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? (3mark)

   - Fall has the highest median, which is expected as weather conditions are most optimal to ride a bike followed by summer and winter.
   - 2019 has a higher median than 2018, which might be due to the fact that bike rents are getting popular and people are getting more aware of the environment
   - The best conditions for renting bikes are clear skies since the temperature is ideal, the humidity is low, and the temperature is lower.
   - Working and Non-Working days have almost the same median
   - People rent more on nonholidays compared to holidays, so the reason might be they are using their own car during vacations or holidays and prefer rentals for office or any work.

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

   A variable with an n level can be represented by an n-1 variable. So, if we remove the first column Then also, we can represent the data. we retain a reference category. This reference category is represented by all dummy variables being zero, indicating that the category not represented by any of the remaining dummy variables is the reference point. It helps to prevent perfect multicollinearity among the variables.

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

   The temperature had the highest correlation coefficient of 0.63

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

   By plotting residual distribution, it came out to be a normal distribution also Examining the correlation matrix or variance inflation factor (VIF) for the independent variables can identify multicollinearity issues

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

   The following are the top 3 features contributing significantly towards explaining the demands of shared bike: year(0.2341), holiday(-0.0963), and windspeed(-0.1481)

General Subjective Questions

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

Linear regression is a supervised learning algorithm used for predicting a continuous target variable based on one or more independent variables. It assumes a linear relationship between the independent variables and the target variable. The goal of linear regression is to find the best-fitting line that minimizes the difference between the predicted and actual values.

Let's dive into the steps involved in the linear regression algorithm:

1. Data Preparation: Gather the dataset consisting of the target variable (also called the dependent variable) and one or more independent variables (also called features). Ensure the data is cleaned, missing values are handled appropriately, and relevant preprocessing steps like feature scaling or encoding categorical variables are performed.

2. Model Representation: Linear regression assumes a linear relationship between the independent variables and the target variable. The general form of a linear regression model is given by the equation:

   y = b0 + b1*x1 + b2*x2 + ... + bn*x

   where:

   - `y` is the target variable.

   - `b0` is the intercept (the value of `y` when all independent variables are zero).

   - `b1, b2, ..., bn` are the coefficients or slopes associated with the independent variables `x1, x2, ..., xn`.

3. Cost Function: Linear regression uses a cost function to measure the difference between the predicted values and the actual values in the training data. The most common cost function is the Mean Squared Error (MSE), which sums the squared differences between the predicted and actual values. The objective is to minimize the cost function by finding the optimal values for the coefficients.

4. Gradient Descent: Gradient descent is an iterative optimization algorithm used to minimize the cost function. It updates the coefficients by taking steps proportional to the negative gradient of the cost function. This process continues until convergence or a predefined stopping criterion is met.

5. Model Training: Initialize the coefficients with arbitrary values or zeros. Use the training data to fit the linear regression model by iteratively updating the coefficients using the gradient descent algorithm. The learning rate, which determines the size of the steps taken during gradient descent, is an important parameter to consider.

6. Model Evaluation: Once the model is trained, evaluate its performance on a separate validation or test dataset. Common evaluation metrics for linear regression include R-squared (coefficient of determination), adjusted R-squared, root mean squared error (RMSE), mean absolute error (MAE), and others. These metrics assess how well the model fits the data and how accurately it predicts the target variable.

   7. Prediction: Use the trained model to make predictions on new, unseen data by applying the learned coefficients and the linear regression equation.

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

Anscombe's quartet is a set of four data sets that have nearly identical summary statistics but appear very different when graphed. The quartet was created by Francis Anscombe in 1973 to illustrate the importance of graphing data before conducting statistical analysis.

The four data sets in Anscombe's quartet are:

- Data set 1: (1,3),(2,5),(3,7),(4,9),(5,11)
- Data set 2: (1,1),(2,2),(3,3),(4,4),(5,5)
- Data set 3: (1,4),(2,12),(3,20),(4,28),(5,36)
- Data set 4: (1,8),(2,8),(3,8),(4,8),(5,8)

As you can see, all four data sets have the same mean, median, standard deviation, and correlation coefficient. However, when you graph the data sets, you can see that they are very different. Data Set 1 is a linear relationship, data set 2 is a quadratic relationship, data set 3 is a curvilinear relationship, and data set 4 is a constant relationship.

Anscombe's quartet illustrates the importance of graphing data before conducting statistical analysis. Just because two data sets have the same summary statistics does not mean that they are the same. The only way to truly understand the relationship between two variables is to graph the data.

Some of the key takeaways from Anscombe's quartet:

- Summary statistics can be misleading.
- It is important to graph data before conducting statistical analysis.
- The shape of the relationship between two variables is more important than the summary statistics.

