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)**

   Answer: Following are the categorical variables and there analysis-

   1. Season: Most of the bookings were happening on Fall season whereas spring is having less booking. This variable can be good predictor for dependent variable.
   2. Mnth: Most of the bookings were happening in the month of June, July, August, September & October while janauary is having less. This variable can be good predictor for dependent variable.
   3. Weathersit: Most of the booking were happening in Clear Weather. This variable can be good predictor for dependent variable.
   4. Holiday: The data is clearly biased, this indicates that this CANNOT be a good predictore for dependent variable.
   5. Weekdays: This variable show very close trend. The variable can have some or no influence towards the predictor.
   6. Working Days: Most of the booking were happening on working day. This variable can be good predictor for dependent variable.

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

   Answer: By default, most libraries create n dummy variables for n categories, including the first category, but we can set drop_first=True to drop the first category and create only n-1 dummy variables. This means that the first category becomes the baseline category, and the remaining n-1 categories are compared to it.

   Therefore, it's important to use drop_first=True during dummy variable creation to avoid the dummy variable trap, make the model more interpretable, and reduce computational complexity.

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

   (1 mark)

   Answer: " 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)**

   Answer: Following assumptions are done to validate the assumptions of Linear Regression:

   1. Normality of error term – error terms are normally distributed.
   2. Multicollinearity Check – We could find that no multicollinearity existed between predictor variables, as all the values are below the range below 5.

3. Linear Relationship Validation – we cold see linear relationship between “temp” and “atemp” with the predictor “cnt”.

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

Answer*:* The top three predictor variables are -

1. **Tempreture(temp)** - with the coefficient value 0.479457 indicates that the increase in temp variable increases the bike booking by 0.479457.
2. **Light Snow(Light_snowrain)** - with the coefficient value -0.285587 indicates that the increase in yr variable decreases the bike booking by 0.285587.
3. **Year(yr)** - with the coefficient value 0.234391 indicates that the increase in yr variable increases the bike booking by 0.234391.

General Subjective Questions

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

Answer: Linear regression is a statistical method used to model the relationship between a dependent variable (Y) and one or more independent variables (X). The goal of linear regression is to find the line of best fit that minimizes the sum of the squared errors between the predicted values and the actual values.

The linear regression algorithm can be explained in the following steps:

1. Data collection: Gather the data for the dependent variable (Y) and independent variable(s) (X).
2. Data preprocessing: Clean and preprocess the data, which may involve removing missing values, normalizing the data, and removing outliers.
3. Split the data: Divide the data into training and testing sets. The training set is used to train the model, and the testing set is used to evaluate its performance.
4. Model training: Fit a line to the data by finding the coefficients that minimize the sum of the squared errors. The line is of the form Y = mX + b, where m is the slope of the line and b is the intercept.
5. Model evaluation: Use the testing set to evaluate the performance of the model. The most common evaluation metrics for linear regression include mean squared error (MSE), root mean squared error (RMSE), and R-squared.
6. Model deployment: Once the model has been trained and evaluated, it can be deployed to make predictions on new data.

There are two types of linear regression: simple linear regression and multiple linear regression. Simple linear regression involves one independent variable, while multiple linear regression involves two or more independent variables.

Linear regression can be performed using various techniques, such as ordinary least squares (OLS), gradient descent, and stochastic gradient descent (SGD).

Overall, linear regression is a powerful tool for predicting the values of a dependent variable based on one or more independent variables, and is widely used in various fields such as economics, finance, and social sciences.

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

Answer: Anscombe's quartet is a set of four datasets that have identical statistical properties, yet appear very different when graphed. These datasets were first introduced by the statistician Francis Anscombe in 1973 to illustrate the importance of graphing data in statistical analysis and to demonstrate the limitations of relying solely on summary statistics.

Each of the four datasets contains 11 pairs of x and y values, and the summary statistics for each dataset are identical: the mean and variance of x and y, the correlation coefficient between x and y, and the linear regression equation between x and y. However, when plotted, the datasets display distinct differences in their patterns and relationships.

For example, the first dataset appears to have a linear relationship between x and y, with a clear positive correlation. The second dataset, however, is not linear but rather appears to have a curved relationship between x and y. The third dataset has an outlier that has a significant effect on the linear regression line, and the fourth dataset has a very strong relationship between x and y, but is highly influenced by a single outlier.

The significance of Anscombe's quartet lies in its ability to demonstrate the limitations of relying solely on summary statistics such as mean, variance, and correlation coefficient. Graphing the data provides a visual representation of the relationships and patterns within the data that cannot be captured by summary statistics alone.

Therefore, Anscombe's quartet highlights the importance of data visualization and exploration in statistical analysis, as well as the potential pitfalls of relying solely on summary statistics to understand and analyze data.

