# 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?

    - Fall season shows more users as compared to others, Also overall users are much higher in 2019 compared to previous year
    - Peak season is around mid of the year, and bookings are less towards starting and ending of the year
    - Clear weather has more customers followed by Misty.
    - Thurs, Fri & Sat show high number of customers & Sun shows low number of customers probably due to holiday and people spending more time at home
    - Bookings are higher on non-holidays as people would commute more on those days
    - Bookings are higher on working days as people would commute more on those days

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

    - It is important to remove the extra column which is created during dummy variable creation. For example, In the case of 3 types of values for a categorical column, if the value is not A or B , it is obviously C. So a 3rd variable for C is not required.

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

    - Temp variable has highest correlation with target variable.

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

    - Normality of error terms
        - Plotting error terms and checking distribution is normal
    - Multicollinearity check
        - Checked correlation between variables using heatmap
    - Linear relationship validation
        - Checked using CCPR plot
    - Homoscedasticity
        - Plotted scatter plot of Residual values
    - Independence of residuals
        - Using Durbin-Watson value

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

- Temp
- Light_snowrain
- year

# General Subjective Questions

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

Answer:
Linear regression, in essence, is a statistical framework employed to examine the linear association between a dependent variable and a specified set of independent variables. This notion of linearity implies that when one or more independent variables experience alterations, whether in an upward or downward direction, the dependent variable will also respond accordingly, mirroring these changes. This relationship can be expressed mathematically through the following equation:

$Y = mX + c$

In this equation:
- Y represents the dependent variable under scrutiny for prediction.
- X stands for the independent variable employed for predictive purposes.
- m denotes the slope of the regression line, signifying the impact of X on Y.
- c is a constant known as the Y-intercept, indicating that when X equals 0, Y will equal c.

Furthermore, this linear relationship can manifest itself in two distinct ways:
- Positive Linear Relationship: This type of linear relationship emerges when both the independent and dependent variables increase in tandem.
- Negative Linear Relationship: Conversely, a negative linear relationship occurs when an increase in the independent variable corresponds to a decrease in the dependent variable.

In the context of Linear Regression, several assumptions regarding the dataset come into play:
1. Multi-collinearity: The model assumes minimal or no presence of multi-collinearity within the data. Multi-collinearity arises when the independent variables or features exhibit interdependencies.
2. Auto-correlation: Another assumption involves minimal or no auto-correlation within the data. Auto-correlation arises when there is a dependency between residual errors.
3. Relationship between variables: Linear regression expects that the relationship between response and feature variables adheres to a linear pattern.
4. Normality of error terms: Error terms are expected to follow a normal distribution.
5. Homoscedasticity: Residual values should not display any discernible patterns.

2. Explain the Anscombe’s quartet in detail. (3 marks)
Answer:

Anscombe's quartet is a set of four small datasets that have nearly identical simple descriptive statistics but exhibit strikingly different patterns when graphically visualized. This collection of datasets was created by the statistician Francis Anscombe in 1973 to illustrate the importance of data visualization and the limitations of relying solely on summary statistics like means and variances. Anscombe's quartet serves as a powerful reminder that examining data graphically can reveal nuances and relationships that might be missed when only considering summary statistics.

Here are the details of Anscombe's quartet, which consists of four datasets (I, II, III, and IV):

1. Dataset I:
   - x-values: [10.0, 8.0, 13.0, 9.0, 11.0, 14.0, 6.0, 4.0, 12.0, 7.0, 5.0]
   - y-values: [8.04, 6.95, 7.58, 8.81, 8.33, 9.96, 7.24, 4.26, 10.84, 4.82, 5.68]

2. Dataset II:
   - x-values: [10.0, 8.0, 13.0, 9.0, 11.0, 14.0, 6.0, 4.0, 12.0, 7.0, 5.0]
   - y-values: [9.14, 8.14, 8.74, 8.77, 9.26, 8.10, 6.13, 3.10, 9.13, 7.26, 4.74]

3. Dataset III:
   - x-values: [10.0, 8.0, 13.0, 9.0, 11.0, 14.0, 6.0, 4.0, 12.0, 7.0, 5.0]
   - y-values: [7.46, 6.77, 12.74, 7.11, 7.81, 8.84, 6.08, 5.39, 8.15, 6.42, 5.73]