Anscombe's quartet is a classic example of the importance of data visualization in statistics. By graphing data, we can gain a deeper understanding of the relationship between variables and avoid making incorrect conclusions based on summary statistics alone.

3. What is Pearson’s R? (3 marks)

. Pearson's R is a statistical measure that is used to quantify the strength and direction of the linear relationship between two variables. It is a number between -1 and 1, where:

- A value of 1 indicates a perfect positive correlation, meaning that as one variable increases, the other variable also increases.
- A value of -1 indicates a perfect negative correlation, meaning that as one variable increases, the other variable decreases.
- A value of 0 indicates no correlation, meaning that there is no relationship between the two variables.

Pearson's R is calculated by taking the covariance of the two variables and dividing it by the product of their standard deviations. The covariance is a measure of how much the two variables vary together, and the standard deviation is a measure of how much each variable varies from its mean.

Pearson's R is a commonly used measure of correlation because it is relatively easy to calculate and interpret. However, it is important to note that Pearson's R is only appropriate for measuring linear relationships. If the relationship between the two variables is not linear, then Pearson's R may not be an accurate measure of the strength of the relationship.

some examples of how Pearson's R can be used:

- To determine the strength of the relationship between height and weight in a population of adults.
- To determine the strength of the relationship between income and education level in a

population of workers.

- To determine the strength of the relationship between test scores and hours of study in a group of students.

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

Scaling is the process of transforming the values of features in a dataset to a common scale. This is done to facilitate data analysis and modeling, and to reduce the impact of different scales on the accuracy of machine learning models.

There are two main types of scaling:

- Normalized scaling: In normalized scaling, each feature is divided by its maximum value. This ensures that all features have a range of 0 to 1.
- Standardized scaling: In standardized scaling, each feature is subtracted from its mean and then divided by its standard deviation. This ensures that all features have a mean of 0 and a standard deviation of 1.

Scaling is performed for a number of reasons, including:

- To make features with different scales comparable. For example, if one feature is measured in centimeters and another feature is measured in kilograms, then it is not possible to compare them directly. Scaling them to a common scale, such as centimeters or kilograms, allows them to be compared.
- To improve the accuracy of machine learning models. Machine learning models are often trained on data that has been scaled. This is because scaling can help to reduce the impact of outliers and noise on the model.
- To make data visualization easier. Scaling can make it easier to visualize data by making it easier to see the relationships between different features.

The main difference between normalized scaling and standardized scaling is that normalized scaling does not centre the data, while standardized scaling does. This means that normalized scaling will not affect the mean of the data, while standardized scaling will. In general, standardized scaling is preferred over normalized scaling because it is more robust to outliers. However, normalized scaling can be used if it is important to preserve the mean of the data.

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

The Variance Inflation Factor (VIF) is a measure of multicollinearity in a regression model. Multicollinearity occurs when two or more independent variables are highly correlated with each other. When VIF is infinite, it means that one of the independent variables in the model is perfectly correlated with another independent variable. This can happen when two variables are measuring the same thing or when one variable is a linear function of another variable. There are a few reasons why VIF might be infinite. One reason is that the data may be poorly designed. For example, if the same variable is measured multiple times in the same dataset, then the VIF for that variable will be infinite. Another reason is that the model may be too complex. If there are too many independent variables in the model, then it is possible that some of the variables will be perfectly correlated with each other.

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 is a graphical method for comparing two probability distributions. It is a scatterplot where the quantiles of one distribution are plotted against the quantiles of another distribution. The quantiles of a distribution are the values that divide the distribution into equal parts. For example, the 25th percentile is the value that divides the distribution into two parts, such that 25% of the data is below the value and 75% of the data is above the value.
In linear regression, a Q-Q plot can be used to assess the assumption of normality. The assumption of normality states that the residuals (the difference between the observed values and the predicted values) are normally distributed. If the residuals are not normally distributed, then the linear regression model may not be accurate.
To create a Q-Q plot for linear regression, you will need to:

1. Calculate the residuals from the linear regression model.
2. Calculate the quantiles of the residuals.
3. Plot the quantiles of the residuals against the quantiles of a standard normal distribution.

If the residuals are normally distributed, then the Q-Q plot will be a straight line. If the residuals are not normally distributed, then the Q-Q plot will not be a straight line.
Here are some of the things that can cause the residuals to not be normally distributed:

- Outliers: Outliers are data points that are very different from the rest of the data. Outliers can cause the residuals to not be normally distributed.
- Non-linearity: If the relationship between the independent and dependent variables is not linear, then the residuals will not be normally distributed.
- Heterogeneity of variance: If the variance of the residuals is not constant, then the residuals will not be normally distributed.

If the residuals are not normally distributed, then you may need to transform the independent or dependent variables, or use a different regression model.