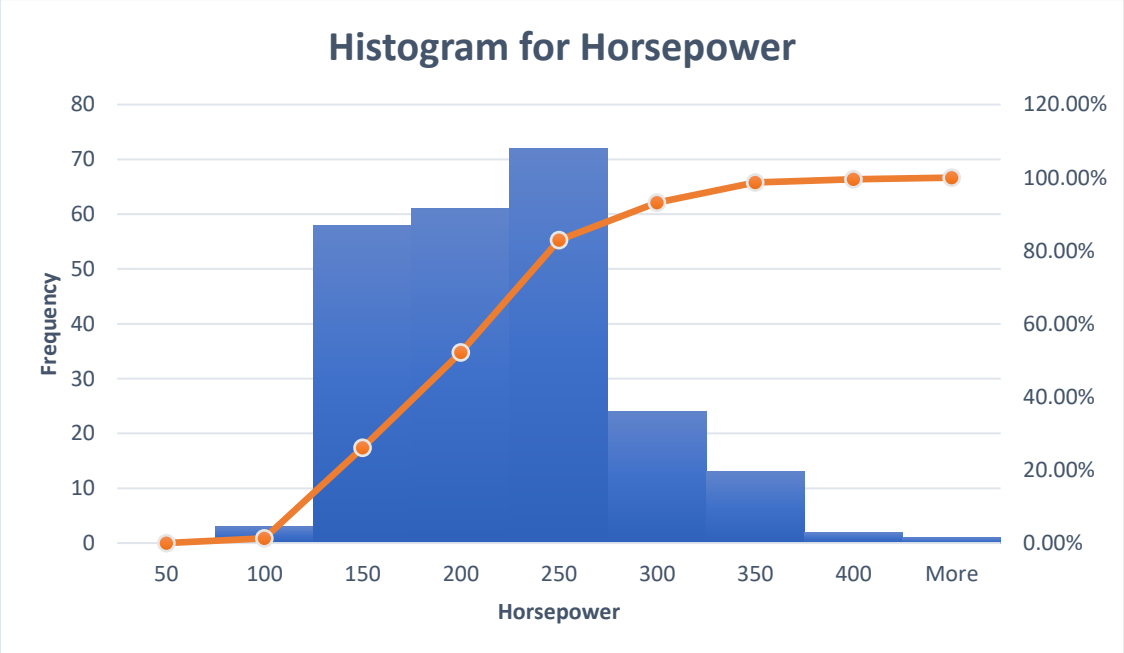
Project 6: Horsepower and Retail Price

1. Introduction

Most households in the US own at least one car as transportation is an important part of American daily life. There are several factors that affect care retail price, including engine size, number of cylinders, horsepower, city mpg, highway mpg, weight, length, and width. Here, we will focus on how horsepower affects the retail price. We will make the assumption that the higher the horsepower, the higher the retail price, so that on average, a higher horsepower vehicle will cost more. My reasoning is that people are willing to pay a higher price for a vehicle with more power.

2. Describing the Data:

Both the horsepower and retail price histogram are both right-skewed because the means are to the right of the medians. There are higher frequencies for lower horsepower and retail price. Below is shown the histograms for horsepower and retail price, with cumulative percentage shown with the orange line.



3. Empirical Rule

The x distribution satisfies the empirical rule within 1 standard deviation as it exceeds 68% at 68.80%. However, the y distribution doesn't satisfy the empirical rule within 1 standard deviation as it is less than 68% at 79.06%. Both the x and y distributions satisfy the empirical rule within 2 standard deviations because at least 95% of the data falls within 2 standard deviations of the mean. 96.58% of the x distribution and 95.73% of the y distribution falls within 2 standard deviations. However, the x and y distributions don't satisfy the empirical rule within 3 standard deviations because less than 99.7% of the data is within 3 standard deviations. 99.57% of the x distribution and 98.72% of the y distribution falls within 3 standard deviations.

Range	Min	Max	Freq	True Percentage	Satisfy Empirical Rule?
X distribution (Horsepower)					
Within 1 Sd	135.76	263.83	161/234	68.80%	Yes
Within 2 Sd	71.73	327.87	226/234	96.58%	Yes
Within 3 Sd	7.70	391.90	233/234	99.57%	No
Y distribution (Retail Price)					
Within 1 Sd	13871.79	45641.93	185/234	79.06%	No
Within 2 Sd	-2013.29	61527.00	224/234	95.73%	Yes
Within 3 Sd	-17898.4	77412.08	231/234	98.72%	No

This is a table summarizing the true percentages and whether they satisfy the empirical rule for each range of standard deviations.

