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)**

Solution:

After conducting an analysis on the categorical columns using boxplots and bar plots, the following observations can be inferred:

- The fall season appears to have generated more bookings. Additionally, for each season, there has been a significant increase in booking counts from 2018 to 2019.
- The months of May, June, July, August, September, and October have witnessed the highest number of bookings. The trend shows an increase in bookings from the beginning of the year until the middle of the year, followed by a decrease towards the end of the year.
- It is evident that clear weather conditions attract more bookings, which aligns with common expectations.
- Thursdays, Fridays, Saturdays, and Sundays exhibit a higher number of bookings compared to the earlier days of the week.
- Non-holiday periods tend to have fewer bookings, which is reasonable as people may prefer to stay at home and spend time with their families during holidays.
- The number of bookings appears to be relatively consistent between working days and non-working days.
- The year 2019 has seen a higher number of bookings compared to the previous year, indicating positive progress in terms of business growth.

These observations provide valuable insights into the booking patterns and preferences, helping to understand customer behaviours and inform decision-making processes for the business.

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

Solution:

As we know we can represent a column having ‘n’ categories we can represent the categories by using ‘n-1’ dummies, pd.get_dummies will create ‘n’ dummies by

Using `drop_first=True’ It ensures that one dummy variable is dropped as a reference category, making the remaining dummy variables independent and providing clear comparisons between categories.

during dummy variable creation is important to avoid multicollinearity issues, improve model interpretability, and reduce model complexity.

For example, if we have a column called "students" with 3 categories: 'Class A', 'Class B', and 'Class C', using pd.get_dummies without drop_first=True would result in 3 dummy columns. However, by using drop_first=True, the first dummy column ('Class A') is dropped, and we can represent these categories with 2 dummies instead of 3. This approach simplifies the model, avoids redundancy, and facilitates clearer comparisons between categories.

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

Solution:

'temperature' is the variable which has the highest correlation in the target variable.

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

Solution:

I have assessed the validity of the Linear Regression Model based on the following five assumptions:

1. **Normality of error terms:**

- The assumption states that the error terms (residuals) should be normally distributed.
- Normality ensures that the residuals have a symmetrical distribution around zero with a constant variance.
- To verify this assumption, I examined the distribution of residuals and checked if they approximately follow a bell-shaped curve.

2. **Multicollinearity check:**

- Multicollinearity refers to the presence of high correlations between predictor variables in the regression model.
- Detecting and addressing significant multicollinearity is crucial because it can lead to unstable parameter estimates and difficulties in interpreting individual variable effects.
- I assessed multicollinearity by calculating the correlation matrix and evaluating the variance inflation factor (VIF) values to identify highly correlated predictors and eliminate them.

3. **Validation of linear relationship:**

- This assumption assumes a linear relationship between the predictor variables and the response variable.
- I examined this relationship through scatter plots and by assessing the linearity of residuals with respect to predicted values.
- A clear linear pattern in scatter plots or random scatter of residuals around the line indicates a linear relationship.

4. **Homoscedasticity:**

- Homoscedasticity assumes that the residuals have a constant variance across all levels of the predictor variables.

- Violations of homoscedasticity can result in heteroscedasticity, where the spread of residuals varies across different ranges of predictors.
- I assessed homoscedasticity by plotting residuals against predicted values and checking for any discernible patterns or trends.

5. **Independence of residuals:**
- This assumption assumes that the residuals are independent of each other, indicating no correlation or autocorrelation.
- Autocorrelation occurs when the residuals at one point in time are correlated with residuals at another point, suggesting a pattern or dependence.
- auto correlation is also evaluated on the correlation matrix.

By evaluating these assumptions, I ensured the validity and reliability of the linear regression model and made accurate interpretations of the results.

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

Solution:

The demand for shared bikes is primarily influenced by three significant features:

- Temperature
- Winter season
- September

These three factors have been found to contribute significantly to explaining the variations in the demand for shared bikes.

General Subjective Questions

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

Solution:

Linear regression is a supervised machine learning algorithm used for predicting continuous numeric values. It establishes a linear relationship between the input features (independent variables) and the target variable (dependent variable) by fitting a straight line that best represents the data. The goal of linear regression is to find the best-fitting line that minimizes the differences between the predicted values and the actual values.

