INDIVIDUAL PROJECT

Data Options: Dataset 7 – Cars Retail Price.
Name: Richard Veloz Salazar
BA 240
02/11/2024

Comparing Retail Price to Highway MPG.

1. **INTRODUCTION:**
   Highway MPG (miles per gallon) is a crucial metric used to evaluate the fuel efficiency of vehicles, representing the distance a car can travel on a gallon of fuel under highway driving conditions. This metric is significant not only for environmental reasons but also for economic considerations, as vehicles with higher MPG ratings typically save drivers money on fuel costs over time.

   In the context of data analysis, exploring the relationship between highway MPG and the price of a vehicle can provide valuable insights for consumers, manufacturers, and policymakers alike. Understanding how the price of a vehicle correlates with its fuel efficiency can inform decisions regarding vehicle purchases, environmental regulations, and technological advancements in the automotive industry.

   Before delving into the data analysis, it's essential to make some assumptions:

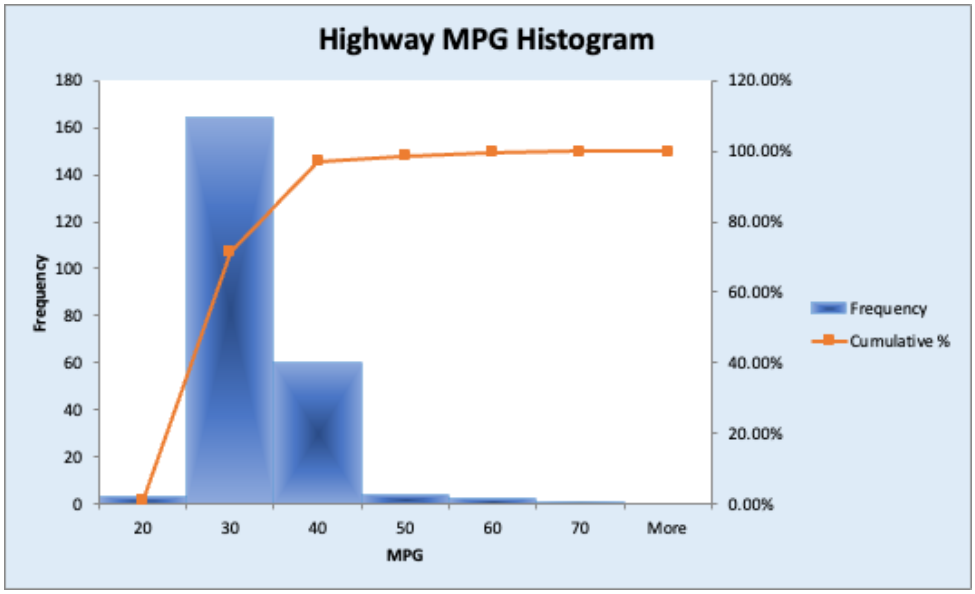
   - Higher Price, Higher MPG Expectation:
     Generally, there's an expectation that vehicles with higher price tags would offer better fuel efficiency due to potentially more advanced technology and engineering. However, luxury or performance vehicles might defy this trend, prioritizing power over efficiency.

   - Variability Across Vehicle Types:
     Different types of vehicles (e.g., sedans, SUVs, trucks) may exhibit different MPG-price relationships. For instance, SUVs and trucks, often associated with larger engines and heavier builds, might have lower MPG compared to compact cars, even within similar price ranges.

   - Technology Advancements Impact:
     Recent technological advancements, such as hybridization and electric propulsion, may disrupt traditional MPG-price relationships. Electric vehicles (EVs) can offer high MPG equivalents, but their initial purchase prices can be higher due to battery costs.

   By considering these assumptions, we can approach the data analysis with a nuanced perspective, anticipating potential trends and outliers based on the interplay between vehicle price and highway MPG.

2. DESCRIBE THE DATA:



- **Y-Variable Histogram: (Retail Prices)**
  Graph is skewed to the right.
- **X-Variable: (MPG)**
  Graph is skewed to the right.

3. Analyze whether the x and y distributions satisfy the empirical rule (Yes or No, explain why). **Show details such like the range of within 1 standard deviation, within 2 standard deviation and within 3 standard deviation and the corresponding true percentage falling in these ranges.**

VARIABLE X:

-Within 1 Standard Deviation? The range within 1 Standard Deviation is 185 and the total value is 234 so 185/234 = 0.7906. Yes, it satisfies the Empirical Rule because 79.06% is bigger than 68%

-Within 2 Standard Deviation? The range within 2 Standard Deviation is 224, so 224/234=0.957. Yes, it satisfies the Empirical Rule because 95.70% is bigger than 95%.

-Within 3 Standard Deviation? The range within 3 Standard Deviation is 230, so 230/234=0.9829. No, it does not satisfy the Empirical Rule because 98.29% is less than 99.70%.

VARIABLE Y:

-Within 1 Standard Deviation? The range within 1 Standard Deviation is 185 and the total value is 234 so 185/234 = 0.7906. Yes, it satisfies the Empirical Rule because 79.06% is bigger than 68%

-Within 2 Standard Deviation? The range within 2 Standard Deviation is 224, so 224/234=0.957. Yes, it satisfies the Empirical Rule because 95.70% is bigger than 95%.