4. Outliers

The outliers for the x distribution are those outside of 2 standard deviations, which is under 72 and above 327. The 8 outliers are: 330, 340, 340, 340, 390, 390, 349, and 493. The outliers for the y distribution are under -2,012 and above 61,527. The 10 outliers are: 69190, 69195, 73195, 63120, 68995, 74995, 94820, 128420, 74320, and 86970. Outliers that are within 2 and 3 standard deviations are considered normal outliers and those that are outside 3 standard deviations are considered extreme outliers.

	X (Horsepower)	Y (Retail Price)
Normal Outliers	330, 340, 340, 340, 390, 390, 349	69190, 69195, 73195, 63120, 68995, 74995, 74320
Extreme Outliers	493	94820, 128420, 86970

5. Descriptive Statistics

	x	**z-scores**	**y**	**z-scores**
Mean	199.7991	0	29756.86	0
Median	200	0.003137	26007.5	-0.23603
Mode	200	0.003137	13270	-1.03788
Standard Deviation	64.03424	NA	15885.07	NA
Min	73	-1.98018	10280	-1.22611
25 percentile	150	-0.7777	19161.25	-0.66702
75 percentile	232	0.502869	36831.25	0.445348
Max	493	4.578813	128420	6.21106

6. Simple Linear Regression Output and Plots:

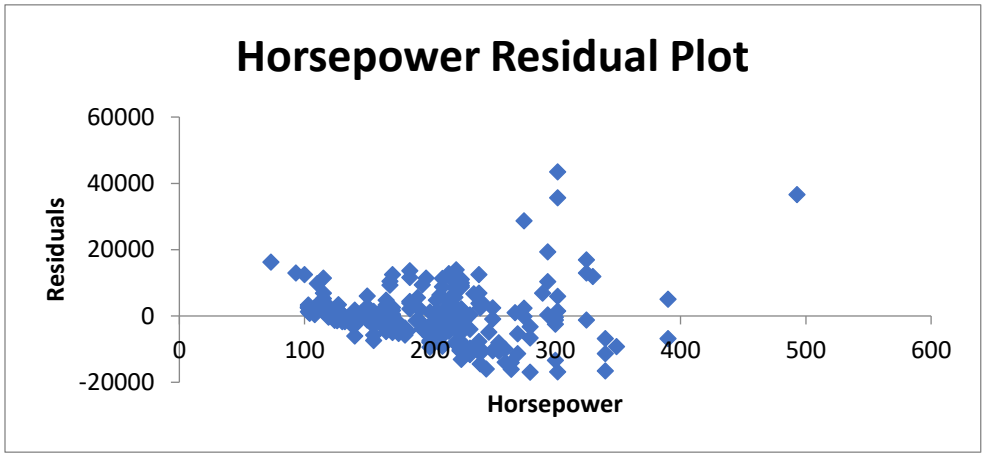
SUMMARY OUTPUT

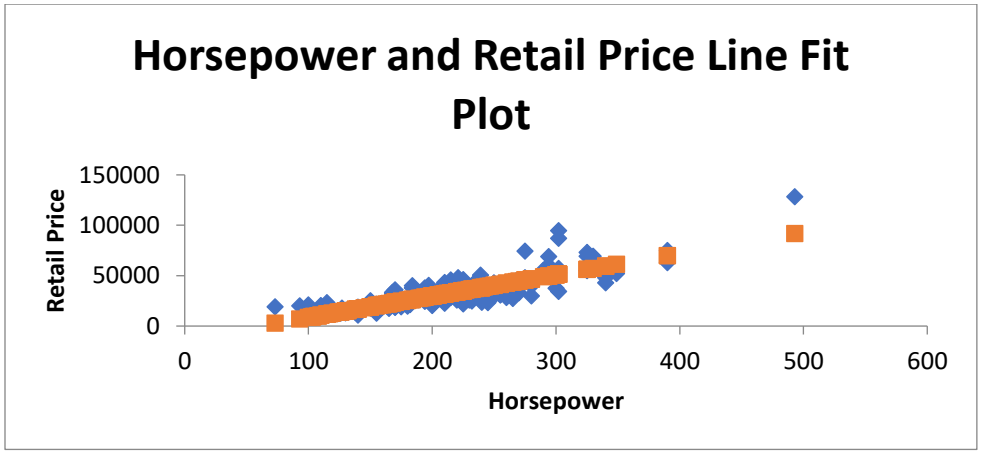
Regression Statistics	
Multiple R	0.853122427
R Square	0.727817876
Adjusted R Square	0.726644677
Standard Error	8305.254852
Observations	234

