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 done analysis on categorical columns using the boxplot and bar plot. (In the code And notebook file)

   Here are unique category columns and their impact.

   Year: 2019 saw a higher number of bookings compared to the previous year, indicating positive progress in terms of business.

   Weather: Clear weather attracts more bookings.

   Season: Fall season appears to have garnered more bookings.

   Month: A rising trend from the beginning of the year until mid-year, followed by a decline towards the end of the year.

   Working day: Equal booking whether it is a working day or no – working day.

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

   Answer:

   Utilizing drop first=True during dummy variable creation is important primarily to avoid multicollinearity issues in regression analysis. Multicollinearity occurs when predictor variables in a regression model are highly correlated with each other. By dropping one of the dummy variable columns, you effectively create a reference category, which serves as a baseline for comparison with the other categories.

   When all dummy variables are included without dropping one, perfect multicollinearity can occur because the information about one category is perfectly redundant with the information about the other categories. This can lead to numerical instability in estimation and interpretation issues.

   Additionally, dropping the first category can also help in reducing the dimensionality of the feature space, making the model more parsimonious and potentially improving computational efficiency.

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

   Answer:

   The variables 'temp' and 'atemp' exhibit the strongest correlation with the target variable.

   From the pair plot graph and correlation matrix graph in notebook

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 and all the validation steps are clearly mentioned in Jupiter notebook file.

   Linear Relationship: Using Pair plot and correlation graphs, I have validated that there is a linear relationship with our target variables.

   Multicollinearity check: Using Vif, I have made sure that Vif should be in range of all the variables to check Multicollinearity in linear regression.

   Normality of error terms: Error terms are normally distributed and there is no pattern among them.

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 theshared bikes are temp, yr and windspeed.

General Subjective Questions

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

Linear regression is a fundamental statistical method used for modeling the relationship between a dependent variable (target variable) and one or more independent variables (predictor variables). It assumes a linear relationship between the predictor variables and the target variable. Here's a detailed explanation of the linear regression algorithm:
Assumption of Linearity: Linear regression assumes that the relationship between the independent variables x1,x2,…,xn and the dependent variable y is linear, meaning that the change in y is proportional to the change in each xi, with constant coefficients.
Simple Linear Regression: In simple linear regression, there is only one independent variable (x). The relationship between x and y can be represented by the equation of a straight line:
Y = constant + coefficient *X1 + error
Multiple Linear Regression: In **Multiple** linear regression, there can be many independent variable (x). The relationship between x1,x2,…,xn and y can be represented by the equation of a straight line:
Y = constant + coefficient1 *X1 + coefficient2 *X2+ coefficient3 *X3+ coefficient4 *X4+… error
Fitting the Model: The goal of linear regression is to estimate the coefficients (constant and coefficients i.e. slopes) that minimize the difference between the observed and predicted values of *y*. This is often done using the method of least squares, which minimizes the sum of the squared differences between the observed and predicted values.
Interpretation of Coefficients: The coefficients represent the average change in the dependent variable (*y*) for a one-unit change in the corresponding independent variable, holding all other variables constant.
Model Evaluation: Linear regression models can be evaluated using various metrics, such as *R*2 (coefficient of determination), adjusted *R*2, root mean squared error (RMSE), and mean absolute error (MAE), among others. These metrics assess the fitness and the predictive performance of the model.
Assumptions of Linear Regression: It's important to check the assumptions of linear regression, including linearity, independence of errors, constant variance of errors (homoscedasticity), and normality of errors. Violations of these assumptions can lead to biased estimates and unreliable predictions.
In practice, linear regression is widely used for predictive modeling, forecasting, and understanding the relationships between variables in various fields such as economics, finance, social sciences, and machine learning.

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, such as means, standard deviations, and correlation coefficients, but are vastly different when plotted visually. It was created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data before drawing conclusions. Here's a detailed explanation of Anscombe's quartet:

Description of the Quartet: Anscombe's quartet consists of four datasets, each containing 11 (x, y) pairs of data points.

Dataset Characteristics: Despite having similar summary statistics, each dataset has distinct characteristics:

- Dataset I: This dataset follows a linear relationship between x and y. It is well-suited for linear regression analysis.
- Dataset II: This dataset also follows a linear relationship but has an outlier that significantly affects the regression line.
- Dataset III: This dataset forms a non-linear relationship, resembling a quadratic curve. It showcases the importance of considering non-linear relationships.
- Dataset IV: This dataset consists of several groups of data points with identical x values but varying y values. It emphasizes the impact of influential points on regression analysis.

Visual Representation: When plotted, each dataset reveals unique patterns:

- Dataset I: The data points form a clear linear trend, and linear regression provides an appropriate model.
- Dataset II: Despite the outlier, the majority of points align along a linear trend, but the outlier significantly affects the regression line.
- Dataset III: The data points follow a non-linear pattern, indicating that a linear regression model would not be appropriate.
- Dataset IV: Although the summary statistics suggest a linear relationship, the presence of distinct groups of data points challenges the validity of linear regression.

