BUAN 6356 – Business Analytics with R

Used Car Price Prediction

Group 11

Parth Shah, Siddesh Iyer, Akshat Bhandari, Sai Gagan Shakamuri, Aralyn Tran

Executive Summary

The automotive industry represents around 3% of the United States' total GDP. With a used-car market more than double the size of the new-car segment and exhibiting faster growth, it is a significant area of interest. According to McKinsey's auto retail micro-market model, Americans purchase approximately 39.4 million used cars annually compared to 17.3 million new vehicles (2018), and used-vehicle sales are expected to grow more rapidly than new-vehicle sales over the next five years. This paper aims to predict used car selling prices by employing various data analytics models and pinpointing key factors influencing used car prices. We selected the "Used Cars Dataset" from Kaggle for this project due to its comprehensive, clean, and up-to-date information. The data is examined using an array of predictive models, including linear regression and decision tree models. Ultimately, we draw a conclusion by identifying the most accurate model for price prediction.

1. Objective

Valuation specialists at Edmunds.com report that the average pre-owned vehicle with 100,000-109,999 miles on a dealer's lot increased in value by 31% over the past year, from $12,626 in June 2020 to $16,489 in June 2021. The ongoing microchip shortage has led to slowed or halted production for numerous new-vehicle lines, resulting in limited supply and driving more U.S. consumers, including rental companies, towards the used-car market. Consequently, heightened demand combined with a restricted supply of available vehicles (fewer new models sold translates to fewer trade-ins on used vehicle lots) has caused prices to soar across the board. To succeed in this market, investors, OEMs, and retailers must understand their used-car customer base, identify their priorities, recognize which companies are meeting those needs, and determine how the industry can exceed customer expectations to secure their business and loyalty. Customer preferences may differ by segment and location, necessitating that used-car retailers pinpoint specific customer characteristics and develop strategies to attract and sell to them effectively.

Pricing analytics can be instrumental in helping both customers and companies navigate this dilemma by evaluating the market price of a vehicle before purchasing or selling it.

Consequently, this project aims to achieve two main objectives:

- Estimate the price of a used car by considering a range of features, drawing from historical data.
- Gain deeper insight into the most significant features that contribute to determining the price of a used vehicle.

2. Data Description

The data that is used for this project is downloaded from Kaggle (Data Link). This is secondary dataset updated monthly using web scraping techniques from US Craigslist posts (the world's largest collection of used vehicles for sale). We are using this dataset because it has more records, more attributes, and includes clean information on location and time data.

The Database consists of 426,880 rows and 26 features, one of which is the continuous dependent variable ("price") that we want to predict. The large amount of data helps us to improve the accuracy of the model. The attributes can be easily related with each other, and we can easily infer the desired objective. There are some considerable of number of missing values therefore data cleansing is required, but other than this dataset has all that we need to build the model that we desire.

3. Data Preprocessing

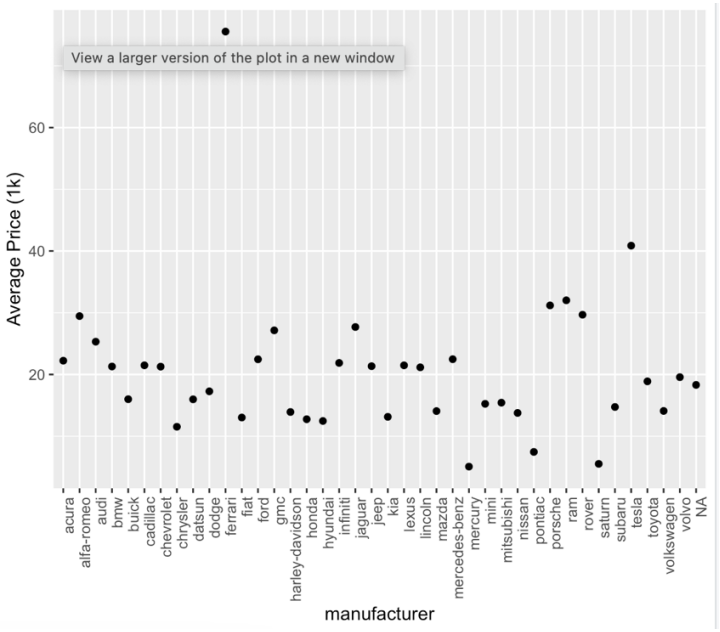
Upon importing the data, it became apparent that the dataset contained numerous missing and inconsistent values. It also has many features that are irrelevant and contain no information to help us in our analysis process. We determined that data cleansing was necessary to derive meaningful conclusions and took the following steps to clean the dataset. This process effectively removed many outliers resulting from poor data quality:

