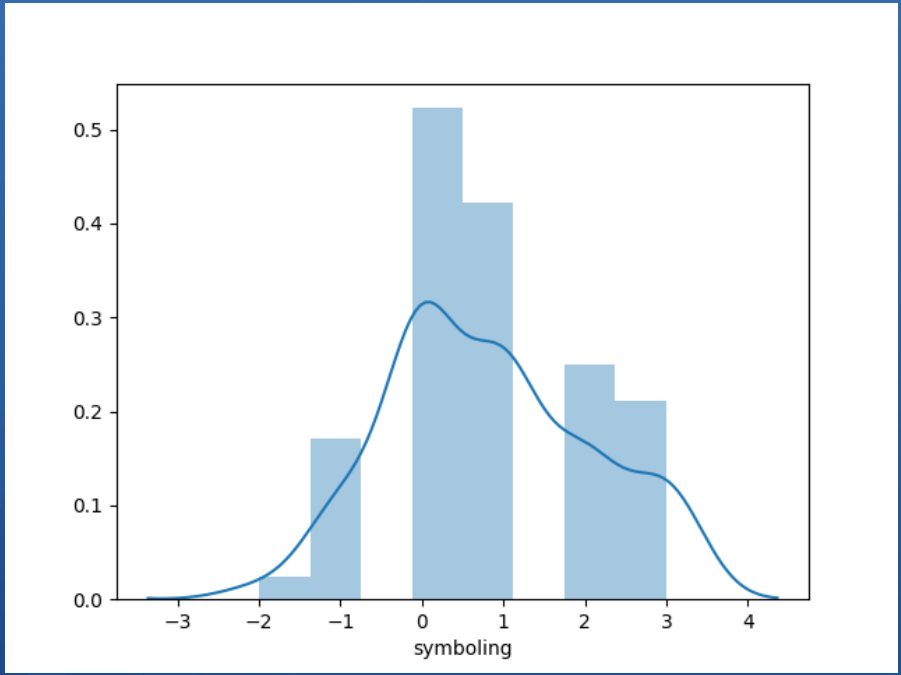
Car Pricing Prediction

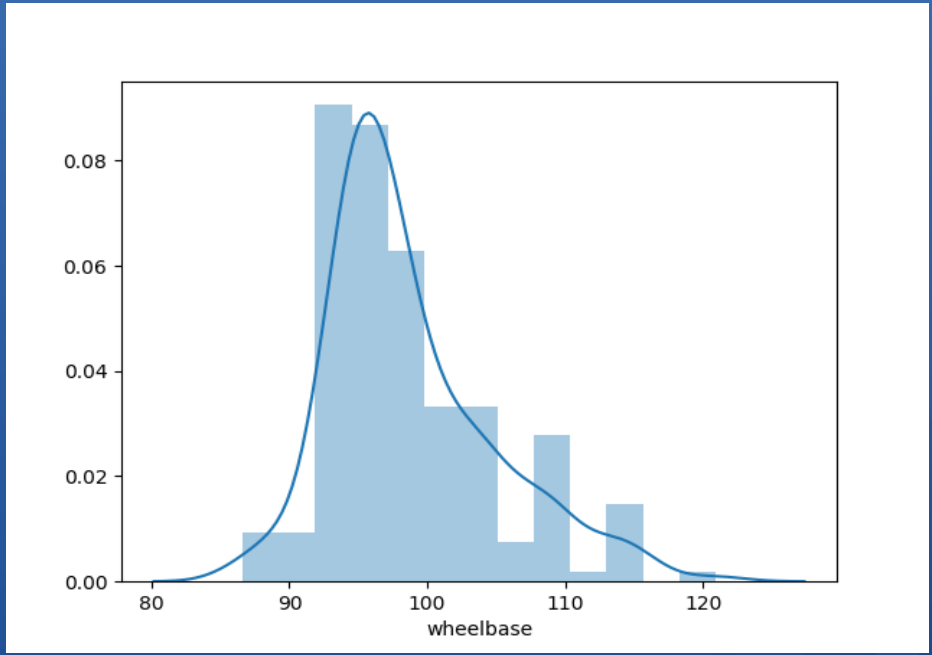
- The objective is to identify the best possible way for the car manufacturing companies, to come up with the accurate price tag for their respective models.
- The data being used in the analysis

Methodology

Data understanding and exploration
Data cleaning
Data preparation
Model building and evaluation

