Project Step 1

Kiran Komati

5/16/2021

Introduction

The topic I chose for research is the price prediction of used cars based on the different attributes of a car such as the brand, year of purchase,selling price, no. of kms driven etc., The data set I chose is carDekho's dat set from Kaggle and it has 3 sets of data.Data science can help us in estimating the true value of the car there by helping the sellers to decide if it is a good buy or not.

Research questions

1. Which factors affect the price of the car.
2. Are there any factors that have little to no affect on the price of the car.
3. Are there are factoros that have negative correlation with the price.
4. As anyone expects, is the year of purchase heavily correlated with the price of the car?
5. Perform hierarchical regression Analysis on the data set with selling price as the dependent variable and the no. of km driven as the predictor and adding few other fields that can help the overall accuracy of the model.

Approach

I'll see if we can bring the dat set into a single data frame for easy analysis and perform the transformations as necessary based on the data quality. I'll validate if the data has any NA values and replace or filter them as necessary.I'm planning to use the approaches that we learned in the course so far such as covariance, correlation, R squared to identify the relations of the variables and once identifed, i will use the linear regression to come up with a model that can predict the selling price of the used car and the predictors. car.

How your approach addresses (fully or partially) the problem.

The approach i'm planning to take makes sure that the data is good and that we are not just working on any random data. The values that we are planning to calculate helps in identifying how strongly or weakly the fields are correlated.

Data

The data is from https://www.kaggle.com/nehalbirla/vehicle-dataset-from-cardekho. This datset has information about the used cars listed in the cardekho website from India. The source data was compiled around

7 months back. This has a total of 9 columns. Initial analysis shows that there is no missing data in the dataset.

Required packages

1. tidyverse
2. dplyr
3. car
4. ggplot
5. rmarkdown
6. tinytex
7. QuantPsyc
8. knitcitations

Plots and Table Needed

1. Histogram to check if the data is normally distributed.
2. Scatterplots
3. tables such as summary tables, correlation tables, covariance tables.
4. Confidence intervals
5. anova and Durbin Watson tests

Questions for future steps

1. Are there any cleaning required for the data.
2. Thorough Analysis for the missing data.
3. Does the dataset have any categorical data , if yes how are we planning to handle it.
4. Which method best helps us in predicting the prices.
5. What functions can be used to identify the relationship between the variables.