Project Report(Regression)

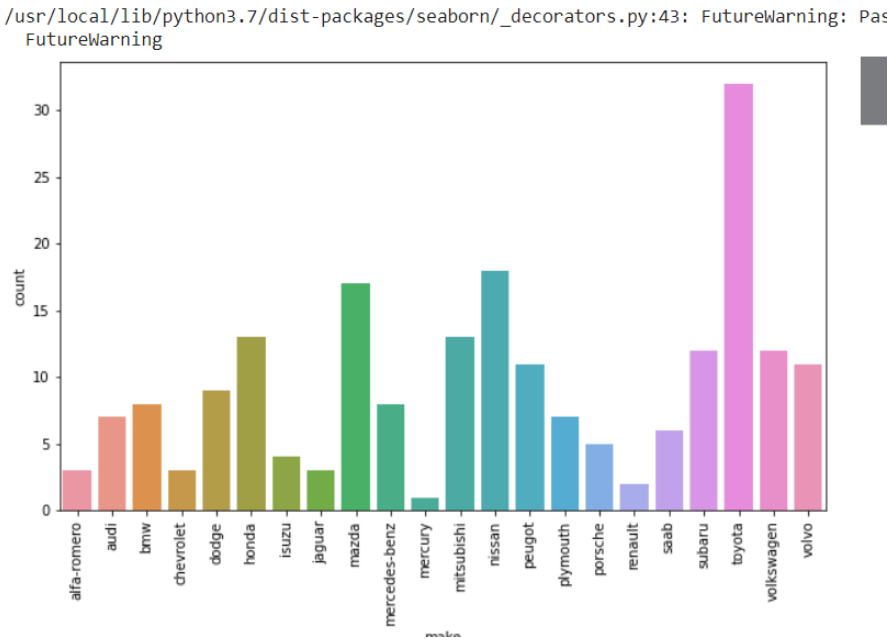
Overview:

The given dataset is a regression dataset containing 206 samples with 25 features. Predict the target price of the car.

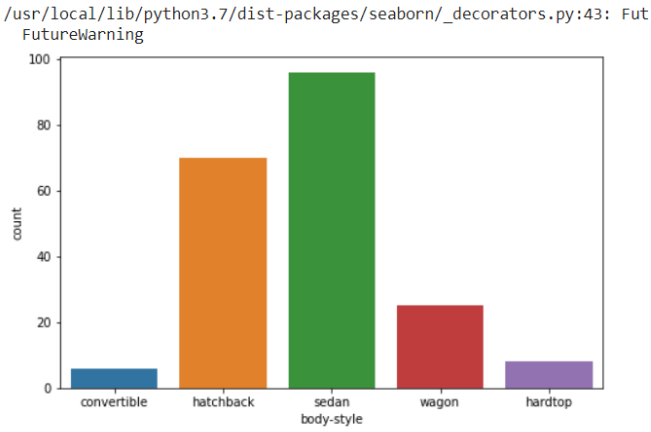
First import required libraries for pre-processing and then load the dataset using Pandas. The dataset contains different datatypes.

Data Analysing:

Finding how many cars each of company (make) possess in the given dataset.



Checking the distribution of cars in data based on their body style using count plot.



By using describe we analysed the numeric features mean ,standard deviation.

There are no null values in th data but there are some attributes filled with '?',which are to be handled.

Data Pre-processing:

1. Converted '?' values in data into Nan values.

2. Changed the datatype of the numeric columns from object to float, in order to perform operations flawlessly.

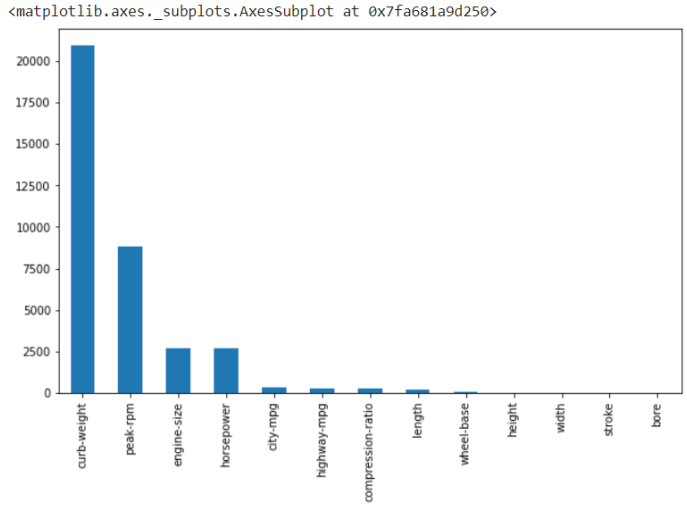
a. Filled the missing values in columns num-of-doors ,bore, stroke by their frequent values.

b. Filled the missing values in columns peak-rpm, horse-power by their mean values.

c. Filling missing values in price column(target) may effect the target distribution ,hence dropped the rows containing missing price values.

d. Now for column normalized-losses there are 37 missing values. Because of smaller number instances in the data, if they are replaced by any the normalized-losses column will be skewed by increasing value counts of particular value. Hence the feature is dropped.

