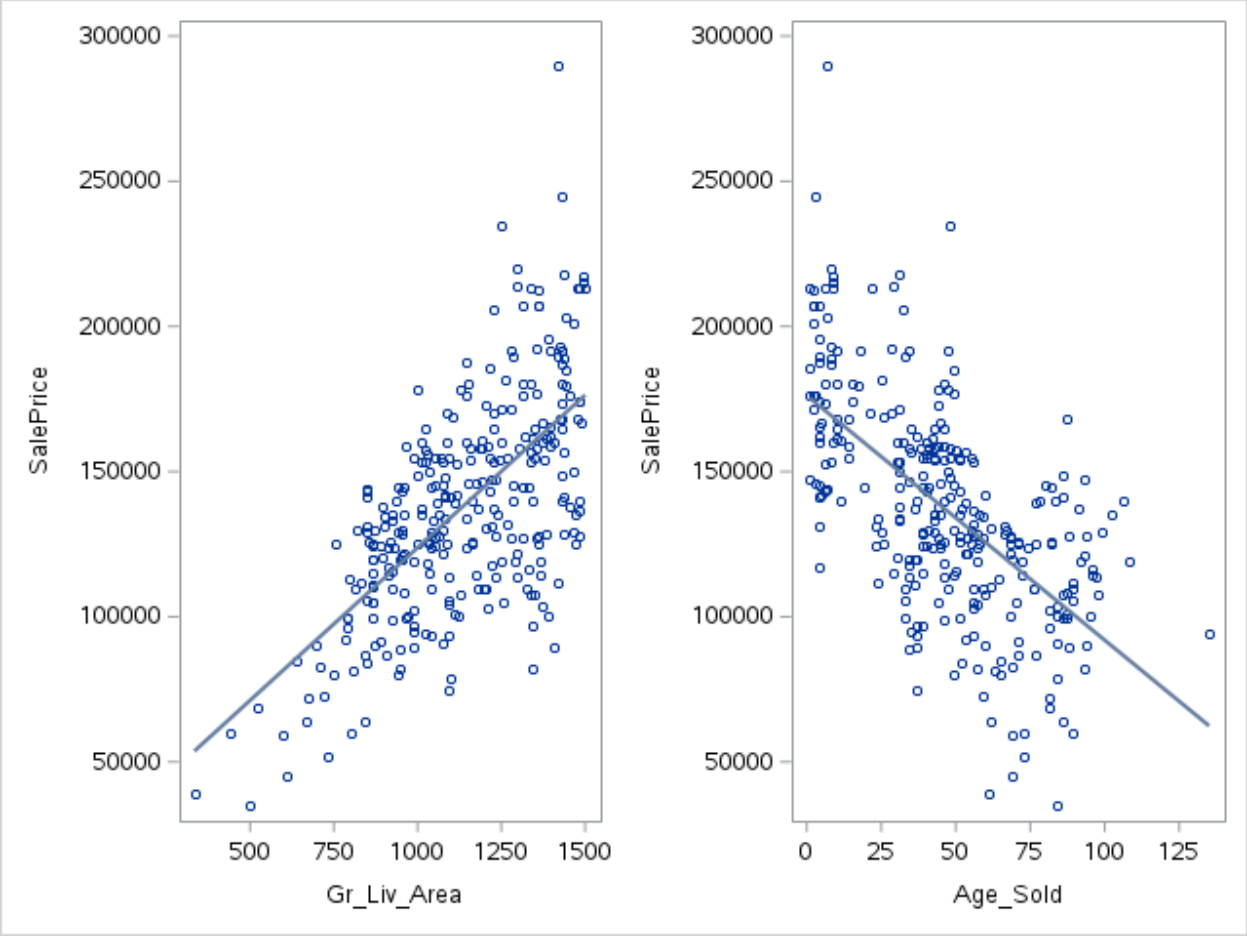

Figure 5. Scatterplot of ground living area and age of the house with sale price as output and a regression line.

