1. Remove unwanted columns that are not necessary for prediction like names, ids, lot no., etc. Any data that is specific to identification of that particular data.
2. Remove duplicates by using drop\_duplicates(df) from pandas library.
3. Handle missing values:
   1. If less than 5% data is missing for a column, ignore those rows
   2. For categorical data use mode
   3. For numerical data: plot boxplot, if too many outliers for a particular data use median, else use mean to fill the missing values
4. Using heatmap and select k best features, find if we need to remove a column.
5. Plot boxplot and check which columns to remove outliers from.
6. Apply encoding on categorical data:
   1. One hot encoding
   2. Ordinal Encoding
   3. Binary Encoding
   4. Label Encoding
7. Split the data into train and test.
8. Remove outliers from training and testing from columns we got from point 5.
9. Split the data into xtrain, xtest and ytrain, ytest.
10. Find which of the numeric columns to apply standardisation and normalisation and apply on train and test both.
11. Fit your regression model using xtrain and ytrain
12. Get predicted values using fitted regression model.
13. Plot the y\_pred vs y\_test(convert y\_test into numpy array while plotting)
14. Calculate the median absolute percentage error(MDAPE) using y\_pred and y\_test
15. Calulate accuracy = 1-MDAPE

Load csv with df = pd.read\_csv("/content/drive/MyDrive/car\_prices.csv.zip", on\_bad\_lines='skip')

For 70% of necessary/mandatory data missing in a row use the function

Read on whether to remove outliers column by column or whole at once

Kfold splitting

Gridsearchcv

Depending on data, refer excel sheet to choose regression model

What all to do on test data to predict