# Required Answers

Ans1. The approach consisted of cleaning the dataset as it had data with columns which were redundant. Also the formatting of data needed a slight touch up. The model was built according to the problem's definition, analysing all the outliers and understanding the need of dependent variables to be interpreted with abundantly present columns with classified values. The model best fit was XGBoost Tree algorithm but here I only performed basic regression model as the problem demanded to be more centric on outliers and analysis.

Ans2.

a. The model with MLR at max give and MSE of 0.62. These models do not work a perfect fit for the data but do provide us with some insights of how the model should be developed with further evaulationand correction of metrics. I could determine that the best possible models were either Ordered Logits if we used regression or XGBoost Tree algorithm.

b. It will work well for sure if we give some time to tuning the hyperparameters as the model is very sensitive and not thoroughly correlated with all the variables.

c. I used MSE and R Square for predictive accuracy.

d. There were numerous issues I faced during the outlier analysis and exploratory data analysis as the model had no other variables than “odometer_readings” and “year” that provided a good correlation with the dependent variable of the model. Neither negative nor positive. The data also had the discrete rating column which can be solved using numerous ways. So, reaching to a conclusion of XGTree Boost being the best choice for this model required some effort.

e. We could say that the engines with Hybrid fuel_type were with the least ratings. The year and dependent variable had a positive correlation, that means when the year increased so did the values of the rating. The odometer reading and dependent variable had a direct negative correlation, that means when the odometer readings increased,the value of ratings decreased. These two variables ‘year’ and ‘odometer_readings’ were highly correlated with the dependent variable and hence the most important predictors.

Ans3. There could have been numerous data about the wear and tear of the engine which could have been used but also numerical data like the life of the engine and fuel type related life is also detectable as diesel engines have a lesser life than petrol engines, number of accidents, number of services etc. Some of these variable could have acted as constraints or good predictors but with a good and strong correlation for determining the output.

Ans4.I would have used more time to identify more anomalies in data and specifically used that time to build a good model for the data as it took me a little more than usual time to determine the best possible methods to find the prediction model. I came up with the conclusions that Ordered Logits for Regression and XGBoost Tree algorithms were best fit for creating this model. But, lack of a few more minutes to an hour or so couldn't let me finish with that deterministic approach of mine. Therefore, with more time I would have created a better model using the above mentioned algorithms and then would have used the time to evaluate the model and refine it using the GridSearchCV and probably would have also used Lasso and Ridge models for variable selection and regularization.