Scatterplot on the left shows that there is a positive linear association between above-ground living area and the sale price of the house, which means that increase in living area leads to increase in price. The right scatterplot shows a negative linear association between the age of the house and the sale price, which makes sense since houses usually lose value with increase in age.

Top of the Document

Creating a model for the output sale price using features ground living area and age of the house.

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	68498	7031.03707	9.74	<.0001
Gr_Liv_Area	1	89.32657	5.47233	16.32	<.0001
Age_Sold	1	-696.90431	46.33458	-15.04	<.0001

Table 8. Hypothesis testing for influence of living area and age of house on the sale price.

P-value for the t-test of our model shows that both features are statistically significant in explaining the change in the sale price. The table also shows the parameter estimates for the features of our model.

Root MSE	21602	R-Square	0.6725
Dependent Mean	137525	Adj R-Sq	0.6703
Coeff Var	15.70759		

Table 9. Table showing the model's relevance.

Looking at the R-Square number we can see that the model can be used to explain 67% of the sale price. Considering that the dataset has 34 features, our model is quite good since it uses only two features to explain two thirds of the sale price.

Top of the Document

Testing the normality of residuals for the model using a histogram.

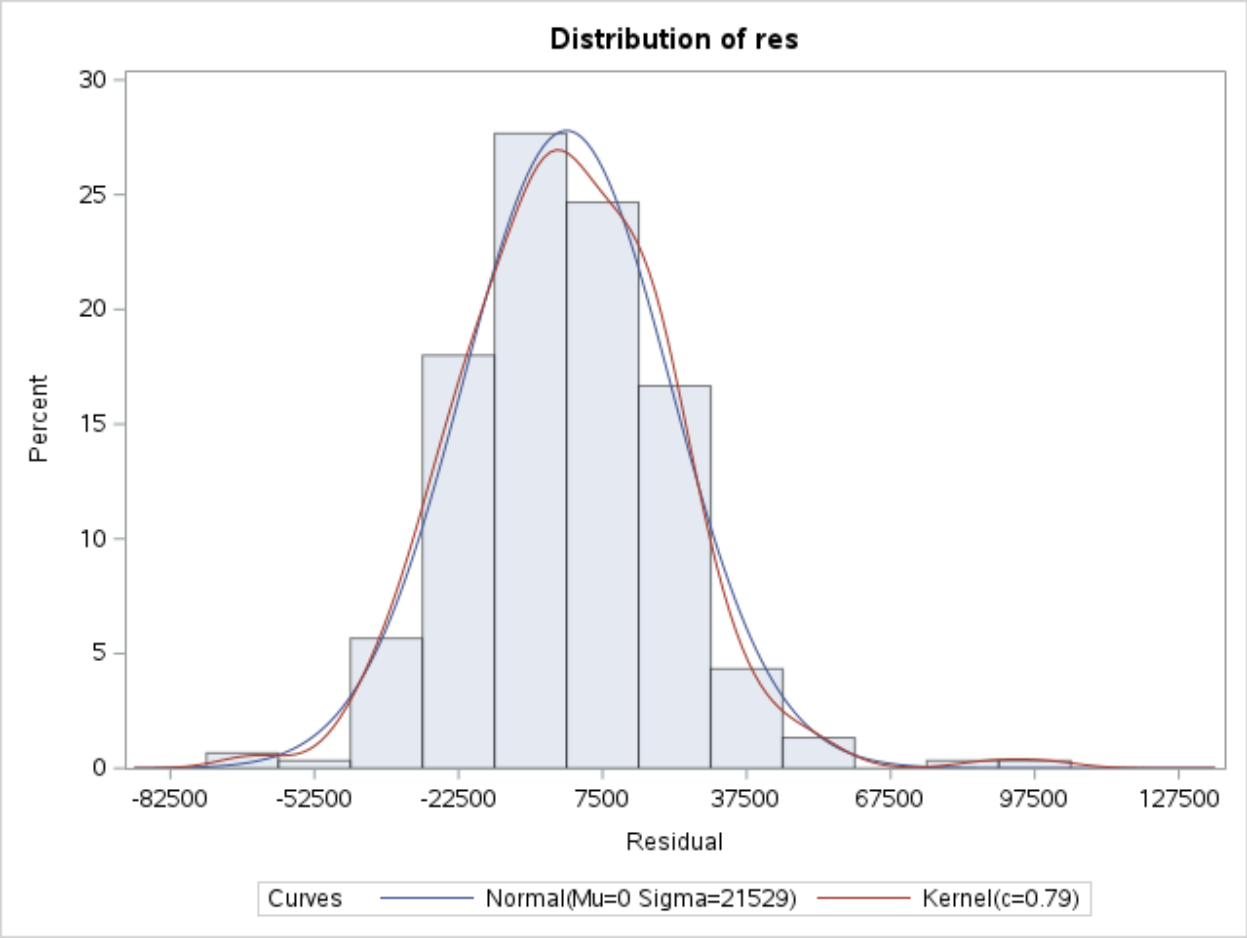

Figure 6. Histogram plot of the model residuals.

Plotting the residuals using a histogram to view their distribution. Residual data is close to being normally distributed, with normal and kernel density curves nearly the same.

Top of the Document

Testing the normality of residuals for the model using a Q-Q plot.

