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:

   I have conducted an analysis on categorical columns utilizing box plots and bar plots. The visualizations reveal the following insights:

   - Bookings during the fall season have noticeably increased, and across all seasons, there is a significant rise in booking counts from 2018 to 2019.
   - The majority of bookings occurred in May, June, July, August, September, and October. The trend displays an increase from the beginning of the year until mid-year, followed by a decrease towards the end of the year.
   - Clear weather conditions correlated with higher booking numbers, which aligns with expectations.
   - Thursday, Friday, Saturday, and Sunday experienced a higher volume of bookings compared to the weekdays.
   - During non-holiday periods, the number of bookings is relatively lower, which is reasonable as people may prefer to stay home and spend time with family during holidays.
   - Booking frequencies appear to be almost equal between working days and non-working days.
   - There is a noticeable increase in bookings in 2019 compared to the previous year, indicating positive progress in terms of business.

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

   Answer:

   When creating dummy variables in a categorical feature using one-hot encoding, the drop_first=True parameter is important for avoiding multicollinearity issues in linear regression models.

   It helps with the following:

   - Avoiding the Dummy Variable Trap
   - Multicollinearity in Linear Regression
   - Interpretability of Coefficients
   - Efficiency and Model Simplicity

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

   Answer:

   The **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:

   I have validated the assumption of Linear Regression Model based on below 5 assumptions –

   - Multicollinearity check
     - There should be insignificant multicollinearity among variables.
   - Normality of error terms
     - Normally distributed error terms
   - Homoscedasticity
     - There should be no visible pattern in residual values.
   - Linear relationship validation
     - Linearity should be visible among variables

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 3 features contributing significantly towards explaining the demand of the shared bikes –
   - Temp
   - Sep
   - Winter

General Subjective Questions

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

Answer:

Linear regression can be described as a statistical model that examines the linear association between a dependent variable and a provided set of independent variables. In the context of linear regression, a linear relationship signifies that alterations in the values of one or more independent variables, whether increasing or decreasing, correspondingly lead to changes in the value of the dependent variable.

Mathematically the relationship can be represented with the help of following equation –

$Y = mX + c$

Here, Y is the dependent variable we are trying to predict.

X is the independent variable we are using to make predictions.

m is the slope of the regression line which represents the effect X has on Y c is a constant,

known as the Y-intercept. If X = 0, Y would be equal to c.

Linear regression is of the following two types –

- Simple Linear Regression
- Multiple Linear Regression

Assumptions -

The following are some assumptions about dataset that is made by Linear Regression model –

- **Auto-correlation –**
  - Another assumption Linear regression model assumes is that there is very little or no auto-correlation in the data. Basically, auto-correlation occurs when there is dependency between residual errors.
- **Multi-collinearity –**
  - Linear regression model assumes that there is very little or no multi-collinearity in the data. Basically, multi-collinearity occurs when the independent variables or features have dependency in them.
- **Relationship between variables –**
  - Linear regression model assumes that the relationship between response and feature variables must be linear.

- **Normality of error terms –**
    - Error terms should be normally distributed
- **Homoscedasticity –**
    - There should be no visible pattern in residual values.

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

Answer:

Anscombe's quartet is a set of four datasets that have nearly identical simple descriptive statistics, yet they have very different distributions and appear very different when graphed. This quartet was created by the statistician Francis Anscombe in 1973 to emphasize the importance of graphing data before analyzing it and to demonstrate the impact of outliers on statistical properties.

Description of the Datasets:

1. **Dataset I:**
    - Characteristics:
        - Linear relationship between X and Y.
        - No outliers.
    - Statistical Properties:
        - Close to a perfect linear relationship ($y=3x+2$).
        - Similar mean, variance, and correlation between X and Y.
2. **Dataset II:**
    - Characteristics:
        - Non-linear relationship between X and Y.
        - Outlier in the pair (10, 9).
    - Statistical Properties:
        - Similar mean, variance, and correlation between X and Y.
        - Correlation coefficient is heavily influenced by the outlier.
3. **Dataset III:**
    - Characteristics:
        - Linear relationship with an outlier.
        - Influential outlier significantly impacts regression line.
    - Statistical Properties:
        - Similar mean, variance, and correlation between X and Y.
        - Linear regression line is heavily influenced by a single outlier.
4. **Dataset IV:**
    - Characteristics:
        - Non-linear relationship between X and Y.
        - One outlier.
    - Statistical Properties:
        - Similar mean, variance, and correlation between X and Y.
        - Regression line is largely influenced by the outlier.