Features analysis

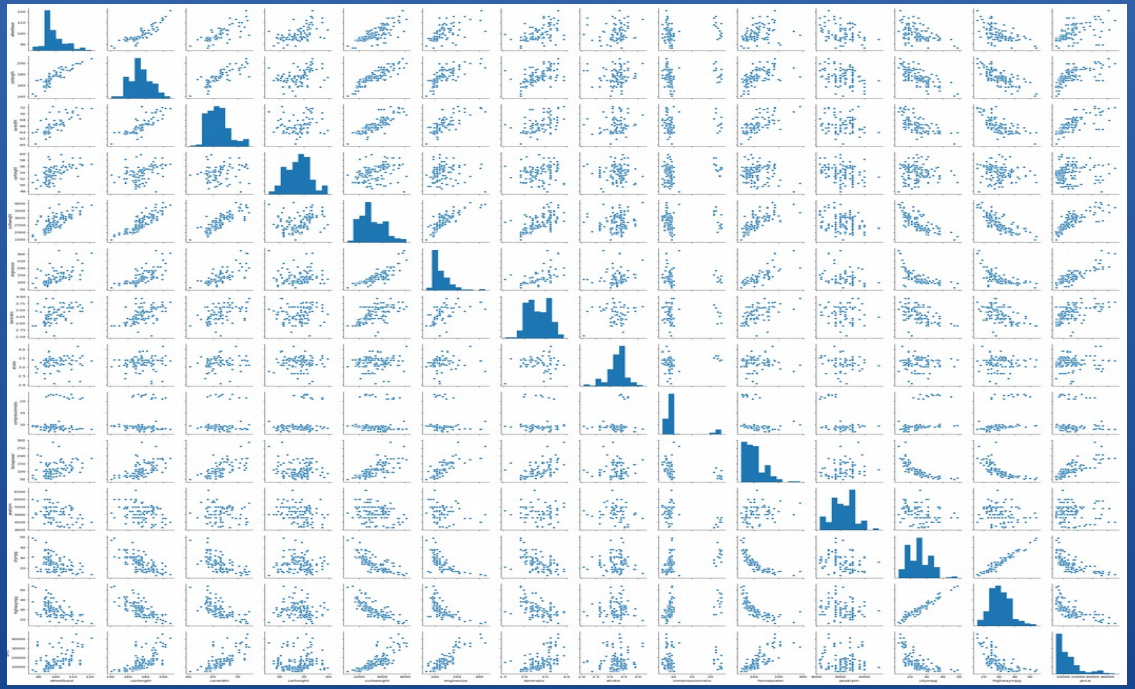




This way we can individually analyse each feature and can have a idea of what is the trend.

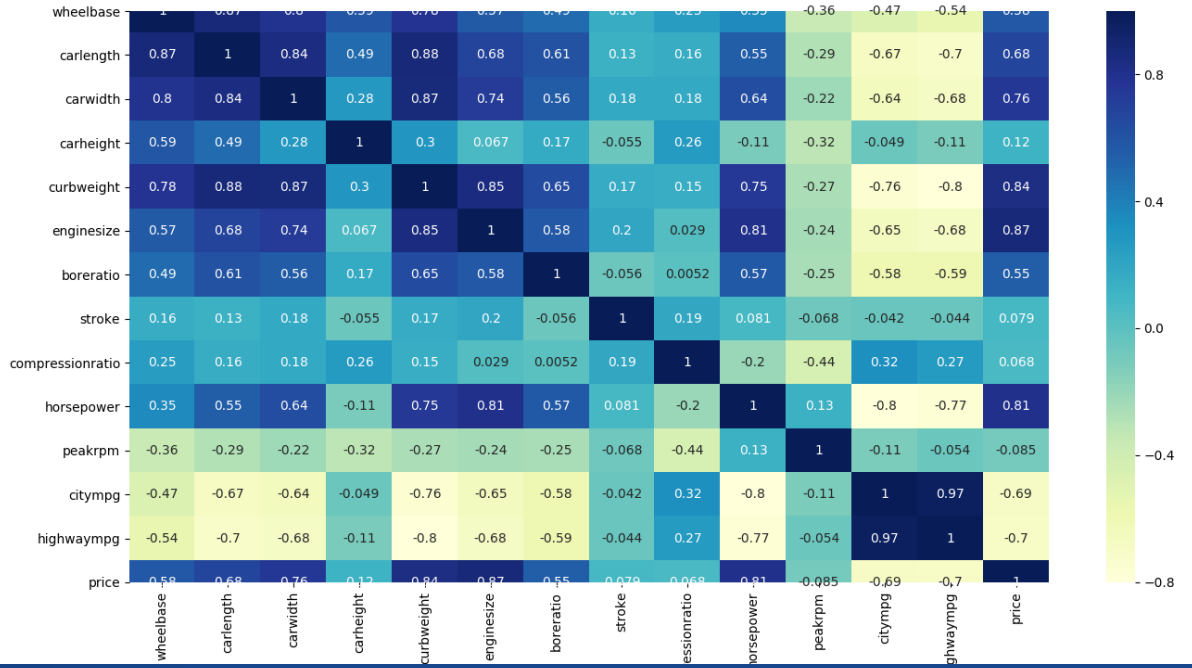
Data Exploration

Scatter plot can be useful to check whether the feature is linearly related to the target variable and in that way we can opt for linear regression according to the justification of trends.