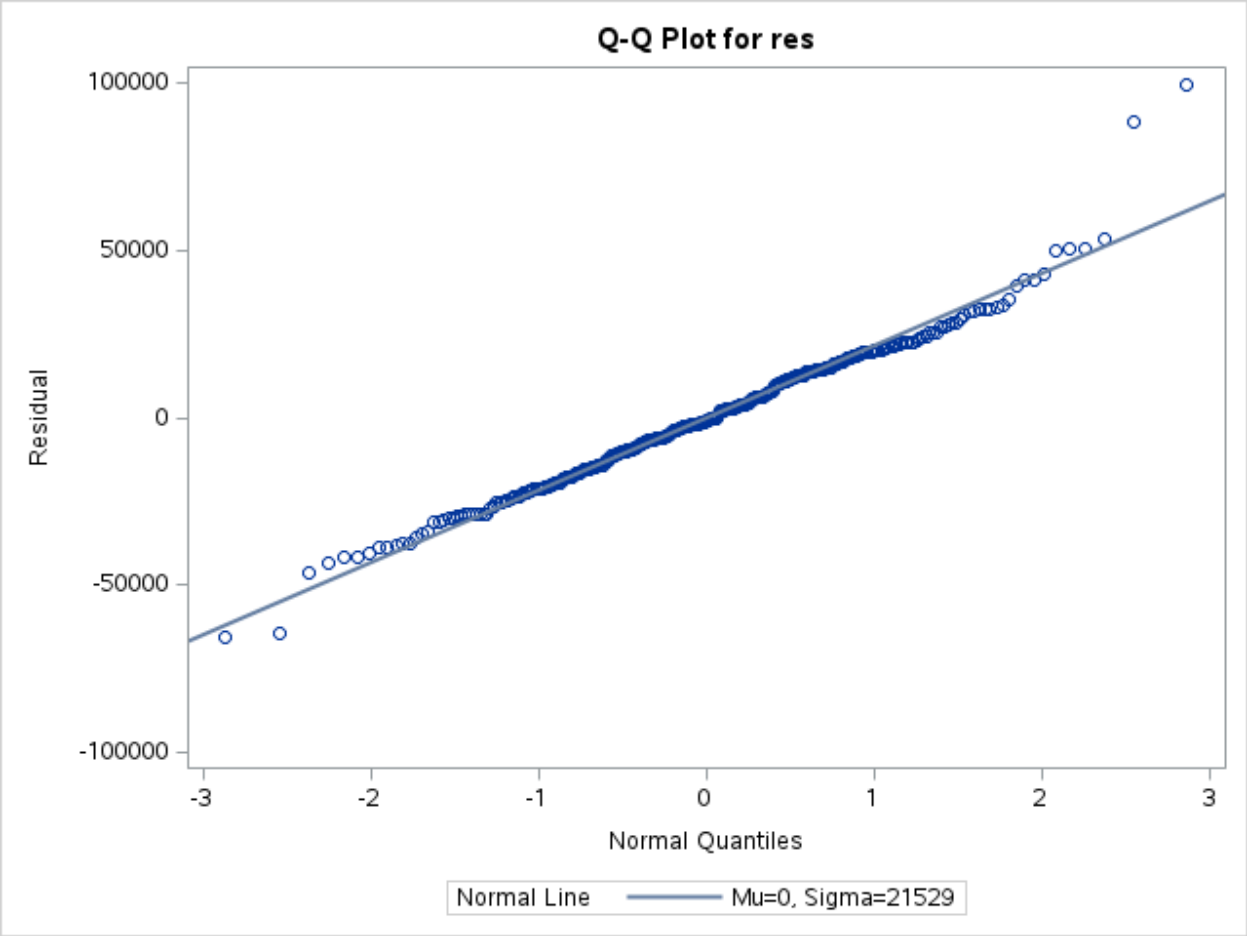

Figure 7. Q-Q Plot of residuals is a test of normality.

Q-Q plot shows that residual data fall on the normal line, with exception of a few points at the ends. We can conclude that the residuals come from a normal distribution.

Top of the Document

Analyzing the predictions made by the model.

Obs	SalePrice	Gr_Liv_Area	Age_Sold	Predicted
1	189000	1434	8	191017.42
2	162000	1395	37	167323.46
3	161500	1342	42	159104.63
4	82500	708	69	83655.17
5	161500	1342	42	159104.63

Table 10. Comparison of actual sale price to the predicted sale price.

This model makes good predictions of real-world data, with guesses usually within $2000 of the actual price. Adding more features would improve model's predictions, and making predictions which are even closer to the actual price.

Top of the Document

ANOVA and ANCOVA

This section uses the data from a sample of 1993 model cars from the *1993 Cars Annual Auto Issue* published by *Consumer Reports* and from *Pace New Car and Truck 1993 Buying Guide*.

Analysis of variance (ANOVA) is used to compare miles per gallon on the highway for different types of vehicles.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	6743.47900	1348.69580	77.64	<.0001
Error	422	7331.03268	17.37212		
Corrected Total	427	14074.51168			

Table 11. The model's dependent variable is MPG (Highway) and independent variable is vehicle type.

The table above shows that the ANOVA model is statistically significant. The hypothesis testing returned an F-value of 77.64, and a very low p-value of less than 0.0001. In conclusion, the vehicle type is a relevant factor in determining the vehicle's miles per gallon.

R-Square	Coeff Var	Root MSE	MPG_Highway Mean
0.479127	15.52701	4.167987	26.84346

Table 12. Table showing the model's relevance.

R-Square value states that we can explain 48% of the miles per gallon using this model, which isn't great. Adding more features would improve the model's relevance in explaining a vehicle's miles per gallon.

Top of the Document

Plotting the miles per gallon (highway) data by the vehicle type.

