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?**

Answer: Post performing the EDA, I can infer on the below points:

- With rise in Temperature, the number of bike rental increases.
- Similarly, with rise in Feel of Temperature, the number of bike rental increases
- Slight decline in number of bike rental with rise of Humidity where majorly Humidity ranges roughly from 40 to 80
- With rise of wind speed, there is a decline in number of bike rental
- Majority of bike rental happens where the weather is Clear, Few clouds, partly cloudy and the rental gradually decreases as the weather deteriorates.
- During Heavy rainfall/snow there is no bike rentals
- There is a steep rise in number of bike rentals in 2019 compared to 2018 to be precise post March 2019
- Majority of bike rental has been observed in fall season and least during "spring"

2. **Why is it important to use drop first=True during dummy variable creation?**

Answer: During dummy variable creation, "n" number of categorical feature is created. These n categorical variables are collinear, i.e. one dummy variable can predicted by combination of collinear variables. This can lead to issue with the coefficient and multi-collinearity and the model can become unstable and performance can be impacted.

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

Answer: With 'cnt' i.e. total bike rental, 'temp' and 'atemp' (temperature and feel of temperature), suggest positive correlation.

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

Answer: Below assumptions has been validated after building the model:

- As the VIF value of is within threshold (i.e. < 5), we can say there is no issue of multi-collinearity

- As the p-value is within threshold (i.e. < 0.05), we can reject the null hypothesis, i.e. the variable is statistically significant and there is a strong evidence that the variables has a relationship with dependent variables.
- Based on the regplot, between residual and predicted values, the linearity assumption is satisfied.

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

Answer: Based on the p-vale (i.e. p-vale less than 0.05), features contributing significantly are **temp, weathersit_Light Snow-Rain, yr, season_spring, weathersit_Mist, weekday, season_winter, season_summer**.

Now, higher the coefficient, higher is the contribution factor:

Based on that **temp (**i.e. **Temperature in °C)** has coefficient value of **0.4973**, **weathersit_Light Snow-Rain (**i.e. **Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds)** has coefficient value of **-0.2807** and **yr (**i.e. **year)** has coefficient value of **0.2461** are the top 3 features contributing significantly towards explaining the demand of the shared bikes.

General Subjective Questions

1. Explain the linear regression algorithm in detail

Answer: Linear regression is a machine learning algorithm used for predicting a continuous dependent variable based on one or more independent variables also called features. It is a supervised learning where the algorithm learns from a set of input data also called the training data to predict the output.

Linear regression assumes linear relationship between the dependent variable "y" and the independent variable(s) "x".

For single feature (Simple Linear Regression), the expression is:

$y = \beta_0 + \beta_1 x + \epsilon$, where β_0 is the intercept, β_1 is the coefficient (slopes/gradient) for feature x and ϵ is error term (difference between actual and predicted values)

For multiple feature (Multiple Linear Regression), the expression is:

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots\ldots\ldots + \beta_n x_n + \epsilon$, where β_0 is the intercept, $\beta_1, \beta_2, \ldots, \beta_n$ are the coefficient (slopes/gradient) for each feature respectively and ϵ is error term (difference between actual and predicted values).

The goal of linear regression is to find the values of $\beta_0, \beta_1, \beta_2, \ldots, \beta_n$ that minimizes the error between the actual and predicted values.

Train the model: We need to find the coefficients $\beta_0, \beta_1, \beta_2, \ldots, \beta_n$. This can be achieved by using the methods like **Ordinary Least Squares (OLS)**.

Once the relationship has been determined we use metrics such as **R-squared (R^2)** to evaluate the performance of a linear regression model.

R-squared (R^2): Proportion of the variance in the dependent variable that is predictable from the independent variables.

While determining the Linear Regression following assumptions are made:

