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)

Categorical values need to be imputed depending on the impact on the dependent variable. Categorical variables can be imputed using mode.

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

It is important to use drop_first = True while creating the dummy variables. It helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables. Hence if we have categorical variable with N distinct values, then we need only N-1 columns to represent the dummy variables.

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

Temp variable has the highest correlation with the target variable.

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

The R-squared value for training data set is 79.6% and Adjusted R-squared on the train data set is 79.3% . The assumption of linear regression model satisfy following conditions:

- Linear relationship between X and Y
- Error terms are normally distributed (not X, Y)
- Error terms are independent of each other
- Error terms have constant variance (homoscedasticity)

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

Following are the top 3 features have impact on the final model :

- Temperature : with coefficient 0.42
- Weather C [3-Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds]: with coefficient 0.24
- Year: with coefficient 0.23

General Subjective Questions

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

Linear Regression is an ML algorithm used for supervised learning. Linear regression performs the task to predict a dependent variable(target) based on the given independent variable(s). So, this regression technique finds out a linear relationship between a dependent variable and the other given independent variables. Hence, the name of this algorithm is Linear Regression.

In a typical 2-d graph, on X-axis is the independent variable and on Y-axis is the output. The regression line is the best fit line for a model. And our main objective in this algorithm is to find this best fit line.

Advantages :

- Linear Regression is simple to implement.
- Less complexity compared to other algorithms.
- Linear Regression may lead to over-fitting but it can be avoided using some dimensionality reduction techniques, regularization techniques, and cross-validation.

Dis-Advantages :

- Outliers affect this algorithm badly.
- It over-simplifies real-world problems by assuming a linear relationship among the variables, hence not recommended for practical use-cases.

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

Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties. Once Francis John "Frank" Anscombe who was a statistician of great repute found 4 sets of 11 data-points in his dream and requested the council as his last wish to plot those points. After that, the council analyzed them using only descriptive statistics and found the mean, standard deviation, and correlation between x and y.

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

In statistics, the Pearson correlation coefficient (PCC), also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or the bivariate correlation, is a measure of linear correlation between two sets of data. It is the covariance of two variables, divided by the product of their standard deviations; thus it is essentially a normalised measurement of the covariance, such that the result always has a value between −1 and 1.

The Pearson's correlation coefficient varies between -1 and +1 where:

- r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
- r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
- r = 0 means there is no linear association
- r > 0 < 5 means there is a weak association
- r > 5 < 8 means there is a moderate association
- r > 8 means there is a strong association

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

Scaling 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 modelling. 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.

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).
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.

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.

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

   If there is perfect correlation, then VIF = infinity. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.
   An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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

Q Q Plot (Quantile-Quantile plot) are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q Q plot is to find out if two sets of data come from the same distribution. A 45 degree angle is plotted on the Q Q plot; if the two data sets come from a common distribution, the points will fall on that reference line.

If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x. Q–Q plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

A Q–Q plot is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.