Here's a detailed explanation of the linear regression algorithm:

1. **Assumptions**:
   - Linearity: Assumes a linear relationship between the independent variables and the target variable.
   - Independence: Assumes that the observations are independent of each other.
   - Homoscedasticity: Assumes that the variance of the errors is constant across all levels of the independent variables.
   - Normality: Assumes that the errors are normally distributed.

2. **Simple Linear Regression:**
   Simple linear regression deals with one independent variable (X) and one dependent variable (Y). The equation of a simple linear regression can be represented as:
   $Y = b0 + b1*X + \varepsilon$
   - Y: Dependent variable (target variable)
   - X: Independent variable (input feature)
   - b0: Intercept (the value of Y when X is zero)
   - b1: Slope (the change in Y for a unit change in X)
   - ε: Error term (residuals)

3. **Multiple Linear Regression:**
   Multiple linear regression extends the concept of simple linear regression to multiple independent variables. The equation can be represented as:
   $Y = b0 + b1*X1 + b2*X2 + ... + bn*Xn + \varepsilon$
   - X1, X2, ..., Xn: Independent variables (input features)
   - b1, b2, ..., bn: Coefficients corresponding to each independent variable
   - ε: Error term (residuals)

4. **Model Training:**
   The linear regression model is trained by estimating the coefficients (b0, b1, ..., bn) that minimize the sum of squared residuals (the difference between the predicted and actual values). This process is typically done using optimization techniques like Ordinary Least Squares (OLS) or gradient descent.

5. **Model Evaluation:**
   The trained model's performance is evaluated using various metrics such as mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), and R-squared (R2) score. These metrics assess how well the model fits the data and predicts the target variable.

6. **Making Predictions:**
   Once the model is trained and evaluated, it can be used to make predictions on new, unseen data. Given the values of the independent variables, the model calculates the predicted value of the dependent variable using the learned coefficients.

Linear regression is a simple yet powerful algorithm widely used for tasks such as trend analysis, forecasting, and understanding the relationship between variables.

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

Solution:

Anscombe's quartet is a set of four datasets that have identical statistical properties but exhibit distinct graphical patterns. It was introduced by the statistician Francis Anscombe in 1973 to highlight the importance of data visualization and to caution against relying solely on summary statistics.

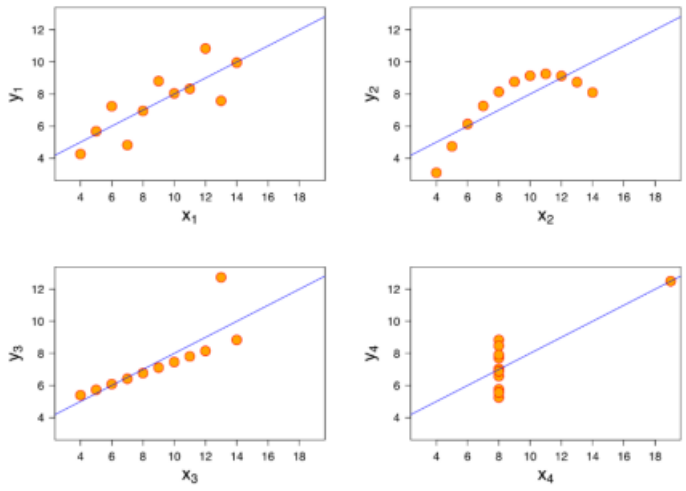
The four datasets in Anscombe's quartet consist of eleven pairs of x and y values. Despite having the same mean, variance, correlation, and linear regression parameters, the datasets have different distributions and relationships between the variables. Here is a detailed description of each dataset:

1. **Dataset I:**
   - x: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5
   - y: 8.04, 6.95, 7.58, 8.81, 8.33, 9.96, 7.24, 4.26, 10.84, 4.82, 5.68
   - Relationship: Approximately linear

2. **Dataset II:**
   - x: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5
   - y: 9.14, 8.14, 8.74, 8.77, 9.26, 8.10, 6.13, 3.10, 9.13, 7.26, 4.74
   - Relationship: Approximately linear, but with a slight upward curvature

3. **Dataset III:**
   - x: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5
   - y: 7.46, 6.77, 12.74, 7.11, 7.81, 8.84, 6.08, 5.39, 8.15, 6.42, 5.73
   - Relationship: Non-linear, with an apparent quadratic trend

4. **Dataset IV:**
   - x: 8, 8, 8, 8, 8, 8, 8, 19, 8, 8, 8
   - y: 6.58, 5.76, 7.71, 8.84, 8.47, 7.04, 5.25, 12.50, 5.56, 7.91, 6.89
   - Relationship: A single outlier significantly affects the linear regression line



The significance of Anscombe's quartet lies in challenging the notion that summary statistics alone are sufficient for understanding and analysing data. Despite having identical statistical properties, each dataset demonstrates unique patterns when visualized. This highlights the importance of data visualization in gaining insights, identifying outliers, assessing relationships, and verifying assumptions.