Key Observations:

1. **Descriptive Statistics:**
    - All four datasets have nearly identical mean, variance, correlation, and regression coefficients.
2. **Graphical Representation:**
    - Graphs (scatter plots) reveal substantial differences in the data distribution and relationships between variables.
3. **Outliers:**
    - The impact of outliers on the regression line and correlation can be demonstrated.
4. **Importance of Visualization:**
    - Anscombe's quartet highlights the importance of graphing data for a comprehensive understanding.

Implications:

- **Statistical Summary Limitations:**
  - Relying solely on summary statistics may lead to incomplete or misleading interpretations.
- **Graphical Exploration:**
  - Graphical exploration of data is crucial for understanding patterns, relationships, and the presence of outliers.
- **Regression Diagnostics:**
  - Outliers can heavily influence regression results, emphasizing the need for diagnostics.

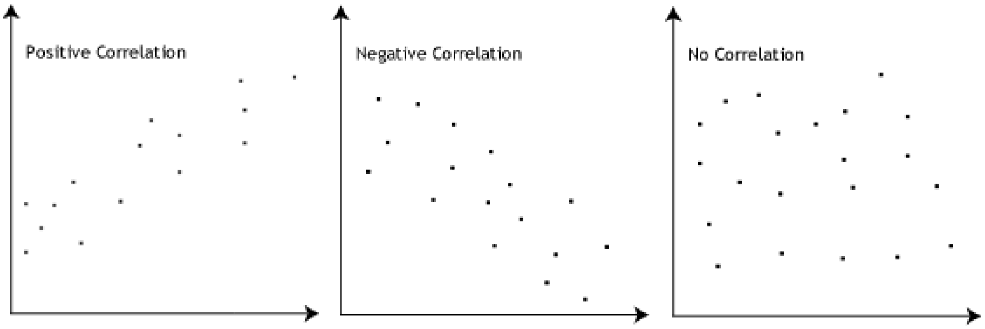
- **Teaching Tool:**
  - Anscombe's quartet is often used in statistical education to emphasize the impact of visualization on data interpretation.

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

Answer:

Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

The Pearson correlation coefficient, r, can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases. This is shown in the diagram below:



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

Answer:

Feature Scaling is a method used to normalize the independent features within a dataset to a consistent range. This process, typically performed during data pre-processing, is crucial for managing significant variations in magnitudes, values, or units. Without feature scaling, machine learning algorithms may assign greater weight to larger values and treat smaller values as if they are relatively smaller, irrespective of their unit of measurement.

S.NO.	Normalized scaling	Standardized scaling
1.	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.

2.	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3.	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
4.	It is really affected by outliers.	It is much less affected by outliers.
5.	Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.

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) measures the extent to which the variance of an estimated regression coefficient increases due to collinearity with other predictor variables in the model. VIF values help assess the multicollinearity among independent variables.

A situation where the VIF becomes infinite typically occurs when there is perfect multicollinearity in the dataset. Perfect multicollinearity means that one or more independent variables in the regression model can be exactly predicted from the others. This perfect predictability leads to mathematical instability when calculating the VIF.

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 (Quantile-Quantile) plot is a graphical tool used in statistics to assess whether a dataset follows a particular theoretical distribution. In the context of linear regression, Q-Q plots are often employed to check the normality of residuals.

Importance in Linear Regression:
In the context of linear regression, the Q-Q plot is specifically used to assess the normality of residuals (the differences between observed and predicted values). The normality assumption is crucial for the validity of statistical inferences and hypothesis tests in linear regression. Here's why Q-Q plots are important in this context:

1. **Normality of Residuals:**

   One of the assumptions of linear regression is that the residuals should be normally distributed. If the residuals are normally distributed, it implies that the errors have constant variance and are unbiased.

2. **Detecting Departures from Normality:**

   A Q-Q plot allows analysts to visually inspect whether the residuals deviate from a normal distribution. Departures from normality may indicate issues such as outliers, nonlinearity, or heteroscedasticity.

3. **Validity of Statistical Inferences:**

   Normality of residuals is essential for valid hypothesis testing, confidence intervals, and other statistical inferences based on regression analysis. Deviations from normality can affect the accuracy of these inferences.