Data Preprocessing & Insights:

- Convert Inspection Time to Python datetime format
- Calculate age in months for the engine using Inspection Year/Month and Registration Year/Month
- Calculate dummy variables for multiple readings of the engine variables viz. *engineTransmission_engineOil_cc_value, engineTransmission_engine_cc_value, engineTransmission_coolant_cc_value, engineTransmission_engineSound_cc_value, engineTransmission_clutch_cc_value, engineTransmission_gearShifting_cc_value* etc.
- One dummy variable for each unique value of the above-mentioned variables. Multiple dummy variables for the same variables can have a value of 1.
- Create dummy variables for categorical variables like *fuel_type, engineOilLevelDipstick, exhaustSmoke, engineBlowByBackCompression*
- Very small proportion (<10%) of cars have small ratings (<2.5). 4 is the most common rating; around 40% records have that rating.
- Very few cars have an age greater than 15 years or 180 months. Around 16% (maximum for an age bin) are new cars with age < 20 months
- Around 80% cars have travelled <100,000 kms

Methods:

- Created multiple dummy variables for the multiple value categorical variables and normal categorical variables
- Since number of unique Ratings is fixed and is also given at intervals of .5, formulated this as a classification problem
- 2:1 split for train-test data; no outliers removed as removing them worsened the MAE
- Xgboost model *(n_estimators = 300, max_depth = 5, learning_rate =0.05)* fitted to the train set and predictions made on the test set. Other parameters tried but they didn't improve the result.
- Mean Absolute Error *(mean(abs(ypred - ytest))* used as a performance metric
- Other methods tried – Multi-class Logistic Regression (MAE = 1.16), XGBRegressor (MAE = 0.44), XGBClassifier with additional options (MAE = 0.37 and 0.38)
- Multi-class logistic regression being a linear model (although very weak) is more interpretable. Can be used to get a sense of how each variable impacts (increases or decreases the chances of a particular prediction) the ratings prediction

Model Results:

- A MAE of ~0.35 was obtained indicating that predicted ratings differ by around .35 from the actual ones

- *engineTransmission_engineSound_cc_value_TN (Timing & Tappet Noise), engineTransmission_engineOil_cc_value_Dirty, engineTransmission_engine_cc_value_Repaired, engineTransmission_coolant_cc_value_Dirty* are top-5 important features

- The trained finalized model saved as a pickle file and is attached

Possible Next Steps:

- Feature Selection
    - Keep only top-10 or 15 important variables
    - Remove highly correlated variables
- Other models
    - Multiclass SVM as the data is sparse because of lot of dummy variables
    - kNN with Gower distance as the distance measure
- Thorough validation & Fine tuning
    - Cross Validation
    - Grid Search Optimization for Xgboost parameters