**Introduction**

The automobile industry in India, currently manufactures 26 million vehicles, out of which the passenger vehicle segment is a large proportion. The ever-increasing demand for the passenger vehicles has made people not only spend money in buying new vehicles but also in pre-owned passenger vehicles. In the pre-owned market, it is important for both the buyer and seller to estimate the selling price of a used vehicle in an efficient manner. This document proposes a literature review for creating a tool that can predict the selling price of a vehicle by using some of the statistical techniques. This proposal includes information about the dataset like the variables, type of these variables, number of observations, the questions that I would like to answer using the dataset, the main goal of the project and the type of statistical analysis that I would be performing on the dataset.

**About the Dataset**

The dataset is sourced from a popular website called Kaggle released by an Indian company named “Cardekho”. The original dataset includes 13 feature and 8,128 observations, but this is planned to be reduced to 5 variables where 4 are predictor variables and 1 is the target variable and the total number of observations ill be 901 (which is a random sample of

13% taken from the original dataset). The 5 featured are, vehicle brand, model year, cubic capacity, Kilometre driven and the selling price of each vehicle which will be the target variable

|  |  |  |
| --- | --- | --- |
| **Variables** | **Continuous/Discrete/Categorical** | **Predictor/Response** |
| vehicle\_brand | Categorical | Predictor |
| year (model) | Discrete | Predictor |
| cubic\_capacity | Continuous | Predictor |
| km\_driven | Continuous | Predictor |
| selling\_price | Continuous | Response |

**Questions planned to be answered**

* Whether a luxury car manufactured by a brand like Mercedes, has a higher selling price than a usual Indian car brand like Maruti?
* Does an old car have a lower value in the resale market? (Except for classic vehicles)
* How does the engine specification of a car affect its selling price?
* Does the resale price of the car drop as the distance covered as it increases?

Finding answers for the above questions will help me identify the casual relationship between the variables. Having a clarity in this will help me develop a better multiple linear regression model that fits the data and will allow me to check if the fitted model has a higher explanatory power which could be interpreted from the computed R2 value. Having done the regression analysis for the above points, I will be able to predict the selling price for the used Indian cars.

**Objective**

The main goal of the project is to predict the price of a used car based on 5 regressors, mentioned above.

**Statistical Analysis**

The

* Please describe your dataset.
* Provide your variables and determine what types of variables you will be using (continuous, categorical). Is there a response variable?
* Are there factors/predictor variables?
* Provide the total number of observations in your dataset.
* Try to have a least 50 observations, but no more than 1000. (suggested guidelines)
* In your proposal, identify some questions you will be able to answer through your analysis and **why you are trying to answer them.**
* What are your objectives?
* What type of statistical analysis do you plan on performing (ex. Two-sample t-test, Two-Factor ANOVA, Multiple Regression, etc.). This proposal should be one page.