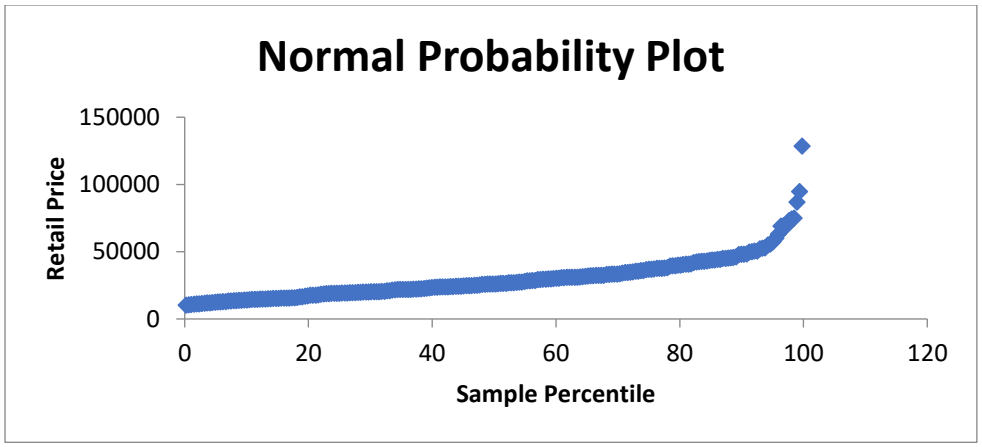
ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	4.28E+10	4.28E+10	620.3705	1.70114E-67
Residual	232	1.6E+10	68977258		
Total	233	5.88E+10			

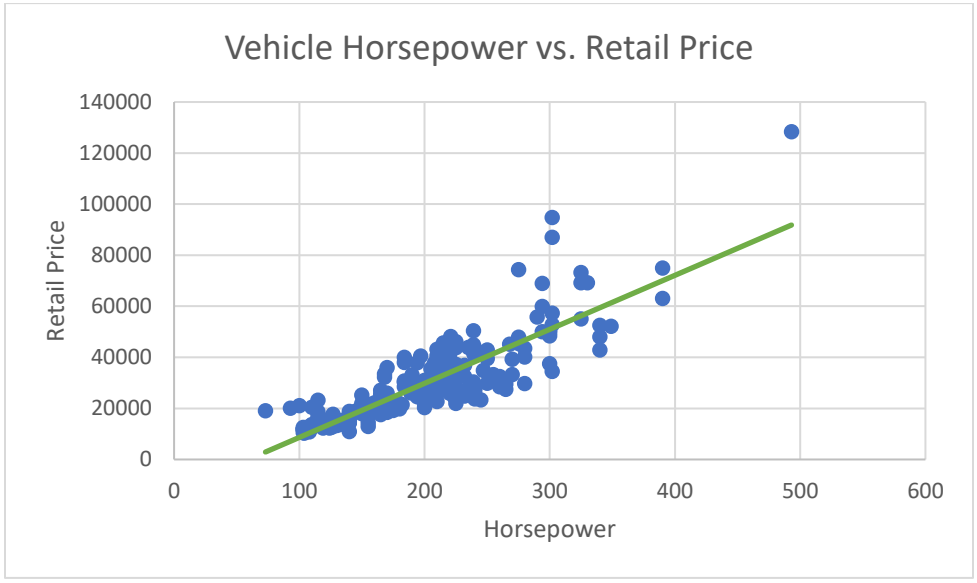
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-12527.70882	1782.386	-7.02862	2.31E-11	-16039.4	-9015.98	-16039.4	-9015.98
X Variable 1	211.6353788	8.496943	24.90724	1.7E-67	194.8943	228.3764	194.8943	228.3764