a. Linearity
b. Homoscedasticity
c. Independence
d. Normality

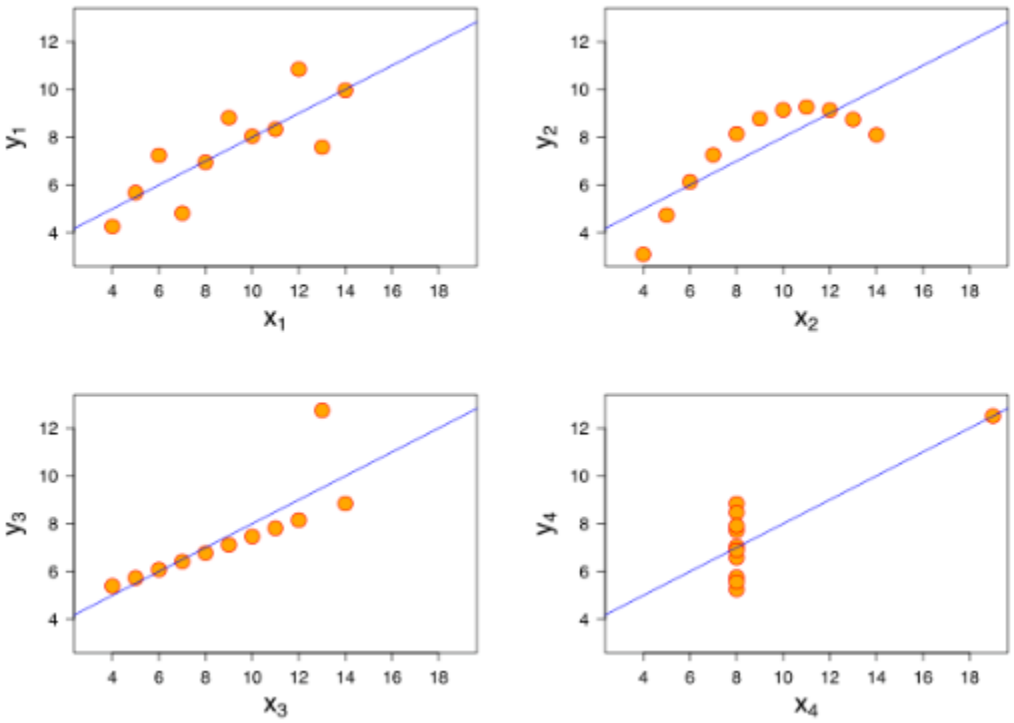
2. Explain the Anscombe's quartet in detail.

Answer: Anscombe's quartet is a famous statistical example introduced by the philosopher and statistician Francis Anscombe in 1973. The quartet consists of four datasets that have nearly identical simple descriptive statistics (such as means, variances, and correlation coefficients) but differ greatly in their distributions and graphical representations. This example was designed to illustrate the importance of visualizing data before performing statistical analysis, as it shows that summary statistics alone can be misleading.

Anscombe's quartet consists of four different datasets, each with 11 data points. Despite having identical summary statistics, they are visually distinct:

- **Dataset I**
  - This dataset represents a classic linear relationship.
  - When plotted, it shows a clear linear trend with a positive slope.
- **Dataset II**
  - This dataset also appears to have a linear relationship, but there is a notable outlier at the top right corner.
  - The outlier heavily influences the linear fit, making it seem like there is a strong relationship even though the majority of the data points don't follow the trend closely.
- **Dataset III**

    - This dataset has a nonlinear relationship. The points are fitted to a curve rather than a straight line, specifically following a quadratic trend.
    - The relationship between the variables is not linear, which would be missed if only linear regression were applied.
- **Dataset IV**
    - This dataset has a linear trend, but it is designed to illustrate the effect of an outlier in a different way. The data points form a vertical line with a single point far from the others.
    - This outlier is influential in the regression fit but doesn't follow the same pattern as the other data points.



Summary Statistics

For each of the four datasets, the following statistics are identical:

- **Mean of x**: 9
- **Mean of y**: 7.5
- **Variance of x**: 11
- **Variance of y**: 4.12
- **Correlation between x and y**: 0.816
- **Regression line**: y=3+0.5xy = 3 + 0.5xy=3+0.5x

Importance of Visualization

Anscombe's quartet highlights several key points about statistical analysis:

- **Summary Statistics Can Be Deceptive**: The identical summary statistics for all four datasets show that relying solely on these numbers can be misleading. The nature of the data and the presence of outliers can affect the interpretation significantly.
- **Graphical Representation Is Crucial**: By plotting the data, one can see patterns, trends, and anomalies that summary statistics alone might obscure. Each dataset in the quartet looks different when graphed, revealing important aspects of the data's structure.
- **Context Matters**: Understanding the context of the data and the underlying distribution is essential for proper analysis and interpretation. Visualizations help in identifying patterns and making more informed conclusions.

In summary, Anscombe's quartet serves as a powerful illustration of why it's important to visualize data, as different datasets with the same statistical summary can convey very different stories.