4. Dataset IV:
   - x-values: [8.0, 8.0, 8.0, 8.0, 8.0, 8.0, 8.0, 19.0, 8.0, 8.0, 8.0]
   - y-values: [6.58, 5.76, 7.71, 8.84, 8.47, 7.04, 5.25, 12.50, 5.56, 7.91, 6.89]

Despite having similar means, variances, and correlation coefficients (approximately 0.816 for all four datasets), the data points in each dataset follow different relationships when graphed. This phenomenon is highlighted in the scatterplots and regression lines for each dataset:

- Dataset I shows a roughly linear relationship between x and y.
- Dataset II follows a linear relationship with one outlier that heavily influences the regression line.
- Dataset III is not linear and exhibits a curve.
- Dataset IV has a clear linear relationship except for an outlier that significantly affects the regression line.

The key takeaway from Anscombe's quartet is that summary statistics alone cannot capture the full story of a dataset. Visualization tools, like scatterplots and regression lines, provide valuable insights into the underlying data patterns, helping researchers and analysts make more informed decisions about their data. This quartet serves as a reminder of the importance of exploring data graphically to understand its behavior comprehensively.

3. What is Pearson’s R?

Answer:
Pearson's correlation coefficient, often denoted as "r" or "Pearson's R," is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. It assesses how well the values of one variable can be predicted by the values of another variable in a linear fashion.

Here are some key points about Pearson's correlation coefficient (r):

1. **Range of Values:** The value of Pearson's r ranges from -1 to 1.
   - An r value of 1 indicates a perfect positive linear relationship, meaning that as one variable increases, the other also increases proportionally.
   - An r value of -1 indicates a perfect negative linear relationship, meaning that as one variable increases, the other decreases proportionally.
   - An r value of 0 suggests no linear relationship between the two variables.

2. **Direction:** The sign of the correlation coefficient (positive or negative) indicates the direction of the linear relationship:
   - Positive r values indicate a positive linear relationship.
   - Negative r values indicate a negative linear relationship.

3. **Strength:** The magnitude (absolute value) of r indicates the strength of the linear relationship. The closer the absolute value of r is to 1, the stronger the linear relationship.

4. **Assumptions:** Pearson's correlation coefficient assumes that the relationship between the two variables is linear, that both variables are normally distributed, and that there is homoscedasticity (constant variance) in the data. It is also sensitive to outliers, meaning that extreme data points can have a significant impact on the correlation coefficient.

$$r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}}$$

Pearson's correlation coefficient is widely used in various fields, including statistics, social sciences, economics, and many scientific disciplines, to assess the degree and direction of association between two continuous variables. It helps researchers and analysts understand whether there is a linear relationship between variables and to what extent they are related.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Answer:
**Scaling** in the context of data preprocessing refers to the process of transforming data to fit within a specific range or distribution. It is commonly performed on numerical features or variables in a dataset. Scaling is essential in many machine learning and data analysis tasks to

ensure that variables contribute equally to the analysis, prevent issues related to different scales, and improve the performance of certain algorithms. Two common methods of scaling data are **normalized scaling** and **standardized scaling**, and they serve slightly different purposes.

**Why Scaling is Performed:**

1. **Equal Weighting:** Some machine learning algorithms, such as k-means clustering or principal component analysis (PCA), rely on distance measures between data points. Variables with larger scales can dominate the calculations and have a disproportionate influence on the results. Scaling ensures that all variables contribute equally.

2. **Algorithm Convergence:** Many optimization algorithms used in machine learning, like gradient descent, converge faster when the features are scaled. Scaling can help speed up the training process.

3. **Interpretability:** Scaling can make the interpretation of coefficients or feature importance in linear models more straightforward because the coefficients represent the change in the target variable per unit change in the feature.

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

Normalized scaling, also known as min-max scaling, transforms data to a specific range, typically between 0 and 1. It is achieved using the following formula for each data point:

X_new =(X - Xmin)/(Xmax-Xmin)

- **Range:** The transformed data will lie in the range [0, 1].
- **Sensitivity to Outliers:** Normalized scaling is sensitive to outliers because it depends on the minimum and maximum values in the dataset.

