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)**

   Following are the inferences made:

   - Among the seasons, Fall has the highest demand and spring has the lowest.
   - Demand for bike sharing has significantly improved in the year 2019.
   - Across the months, the demand increases from Jan to Sep/Oct, after which there is a drastic fall in demand until Dec.
   - Days with clear weather has the highest demand of bikes followed by misty weather and days with light rain and snow has the lowest demand. There is no demand at all on days with heavy rain/snow/thunderstorm.

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

   When creating dummy variables without drop_first = True, we create 1 additional column more than what is required. Hence, it is used to reduce redundancy of the dummy variables. This also prevents multi-collinearity as well between 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)**

   Columns 'temp' and 'atemp' columns have 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)**

   Once the model is trained, and we have a good enough R-squared and adjusted R-squared values along with low VIF and low p-values for the features, we'll need to the residual analysis for the trained set.

   If the residual values are scattered through out with mean of residuals as 0, then the model is good. This followed by the the check if adjusted R-squared value of the model applied on the test-set is close to that of the trained-set, the model holds good.

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

   1. Spring season
   2. Light rain and snow weather

3. Year

General Subjective Questions

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

The Linear regression algorithm is a machine learning technique which is used to predict a dependent variable based on its historical relationship with other independent variables. The algorithm assumes a linear relationship between the dependent and independent variables and tries to model the relationship based on a linear equation.

y = a1*x1 + a2*x2 + ... + an*xn + b

where

y is the dependent or the target variable

x1, x2, ..., xn are the independent variables or features

a1, a2, ..., an are the coefficients of the independent variables

b is the constant

The coefficients of the independent variables are estimated using the Ordinary Least Squares (OLS) method which tries to minimize the squared differences between the observed and the predicted values of the dependent variable.

Minimize Summation from 1 to n of (y_pred - yi)^2

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

Anscombe's quartet is a set of 4 different datasets which have nearly identical statistical properties such as mean, variance, correlation etc. but when they are plotted, they look very different. This highlights the importance of visualizing the data as datasets with identical summary statistics can have different data distributions/patters.

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

Pearsons' R or Pearson's correlation coefficient is a statistical measure of strength and direction of a linear relationship between two variables. The value of this measure can range from -1 to 1 where 0 represents no linear correlation, -1 represents perfect negative correlation (if one variable increases, the other decreases proportionally and vice versa), and +1 represents perfect positive correlation (if one variable increases, the other increases proportionally and vice versa).

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

Scaling is a method used for adjusting the range of values of the features in a dataset. From a machine learning perspective, scaling ensures that all features contribute equally to the model's performance, especially when they have very difference scales and ranges of values.

Scaling is performed for improving model performance bringing an unbiased contribution of the features to the model. It is also used for handling features with different units.

Following are the key differences between normalized and standardized scaling:

1. Data is scaled to a fixed range from 0 to 1 or in some cases from -1 to 1 in a normalized scaler, whereas in a standardized scaler, the ranged of the scaled data is not fixed; it depends on the distribution of the original data.
2. Normalized scaled data is sensitive to outliers, in which case, the normalized data can be skewed. Standardized scaled data is less sensitive to outliers comparatively, as it is adjusts the spread according to the data.

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

VIF can sometimes be infinite because there might be two or more features considered in the model which may have a perfect correlation between each other. In such a case, R2 becomes 1 and since VIF = 1/(1-R^2), the denominator becomes 0 and thus the VIF value becomes infinite.

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

A Q-Q plot or a Quantile-Quantile plot is a chart which plots the quantiles of the data against the quantiles of the theoretical distribution.

Following are the importance and uses of a Q-Q plot:

1. Evaluating normality of the residuals – If the points are on a straight line, they are normally distributed.
2. Identifying outliers – It helps in identifying the outliers by checking the points which lie outside the straight line.
3. Identifying skewness - If the points bend upwards, the data may be right-skewed; if they bend downwards, the data may be left-skewed.
4. Checking for heteroscedasticity – Deviations from the straight line indicate heterocscedasticity.