-Within 3 Standard Deviation? The range within 3 Standard Deviation is 231, so 231/234=0.9872. No, it does not satisfy the Empirical Rule because 98.72% is less than 99.70%.

4. Identify and **list** all outliers in each distribution (**Both X and Y**) using appropriate methodology and explain why they are outliers. If you have more than 10 outliers in either distribution (X or Y) in your dataset, you can just list out the top 10 outliers.

All outliers were chosen because they have a Z-score higher than 2 or less than -2. Outliers for the x distribution include:
43, 43, 44, 46, 51, 51, 66

Outliers for the Y Distribution include:
63120, 68995, 69190, 69195, 73195, 74320, 74995, 86970, 94820, 128420.

5. Calculate the mean, median, and mode. Finish the following table for the five number summary (Minimum, Q1, median, Q3, maximum) and the z-scores of each.

	x	z-scores	**y**	z-scores
Mean	29.3974359	0	29756.85897	0
Median	28	-0.2606	26007.5	-0.236
Mode	26	-0.6329	13270	-1.0378
Standard Deviation	5.372013977	NA	15885.0723	NA
Min	19	-1.935	10280	-1.226
25 percentile	26	-0.632	19161.25	-0.667
75 percentile	31	0.297	36831.25	0.445
Max	66	6.813	128420	6.211

6. **The Regression**: Show the output and all the plots from Excel from Simple Linear Regression analysis. You can copy and paste from Excel output and plots.

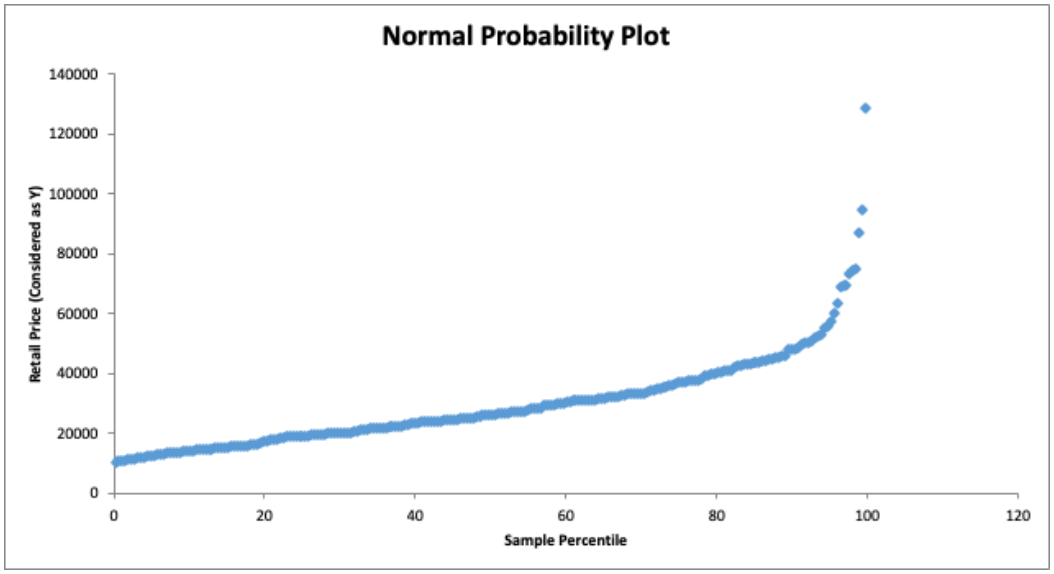
SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.991492066
R Square	0.983056518
Adjusted R Squa	0.982983485
Standard Error	2072.165794
Observations	234

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	57797998547	57797998547	13460.58079	1.9352E-207
Residual	232	996178089.8	4293871.077		
Total	233	58794176636			

