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:

   Given below are the points that we can infer based on the analysis we did on the categorical variables:

   - For the Season variable, there is a high count of bookings for the 'fall' with 'summer' and 'winter' having a slightly lower count. But 'spring' has a very low count when compared to the other three.
   - For the Year variable, there is a major increase in the count of bookings for the year 2019 compared to 2018.
   - For the Month variable, most of the bookings were done during the period May – October. It started at the lowest in Jan and then started increasing till May, kept steady till Oct and then it started decreasing.
   - For the Weather variable, 'Clear' had the highest count for the bookings.
   - For the Holidays, the bookings seem to be less in number when compared to the regular days.
   - There is not much of a difference between the working and non-working days.
   - Even for the Weekdays, there is no major difference in the booking count.

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

   Answer:

   The purpose of using dummy variables is that, for a categorical column with 'n' categories, we would create a dummy column for each of them. Using 'drop_first=True' helps by ignoring the first category and creating columns only for the remaining categories (i.e. n-1).

   e.g. I have a categorical column 'Season' with 4 categories (Spring, Summer, Fall, Winter). When I create dummy variables using 'drop_first=True', then it would create columns for Summer, Fall and Winter. The Spring can be deduced based on when the values of the other three columns is 0.

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

   Answer:

   The numerical variables 'temp' and 'atemp' have the highest correlation with thee target variable.

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

   Answer:

   I have validated the assumptions of Linear Regression Model based on the following:

   - Normality of the error terms
   - Linearity among the variables
   - No auto-correlation
   - Multicollinearity
   - 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)

   Answer:

   - Temp
   - Year
   - Weather

General Subjective Questions

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

   Answer:

   Linear regression may be defined as a form of predictive modelling technique which shows us the linear relationship between a dependent variable and a set of independent variables. It basically means that when the values of the independent variables increase or decrease, the value for the dependent variable would also change.

   The standard equation for it Is given as (Y = mX + C). Here Y is the dependent variable that we are trying to predict. X refers to the independent variable, m is the slope or coefficient (the change in Y for a one-unit change in X). C is basically a constant, it's the intercept of the regression line (value of Y when X = 0).

   Linear regression can be classified into two types:

   - Simple Linear Regression (with one independent variable)
   - Multiple Linear Regression (with more than one independent variable)

   The linear regression model gives a sloped line which describes the relationship with the variables. A linear relationship can be called positive, if there is an increase in the dependent variable due to an increase in the independent variable. If there is a decrease in the

dependent variable while there is an increase in the independent variable, then it's a negative relationship.

For linear regression to give reliable results, certain assumptions must hold:

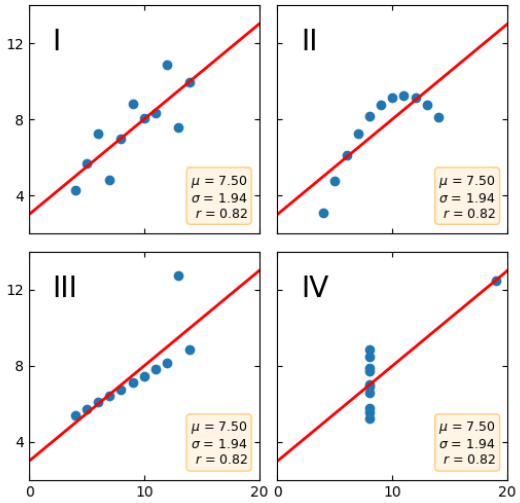
- Linearity: The relationship between the independent and dependent variables must be linear.
- Independence: The observations should be independent of each other.
- Homoscedasticity: The variance of residuals (errors) should be constant across all values of the independent variables (i.e., no heteroscedasticity).
- Normality of Errors: The residuals should follow a normal distribution.
- No Multicollinearity: In multiple linear regression, the independent variables should not be highly correlated with each other.