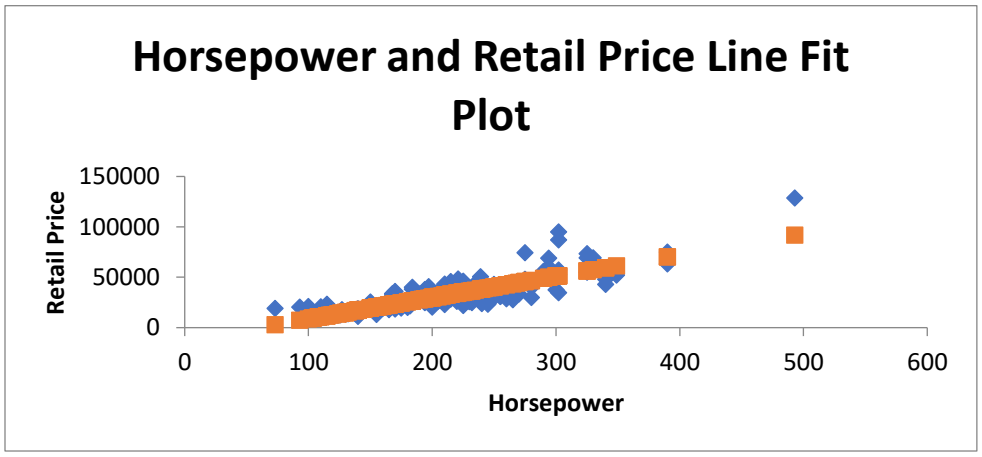


7. Scatterplot



Based on the scatterplot, there appears to be a moderately strong correlation between the horsepower and retail price. As horsepower increases, the retail price generally increases as well.

8. Line Fit Plot



There is a linear relationship between the 2 variables because most of the points fall near and around the line. Based on the descriptive statistics, there are only a few outliers, and it looks like the actual y is close to the predicted y for most x values.

9. Is the Regression Model Significant?

The regression model is significant or important because the Significant F is smaller than 0.05 at 1.70114E-67.

10. Are All Parameters Significant?

All the parameters are significant because the p-value for both parameters are way smaller than 0.05. The intercept p-value is 2.31E-11 and the x variable p-value is 1.7E-67.

11. Mathematical Equation of the Model

The equation will be $Y = 212x - 12528$

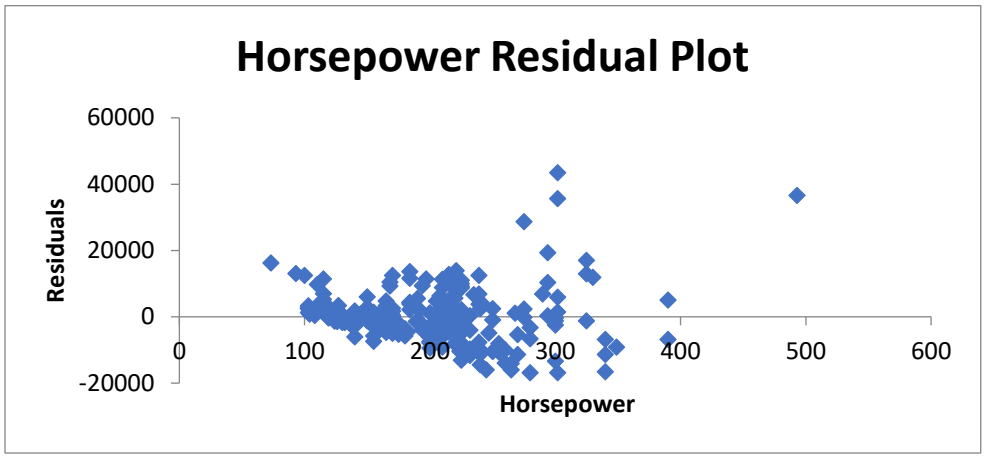
Example 1: If horsepower = 100, then retail price can be estimated to be $8672.