3. What is Pearson's R?

Answer: **Pearson's r**, also known as **Pearson's correlation coefficient**, is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables. Here's a detailed explanation:

Definition and Formula

Pearson's **r** is defined as:

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

where:

- n is the number of data pairs.
- $\sum xy$ is the sum of the products of paired scores.
- $\sum x$ is the sum of the x-values.
- $\sum y$ is the sum of the y-values.
- $\sum x^2$ is the sum of the squares of the x-values.
- $\sum y^2$ is the sum of the squares of the y-values.

Interpretation

- **Value Range**: Pearson's r ranges from -1 to 1.
  - **r=1**: Perfect positive linear relationship; as x increases, y increases proportionally.
  - **r=−1**: Perfect negative linear relationship; as x increases, y decreases proportionally.
  - **r=0**: No linear relationship; changes in x do not predict changes in y linearly.
- **Strength of the Relationship**:

- **0.1 to 0.3** (or -0.1 to -0.3): Weak correlation.
- **0.3 to 0.5** (or -0.3 to -0.5): Moderate correlation.
- **0.5 to 1.0** (or -0.5 to -1.0): Strong correlation.

Assumptions

Pearson' **r** assumes:

- **Linearity**: The relationship between the two variables should be linear.
- **Homogeneity of Variance**: The variability of the data should be consistent across the range of the variables.
- **Normality**: Both variables should be approximately normally distributed (though this is less critical in larger samples).

Use Cases

- **Descriptive Statistics**: Pearson's **r** is used to describe the strength and direction of a relationship between two variables.
- **Predictive Modeling**: It is used in the context of linear regression to determine how well one variable can predict another.

Limitations

- **Nonlinearity**: Pearson's **r** does not capture non-linear relationships. For example, it might be close to zero even if the relationship is strong but non-linear.
- **Outliers**: Pearson's **r** can be heavily influenced by outliers, which can skew the correlation coefficient significantly.

In summary, Pearson's correlation coefficient is a valuable tool for assessing linear relationships between two continuous variables, but it should be interpreted with an understanding of its limitations and assumptions.

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

Answer: **Scaling** is a technique used in data processing to adjust the range and distribution of your dataset's features. It's important because many machine learning algorithms work better when all the features are on a similar scale. Here's a rundown of why you might need scaling and the difference between two common types: normalized scaling and standardized scaling.

Why Do We Scale Data?

1. **Boosts Model Performance**: Some algorithms, like those that use gradient descent, can struggle if features have very different scales. Scaling helps these models learn more efficiently.
2. **Equalizes Feature Impact**: For algorithms that rely on distance calculations (like k-nearest neighbors), features with larger ranges can disproportionately affect the outcome. Scaling ensures every feature contributes equally.
3. **Speeds Up Training**: Scaling can help models train faster by providing a more uniform gradient, which makes it easier for the model to find the best parameters.

Types of Scaling

1. **Normalized Scaling (Min-Max Scaling)**:
   - **What It Does**: This method adjusts your data to fit within a specific range, like 0 to 1. It uses this formula:

$$x' = \frac{x - \min(X)}{\max(X) - \min(X)}$$

   - **Pros and Cons**:
     - **Pros**: Keeps values within a set range, which can be useful for certain algorithms.
     - **Cons**: Can be thrown off by outliers because the min and max values can be skewed.
2. **Standardized Scaling (Z-score Normalization)**:
   - **What It Does**: This method changes your data based on its mean and standard deviation. It uses this formula:

$$x' = \frac{x - \mu}{\sigma}$$

     where μ is the mean and σ is the standard deviation.

   - **Pros and Cons**:
     - **Pros**: Transforms data to have a mean of 0 and a standard deviation of 1, which is useful if your data has outliers or needs to maintain its distribution properties.
     - **Cons**: Does not limit the range of values, so data isn't bounded.

Key Differences

- **Range**: Normalized scaling keeps data within a specific range (like 0 to 1), while standardized scaling adjusts data to have a mean of 0 and a standard deviation of 1 but doesn't set a fixed range.
- **Outlier Sensitivity**: Normalized scaling can be affected by outliers because they affect the minimum and maximum values. Standardized scaling is more robust to outliers since it uses mean and standard deviation.

- **Use Case**: Normalized scaling is handy when you need values within a certain range. Standardized scaling is better if you want to keep the data's overall distribution intact, especially when dealing with algorithms that assume normally distributed data.

