**Project Problem Part - 1 – Linear Regression**

1. Should Outlier treatment be done for this question?  
   *Ans: An observation is considered to be an outlier if that particular observation has been mistakenly captured in the data set. Treating outliers sometimes results in the models having better performance but the models lose out on generalization. So, a good way to approach this would be to build models with and without treating outliers and then report the results. On the other hand, it is perfectly fine if you are building your models only once, i.e. either after treating or not treating the outliers.*
2. Should both ‘sklearn’ as well as ‘statsmodels’ be used for building the Linear Regression model?  
   *Ans: It is perfectly fine to use either of these libraries to build the Linear Regression model and report the output in the business report. However, you can choose to build the Linear Regression model using both the libraries.*
3. What methods are suitable for dealing or checking the high multicollinearity in the data?  
   *Ans: Multicollinearity is the presence of a strong correlation between the independent variables. Using correlation heatmap is a way to visually understand the degree of correlation present within the various variables. Pair plots are also a great way to understand how the various variables (taken two at a time) behave with each other. Lastly, you can check the Variance Inflation Factor (VIF) values of the various independent variables and draw a conclusion about the multicollinearity problem. If you feel that there is still some extent of multicollinearity present in the model (which will be present even after all of these treatments), you can always perform Principal Component Analysis. But do remember that the interpretability of the Linear Regression model becomes tedious when we fit the new independent variables (after performing PCA) into the Linear Regression model.*
4. What kind of encoding procedure needs to be used in order to encode the Categorical variables?  
   *Ans: For the ordinal categorical variables, please do make sure to encode the variables in a way which will highlight the rank order that is mentioned in the problem statement. If there is any categorical variable which is nominal, please do proceed with dummy variable encoding if that is going to help in the model building procedure. Please do check the documentation of a particular function from any libraries.*
5. Can any kind of transformation be used for any of the variables before fitting them in the Linear Regression model?  
   *Ans: Yes, please do go ahead with appropriate transformations which might make the association between independent and dependent variables linear in nature. After plotting to scatter plots between the various independent variables and the dependent variable, you can take a decision whether a transformation is needed. But it is in no way mandatory to perform a mathematical transformation of a variable before fitting them into a linear regression model.  
   If a particular mathematical transformation has been applied to the dependent variable and the model has been trained using the transformed variables, please do make sure to take the inverse transformation while calculating and reporting the Root Mean Squared Error value. Please be careful about the interpretation of the model after the transformation.*
6. Should we scale the data and then split the data into training and test set or vice-versa?  
   *Ans: If you are looking to scale the data for a non-regularized parametric algorithm, it is better to split the data into train and test and then proceed with the scaling operation. But a quick question here would be to stop and think whether scaling is necessary for the Linear Regression model.*
7. Is there a particular range for Root Mean Squared Error (RMSE) or does it depend on the magnitude and the scale of the variables in the model?  
   *Ans: RMSE does have any pre-defined range within which it has to lie. It can take any values based on the dependent variable.*
8. What method is to be used for the imputation of missing or null values?  
   *Ans: In this case, the mean or median values can be used for imputing the missing or null values of the continuous numerical variables and the mode can be used for imputing the missing or null values in the categorical variables. However, please do feel free to try out any other imputation method that might seem more appropriate but do provide adequate explanations in the business report for the kind of imputation technique used.*

**Project Problem Part - 2 – Logistic Regression and Linear Discriminant Analysis**

1. Should Outlier treatment be done for this question?  
   *Ans: An observation is considered to be an outlier if that particular observation has been mistakenly captured in the data set. Treating outliers sometimes results in the models having better performance but the models lose out on generalization. So, a good way to approach this would be to build models with and without treating outliers and then report the results. On the other hand, it is perfectly fine if you are building your models only once, i.e. either after treating or not treating the outliers.*
2. In this problem, can a new variable be created/engineered based on the given independent variables?  
   *Ans: Yes, new variables can be derived/created/engineered based on the given independent variables. Clubbing a few categories of the categorical variable is one way to create new variables. However, any kind of feature engineering can be done if the proper explanation of the same is given in the business report. But it is not at all mandatory to perform feature engineering for this project.*
3. If the model accuracy (or other model performance parameters) of the Logistic Regression model and the Linear Discriminant Analysis model is less, how do we improve it?  
   *Ans: To improve the model performance, hyper-parameters can be tweaked until a suitable combination of hyper-parameters is found which gives the most optimum performance. Some kind of feature engineering (like outlier treatment or encoding of categorical variables amongst other things) can be looked at to improve the model performance.*
4. Should scaling of the variable be done before fitting the Logistic Regression model and the Linear Discriminant Analysis model?  
   *Ans: No, the data should not be scaled before fitting into either of the Logistic Regression model or the Linear Discriminant Analysis model. The same has been mentioned in the rubric for this particular question.*
5. Do we need to compare the model performance for both training and test for both Logistic Regression and Linear Discriminant Analysis?  
   *Ans: Yes, the comparison of model performance over both training and test is necessary. This tells us how well the model generalizes on unseen data. All possible model performance measures should be calculated and compared on the training and test data across both Logistic Regression and Linear Discriminant Analysis.*
6. What kind of business recommendations and business insights should be put in the report?  
   *Ans: It is expected that the EDA results and the model results are summarized and put in a proper format. Any additional pointers that might be necessary for the business to understand from the data. The models need to be evaluated in terms of the various model performance parameters both on the training and the test set.*