Sean McLean

ALY 6010

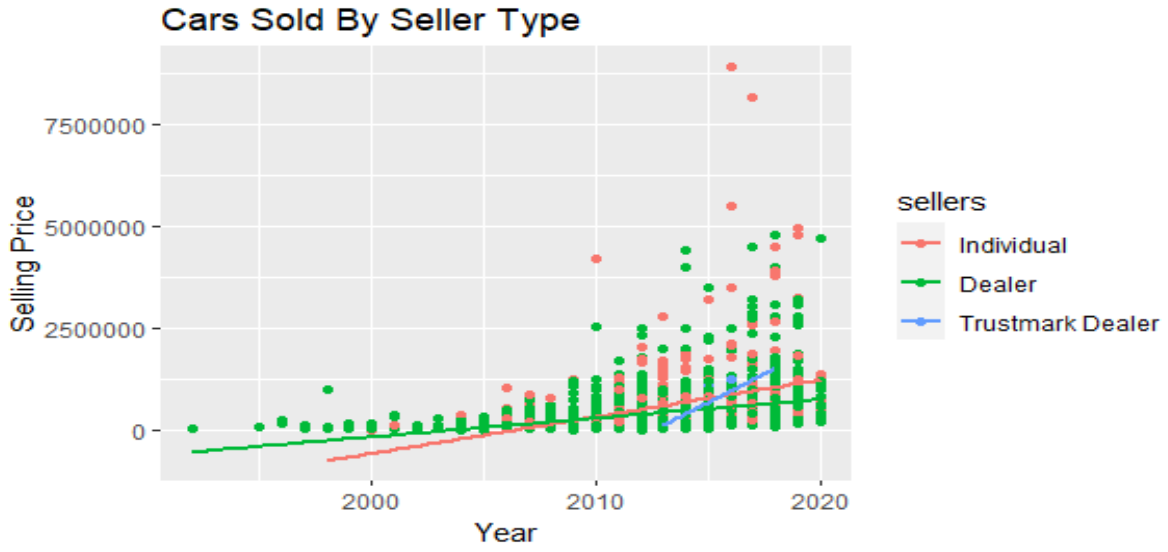
Module 6 – R Practice

Introduction

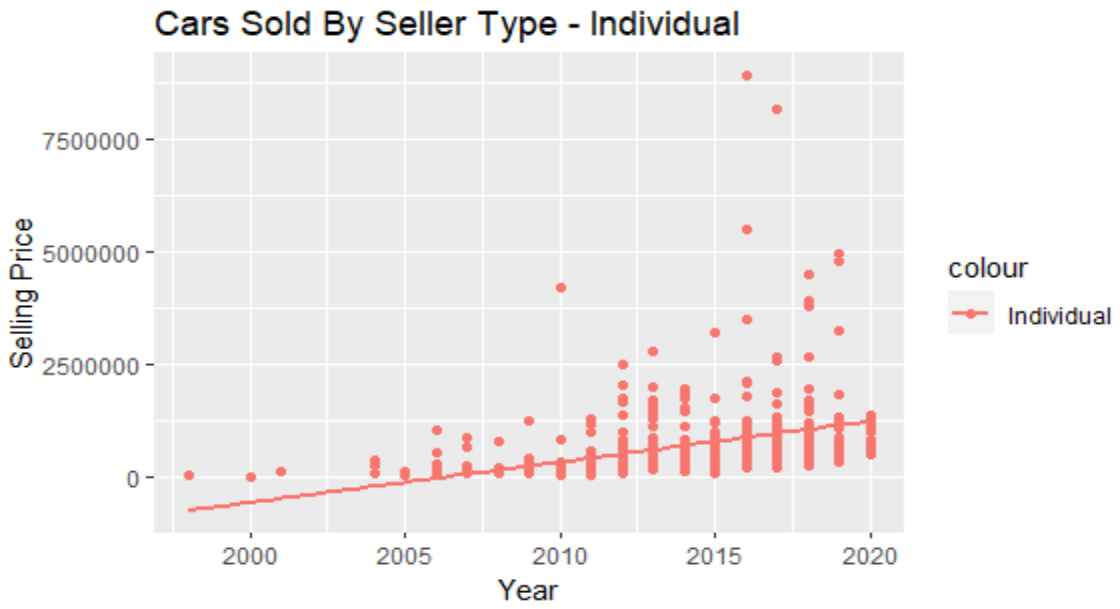
The R report summarizes car seller behavior among car sales that were recorded for the 2022 year. The data set uploaded into R consists of eight variables and 4340 entries of car sales from that year. The eight variables in the data set are the name of the vehicle, the year of the car, the selling price of the car, how many kilometers the car had when it was sold, the type of fuel the car uses, the type of seller for the vehicle, the transmission contained in the car, and how many owners the car had at the point of purchase. To analyze one aspect of the seller behavior of car sales, the seller type is being used as the dummy variables to study the relationship between selling price and the year of the car. The selling price will be the dependent variable and the year will be the independent variable for studying the relationship between the three variables.