Contd.

It can be seen that it’s really hard to interpret from the pairwise plots, which pushes us for the heat where we can see the correlation between the features.



Data Cleaning

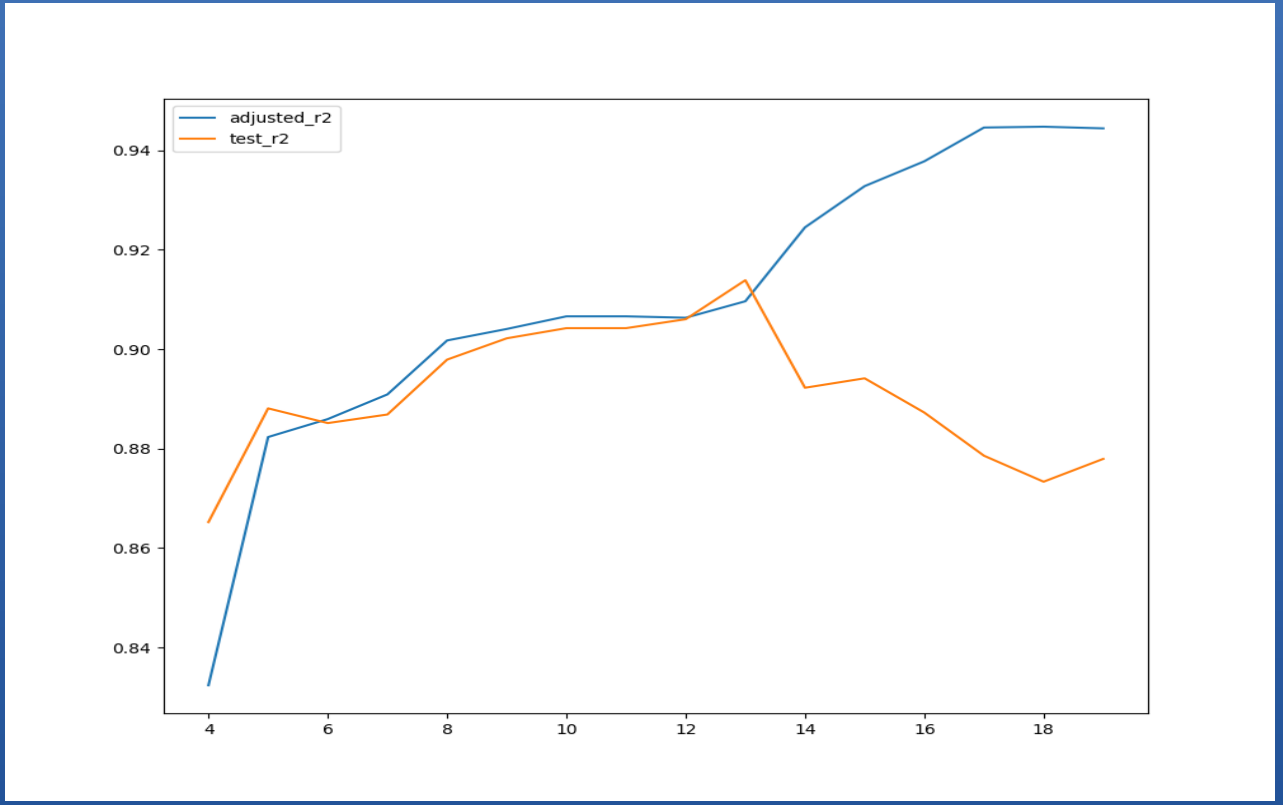
- With the help of split appropriate carname is extracted. Then the orignal feature pertaining to the name of car has been dropped and updated one is added.
- Misspelled data are corrected.
- As per the requirements some of the data are modified with their data types.