- Remove 9 features that contain no relevant information, features were considered are: 'region','price','year','model','condition','cylinders','fuel','odometer','title_status','transmission','drive','size','type','paint_color','state','posting_date','manufacturer'
- Several records show the pricing from $0 to above $3M, we narrow down the price to the range from $1k to $100k
- Same with odometer, some records show 0 miles up to millions of miles. Those are unrealistic so we give a range of over 0 miles and under 200,000 miles
- We create a new feature called Age for the Year of the vehicles, and only consider age less than 50 years old
- We checked for missing values in the dataset and exclude row having NULL Age and NULL Odometer
- Next we check for duplicates, convert cylinders to numeric values, and remove models having less than 50 entries
- For categorical feature condition has 5 levels from excellent to salvaged, we create a new feature called new_condition that contain only 4 levels new, good, and fair_salvaged and N/A
- Lastly, we assign N/A to all missing values and convert categorical variables to factors
- After the data cleaning and preprocessing steps, the resulting dataset contains 347543 rows and 22 features

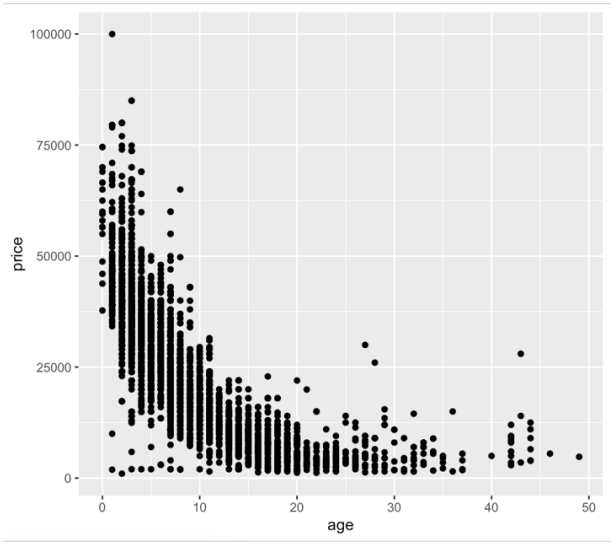
4. Exploratory Analysis

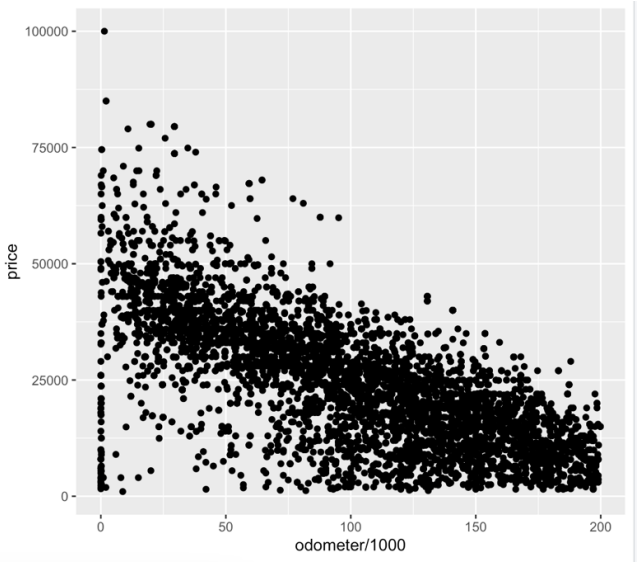
Next, we work on furthur exploratory analysis on the data. First, we want to see who are the top manufacturers and their average prices.

	manufacturer	avg_price_1k	max_price_1k	min_price_1k	count
1	ferrari	75.545833	99.999	2.034	36
2	tesla	40.866514	95.000	1.004	809
3	ram	32.012254	99.981	1.029	14496
4	porsche	31.187804	98.995	1.200	1213
5	rover	29.680247	96.995	1.089	1831
6	alfa-romeo	29.459267	62.950	2.000	828
7	jaguar	27.688613	75.000	1.100	1828
8	gmc	27.146165	90.000	1.030	13595
9	audi	25.314133	99.991	1.100	6868
10	mercedes-benz	22.475060	99.995	1.051	9481