Choosing the right type of scaling depends on your model's needs and the nature of your data.

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

Answer: The Variance Inflation Factor (VIF) measures how much the variance of an estimated regression coefficient increases due to multicollinearity in the model. If the VIF for a particular predictor is very high or even infinite, it indicates a severe problem with multicollinearity. Here's why this can happen:

Reasons for Infinite VIF

- **Perfect Multicollinearity**:
    1. **Definition**: Perfect multicollinearity occurs when one predictor variable in the regression model is a perfect linear combination of one or more other predictor variables. This means that the predictor can be exactly predicted from other predictors.
    2. **VIF Implication**: In cases of perfect multicollinearity, the correlation between the predictor and other predictors is 1 or -1, leading to an infinite or undefined VIF because the formula for VIF involves division by zero. Specifically, if $\mathbf{R^2}$ (the coefficient of determination) is 1 for the regression of a predictor on all other predictors, then $\mathbf{1/(1-R^2)}$ results in an infinite VIF.
- **Redundant Predictors**:
    1. **Definition**: Redundant predictors are variables that provide no new information beyond what is already provided by other variables in the model.
    2. **VIF Implication**: If a predictor is redundant and perfectly correlated with other predictors, its VIF can become extremely high or infinite.

Understanding VIF Formula

The VIF for a predictor X_i is calculated as:

where R_i^2 is the coefficient of determination of the regression of X_i on all the other predictors. If R_i^2=, which means the predictor is perfectly predicted by the other predictors, the denominator becomes zero, leading to an infinite VIF.

$$\text{VIF}(X_i) = \frac{1}{1 - R_i^2}$$

Practical Implications

- **Model Interpretation**: An infinite VIF indicates that there's a fundamental problem with the model's predictors being linearly dependent. This makes it difficult to determine the individual effect of each predictor because their effects are indistinguishable from one another.
- **Solutions**: To address infinite VIF values, consider removing one or more of the collinear predictors from the model. Another approach is to use techniques such as Principal Component Analysis (PCA) to reduce dimensionality and remove multicollinearity.

In summary, infinite VIF values are a clear sign of perfect multicollinearity or redundant predictors, which need to be addressed to ensure a reliable and interpretable regression model.

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

Answer: The A **Q-Q plot** (Quantile-Quantile plot) is a graph that helps you see if your data follows a specific theoretical distribution, like the normal distribution. Here's how it works:

- **Quantiles**:
    1. **Theoretical Quantiles**: Values you would expect from a theoretical distribution (e.g., normal distribution).
    2. **Sample Quantiles**: Values from your actual data.
- **Plotting**:
    1. You plot the sample quantiles on the x-axis.
    2. You plot the theoretical quantiles on the y-axis.
- **Interpreting the Plot**:
    1. If your data follows the theoretical distribution, the points should fall roughly along a straight line (often a 45-degree line).
    2. If the points deviate significantly from this line, your data does not follow the theoretical distribution.

Use a Q-Q Plot in Linear Regression

In linear regression, a Q-Q plot is used to check one of the key assumptions of the model: that the residuals (errors) should be normally distributed. Here's why this matters:

- **Assumption of Normality**:

    1. **Why It's Important**: For the linear regression results (like p-values and confidence intervals) to be valid, the residuals should ideally follow a normal distribution.
    2. **How Q-Q Plot Helps**: By plotting the residuals' quantiles against the normal distribution quantiles, you can visually check if the residuals are normally distributed.
- **Model Diagnostics**:
    1. **Spotting Outliers**: Points that don't fit the straight line in the Q-Q plot might be outliers or anomalies in your residuals.
    2. **Evaluating Fit**: If the Q-Q plot shows significant deviations from the straight line, it might indicate that your model isn't capturing some aspects of the data.
- **Improving Your Model**:
    1. **Transformations**: If your residuals are not normal, you might need to transform the dependent variable (like taking the log or square root) to better meet the assumption.

Importance of a Q-Q Plot

1. **Valid Inferences**:
    - Ensures that the conclusions and predictions made from your linear regression model are based on valid assumptions.
2. **Model Refinement**:
    - Helps in identifying issues with the model and guiding improvements for better accuracy.
3. **Diagnostic Check**:
    - A simple yet effective visual tool to check the distribution of residuals and validate the model's assumptions.

In essence, a Q-Q plot is a valuable diagnostic tool in linear regression that helps ensure the residuals are normally distributed, thereby validating the reliability of the model's results.