Checked the importance of numeric data using Chi-squared coefficients.
Plotted the barplot with important features from left to right(i.e,descending)



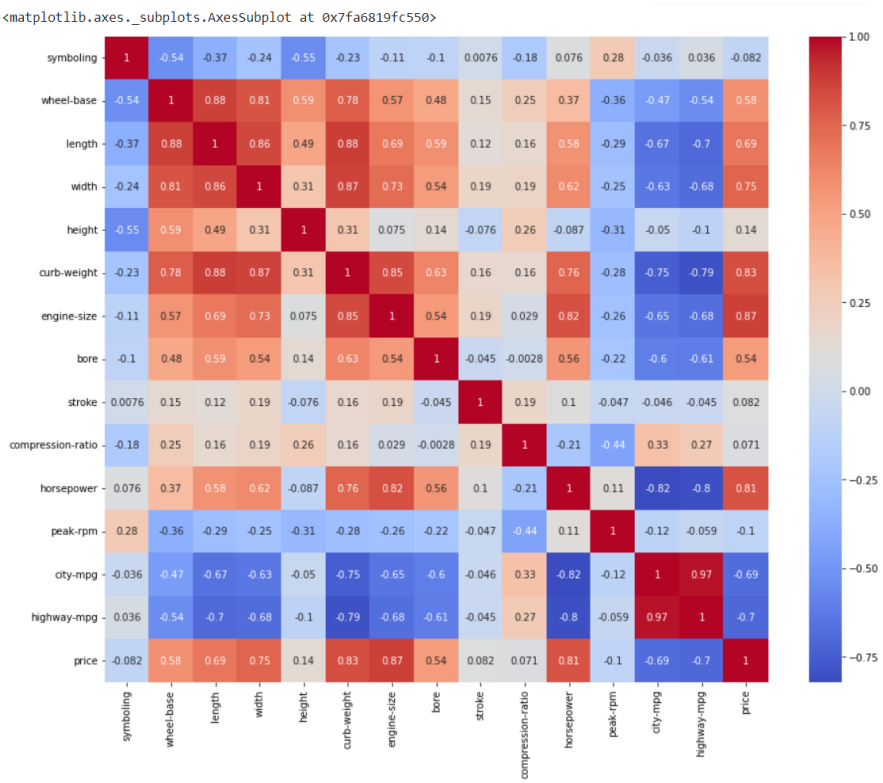
From the above plot we got the important numeric features in the data, which are curb-weight , peak-rpm, engine -size, horse-power.

Plotted a heatmap to gather the correlation of the numeric features using Pearson's Correlation coefficient.

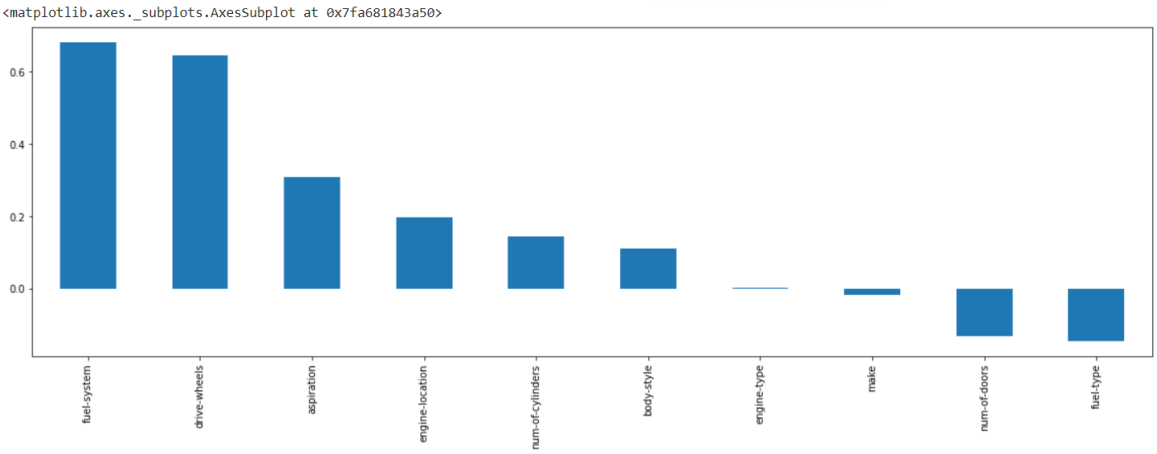
From the below plot:

- The features length, width, wheel-base, curb weight are highly correlated with each other.
- compression-ratio and stroke doesn't have any correlation with the price.
- engine-size and horse power are highly correlated with each other.
- city-mpg and highway-mpg are also high correlated.