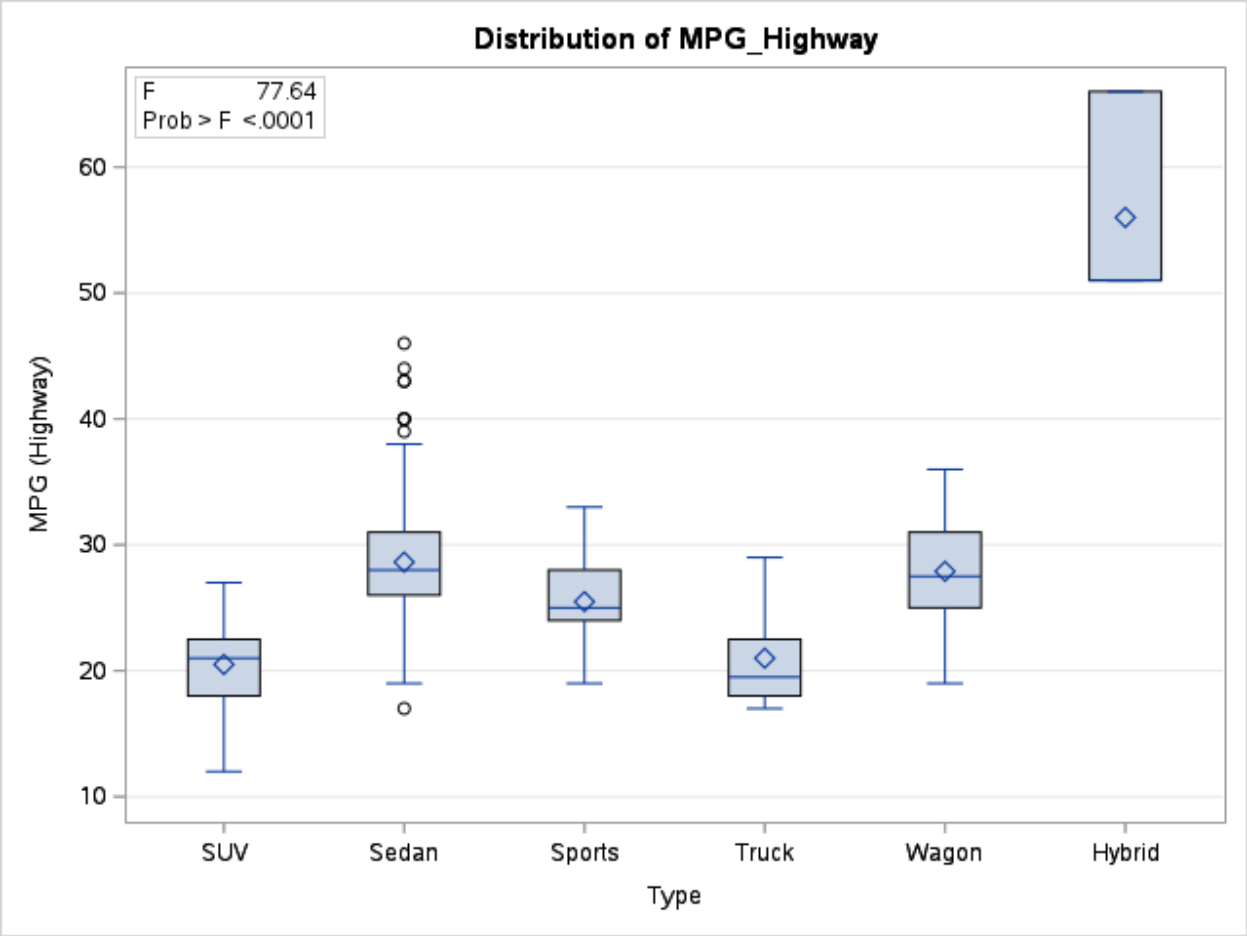

Figure 8. Boxplot of MPG Highway by the type.

Boxplot shows that Hybrid type cars have by far the best miles per gallon on the highway. All other cars are within 20-30 miles per gallon range, except for trucks which give slightly less than that range.

Top of the Document

Least squares means analysis for miles per gallon (highway) per car type.