Findings

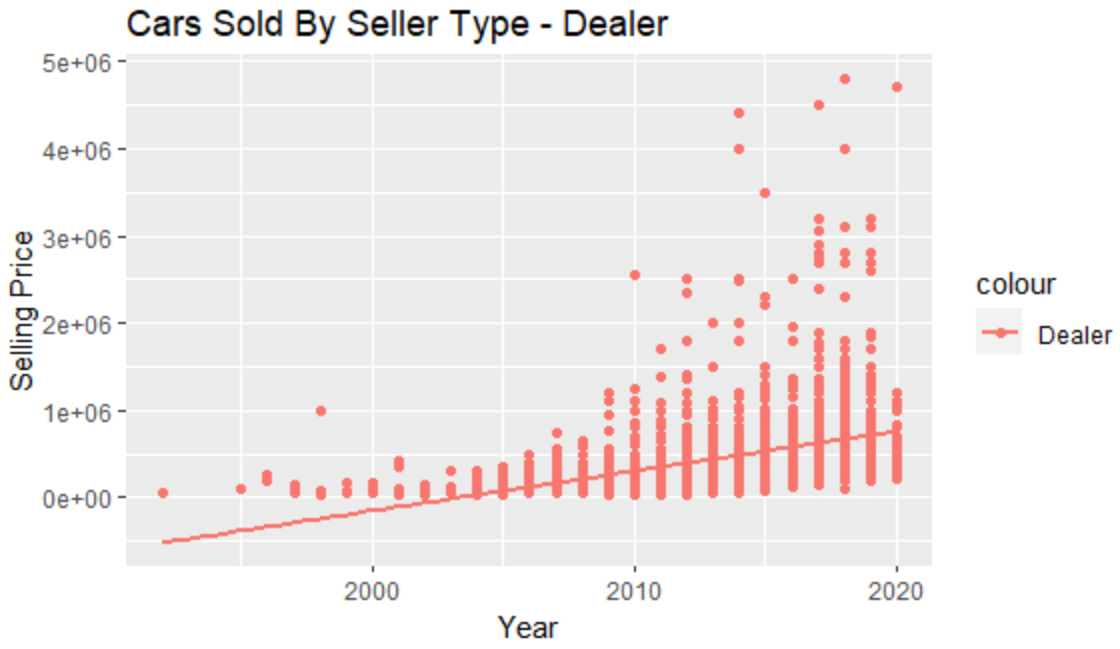
There are three subsets that were created from the aforementioned variables and each contain a regression line. Comparing the correlation coefficient between the variables showed a positive relationship between the selling price and year of the car. This could indicate that the selling price of the car could be higher depending on the year of the vehicle. A regression analysis of the data when looking at the selling price variable does not explain much about its variance, and it has a very low p-value overall. The following scatterplot shows the relationships between the three types of car sellers and each have evidence of having higher selling prices as the car gets newer. The regression lines in the scatterplot are relevant in that provides more clarity in what kind of relationship has been developed between the dependent variable and independent variable and how they are different between the categorical variables. It also has an immense impact in that it provides patterns and trends between the variables in the scatterplot.