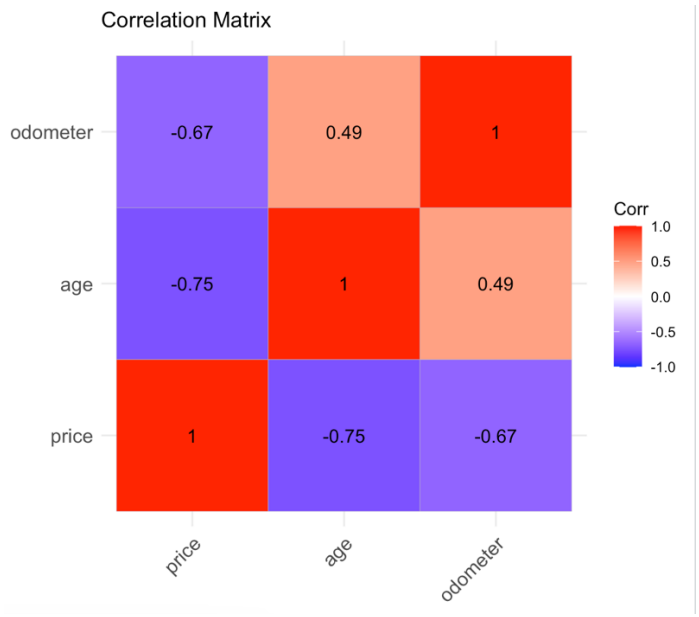


After that we look at the numerical variables age and odometer.