Hence from above we can drop length,width,wheelbase,compression-ratio,stroke,horsepower,city-mpg from the data without effecting the target variable.



Below is a bar-plot plotted representing the increasing order of the categoric feature importance which is found out by the spearman's correlation method.



From the above plot the least values plotted at engine-type, make. Hence we can drop these features.

3.Hence by considering the above correlations we dropped the following columns from data:

columns=['symboling','make','engine-type','length','width','wheel-base','horsepower','stroke','compression-ratio','city-mpg','engine-location']

4. converted the categoric features into dummies by pandas dummy(or onehotencoding) and stored in new data-frame.
Now joined the data-frames dummies and numeric data.
5. From box plot we found there are outliers in curb-weight, peak-rpm, highway-mpg.
Using Inter Quantile range method remove outliers i.e,

Q1=data for quantile 0.25

Q3= data for quantile 0.75

IQR=Q3-Q1

Upperlimit= Q3+(1.5*IQR)

Lowerlimit=Q1-(1.5*IQR)

The data points beyond these limits are outliers which are dropped from columns mentioned before.

6. Spliting the data - Using train_test_split we split 20% of data as test data and remaining for training.

7. Feature Scaling

Features are scaled using standard scaler because when we perform certain algorithms based on Euclidean or Manhattan distance there will be deflection in the model because of the farthest data points. Hence we scaled them by standardisation

New_data=(data-mean)/standard deviation

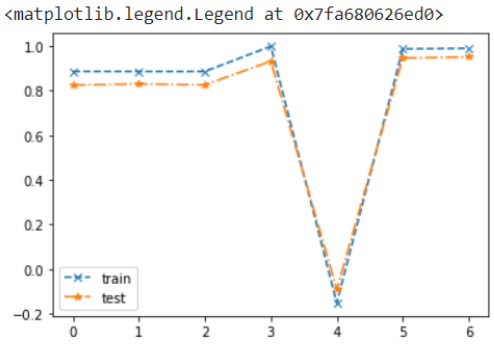
Modeling :

By using different regression models train the data to get better scores. These includes:

Linear Regression, Lasso(alpha=0.01),Ridge, Decision Tree Regressor, SVR, Random Forest Regressor, Gradient Boosting Regressor.

Now trained the data using all the above models with their default parameters.

The scores obtained for train and test data plotted in order to select better model.



Here,

0-Linear Regression, 1- Lasso(alpha=0.01),2-Ridge, 3-Decision Tree Regressor, 4-SVR, 5-Random Forest Regressor, 6-Gradient Boosting Regressor.

Therefore the better model we obtained is Gradient Boosting Regressor.

Now tune the parameters of the model to gain better accuracy.

Tuning the parameters of model:

Using Gridsearch CV parameters of gradient boosting regressor are tuned as below.

```
#Hence tune the parameters of gradient boosting regressor for further better scores.

from sklearn.model_selection import GridSearchCV
parameters={'loss':['squared_error','absolute_error','huber','quantile'],'learning_rate':[0.1,0.2,0.3,0.9],
            'n_estimators':[20,50,100],'max_depth':[3,5],'criterion':['friedman_mse','squared_error']}
gs=GridSearchCV(GradientBoostingRegressor(),param_grid=parameters)
gs.fit(X_train,y_train)
```

```
GridSearchCV(estimator=GradientBoostingRegressor(),
             param_grid={'criterion': ['friedman_mse', 'squared_error'],
                         'learning_rate': [0.1, 0.2, 0.3, 0.9],
                         'loss': ['squared_error', 'absolute_error', 'huber',
```

Like wise parameters are tuned and got the best parameters as:

{'criterion': 'squared_error', 'learning_rate': 0.2, 'loss': 'absolute_error', 'max_depth': 5, 'n_estimators': 100}

Final-Model:

By using above parameters we build a final model for the data .

We got the metrics values as:

```
R2-Score:  0.9411341071946313
Mean-Absoute-Error:  1278.896068220435
Mean-Squared-Error:  3532759.112598464
```

Feature importance:

The importance of different features are plotted in bar plot during RFE.

From the above correlation concepts discussed in Pearson's correlation method, Spearman's Correlation method. The important features for the dataset are:

a. Curb-weight, engine-size are important, because of having correlation values with regarding target variable price. Bore also posses the positive correlation with Price.

b. From categorical features fuel-system, drive-wheels have good importance plotted in spearman's method.