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)

   The categorical variables which have been feature-engineered to create the final model are as follows. Also, the coefficients have been provided –

   ```
   yr             0.2369
   holiday       -0.0722
   spring        -0.1472
   Light Rain    -0.2453
   ```

   Thus, their effect on the dependent variable 'cnt' can be inferred from the coefficients.

2. Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)
   The dummy variables needed for a categorical variable with n levels is n-1. Hence, we use **drop_first=True** so that the first variable is dropped and we get n-1 variables.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)
   The variables 'temp' & 'atemp' have 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)
   One of the assumptions of Linear Regression is that the residuals should be normally distributed. Hence, we plotted a histogram (distplot) of the residuals and validated the assumptions of Linear Regression.

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

General Subjective Questions

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

   Linear regression is a machine learning algorithm that finds the best linear-fit relationship on any given data, between independent and dependent variables. It is mostly done by the Sum of Squared Residuals Method.

The assumptions of linear regression are:

- There is a linear relationship between the dependent and independent variables.
- It is assumed that the error terms, ε (i), are normally distributed.
- It is assumed that the residuals have a mean value of zero, i.e., the error terms are normally distributed around zero.
- It is assumed that the residual terms have the same (but unknown) variance, σ2.
- It is assumed that the residual terms are independent of each other, i.e. their pair-wise covariance is zero.

•The independent variables are linearly independent of each other, i.e. there is no multicollinearity in the data.

To see if linear regression is suitable for any given data, a scatter plot can be used. If the relationship looks linear, we can go for a linear model.

The Null and Alternate Hypothesis used in the case of linear regression, respectively, are:

$\beta 1=0$

$\beta 1\neq 0$

A linear regression model is quite easy to interpret. The model is of the following form:

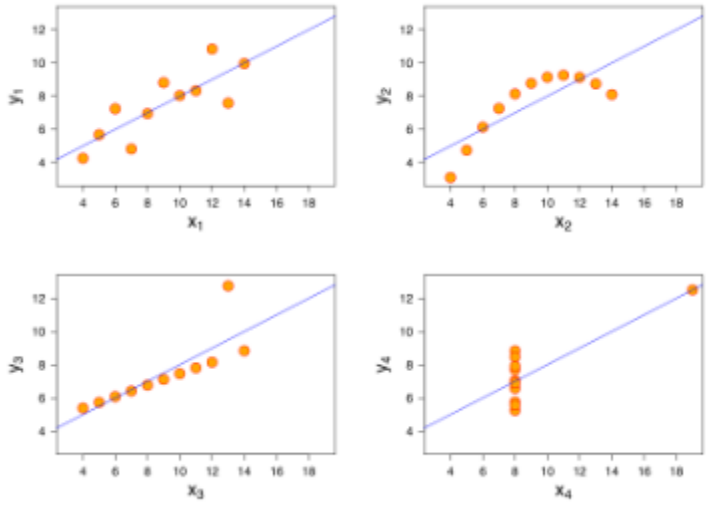
$$y=\beta 0+\beta 1X1+\beta 2X2+...+\beta nXn$$

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

Simple linear regression has quite a few shortcomings:

- It is sensitive to outliers
- It models the linear relationships only
- A few assumptions are required to make the inference

These phenomena can be best explained by the Anscombe's Quartet, shown below:



As we can see, all the four linear regression are exactly the same. But there are some peculiarities in the datasets which have fooled the regression line. While the first one seems to be doing a decent job, the second one clearly shows that linear regression can only model linear relationships and is incapable of handling any other kind of data. The third and fourth images showcase the linear regression model's sensitivity to outliers. Had the outlier not been present, we could have gotten a great line fitted through the data points. So we should never ever run a regression without having a good look at our data.

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

Pearson's R correlation coefficient is the correlation coefficient designed for linear relationships and it might not be a good measure for if the relationship between the variables is non-linear. Although Pearson's R might give a correlation coefficient for non-linear relationships, it might not be reliable. For example, the correlation coefficient given for the relationship

y=X3 for 100 equally separated values between 1 and 100, Pearson's R≈0.91.

For the relationship y=X10 for the same data points, Pearson's R≈0.66.

It might happen that even for a non-linear relationship, the Pearson's R value might be high, but it is simply not reliable.

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

It is a step of data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm. Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units' hence incorrect modeling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

Normalization/Min-Max Scaling:

It brings all of the data in the range of 0 and 1. sklearn.preprocessing.MinMaxScaler helps to implement normalization in python.

$$\text{MinMax Scaling: } x = \frac{x - min(x)}{max(x) - min(x)}$$

Standardization Scaling:

Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).

$$\text{Standardisation: } x = \frac{x - mean(x)}{sd(x)}$$

sklearn.preprocessing.scale helps to implement standardization in python.

One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

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

The formula for VIF = 1/ (1-R^2). Hence, the value for VIF is infinite when the value of denominator i.e. (1-R^2) is 0. This means that the value of R is 1. When the variance of one variable is completely explained by other variables, then the value of R is 1 and VIF is infinite.

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 a scatter plot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight. Here's an example of a normal QQ plot when both sets of quantiles

truly come from normal distributions.