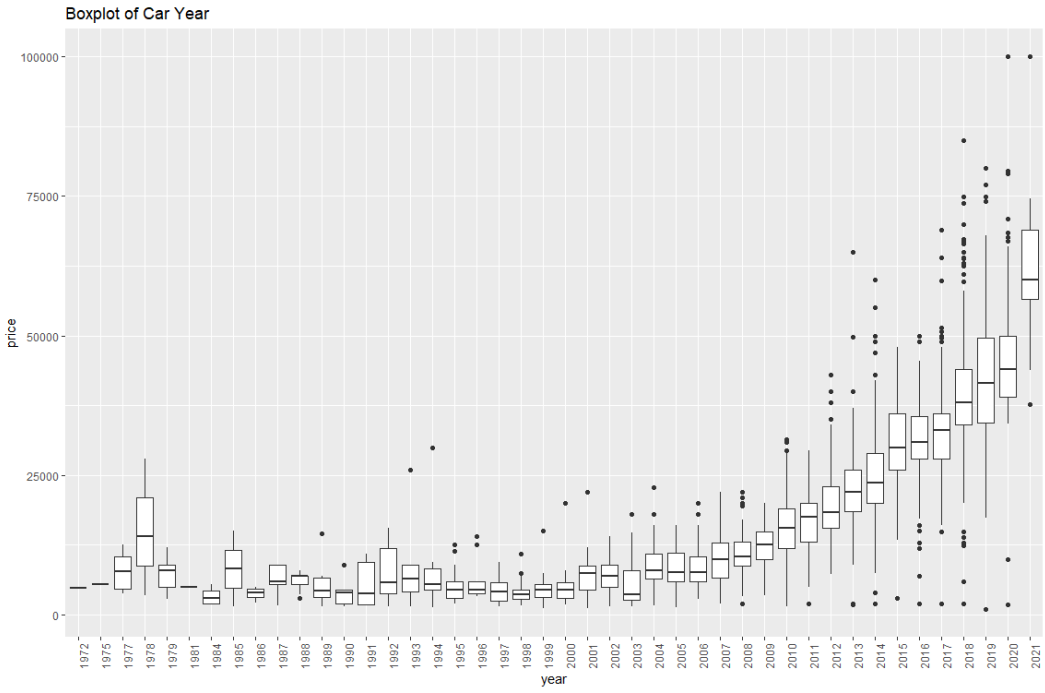


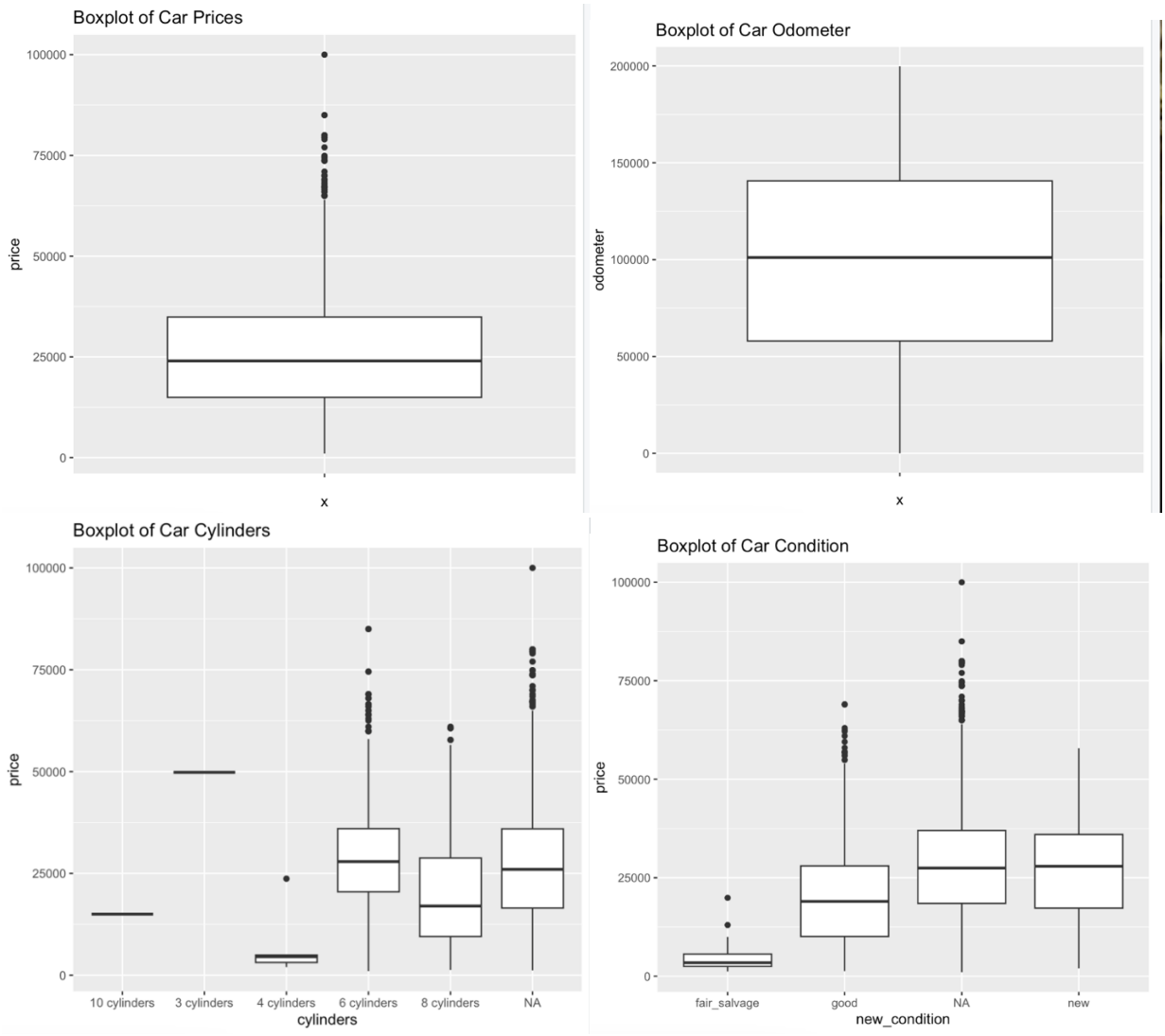


The analysis reveals negative correlations between the odometer and price, as well as between the price and age of the car. These correlations make intuitive sense since as the car ages or accumulates more mileage, its value generally decreases, leading to a negative relationship between these variables.