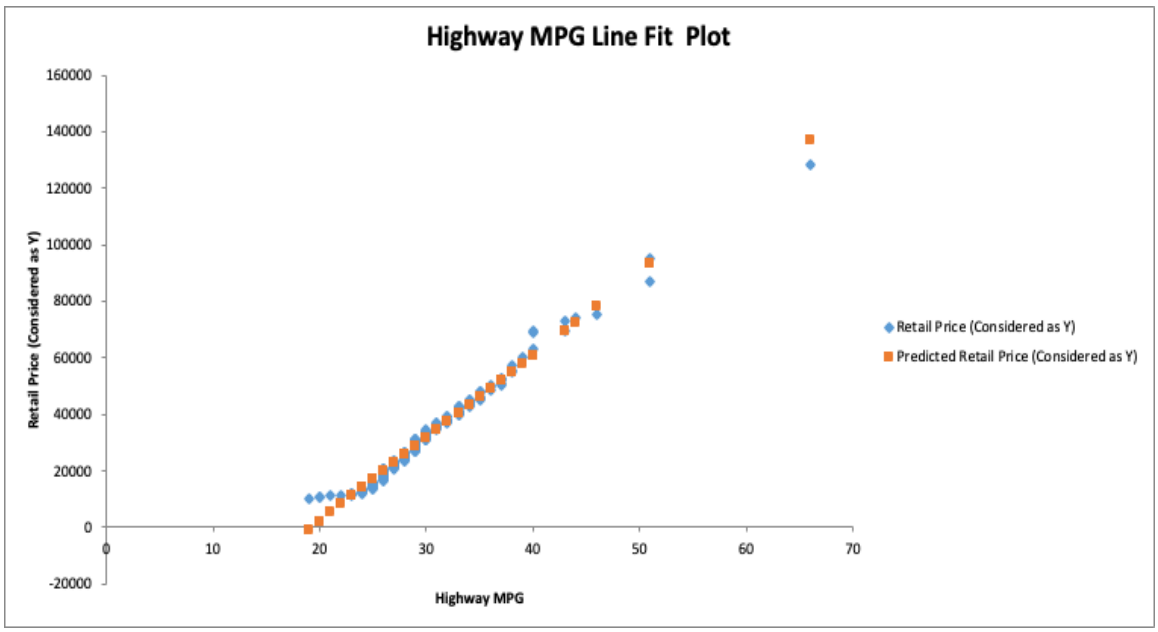
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-56431.92581	755.1298283	-74.73142193	2.6355E-164	-57919.71427	-54944.13736	-57919.714	-54944.13736
Highway MPG	2931.847018	25.27024228	116.0197431	1.9352E-207	2882.058527	2981.63551	2882.05853	2981.63551



7. **The Regression**: Create a scatter plot of your independent variable against the dependent variable using Excel. Make sure your dependent variable is y and your independent is x on the graph. Write a paragraph about your finding in the scatter plot.

This scatter plots indicates that the majority of cars have an average MPG falling within the range of 20 to 40.

8. **The Regression:** Display the 'Line Fit Plots' from the Simple Linear Regression Output. Is there a linear relationship between these two variables from the plot? Explain why?



There's a relationship between the MPG and cost of a car. MPG of a car increases as the cost goes down.

9. **The Regression: Is this regression model is important/significant? Why or why not?**
   The model is significant/important because the significance F is 1.9352E-207 which is smaller than 0.05.

10. **The Regression:** Are all parameters important/significant? Why or why not?
    The whole parameter is important/significant because P value for the Intercept is 2.6355E-164 and the P value for the X variable is 1.9352E-207, which both are significantly smaller than 0.05.

11. **The Regression: Show the Mathematical Equation of this model.**
    The mathematical equation for this regression would be
    y = -56431.92581 + 2931.847018x
    -**Example 1:** y = -56431.92581 + 2931.847018(20). So in this case we plugged the example value x(20) to the equation, and we got the answer as 2205.01455, which is the value we predict Y to be.

    -**Example 2:** y = -56431.92581 + 2931.847018(25) In this case we plugged in the number 25 as our X value, and after calculation, the y value was equal to 16864.24964, which is supposed to be predicted value of Y using our equation.

12. **The Regression**: Is this model a reliable predictor of y? Explain how much of variation is explained. Do you think there is a strong correlation and explain why or why not.

-The amount of variation explained by the regression is 98.3%, which is relatively high. The model is relatively reliable.

-A high R-squared value suggests a strong correlation between the independent and dependent variables. In this scenario, with such a high percentage of variation explained, it's likely that there is indeed a strong correlation between highway MPG and the price of a vehicle. This suggests that as highway MPG increases, the price of the vehicle tends to decrease, or vice versa.

13. **The regression:**
    Write a paragraph for the 4 assumptions and explain.

Assumption #1(Mean of 0): This was a bit hard to tell, but from the Residual Plot it looks like there are more dots on the bottom of the plot than the top, so it is not symmetric to 0, which means it does not satisfy the assumption 1. Assumption #2(Constant Variance): From the Residual Plot, we can tell those two parallel lines cannot be fit because the shape looks like x is getting larger, so that means the dots are not constant, which violates or does not satisfy the Constant Variance Assumption. Assumption #3(Independence): From the Residual Plot, the plot does not look like anything, it does not have any pattern, so it satisfies Assumption #3. Assumption #4(Normality): From the Normal Probability Plot, we can see that the graph is curvier rather than straight, so it does not satisfy the 4th assumption.

14. Summary:

In summary, our analysis of the relationship between highway MPG and vehicle price revealed several key findings. Through data visualization and regression analysis, we observed a clear negative correlation between highway MPG and vehicle price, indicating that vehicles with higher MPG tend to have lower prices, and vice versa. The regression model produced a high R-squared value of 98.3%, suggesting that approximately 98.3% of the variation in vehicle price can be explained by highway MPG. This indicates a strong relationship between these variables, making the model a reliable predictor of vehicle prices based on fuel efficiency. From this project, we've learned the importance of considering multiple factors when analyzing relationships between variables, as well as the potential impact of technological advancements and market trends on these relationships. Moving forward, further improvements to the model could involve incorporating additional variables such as vehicle type, engine size, or geographic location to provide a more comprehensive understanding of the factors influencing vehicle prices.