Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

Answer: I got the following relationship between the target ('cnt') and predictor variables where 'sat' and 'clearWithFewClouds' (2 out of 3) predictor variables are categorical in nature. The co-efficient of these variables is low compared to 'registered' numerical variable hence can conclude that categorical variables definitely have effect but not as high as numerical variable 'registered'.

cnt(demand) = -0.0132 + 0.9862*registered + 0.1054*sat + 0.0241*clearWithFewClouds.

2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)

Answer: Though I have not used this parameter when getting dummy variables for reasons stated in the file 'bpatil_BikeSharingAssignment.ipynb' but it is important to drop one variable to avoid multicollinearity i.e. high correlation of predictor variables between themselves. If there are n states of categorical variables then n-1 dummy variables are enough to define all states of the said categorical variables

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Answer: The numerical variable 'registered' seems to have highest correlation with the target variable 'cnt' as per the pair plot details.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Answer: Did the residual analysis and found out that the distribution of error\residue is normally distributed and the mean of residues is centered very close around zero.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Answer: I got the following relationship based on Liner Regression model indicating the 'registered', 'sat'(weekday) and 'clearWithFewClouds'(weathersit) are the 3 features contributing significantly to the demand ('cnt') of the shared bikes.

cnt(demand) = -0.0132 + 0.9862*registered + 0.1054*sat + 0.0241*clearWithFewClouds.

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Answer: The Linear Regression algorithm tries to find the best fit line (relationship between target and predictor variables) using the training dataset such that the relationship is so generic that it is able to predict to a high accuracy the target variable value based on input test or any other dataset. The best fit here indicates that the difference(residue\error) between predicted value using linear regression model and actual value of the target value is nearly equal to zero. All of this can be mathematically put as residual\error values are normally distributed (bell curve shape) with the error mean being almost or very close to zero. The R^2 value also known as R Squared value of the model indicates the ability of the model to explain the percentage of the variance of the target data based on predictor variables using the LR model. The higher the R^2 value, the better is the efficacy of the model.

2. Explain the Anscombe's quartet in detail. (3 marks)

Answer: It is nothing but analysing the data visually using graphs and charts to find any trends and outliers in the dataset as part of the exploratory data analysis done at the initial stages of Linear Regression before building the actual model. The EDA done as part of Anscombe's quartet gives out cues on what the final model might look like and also if the dataset has the necessary qualities to fit a linear regression model or not.

3. What is Pearson's R? (3 marks)

Answer: The Pearson's r is a co-efficient whose value can range between -1 to +1. If this value is negative or between say -1 and 0 for a pair of variables means that increase in one variable leads to decrease in another variable. If it is between 0 and +1 indicates that increase in value of one variable leads to increase in value of the other variable and vice-versa. If this value is zero indicates that the two variables do not have any impact/correlation with each other and increase or decrease in value of one variables does not have any impact\trend on value of other variable.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Answer: Scaling is performed to bring the value ranges of the variables in similar range hence making the situation ideal for the Linear Regression analysis where Gradient Descent is used to find the best variables for linear regression model. It helps in faster elimination of unwanted variables leading to faster or better performance of the model. Normalization and Standardization are two methods used for scaling with Normalization being the most preferred method which is based on min and max value of the variables.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Answer: The higher the VIF(Variable Inflation Factor) value, the more easily the said variable is definable with the other predictor variables. Infinity value for VIF for a given variable indicates

that it has perfect correlation with the other predictor variables and hence a very apt candidate to be dropped from model for avoiding multicollinearity. The formula for VIF is 1/(1-R^2) and when the R^2 or the R squared value is 1, the VIF get infinity as value. R^2 value of 1 is practically impossible and may lead to a case of overfit model where the model fails to generalise and make good predictions on any data other than training dataset.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)

Answer: The Q-Q plots are used to plot the residual values and find out if they are normally distributed with mean around zero which is the basic condition for the model to qualify to be used in predictions using linear regression algorithms. The Q-Q plot may give misleading inferences when the sample size is small.