We also want to look at outliers



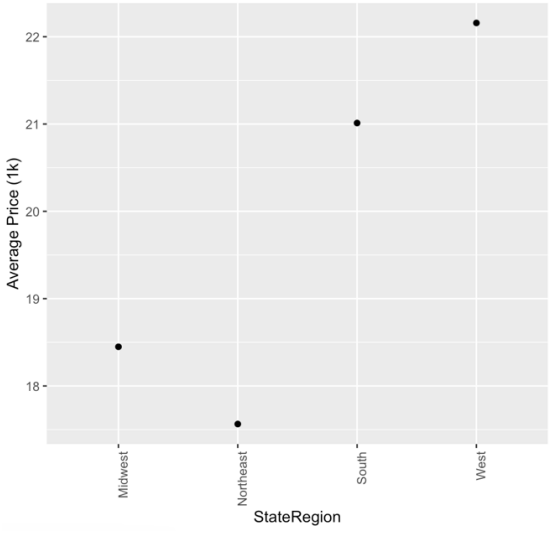


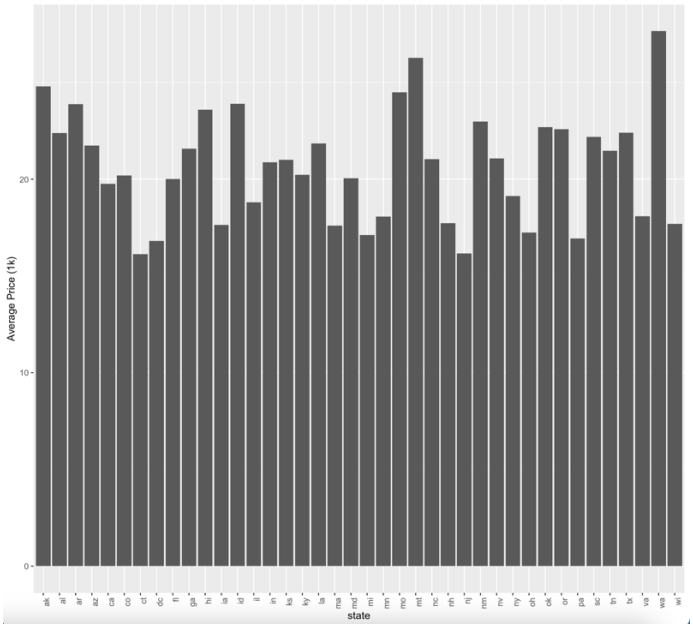
As part of our analysis, we also sought to determine the average price by region

	StateRegion	avg_price_1k	max_price_1k	min_price_1k	count
1	West	22.15774	655.000	1.003	103585
2	South	21.01159	990.000	1.004	109280
3	Midwest	18.44883	566.567	1.002	77568
4	Northeast	17.56410	400.123	1.050	57665

row names

	state	avg_price_1k	max_price_1k	min_price_1k	count
1	wa	27.64870	239.900	1.004	10668
2	mt	26.26549	163.540	1.057	5347
3	ak	24.78816	145.000	1.100	3246
4	mo	24.49843	299.500	1.200	3543
5	id	23.90541	425.000	1.100	7223
6	ar	23.87726	163.540	1.150	2902
7	hi	23.59335	289.995	1.035	2404
8	nm	22.97091	163.540	1.200	3393
9	ok	22.69373	777.777	1.200	5134
10	or	22.58478	299.500	1.003	11266





5. Model Building and Evaluation

Based on the correlation plot, we figured that Age of the car and Miles driven by the car would determine the price of the used car. Additionally, based on the graph (x) we can say that the car model plays a vital role in determining the price.

To arrive at the final model that we should use, we created multiple regression models and evaluated their R-Squared value (variance explained by the model).

Initially, we decided to build a linear regression model specific to a car model to predict the car prices with Age and Odometer as the input variables/features. In our model, we have odometer_1k as a variable for 1000 Miles driven by a car.

Below are the steps/modifications we carried out to identify the optimal (final) model

1. Base model with which we started -
   linearreg <- lm(price ~ age+odometer_1k , data = d_subset_check)
   Adjusted R-squared = 0.6832

2. While evaluating the performance of this model, we figured that the extreme values (top and bottom) in our dataset did not make sense. For instance, a new Jeep Wrangler costs $92k (top-tier variant and add-ons). But in our dataset, we saw outliers having prices as low as $1k and $435k. This could be due to data entry at the original source and/or modified car. Since we can not accommodate these factors in our model, we decided to remove top and bottom 2 percentile data points while builiding the model.
   Adjusted R-squared post removing the extreme data points = 0.7106