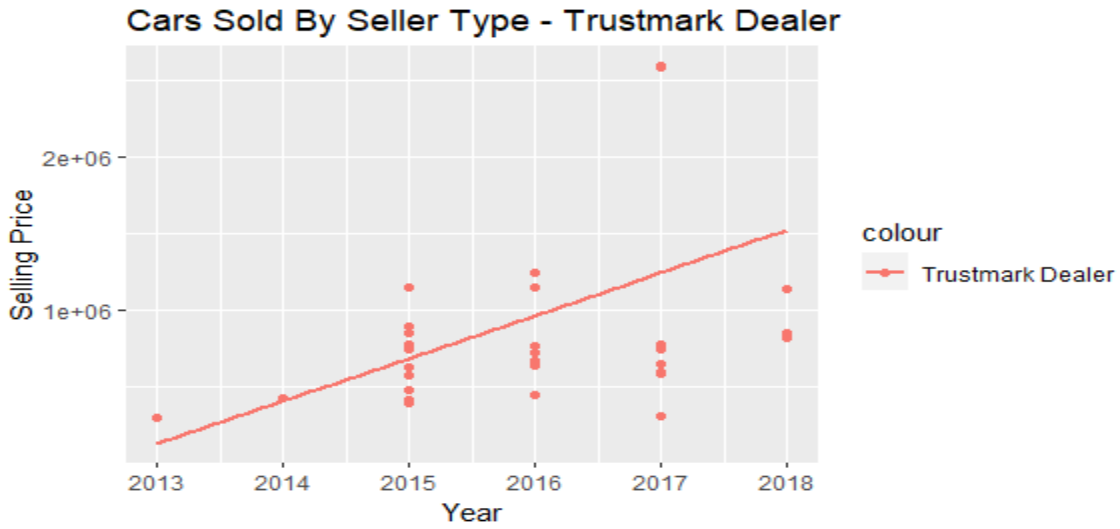
The second question in the assignment looks at each subset of the variables and puts them into individual scatterplots. This proves to be an effective way at looking at the dummy variables in that it allows for analyzing a target seller type on its own for certain behaviors. The scatterplot condenses down to the range of the dependent and independent variables instead of the multiple regression model which is bigger and therefore shows all the categorical variables in one plot. It becomes a lot easier as well to look at the data as it prevents cluttering that the multiple regression model provides. The individual seller in the first scatterplot below has a positive relationship with newer cars being sold at gradually higher selling prices. There are a few outliers in the plot and the majority of the cars are from the year 2005 to the present day.



The cars that are sold by a dealer also shows a positive relationship but with more outliers than the individual seller. The difference between the two sellers is that the individual seller has a few more sells at higher prices which is why the selling price range for the dealer plot is condensed in size. More older cars are also being sold by the dealer than the individual that are lower in selling price than the newer vehicles.



There is much less data from trust mark dealers but does show a linear relationship in that the newer vehicles are being sold at slightly higher selling prices. There is really only true outlier in the plot and the range of the vehicles is much smaller than the other types of sellers. The majority of the data points are far from the regression line which could indicate prediction errors and high residuals.



Conclusion

The dummy variable when compared with the dependent and independent variable in the car's dataset from a multiple regression line scatterplot and subset plots provide interesting insights into car selling behaviors. The selling price of a car will be higher if the car is newer

regardless of what type of seller is involved. Dealer sellers are similar but will tend to sell cars that are older as well, and the trust mark dealers plot shows weak relationships between the other variables in how their cars are sold despite a smaller dataset. Each type of seller shows different qualities and approaches toward how the cars are sold which can provide some input for future car buyers that are looking to purchase a new or used vehicle.

References

Bluman, A. G. (2017). Elementary Statistics: A Step-by-Step Approach. 10th edition. McGraw-Hill Education.

Chat GPT. (2023, December 10th). Default (GPT 3.5). https://chat.openai.com/

Dattatray Khare, Akshay. (2020, June 26th). Car Details Dataset. *Kaggle*. https://www.kaggle.com/datasets/akshaydattatraykhare/car-details-dataset