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: The categorical variables had below inference:
   Season: Spring had less demand and summer, fall & Winter gave an average demand of above 4000
   Yr: The year variable had 2019 with more demand compared to 2018
   Mnth: April to Nov saw a average demand above 4000 and Dec to March saw below 4000 demand
   Weatherist: When it was Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered the demand was below 4000

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

ANSWER: I did the same mistake initially and had a huge VIF values with INF stating it has high correlations. You have to use drop_first to reduce the extra column created during dummy varible. It makes a case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity, so its important to use drop_first=True to avoid perfect correlation

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

ANSWER: atemp – 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)

ANSWER: Assumptions of LR were validated using Residual Analysis, Y_train_pred was calculated and dist plot with residual was made to understand the mean was cantered at 0

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: The top 3 features contributing significantly towards the demand are as follows:

1. atemp – 0.4403
2. yr – 0.2350

3. winter – 0.0788

General Subjective Questions

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

ANSWER: Linear regression shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis).To calculate best-fit line linear regression uses a traditional slope-intercept form. A regression line can be a Positive Linear Relationship or a Negative Linear Relationship. The goal of the linear regression algorithm is to get the best values for a0 and a1 to find the best fit line and the best fit line should have the least error. In Linear Regression, **Mean Squared Error (MSE)** cost function is used, which helps to figure out the best possible values for a0 and a1, which provides the best fit line for the data points. Using the MSE function, we will change the values of a0 and a1 such that the MSE value settles at the minima. Gradient descent is a method of updating a0 and a1 to minimize the cost function (MSE)

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

Anscombe's Quartet can be defined as a group of four data sets which are **nearly identical in simple descriptive statistics**, but there are some peculiarities in the dataset that **fools the regression model** if built. They have very different distributions and **appear differently** when plotted on scatter plots. It was constructed in 1973 by statistician **Francis Anscombe** to illustrate the **importance** of **plotting the graphs** before analysing and model building, and the effect of other **observations on statistical properties**. There are these four data set plots which have nearly **same statistical observations**, which provides same statistical information that involves **variance**, and **mean** of all x,y points in all four datasets. This tells us about the importance of visualising the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc. Also, the Linear Regression can be only be considered a fit for the **data with linear relationships** and is incapable of handling any other kind of datasets.

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

In Statistics, the Pearson's Correlation Coefficient is also referred to as **Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or**

bivariate correlation. It is a statistic that measures the linear correlation between two variables. Like all correlations, it also has a numerical value that lies between -1.0 and +1.0. Whenever we discuss correlation in statistics, it is generally Pearson's correlation coefficient. However, it cannot capture nonlinear relationships between two variables and cannot differentiate between dependent and independent variables.**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. Pearson's Correlation Coefficient is named after Karl Pearson. He formulated the correlation coefficient from a related idea by Francis Galton in the 1880s.

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

   Scaling is performed as variables are at different scale ina. Data set and the coefficients interpretation becomes difficult after the model building. So to bring them all under the same pane we do scalling.

   The difference between normalised scaling and standardized scaling is as follows:
   Normalization: (x-xmin)/(xmax-xmin)
   Standardization: (x-mu)/sigma

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

I did the same mistake initially and had a huge VIF values with INF stating it has high correlations. You have to use drop_first to reduce the extra column created during dummy varible. It makes a case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity, so its important to use drop_first=True to avoid perfect correlation

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

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution. This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

Advantages:
a) 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.