Anscombe's quartet serves as a reminder that relying solely on summary statistics without considering the data's graphical representation can lead to erroneous conclusions and oversimplification. It underscores the need to explore and visualize data to gain a comprehensive understanding of its characteristics and relationships.

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

Solution:

Pearson's R, also known as the Pearson correlation coefficient, is a statistical measure that quantifies the linear relationship between two continuous variables. It is denoted by the symbol "r" and takes values between -1 and 1.

Pearson's R is calculated using the following formula:

r = (Σ((Xi - X_mean) * (Yi - Y_mean))) / (sqrt(Σ(Xi - X_mean)^2) * sqrt(Σ(Yi - Y_mean)^2))

Where:
- Xi and Yi are the individual values of the two variables.
- X_mean and Y_mean are the means of the two variables.
- Σ denotes the summation symbol.

Key characteristics of Pearson's R:

1. Range: The value of r ranges from -1 to 1. A positive value indicates a positive linear relationship, a negative value indicates a negative linear relationship, and a value of 0 suggests no linear relationship between the variables.

2. Strength of Relationship: The magnitude of r indicates the strength of the linear relationship. A value closer to -1 or 1 signifies a stronger linear association, while values closer to 0 represent a weaker relationship.

3. Significance: The statistical significance of Pearson's R can be determined using hypothesis testing. It helps determine whether the observed correlation is statistically significant or occurred by chance. The p-value associated with the correlation coefficient is used to make this determination. A p-value below a chosen significance level (e.g., 0.05) suggests a significant correlation.

4**. Assumptions:** Pearson's R assumes that the relationship between the variables is linear, the variables are normally distributed, and there is homoscedasticity (constant variance of the residuals).

Pearson's R is commonly used in various fields, including statistics, social sciences, finance, and machine learning, to assess the strength and direction of the linear relationship between variables. It provides valuable insights into how two variables are related and aids in making predictions, understanding patterns, and identifying associations in the data.

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

Solution:

Scaling is a preprocessing step in machine learning that involves transforming the features of a dataset to a consistent scale. It ensures that all features have similar ranges, which can be important for certain machine learning algorithms and data analysis techniques. Scaling is performed to address issues related to the magnitude and units of the features, and to prevent features with larger values from dominating the learning process.

The main reasons for performing scaling are:

1. **Comparison of Features:** Scaling allows for a fair comparison between different features that have different units and scales. Without scaling, features with larger values may have a disproportionate influence on the model's learning process or distance-based algorithms.

2. **Gradient Descent Optimization**: Scaling helps in optimizing the performance of gradient descent-based algorithms, which rely on updating weights iteratively. Scaling the features can improve convergence and speed up the learning process.

3. **Regularization Techniques:** Scaling is often required when using regularization techniques such as Ridge regression or Lasso regression. These techniques penalize large weights, and without scaling, the regularization term may disproportionately penalize features with larger values.

There are two common types of scaling methods:

1**. Normalized Scaling (Min-Max Scaling):**

- Normalized scaling, also known as Min-Max scaling, rescales the features to a fixed range, typically between 0 and 1.
- The formula for normalized scaling is:

  X_scaled = (X - X_min) / (X_max - X_min)
- X is the original feature value, X_min and X_max are the minimum and maximum values of the feature, respectively.
- Normalized scaling preserves the original distribution of the data but scales it to a specific range.

2. **Standardized Scaling (Z-score normalization):**

- Standardized scaling, also known as Z-score normalization, transforms the features to have zero mean and unit variance.
- The formula for standardized scaling is:

  X_scaled = (X - X_mean) / X_std
- X is the original feature value, X_mean is the mean of the feature, and X_std is the standard deviation of the feature.
- Standardized scaling centers the data around zero and scales it based on the spread of the data.

The main difference between normalized scaling and standardized scaling lies in the range and distribution of the scaled data. Normalized scaling transforms the data to a fixed range (e.g., 0 to 1), while standardized scaling standardizes the data to have zero mean and unit variance. Normalized scaling is suitable when preserving the original distribution and range is important, while standardized scaling is useful when comparing features with different scales and for algorithms that assume normally distributed data.

