Assignment-based Subjective Questions

Question 1 **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?**

Answer 1 Categorical variables which are part of the final Model are 'year', 'holiday', 'summer', 'winter', 'cloudy', 'lightrain' and 'september'. In final model among all variables 'holiday', 'lightrain' and 'september' are having negative coefficients, which means they are inversely related with the dependent variable. Increase in these variables will lead to decrease in the dependent variable. On other hand 'year', 'summer', 'winter' and 'september' were having positive coefficients and were directly related to the dependent variable.

It can be inferred that due to lightrain and cloudy weather conditions people less likely to travel through bikes. On other hand as year increased from 2018 to 2019 there was great impact on Total Rentals. Also people like to travel through bikes in summer and winter seasons and avoid bike riding in spring season.

Question 2 **Why is it important to use drop_first=True during dummy variable creation?**

Answer 2 While creating dummies for Categorical Variables we use command of drop_first=True. It drops the first column of the created dummy dataframe. When we create dummy for a categorical variable with 'N' categories, we get 'N' columns and each column defines each of the category.

The first column/category among 'N' categories which was dropped can be explained by absence of all other categories. Suppose we create dummies for season categorical variable with spring, fall, summer and winter as season. When we create dummies, we simply dropped spring category. Absence of other categories like fall, summer and winter will automatically explain the model about spring category. Therefore, to avoid overcrowding of variables, we use the command drop_first=True for categorical variables.

Question 3 **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**

Answer 3 Apart from Registered and Casual variables, temp was showing the highest correlation with the target variable (cnt). Registered and Casual are part of target variable therefore, they will have high correlation with target variable. Apart from them temp was showing maximum positive correlation.

Question 4 **How did you validate the assumptions of Linear Regression after building the model on the training set?**

Answer 3 There are four linear Regression assumptions and each was validated by following

1. Linear Assumption between variables – Before building the model, scatter plot and correlation matrix between various variables were seen and found to be linearly related with target variable (cnt). Variance Inflation factor were observed for variables in final model to avoid multicollinearity which can violate this assumption.
2. Target variable continuous in nature – Target variable was describing demand on bikes and was continuous in nature, therefore it also satisfies the linear regression assumption.

3. Errors terms are normally distributed – We checked this by performing univariate analysis of error terms or plotting a distplot/histogram to check the distribution.
4. No Relation between error terms – To observe this, the residuals were plotted using line plot and no trends were observed in the error/residuals.
5. Homoscedasticity – Predicted values were plotted against the Residuals and almost constant variance was observed between error terms.

As all the assumptions of Linear Regression were satisfied, it was concluded that Linear Regression was an appropriate model of this analysis.

Question 5 **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes**

Answer 5 9 variables were found to be part of the final model. As all variables were scaled their value of coefficient was used to determine their contribution towards change in target variable. Among these 9 variables **'temp', 'weather_lightrain', 'yr'** were the top 3 variables. 'temp' and 'yr' were having high positive coefficients and good positive relation with target variable whereas 'weather_lighrain' was having negative coefficient and good negative relation with target variable.

General Subjective Questions

Question 1 **Explain the linear regression algorithm in detail.**

Answer 1 Linear Regression is a type of Supervised Learning. As name suggests 'Linear' means the dependent variable and independent variables should have linear relation and 'Regression' signifies that the target variable being continuous in nature other variables can be used to determine the value of the target variable.

We use following equation for Linear Regression where y is the dependent variable and x_1 to x_n are the independent variables. Similarly where β_o represents the intercept or bias term and β_1 to β_n are coefficients/slopes of the independent variables.

$$y = \beta_o + \beta_1 x_1 \dots \dots . + \beta_n x_n$$

We use independent variables to predict the dependent variable in Linear Regression Following are assumed for Linear Regression: -

1. Independent and Dependent variables should have linear relation.
2. Dependent Variable should be continuous in nature.
3. Error terms are normally distributed.
4. Errors or residuals observed should not be related with each other.
5. Error terms should have constant variance or homoscedastic in nature.

Question 2 **Explain the Anscombe's quartet in detail.**

Answer 2 Anscombe's quartet is a made up of 4 data sets comprising of X and Y values. These 4 datasets were having same statistical properties like variance, standard deviation etc. This quartet signifies the importance of graphical representation of X and Y values. Although best fitted line passing through the 4 data sets was similar in nature, but their graphical representation(scatterplot) was quite different. This shows of importance of graphical visualization before analysis

In Anscombe's quartet there are 4 datasets as follows.

1. First data set have X and Y values simply spread and line passing through them signifies the linear relation.
2. In second dataset X & Y were observed to be not having linear relationship, but when best fitted line was drawn, it was same as line from first dataset.
3. While visualizing third dataset, it can be seen that how slope of the line got effected by the one outlier.
4. In fourth data set there was no relation between X and Y, but due to one data point there was high linear relation was observed between X and Y values.

This illustrates the importance of graphical analysis.

Question 3 **What is Pearson's R?**

Answer 3 Pearson's R is a measure of strength of linear Relation between the variables. It simply tells about the linear change in 1 variable with respect to change in another variable. Its value lies in the range of -1 to 1. It can be calculated by finding the covariance between variables and then dividing them with the product of their standard deviation.

It only tells about the linear relation between the variables. It simply ignores the other types of relations and gives the measure of linear relation only. This can be considered as its limitations.

Question 4 **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**

Answer 4 Scaling means bring all the variables to a common scale to avoid one variable dominating over the others without any loss of information.

It is not necessary that all the variables are present on same scale. While building a model, unscaled variables will lead to their coefficients present on different scales. This makes the interpretations of the coefficients much difficult e.g. variable present in smaller scale might have large coefficients on other hand variable with larger scale have small coefficient. Suppose they have similar effect on target variable, but their interpretation becomes very difficult and their relation with target variable will not be easy to compare because of their coefficients. Scaling also helps in faster convergence for gradient descent methods.

We can use two types of scaling methods normalization and standardization.

In normalization also know as min max scaling we use max and min values to scale a variable. In this scaling all the variables comes to value between 0 and 1.

$x' = (x - \min(x))/(\max(x) - \min(x))$

In standardization we use mean and standard deviation of the data to scale the features so that all data points follow normal standard distribution. In this scaling features comes in range of -1 to 1.

$x' = (x-\mu)/\sigma$

Question 5 **You might have observed that sometimes the value of VIF is infinite. Why does this happen?**

Answer 5 VIF or variance inflation factor is a measure of relation between independent variables. How an independent variable or combination of independent variables explains the other independent variable.

$VIF = 1/(1-R_i^2)$

R_i^2 represents the r squared value for the prediction of independent variable by other independent variables. If this value becomes 1, that means one independent variable

can be completely explained by one or more other independent variables. When R_i^2 value becomes 1 then VIF value will tend towards the infinity value. It signifies the a direct one to one relation between variables.

Question 6 **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.**

Answer 6 Q-Q plot or Quantile-Quantile plot is a graphical method of comparing two probability distributions or comparing 2 sets of quantiles against each other.

It is used to compare the shapes ,location ,skewness and scale of two distributions graphical means. More often Q-Q plot is used for two theoretical distributions with each other

Q-Q plot provides an assessment using graphical summary.

if we run statistical analysis that assumes that our dependent variable is following a normal distribution, then we can use a Q-Q plot to check this assumption. If 2 quantiles are from same distribution or 2 distributions compared are similar the points plotted on Q-Q plot will always be non-decreasing and plot generated will be a straight line having an equation of straight line passing from 0,0 and slope of 45° (y=x).