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

Answer:

Anscombe's quartet can be defined as a set of four datasets that have nearly identical simple descriptive statistics, such as the mean, variance, and correlation, but have very different distributions and visual patterns when graphed. The purpose of the quartet is to highlight the importance of graphing data before performing statistical analysis. It demonstrates that relying solely on summary statistics can be misleading, as different datasets can have the same statistical properties but vastly different underlying structures.

Given below are four dataset plots which have nearly the same statistical observations, i.e. variance and mean of all x, y points in all four datasets. When we plot these in a graph, we can see that they have the same regression line, but each dataset is telling a different story.



Each of the four datasets in Anscombe's quartet shares the following identical or nearly identical summary statistics:

- Mean of X: 9
- Mean of Y: 7.50
- Variance of X: 11
- Variance of Y: 4.12
- Correlation between X and Y: r = 0.816
- Linear Regression Line: Y=3.00 + 0.500*X

This quartet emphasizes that regression algorithms can be fooled, so its important to perform data visualization to get a better understanding of the datasets.

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

Answer:

Pearson's R, also known as the Pearson correlation coefficient, is a statistic that measures the strength and direction of a linear relationship between two continuous variables. The value of Pearson's R ranges from −1 to 1, where:

- 1 indicates a perfect positive linear relationship.
- -1 indicates a perfect negative linear relationship.
- 0 indicates no linear relationship.

Formula

The Pearson correlation coefficient between two variables X and Y is calculated as:

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

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

Answer:

Scaling is the process of transforming the values of numerical features into a common range or distribution. It is a critical preprocessing step in many machine learning algorithms that rely on the distance between data points (such as gradient descent-based algorithms, k-nearest neighbours, and support vector machines), as well as in algorithms that are sensitive to feature magnitudes.

Given below are the reasons for performing scaling:

- Improves Model Performance: Many machine learning algorithms assume that all input features are on a similar scale. If the features have vastly different scales, the model may

prioritize features with larger ranges over those with smaller ranges, leading to suboptimal results.

- Faster Convergence: Gradient-based optimization methods (e.g., in linear regression, logistic regression, neural networks) benefit from scaled features because it leads to faster and more efficient convergence during training. When the features are on different scales, the optimization process may take longer to converge.
- Consistency in Distance-Based Algorithms: Algorithms like K-Nearest Neighbors (KNN), K-Means, and Support Vector Machines (SVM) use distance metrics like Euclidean distance. If features are not scaled, features with larger values will dominate the distance computation, leading to biased results.
- Prevents Bias in Regularization: In regularization techniques like Lasso and Ridge regression, unscaled features can cause the regularization term to disproportionately penalize certain coefficients, affecting model performance.

Normalization transforms the data to fit within a specific range, typically between 0 and 1 (or -1 and 1). This is done using the min-max scaling method. Whereas Standardization transforms the data to have a mean of 0 and a standard deviation of 1, which makes the feature distribution resemble a standard normal distribution.

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

Answer:

The Variance Inflation Factor (VIF) is a metric used to detect multicollinearity in a set of regression variables. It quantifies how much the variance of a regression coefficient is inflated due to collinearity with other variables. A VIF value that is infinite (or extremely high) indicates that one of the variables is a perfect linear combination of other variables. This perfect multicollinearity causes the variance of the regression coefficient to go to infinity, which is why the VIF becomes infinite.

A very high VIF value 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.

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 (Quantile-Quantile plot) is a graphical tool used to compare the distribution of a dataset with a theoretical distribution, usually the normal distribution. It helps assess whether a dataset follows a particular distribution by plotting the quantiles of the observed data against the quantiles of the theoretical distribution.

Q-Q plot can also be used to determine whether two distributions are similar or not. If they are quite similar you can expect the QQ plot to be more linear. The linearity assumption can best be tested with scatter plots. Secondly, the linear regression analysis requires all variables to be multivariate normal. This assumption can best be checked with a histogram or a Q-Q-Plot.

Importance of Q-Q Plot in Linear Regression

- Validating the Normality Assumption.
- Detecting Model Misspecification.
- Outlier Detection.
- Guiding Transformations.