In summary, scaling is performed to ensure consistency in feature scales, facilitate fair feature comparisons, and improve the performance of machine learning algorithms. Normalized scaling maintains the original distribution and scales the data to a fixed range, while standardized scaling centers the data around zero and scales it based on the spread of the data. The choice between the two depends on the specific requirements and characteristics of the dataset and the machine learning algorithm being used.

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

Solution:

The occurrence of an infinite value of Variance Inflation Factor (VIF) is known as "perfect multicollinearity." It happens when there is an exact linear relationship between one or more independent variables in a regression model. Perfect multicollinearity leads to unstable parameter estimates, making it impossible to compute accurate VIF values for the affected variables.

Here are a few common scenarios that can result in perfect multicollinearity and infinite VIF values:

1. **Duplicate or Redundant Variables:** When two or more variables in the dataset are identical or perfectly correlated, it leads to redundancy in the information they provide. For example, having both "height in inches" and "height in centimeters" as independent variables would introduce perfect multicollinearity.

2. **Data Transformation Issues**: Applying inappropriate data transformations can also result in perfect multicollinearity. For instance, if you convert a continuous variable into categorical bins and include all the bins as independent variables, it can lead to perfect multicollinearity.

3. **Creation of Derived Variables:** When new variables are created from existing variables using mathematical operations, it's essential to avoid introducing perfect multicollinearity inadvertently. For example, if you create a new variable by summing two existing variables, and those two variables are perfectly correlated, it will result in perfect multicollinearity.

4. **Including Interactions or Polynomial Terms**: Interaction terms or higher-order polynomial terms can introduce perfect multicollinearity if they involve variables that are perfectly correlated or linearly related. For instance, including both "age" and "age squared" as independent variables can lead to perfect multicollinearity.

To address the issue of perfect multicollinearity and infinite VIF values, it is necessary to identify and remove the redundant or perfectly correlated variables from the model. This can be done by carefully examining the variables, performing feature selection, or applying dimensionality reduction techniques such as principal component analysis (PCA).

It is important to note that while infinite VIF values indicate the presence of perfect multicollinearity, high VIF values (but not infinite) suggest strong multicollinearity between variables. In such cases, it is advisable to assess the impact of multicollinearity on the model's performance and consider addressing it by removing or transforming the correlated variables.

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

(3 marks)

Solution:

A Q-Q plot, short for quantile-quantile plot, is a graphical tool used to assess the distributional similarity between a given dataset and a theoretical distribution, such as the normal distribution. It plots the quantiles of the observed data against the quantiles of the theoretical distribution, allowing visual comparison and identification of deviations from the expected distribution.

The use and importance of a Q-Q plot in linear regression are as follows:

1. Assessing Normality Assumption:

In linear regression, one of the assumptions is that the residuals (the differences between the observed and predicted values) follow a normal distribution. The Q-Q plot can help evaluate the normality assumption by comparing the quantiles of the residuals with the quantiles expected from a normal distribution. If the points in the Q-Q plot fall approximately along a straight line, it suggests that the residuals are normally distributed. Deviations from the straight line indicate departures from normality.

2. Detecting Skewness and Outliers:

The Q-Q plot can reveal skewness and identify outliers in the data. If the points on the plot deviate from the straight line systematically, it indicates skewness or heavy tails in the distribution. Outliers can be identified as points that deviate significantly from the expected line. This information helps in understanding the distributional characteristics of the residuals and identifying influential data points that may impact the linear regression model.

3. Model Evaluation and Assumption Checking:

The Q-Q plot serves as a diagnostic tool to evaluate the adequacy of the linear regression model. Deviations from the expected line in the Q-Q plot may indicate model misspecification or violation of assumptions. If the Q-Q plot reveals substantial departures from the expected line, it suggests that the linear regression model may not be appropriate for the data, and further investigation or model refinement might be necessary.

4. Comparison of Distributions:

Besides assessing normality assumptions, Q-Q plots can also be used to compare the distributions of two datasets. By plotting the quantiles of two datasets against each other, it becomes easier to compare their distributions, identify differences in skewness, or detect systematic deviations.

In summary, the Q-Q plot is a valuable tool in linear regression for assessing the normality assumption, detecting skewness and outliers, evaluating model adequacy, and comparing distributions. It helps in understanding the distributional characteristics of the residuals and provides insights into potential issues that may affect the validity and reliability of the linear regression analysis.