R Markdown Data science

2023-10-24

## Student Name: Manav Saini

## Student ID: 20608893

## Car dekho vehicle dataset (Source Kaggle)

The data taken is the CarDekho vehicle dataset from Kaggle. I have selected this dataset due to its mix of categorical and numerical variables which would help me deploy methods to predict the price of the car based on the provided variables.

#Read the file and store it in the variable named CD.   
CD = read.csv("/Users/manavsaini/Downloads/CAR DETAILS FROM CAR DEKHO.csv")

### Checking the first few rows.

We observe that the data has a mix of categorical and numerical variables.

# As observed, the data set contains 7 variables of both numeric and categorical.  
head(CD)

## name year selling\_price km\_driven fuel seller\_type  
## 1 Maruti 800 AC 2007 60000 70000 Petrol Individual  
## 2 Maruti Wagon R LXI Minor 2007 135000 50000 Petrol Individual  
## 3 Hyundai Verna 1.6 SX 2012 600000 100000 Diesel Individual  
## 4 Datsun RediGO T Option 2017 250000 46000 Petrol Individual  
## 5 Honda Amaze VX i-DTEC 2014 450000 141000 Diesel Individual  
## 6 Maruti Alto LX BSIII 2007 140000 125000 Petrol Individual  
## transmission owner  
## 1 Manual First Owner  
## 2 Manual First Owner  
## 3 Manual First Owner  
## 4 Manual First Owner  
## 5 Manual Second Owner  
## 6 Manual First Owner

### Checking the attributes using the “str” function in R.

As stated, the data frame has a total of 4340 rows and 8 columns. This presents an oppertunity for utilization for the regression and classification tasks.

str(CD)

## 'data.frame': 4340 obs. of 8 variables:  
## $ name : chr "Maruti 800 AC" "Maruti Wagon R LXI Minor" "Hyundai Verna 1.6 SX" "Datsun RediGO T Option" ...  
## $ year : int 2007 2007 2012 2017 2014 2007 2016 2014 2015 2017 ...  
## $ selling\_price: int 60000 135000 600000 250000 450000 140000 550000 240000 850000 365000 ...  
## $ km\_driven : int 70000 50000 100000 46000 141000 125000 25000 60000 25000 78000 ...  
## $ fuel : chr "Petrol" "Petrol" "Diesel" "Petrol" ...  
## $ seller\_type : chr "Individual" "Individual" "Individual" "Individual" ...  
## $ transmission : chr "Manual" "Manual" "Manual" "Manual" ...  
## $ owner : chr "First Owner" "First Owner" "First Owner" "First Owner" ...

### Converting Categorical to Numerical.

# Create a categorical variable  
#We converted the variables into factors.   
CD$name = as.factor(CD$name)  
CD$fuel = as.factor(CD$fuel)  
CD$seller\_type = as.factor(CD$seller\_type)  
CD$transmission = as.factor(CD$transmission)  
CD$owner = as.factor(CD$owner)

### Potential research questions:

This could be some of the potential research questions:

Regression Analysis:

1. Can we predict the selling price of a used car based on its other attributes like “year,” “km\_driven,” “fuel,” etc.?
2. How does the year of manufacture affect the selling price?

Classification Analysis:

1. Can we classify cars as “Fast Selling” or “Slow Selling” based on attributes like “year,” “selling\_price,” and “km\_driven”?
2. Can we predict the type of owner (“First Owner,” “Second Owner,” etc.) based on other variables?

The target variable has been identified as: **“selling\_price,”.**