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

Answer: Pearson's correlation coefficient (often referred to as Pearson's R or simply the correlation coefficient) is a statistical measure that indicates the extent to which two variables are linearly related. It is denoted by the symbol r and can range from -1 to +1, where -1 indicates

a perfect negative correlation, +1 indicates a perfect positive correlation, and 0 indicates no correlation at all.

Pearson's R is calculated as the covariance between the two variables divided by the product of their standard deviations. The formula for Pearson's R is:

r = (Σ((X - mean(X)) * (Y - mean(Y)))) / (sqrt(Σ(X - mean(X))^2) * sqrt(Σ(Y - mean(Y))^2))

where X and Y are the two variables being compared, and mean(X) and mean(Y) are their respective means.

Pearson's R is widely used in many fields, including social sciences, engineering, and finance, to measure the strength and direction of the relationship between two variables. It is particularly useful in situations where there are multiple variables, as it can help identify which variables are most strongly related to each other.

However, it should be noted that Pearson's R only measures the linear relationship between two variables and does not capture any other types of relationships, such as nonlinear or monotonic relationships. Additionally, Pearson's R can be affected by outliers and does not indicate causation between the two variables.

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

Answer: In data analysis, scaling refers to the process of transforming the features or variables of a dataset to have a specific scale or range.

The purpose of scaling is to ensure that all the features have similar magnitudes, so that they can be compared on the same level and to avoid one feature dominating the others.

Normalized scaling is a type of scaling that transforms the features so that they have a range of values between 0 and 1. This is achieved by subtracting the minimum value of each feature and dividing by the range of the feature (i.e., the difference between the maximum and minimum values). Normalized scaling is useful when the distribution of the data is not known or when the data contains outliers.

Standardized scaling is a type of scaling that transforms the features so that they have a mean of 0 and a standard deviation of 1. This is achieved by subtracting the mean value of each feature and dividing by the standard deviation of the feature. Standardized scaling is useful when the data is normally distributed and when the features have different units of measurement.

The main difference between normalized scaling and standardized scaling is the range of the transformed values. Normalized scaling results in values between 0 and 1, while standardized

scaling results in values that can be positive or negative and have a mean of 0 and a standard deviation of 1. Both types of scaling can be useful depending on the distribution of the data and the type of analysis being performed.

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

(3 marks)

Answer: VIF (Variance Inflation Factor) measures the degree to which the variance of the estimated regression coefficient is increased due to multicollinearity in the data. When VIF is infinite, it indicates that there is perfect multicollinearity between one or more predictors in the model.

Perfect multicollinearity occurs when two or more predictors in the model are perfectly linearly related to each other, which means one predictor can be expressed as a linear combination of the other predictors. When perfect multicollinearity exists, the estimated regression coefficients become unstable, and the VIF for the predictor(s) involved becomes infinite.

For example, if we have a model with two predictors, X1 and X2, and X1 can be expressed as a linear combination of X2, then the model will have perfect multicollinearity. This means that the estimated regression coefficient for X1 will be based on the combined effect of X1 and X2, making the estimated regression coefficient for X1 unstable and the VIF for X1 infinite.

To avoid perfect multicollinearity, it is important to ensure that the predictors used in the model are independent of each other, and not perfectly linearly related. If perfect multicollinearity is identified, one of the correlated predictors should be removed from the model to avoid the unstable and unreliable estimates of regression coefficients.

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

(3 marks)

Answer: A Q-Q plot, short for quantile-quantile plot, is a graphical tool used to compare the distribution of a sample of data to a theoretical distribution. It is a scatter plot with the sample quantiles on one axis and the theoretical quantiles on the other axis.

The purpose of a Q-Q plot is to visually assess if a sample of data comes from a specified distribution, such as a normal distribution. If the data points fall close to a straight line, it suggests that the data comes from the specified distribution. If the points deviate significantly from a straight line, it suggests that the data does not come from the specified distribution.

In linear regression, Q-Q plots can be used to check the assumption of normality of the residuals. The residuals are the differences between the observed values and the predicted values of the

dependent variable. If the residuals are normally distributed, then the assumptions of the linear regression model are met and the model is valid.

To create a Q-Q plot for the residuals, we plot the quantiles of the residuals against the quantiles of a normal distribution. If the residuals are normally distributed, the plot will show the points following a straight line. If there are deviations from the straight line, it suggests that the residuals are not normally distributed and may violate the assumptions of the linear regression model.

In summary, Q-Q plots are useful in linear regression to visually assess the normality of the residuals, which is an important assumption of the model. By checking this assumption, we can ensure that the model is valid and make more accurate predictions.