3. Looking at the relationship between car prices and age (graph y), we can say that the relationship is not exactly linear. The drop in car prices is steep in the initial years and later on the curve flattens. Hence we created another model to capture this information in our model. Adjusted R Square was higher than the base model.
   linearreg2 <- lm(price ~ age+I(age^2)+odometer_1k , data = d_subset_check)
   Adjusted R-squared = 0.7679

4. Additionally, we need to factor in those cases where a car is heavily used in a short period of time and vice-versa. For this, we added interaction variable: age*odometer_1k. Adjusted R Square was higher than the previous model.
   linearreg3 <- lm(price ~ age+I(age^2)+odometer_1k+age:odometer_1k , data = d_subset_check)
   Adjusted R-squared = 0.7705

5. Next, we wanted to check if the car condition plays a role in determining the price or not. Though Adjusted R square was similar to the last model, p-values for various conditions were higher than 0.05 (we got p-values as high as 0.52).
   linearreg4 <- lm(price ~ age+I(age^2)+odometer_1k+age:odometer_1k+condition , data = d_subset_check)

6. Similar to condition, we tried Fuel Type, transmission, number of cylinders, paint_colour, car-drive, type of car, etc. But all of them had coefficient with very high p-value. Which suggests that these variables are insignificant as compared to the variables age and odometer. Hence, we would not be using them in our final Model.

7. Final Model that we decided based on Adjusted R square and Variable's significance-
linearreg3 <- lm(price ~ age+I(age^2)+odometer_1k+age:odometer_1k , data = d_subset_check)

```
> model_regression('f-150', 1,100);

Call:
lm(formula = price ~ age + I(age^2) + odometer_1k + age:odometer_1k,
    data = d_subset_check)

Residuals:
   Min     1Q Median     3Q    Max
-32559  -3247   -333   3304  26579

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     49419.1345   259.5393 190.411   <2e-16 ***
age             -2874.4709    46.5150 -61.797   <2e-16 ***
I(age^2)           46.2880     1.2096  38.266   <2e-16 ***
odometer_1k       -78.6658     2.8556 -27.548   <2e-16 ***
age:odometer_1k     2.1487     0.2577   8.338   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5849 on 6093 degrees of freedom
Multiple R-squared:  0.7707,    Adjusted R-squared:  0.7705
F-statistic:  5119 on 4 and 6093 DF,  p-value: < 2.2e-16

[1] "For a car having age = 1 and Miles Driven (1k) = 100 , Predicted value = $ 38939.24"
```

Model Interpretation:

The coefficients of the linear regression model are as follows:

Intercept is 49419.13, which means that for car having age=0 and Miles driven = 0, the car price would be $49419.

The coefficient of age is -2874.47, which means that the price of the car decreases by $2874.47 for every unit increase in age. We are not considering the effect of age squared (next term).

The coefficient of age squared is 46.29, which means that the effect of age on the price of the car is not linear but quadratic. Price per year will increase by 46.29*year^2.

The coefficient of odometer reading is -78.66, which means that the price of the car decreases by $78.66 for every thousand miles driven.

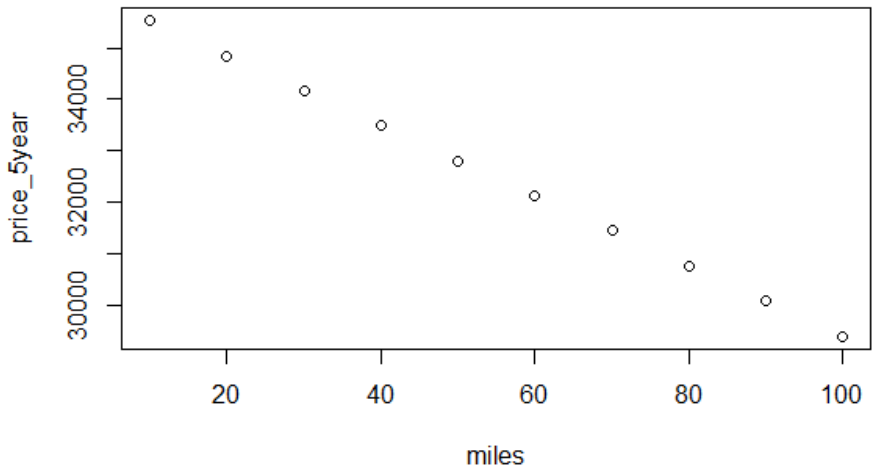
The coefficient of the interaction term between age and odometer reading is 2.15. The effect of odometer reading on the price of the car is moderated by age.