Implications: Anscombe's quartet highlights the limitations of relying solely on summary statistics without visualizing the data. It underscores the importance of exploratory data analysis (EDA) and data visualization techniques in understanding the underlying patterns and relationships within datasets.

Statistical Lessons: The quartet demonstrates that simple summary statistics like means, variances, and correlation coefficients may not capture the complexity of the data. It emphasizes the need for robust statistical methods and critical thinking when analyzing data.

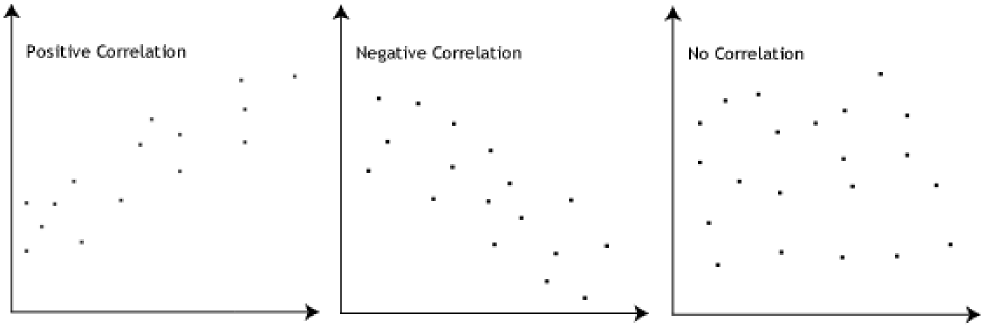
In summary, Anscombe's quartet serves as a cautionary example, reminding researchers and analysts to always visualize their data and not solely rely on summary statistics to draw conclusions. It illustrates the potential pitfalls of overlooking data visualization in favor of numerical summaries.

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 technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

Example**:** If an algorithm is not using feature scaling method then it can consider the value 1000 meter to be greater than 5 km but that's actually not true and in this case, the algorithm will give wrong predictions. So, we use Feature Scaling to bring all values to same magnitudes and thus, tackle this issue.

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:

When the value of the Variance Inflation Factor (VIF) becomes infinite, it indicates perfect multicollinearity between the independent variables in the regression model. Perfect multicollinearity occurs when one or more independent variables can be exactly predicted by a linear combination of other independent variables. In other words, there is a perfect linear relationship between at least two independent variables in the model.

The formula for calculating VIF is:

$VIFi$=1/1−Ri21

Where:

$VIFi$ is the VIF for the ith independent variable.

- Ri2 is the R2 value obtained by regressing the ith independent variable against all other independent variables in the model.

When perfect multicollinearity exists, the R2 value in the denominator of the VIF formula becomes 1, leading to a division by zero and resulting in an infinite VIF value for the variable involved in the multicollinearity.

When perfect multicollinearity exists, the R2 value in the denominator of the VIF formula becomes 1, leading to a division by zero and resulting in an infinite VIF value for the variable involved in the multicollinearity.

Perfect multicollinearity can arise due to several reasons, including:

- **Linear Dependence**: One independent variable is a linear combination of other independent variables.
- **Dummy Variable Trap**: In regression models with dummy variables, perfect multicollinearity can occur when all categories of a categorical variable are included, leading to redundant information.

- **Data Errors**: Errors in data collection or entry can sometimes lead to apparent perfect multicollinearity.

When infinite VIF values are encountered, it is essential to identify and address the underlying multicollinearity issue to ensure the reliability and interpretability of the regression model. This may involve removing redundant variables, combining or transforming variables, or using regularization techniques to mitigate multicollinearity effects.

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
(3 marks)

Answer:

The Q-Q plot compares the quantiles of the sample data to the quantiles of the theoretical distribution, typically plotted on the x-axis and y-axis, respectively.

Importance of Q-Q plot in linear regression:

- **Validity of Inference**: Assessing the normality of residuals is crucial for ensuring the validity of statistical inference in linear regression. Normality assumptions underlie many statistical tests and procedures, such as hypothesis testing, confidence intervals, and model diagnostics. Q-Q plots provide a visual method to evaluate whether these assumptions are met.
- **Model Improvement**: If the Q-Q plot reveals significant departures from normality, it may prompt researchers to explore alternative modeling approaches or consider transformations of the response variable to better meet the assumptions of linear regression.
- **Diagnosing Model Assumptions**: Q-Q plots are part of a suite of diagnostic tools used to assess the assumptions of linear regression models. By examining the Q-Q plot along with other diagnostic plots (e.g., residual plots), researchers can identify potential issues with the model and make appropriate adjustments.

In summary, Q-Q plots are valuable tools in linear regression analysis for assessing the normality of residuals, which is a critical assumption underlying many statistical procedures. By visually comparing the sample data to a theoretical distribution, Q-Q plots help researchers evaluate the validity of inference and identify areas for model improvement.