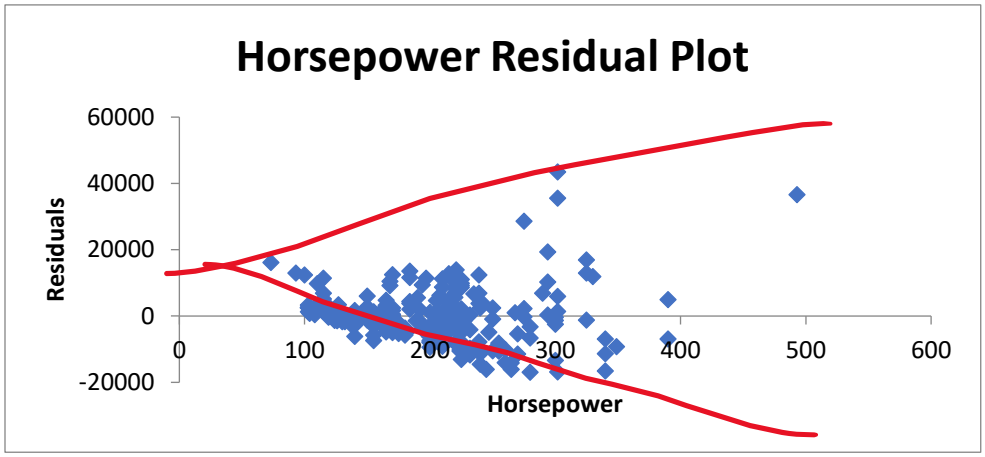
Example 2: If horsepower = 1500, then retail price can be estimated to be $305,472.

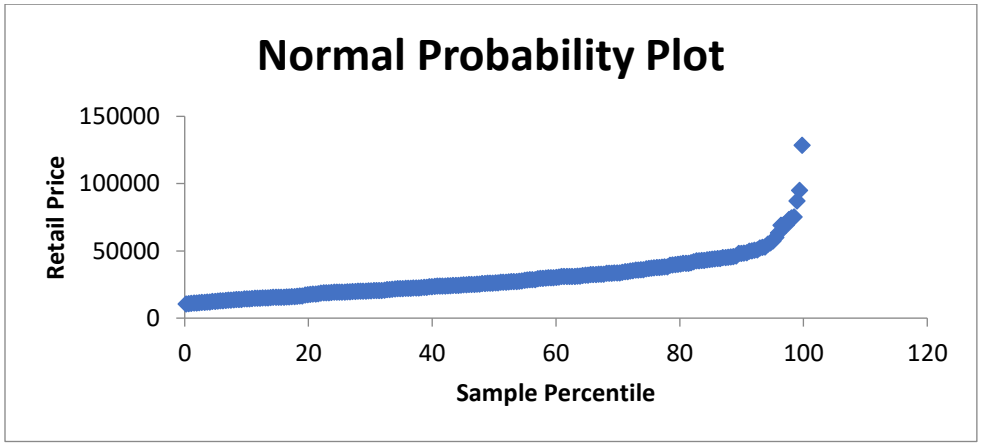
12. Is This Model a Reliable Predictor of Y?

This model is a reliable predictor of y because it has a strong correlation where multiple R is 0.85 and 73.78% of variation is explained in the model. These 2 numbers prove that in general, the retail price is strongly correlated with horsepower.

13. Assumption Checks







It does not satisfy the first assumption of a mean of 0 because the residual plot is not symmetric to 0 since it looks like it is heavier above 0 than it is below 0. It also doesn't satisfy the second assumption of constant variance because it cannot fit 2 parallel lines with equal distance to 0. The lines would intersect somewhere to the left of the plot since the points are more spread out on the right side. The third assumption would be satisfied because there is no clear pattern in the residual plot. For the fourth

assumption, the normal distribution is a curve not a straight line, so this assumption is not satisfied.

14. Summary

In this project, I analyzed the data from the provided Dataset 7 on Car Retail Price, using horsepower as the independent variable and retail price as the dependent variable. I found that horsepower and retail price are strongly correlated where as horsepower increases, retail price also increases, which can be expressed with the mathematical equation $Y = 212x - 12528$. I also created histograms and analyzed the descriptive statistics for both variables, along with testing the empirical rule and outliers. I also took a look at the simple linear regression plots and outputs and a scatterplot, which demonstrated a strong correlation between the two variables. The simple linear regression plots and outputs demonstrated that the regression model and all parameters are significant, and that this model is a reliable predictor of retail price. For example, multiple R is 0.85 and 73.78% of variation showing a high degree of correlation. Finally, I analyzed the residual plot and normal probability plot to check if they satisfied the four assumptions.

From this project, I learned how to prove that a linear regression model is significant or important, and how to use the model to determine the correlation between two variables. I also learned several new functions in Excel, such as implementing Data Analysis and building histograms. Something I would want to improve on in this project is looking further into the residual plot and why the plots are helpful in proving if it satisfies the assumptions.