Type	MPG_Highway LSMEAN	LSMEAN Number
SUV	20.5000000	1
Sedan	28.6297710	2
Sports	25.4897959	3
Truck	21.0000000	4
Wagon	27.9000000	5
Hybrid	56.0000000	6

Table 13. Least Squares Means Adjustment for Multiple Comparisons: Tukey-Kramer

Least squares means model shows that hybrid type cars are well above other vehicle types when it comes to miles per gallon, with trucks giving the fewest miles per gallon.

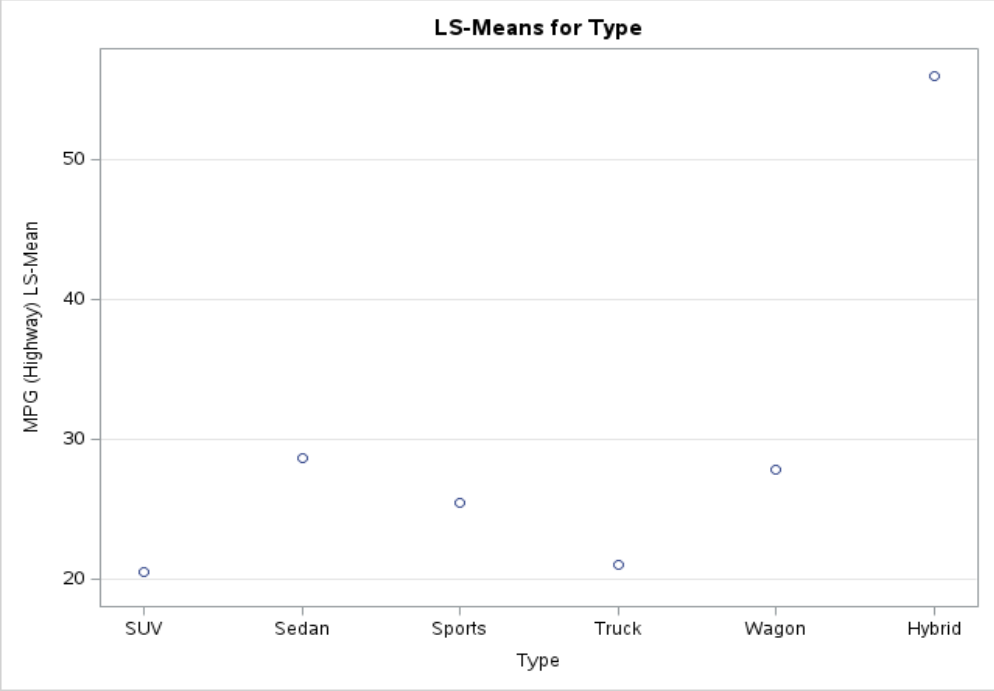

Figure 9. LS-Means plot of mpg (highway) data for each car type.

Plotting the least square data visually shows that hybrid vehicle's miles per gallon (highway) is much higher than each other car type.

Top of the Document

Analysis of covariance (ANCOVA) is used to compare miles per gallon on the highway for different types of vehicles and different horsepower.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	10765.94182	978.72198	123.06	<.0001
Error	416	3308.56986	7.95329		
Corrected Total	427	14074.51168			

Table 14. The model's dependent variable is MPG (Highway) and independent variables are vehicle type and horsepower.

The table above shows the findings of the hypothesis test with an F-value of 123.06 and a p-value less than 0.0001. We conclude that the model for interaction between vehicle type and horsepower is statistically significant.

R-Square	Coeff Var	Root MSE	MPG_Highway Mean
0.764925	10.50594	2.820158	26.84346

Table 15. Table showing the model's relevance analysis.

R-Square value shows that the ANCOVA model is relevant as it can explain 76% of the results.

Top of the Document

Using a scatter plot with regression lines for ANCOVA analysis.

