**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?**

- We can observe that the demand for bikes depends on following variables:
  - yr, holiday, Spring, Light Snow, Mist Cloudy, 3, 5, 6, 8 and 9
  - Demands increases in the month of 3, 5, 6, 8 and 9
- The demad of bike is less in the month of spring when compared with other seasons
- Bike demand in the fall is the highest.
- Bike demand takes a dip in spring.
- Bike demand in year 2019 is higher as compared to 2018.
- Bike demand is high if weather is clear or with mist cloudy while it is low whenthere is light rain or light snow

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

- In machine learning and statistical modeling, it is common to encode categorical variables as binary dummy variables.
- Each category is represented by a separate binary variable that takes a value of 1 if the observation belongs to that category and 0 otherwise.
- When creating dummy variables, it is important to use drop_first=True in order to avoid the problem of multicollinearity, which occurs when one of the dummy variables can be perfectly predicted from the others.
- This means that including all the dummy variables in a model will cause problems with estimation and interpretation of the model's coefficients.

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

- **atemp** and **temp** both have same and highest among all numerical variables

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

- The assumptions of Linear Regression after building the model on the training set is possible by **plotting the scatter plot between the features and the target variables**

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

- The top 3 features contributing significantly are:
  - weathersit_Light_Snow(negative correlation).
  - yr(Positive correlation).
  - month(Positive correlation).

## General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

- Linear regression is a popular statistical algorithm used to model the relationship between a dependent variable and one or more independent variables.
- Linear regression is a simple yet powerful algorithm that can be used for a wide range of applications such as predictive modeling, trend analysis, and forecasting
- The main idea behind linear regression is to fit a line or hyperplane through a set of data points such that the difference between the predicted and actual values of the dependent variable is minimized
- Here are the detailed steps involved in the linear regression algorithm:

    1. ***Data preparation:*** This involves collecting data for the dependent variable and one or more independent variables. Data preparation also involves checking for missing values, outliers, or other anomalies in data.
    2. ***Model selection:*** Model selection involves, selecting a linear regression model that fits your data. There are two main types of linear regression models: simple linear regression and multiple linear regression.
    3. ***Model fitting:*** Once after selecting a linear regression model, you need to fit it to your data. Model fitting involves finding the coefficients of the line or hyperplane that best fits your data. To do this, it use various optimization techniques such as least squares, gradient descent, or normal equations.
    4. ***Model evaluation:*** The performance of model is evalutaed after fitting linear regression model. One way to achieve this is to calculate the residuals (the difference between the predicted and actual values of the dependent variable) and check if they are normally distributed around zero. Model evaluaition also calculate various metrics such as R-squared, adjusted R-squared, root-mean-squared error (RMSE), and mean absolute error (MAE) to measure how well your model fits the data.
    5. ***Model interpretation:*** Finally, it need to interpret the linear regression model. You can use the coefficients of the line or hyperplane to understand the direction and strength of the relationship between the dependent variable and the independent variables. You can also use the model to make predictions on new data points.

2. **Explain the Anscombe’s quartet in detail.**

- Anscombe's quartet is a set of four datasets that have nearly identical statistical properties, but display different patterns when visualized. Anscombe's quartet is to demonstrate the importance of graphical visualization in data analysis.
- The datasets in Anscombe's quartet have the same means, variances, correlation coefficients, and linear regression lines, but when plotted they have very different patterns. This highlights the importance of visualizing data before drawing conclusions based on statistical measures alone.
- In machine learning, Anscombe's quartet is often used to illustrate the importance of data visualization in model selection and evaluation. It shows that even if two datasets have

similar statistical properties, they can still exhibit different patterns that may require different modeling approaches.