Data Preparation

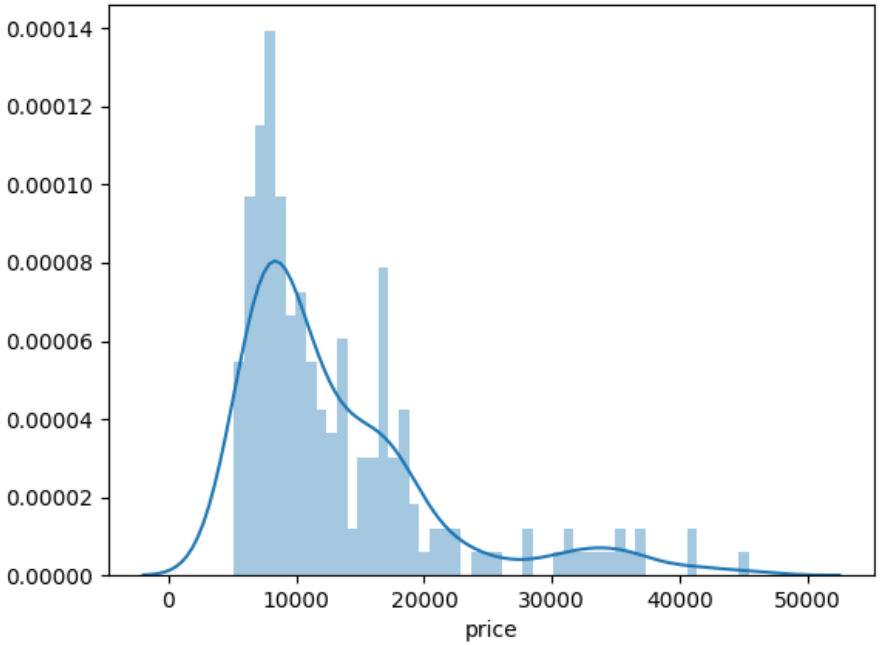
- From all the features present, defined one as the objective and rest as the variables.
- Creating dummies of a feature is a good practice to see relationship of that with the objective in even better way.
- Scaling the data is very much important to remove the unwanted amplification.
- Some part of the data are dectlared as test set and rest are used as training set.

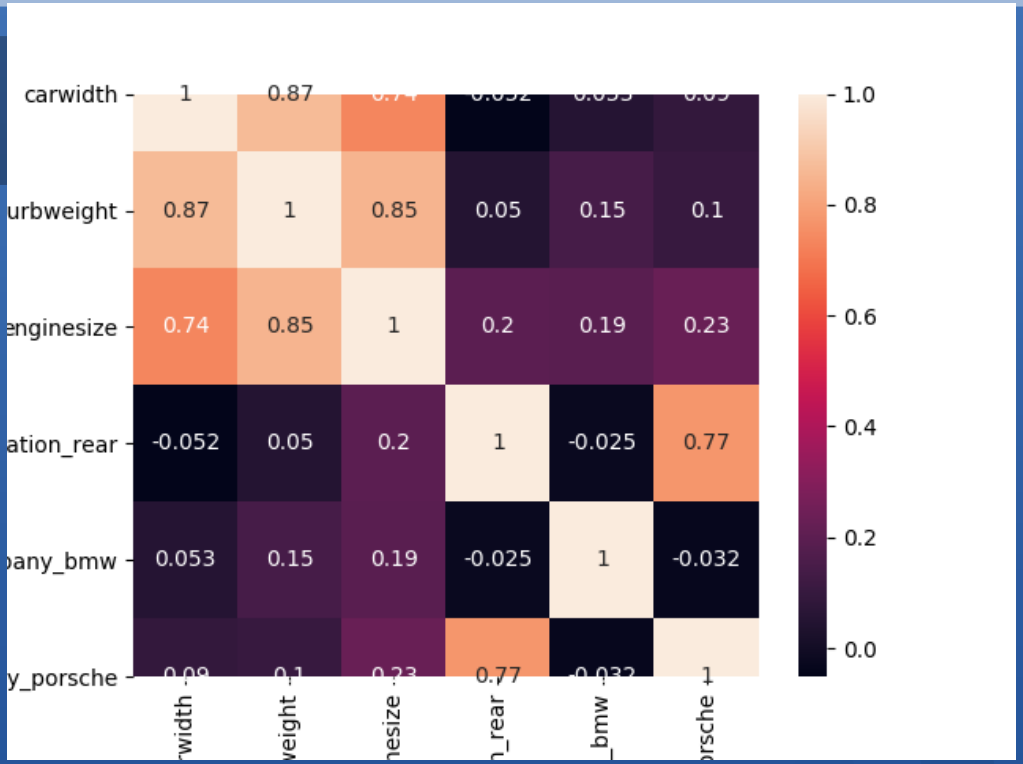
Model Building and Evaluation

- Tried LR with all the features also tried by limiting the number of features with the help of RFE.
- But that gave a whole lot of variation accurace because the adujusted r-squared varies from 93.3 to 88 as we go from 15 to 6 features.
- Now the one way to choose the optimal number of features is to make a plot between n-features and adjusted r-squared and the choose the value of n-feature.



From the above figure we can say that we can choose features between 4 to 12. Let’s choose six features and see the calculation.





This is the final model comprised of six features.

Note that RFE with 6 features is giving about 88% r-squared, compared to 89% with 15 features. Should we then choose more features for slightly better performance

A better metric to look at is adjusted r-squared, which penalises a model for having more features, and thus weighs both the goodness of fit and model complexity. Let's use statsmodels library for this.