**Standardized Scaling (Z-score Standardization):**

Standardized scaling, also known as z-score standardization, transforms data to have a mean of 0 and a standard deviation of 1. It is achieved using the following formula for each data point:

Xnew= X- Xmean/std(X)

- **Mean and Standard Deviation:** The transformed data will have a mean of 0 and a standard deviation of 1.
- **Robustness:** Standardized scaling is more robust to outliers than min-max scaling because it is not affected by extreme values to the same extent.

**Key Differences:**

1. **Range:** The primary difference is the range of the transformed data. Normalized scaling maps data to the [0, 1] range, while standardized scaling centers the data around 0 with a standard deviation of 1.

2. **Sensitivity to Outliers:** Normalized scaling is sensitive to outliers because it depends on the range of data. Standardized scaling is less affected by outliers because it uses the mean and standard deviation.

3. **Use Cases:** Normalized scaling is suitable when you want to preserve the original range of the data and have a specific requirement for the transformed range. Standardized scaling is preferred when you want to standardize data for algorithms that assume a normal distribution or for cases where outliers may be present.

The choice between normalized and standardized scaling depends on the specific requirements of your data and the machine learning algorithm you plan to use.

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

Answer:
Infinite VIF (Variance Inflation Factor) values occur due to perfect multicollinearity or near-perfect multicollinearity among predictor variables in a regression model. Perfect multicollinearity makes it mathematically impossible to calculate VIF, resulting in infinity. High VIF values may also arise from nearly perfect multicollinearity or small sample sizes relative to predictor variables. To address this issue, identify and resolve multicollinearity, consider collecting more data for stability, or apply regularization techniques like ridge or lasso regression to mitigate multicollinearity effects.

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

Answer:
A Q-Q plot, short for Quantile-Quantile plot, is a graphical tool used in statistics to assess whether a dataset follows a particular theoretical distribution, such as the normal distribution. It is a visual representation of how the quantiles (ordered values) of your dataset compare to the quantiles of the theoretical distribution. Q-Q plots are valuable in linear regression and other statistical analyses for several reasons:

****Use of Q-Q Plot:****

1. **Distribution Assessment:** Q-Q plots are primarily used to check if the data follows a specific theoretical distribution. By comparing the quantiles of your dataset to the quantiles of the theoretical distribution, you can visually assess the goodness of fit.

2. **Identifying Departures from Normality:** In linear regression, one common assumption is that the residuals (the differences between observed and predicted values) should follow a normal distribution. A Q-Q plot of the residuals helps you determine if this assumption holds. If the points on the Q-Q plot deviate significantly from the diagonal line, it suggests departures from normality.

3. **Outlier Detection:** Q-Q plots can also help identify outliers or extreme values in your dataset. Outliers may appear as points that deviate substantially from the diagonal line in the plot.

4. **Model Validation:** Q-Q plots are an essential tool in model validation. They allow you to assess whether the residuals of your linear regression model meet the normality assumption. A well-fitting model should result in residuals that closely follow a normal distribution, which is reflected in the Q-Q plot.

**Importance of Q-Q Plot in Linear Regression:**

In linear regression, the Q-Q plot plays a crucial role in the following ways:

1. **Normality Check:** Linear regression assumes that the residuals are normally distributed. A Q-Q plot of the residuals helps you verify this assumption. If the plot shows a roughly straight diagonal line, it indicates that the residuals follow a normal distribution, supporting the validity of your model.

2. **Residual Assessment:** The Q-Q plot allows you to visually identify patterns or deviations in the residuals. For example, if the plot exhibits a significant curve or any non-linearity, it may indicate issues with the model, such as omitted variables or nonlinear relationships.

3. **Outlier Detection:** Outliers in the residuals can be detected through Q-Q plots. Outliers may appear as data points that deviate significantly from the expected straight line, indicating unusual patterns in the residuals.

4. **Model Assumption Testing:** Q-Q plots are a valuable part of a battery of tests and diagnostics used to assess the validity of the linear regression model. They provide a graphical representation that can be easily interpreted alongside other statistical tests.

In summary, Q-Q plots are an important tool in linear regression for assessing the normality of residuals, identifying outliers, and validating model assumptions. By visually inspecting the Q-Q plot, you can gain insights into the quality and appropriateness of your regression model.