3. **What is Pearson’s R**

- Pearson's R, also known as the Pearson correlation coefficient, is a statistical measure that assesses the linear relationship between two continuous variables. It is denoted by the symbol "r" and ranges from -1 to +1, with values closer to -1 or +1 indicating a stronger linear relationship, and values close to 0 indicating a weak or no linear relationship.
- Pearson's R is calculated by dividing the covariance of the two variables by the product of their standard deviations.
- ***The formula for Pearson's R is:***
  - r = (sum of (x - x_mean) * (y - y_mean)) / (sqrt(sum of (x - x_mean)^2) * sqrt(sum of (y - y_mean)^2))
  - where x and y are the two variables of interest, x_mean and y_mean are their respective means, and the sum and square root functions are computed over all observations.
- Pearson's R is commonly used in fields such as psychology, biology, and social sciences to examine the relationship between two continuous variables.

**4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**

- Scaling in linear regression refers to the process of transforming the input features of a linear regression model so that they are on a similar scale.
- Linear regression models rely on the assumption that the input features are on a similar scale, which means that each feature contributes equally to the output variable.
- If the input features have different scales, then the model might give more weight to features with larger scales and less weight to features with smaller scales, which can lead to biased results.
- Normalized scaling and standardized scaling are two common methods for scaling input features in machine learning, including linear regression.
- The main difference between these two methods is how they transform the input features.
- ***Normalized scaling*** involves scaling the input features so that they have a minimum and maximum value within a given range. This range is typically between 0 and 1. The formula for normalized scaling is:
  - X_norm = (X - X_min) / (X_max - X_min)
  - where X is the original value of the input feature, X_min is the minimum value of the feature, and X_max is the maximum value of the feature. Normalized scaling is useful when the distribution of the input features is not necessarily Gaussian or when there are outliers in the data.
- ***Standardized scaling***, on the other hand, involves scaling the input features so that they have a mean of 0 and a standard deviation of 1. The formula for standardized scaling is:
  - X_stand = (X - mu) / sigma

- where X is the original value of the input feature, mu is the mean of the feature, and sigma is the standard deviation of the feature. Standardized scaling is useful when the distribution of the input features is approximately Gaussian, as it preserves the shape of the distribution.

**5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?**

- VIF (Variance Inflation Factor) is a measure of multicollinearity between predictor variables in a linear regression model. It measures how much the variance of the estimated regression coefficients is increased due to collinearity among the predictor variables. A high VIF value indicates a high degree of multicollinearity, which can affect the accuracy and reliability of the linear regression model.
- In some cases, the VIF value can be infinite. This happens when one or more predictor variables are perfectly collinear, meaning they can be expressed as a linear combination of other predictor variables in the model. When this happens, the VIF value for the perfectly collinear variable(s) will be infinite because the variance of the regression coefficient estimate for that variable cannot be computed.
- For example, consider a linear regression model with two predictor variables, X1 and X2. If X1 = 2*X2, then X1 is perfectly collinear with X2, and the VIF value for X1 (or X2) will be infinite. In this case, one of the variables should be removed from the model to avoid the problem of perfect multicollinearity.

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

- ***A Q-Q plot, short for quantile-quantile plot***, is a graphical technique used to compare the distribution of a sample to a theoretical distribution, such as the normal distribution.
- In linear regression, a Q-Q plot is used to assess the normality assumption of the residuals, which is one of the key assumptions of linear regression analysis.
- Residuals are the differences between the observed values and the predicted values from the regression model. A residual plot is a graphical tool to visually examine the distribution of these residuals. A Q-Q plot is a type of residual plot that compares the distribution of the residuals to the expected normal distribution.
- If the residuals are normally distributed, the Q-Q plot will show a straight line. A curved line or deviations from the straight line suggest non-normality of the residuals. Departures from normality can affect the accuracy of statistical inference and hypothesis testing, such as the significance of the regression coefficients, t-tests, and confidence intervals.
- Therefore, a Q-Q plot is important in linear regression because it allows us to evaluate the normality assumption of the residuals, which is necessary for valid statistical inference. If the residuals are not normally distributed, transformations such as logarithmic or square root transformations may be used to normalize the data, or non-linear models may be considered.
- Overall, the Q-Q plot is a useful tool to check the assumption of normality in linear regression and to ensure that the model is appropriate for the data.