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)

While performing EDA, In order to analyze the effect of categorical variables in our dataset two types of graphs are drawn.

1> Observation through Boxplots :-
   a. With the help of box plots we could figure out the variance among medians derived category wise towards our target variable. We have found major variance in the variables “weathersit”, “season”, “yr” etc.
2> Observation through Barplots :-
   a. I have built sub-category wise barplots too to get better clarity and numerical information. The plots represent category wise variation in sum of ‘cnt’ column in percentage for the categorical variables having almost even occurrence (like month, year, season, day etc.); for others (like holyday, workingday etc), I have normalized the occurrence by finding group-wise sum of ‘cnt’ in percentage per unit quantity which clearly shows the influence of that variable on target variable.
3> Changed in the values of R2 and Adjusted R2 when the model was being built:-
   a. It was noticed the each time of addition of the above variables, the R2 and Adjusted R2 were changed significantly.

   The answer is written based on the mathematical analysis done in note book and numerical data is available for reference.

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

Number of dummy variables are created depending on how many different levels are there in a categorical variable and each dummy variable is Boolean in format. So, if the is n dummy variables for a categorical variable, the last one will be the complement set of the union of first (n-1) dummy variables. That means the nth feature is highly correlated with the union of first (n-1) th features; the nth feature is considered automatically at the time of modeling. So, they are dropped, as we will not allow a redundant feature for good model and using drop_first=True helps us minimize the length of code and increase the efficiency.

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

Before dropping any variables ‘registered’ has got highest correlation with the target variable ‘cnt’. After dropping useless variables, it is found that ‘temp’ has the highest correlation with the target variable ‘cnt’.

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

Validation is required to interpret how good the model is. We follow the following steps:

1> Assumptions1-The dependent variable and the independent variables have linear relationship: Checking the value of coefficient of determination (R^2) we can conclude about the assumption. By this we can understand the percent of variance in observation explained by the model. F -value expresses overall how much of the model has improved (compared to the mean) given the inaccuracy of the model

2> Assumptions2-No multicollinearity or very low multicollinearity: The process is completed by checking VIF (Variance Inflation Factor).
   a. If VIF=1, Very Less multicollinearity
   b. VIF<5, Moderate multicollinearity
   c. VIF>5 , Extreme multicollinearity (This is what we have to avoid).

3> Assumptions3-Errors or Residuals must be normally distributed and sum of them is almost zero: We validate this by drawing distribution plot with the residuals.

4> Assumptions4-The variable considered are significant. We check if the p-value is <0.05 or not.

5> Assumptions5-The Error Term should be Homoscedastic. Using point plot, we can validate that.

6> Assumptions6- Error Terms should be independent. The lag plot of the residuals, another special type of scatter plot, suggests whether or not the errors are independent.

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

Based on the final model, the top three features contributing significantly towards explaining the demand of the shared bikes are: 1> “temp” 2>”yr” 3>” weathersit_Light_Snow_Rain” (dummy variables derived from “weathersit”).

General Subjective Questions

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

The linear regression algorithm is a statistical, linear, predictive algorithm which follows regression to set a linear relationship between the dependent and the independent variable. It comes up with a line of best fit, and the value of Y (variable) falling on this line for different values of X (variable) is considered the predicted values.

1> Understanding Different type of Regression (SLR, MLR, Logistic, Step etc)

2> Finding best fitted curve according to different type of regression, for example

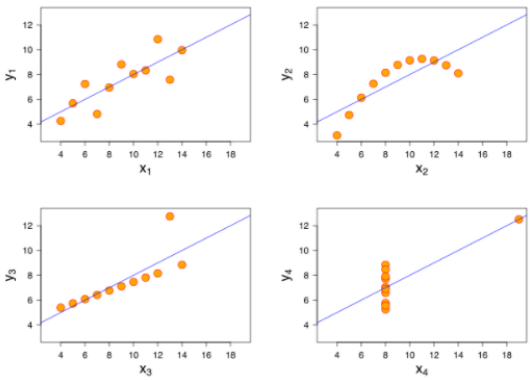
$$Y = b_0 + b_1X + b_2X + \ldots + b_nX \text{ (where n is integer)}$$

For this purpose least square distance is found with the help of gradient descent technology in backend.

3> R-squared, RSS and TSS are calculated.
4> Adjusted R-squared are calculated.
5> P values of variables (independent) are calculated.
6> In order to avoid multicollinearity, variation inflammation factor (VIF) are calculated.
7> F-statistics are done to measure overall performance.
8> Error analysis and Risk analysis are done.

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

In case of linear regression, Anscombe's Quartet (Created in 1973 by Francis Anscombe) is a set of four datasets that have the same mean, variance and correlation but look very different. The four data sets show how important data visualization is when performing exploratory data analytics. The four datasets have the same descriptive statistics, but appear very different when graphed. The exampled graphs are below.



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

In statistics Pearson's R or Pearson's correlation coefficient is the covariance of the two variables divided by the product of their standard deviations. The form of the definition involves a "product moment", that is, the mean (the first moment about the origin) of the product of the mean-adjusted random variables; hence the modifier product-moment in the name. It is formulated below

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2 \sum(y-\bar{y})^2}}$$

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

Scaling data is the technique by which magnitude of data is increased or decreased by a fixed ratio.

The scaling is performed mainly for two reasons.

(1) To keep a balanced synchronization and the interpretability among coefficients of variables used in linear regression equations.
(2) Scaled data increase performance as some back-end process (like Gradient Descent) run faster.
(3) In normalize scaling data are compressed in 0 to 1 scale it is formulated as

$$\frac{x - \min(x)}{\max(x) - \min(x)}$$

In standardized scale data are compressed with a mean zero and standard deviation 1. It is formulated as

$$\frac{x - \text{average}(x)}{\max(x) - \min(x)}$$

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

This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, So the denominator implies 1-1 =0, which lead to 1/(1-R2) infinity. Normally using one variable two times causes this problems.

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

A QQ-Plot is used to visually determine how close a sample is to the normal distribution. The QQ-Plot orders the z-scores from low to high, and plots each value's z-score on the y-axis; the x-axis is the corresponding quantile of a normal distribution for that value's rank. Since the data is normalized, the units correspond to the number of standard deviations away of the data from the mean.

This helps to assess if a set of data has come from some theoretical distribution such as a Normal, exponential or uniform distribution. Also, by this type of plot it can be determined if two data sets belong to the populations with a common distribution. Normally when training and test data set received separately, we can confirm using Q-Q plot that both the data sets are from populations with same distributions an it can be used with sample sizes also b) Many distributional aspects like shifts in

location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.