The model has a good fit, as indicated by the high multiple R-squared, and adjusted R^2 value of 0.77, which means that 77% of the variance in the price of the car is explained by the independent variables age and odometer reading.

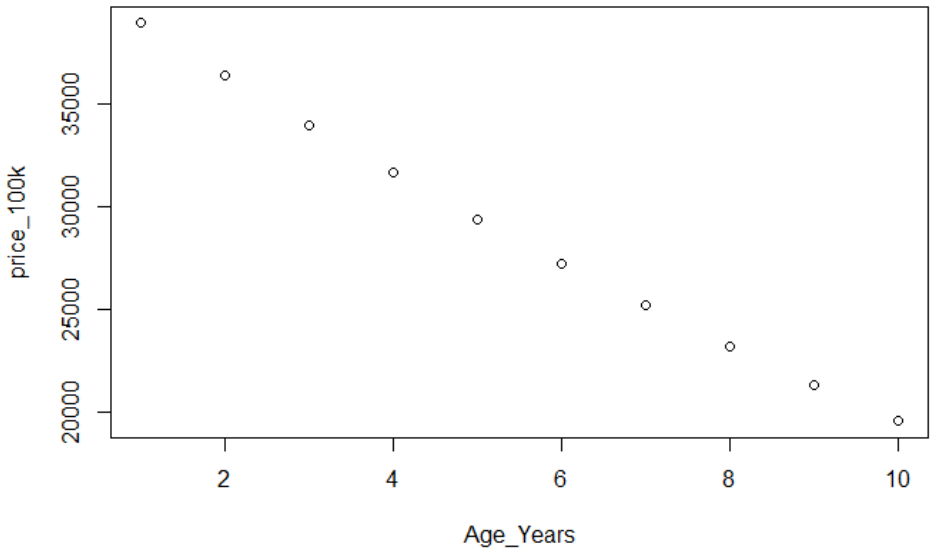
Sample output for the model:

For Car Model ‘f-150’, Age = 1, and Miles Driven = 100k, the predicted price is $38939.

Effect of miles driven when age is constant (=5 Years) for Car Model F-150 on Price:



Effect of age when miles driven is constant (=100k) for Car Model F-150 on Price:



6. Conclusion

Our model suggests that the primary variables for determining the price of a used car are its odometer reading and age. Additionally, the rate of depreciation is higher in the initial years and slows down as the car ages. While this is a good model for estimating a fair price, it is important to note that there may be other factors that the model cannot capture. For example, our model does not take into account any modifications made to the car. Therefore, users should keep these factors in mind while interpreting car prices.

The used car market is a complex one, with many factors influencing the price of a vehicle. In this project, we have explored the use of linear regression (with quadratic terms) to predict used car prices. We have found that the primary variables for determining the price of a used car are its odometer reading and age. Additionally, the rate of depreciation is higher in the initial years and slows down as the car ages.

In addition to the primary variables, there are other factors that can influence the price of a used car which our model can not capture. These include:

- Supply and Demand of the car
- Any modifications made to the car.

Therefore, users should keep these factors in mind while interpreting car prices.
The model can be a good starting point to figure out ballpark price of a car based on its age and odometer reading, but it is important to use judgment and knowledge of the market when making a purchase.

References

- GONGGI, S., 2011. New model for residual value prediction of used cars based on BP neural network and non-linear curve fit. In: Proceedings of the 3 rd IEEE International Conference on Measuring Technology and Mechatronics Automation (ICMTMA), Vol 2. pp. 682-685, IEEE Computer Society, Washington DC, USA.
- Pudaruth, S. “Predicting the Price of Used Cars using Machine Learning Techniques.” (2006).
- Listiani M. 2009. Support Vector Regression Analysis for Price Prediction in a Car Leasing Application. Master Thesis. Hamburg University of Technology.
- McKinsey and Company-Used cars, new platforms: Accelerating sales in a digitally disrupted market
- Car and Driver, news platforms: Used cars are having a moment.
- Forbes.com: Very Used Cars: High-Mileage Models Are Now Selling For Big Bucks
- RegionMapping: https://github.com/cphalpert/census-regions/blob/master/us%20census%20bureau%20regions%20and%20divisions.csv