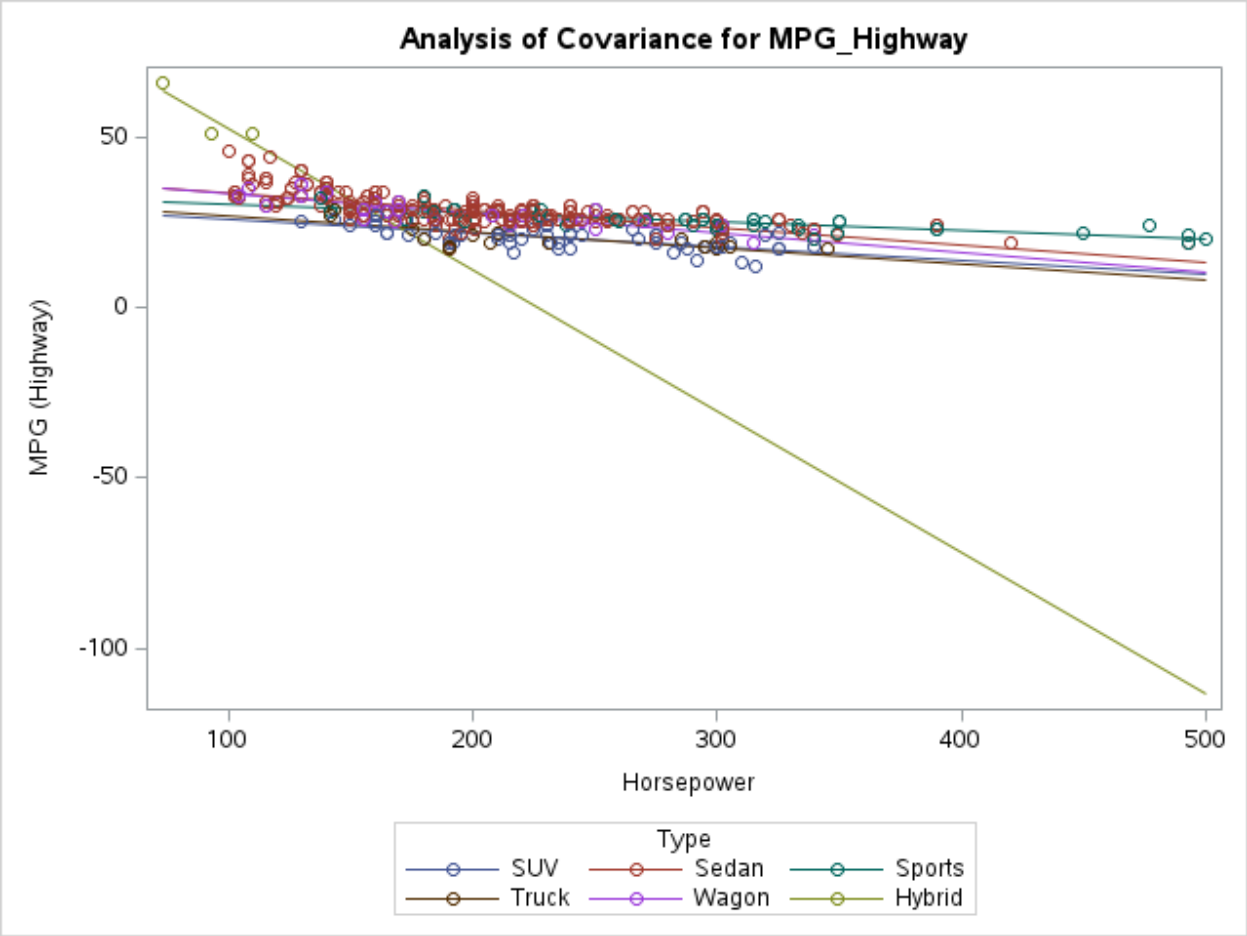

Figure 10. Plot of covariance of horsepower and car type.

The plot above shows that all data seem to follow linear paths, and that all lines are close to being parallel except for the hybrid line which has a significantly different slope. In conclusion, there is interaction between features.

Top of the Document