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?

- Demand for bike rentals is observed more in the summer and the fall season
- People are more likely to rent bikes in in September and October
- Saturday, Wednesday and Thursday are the days where more bikes are rented
- People prefer to rent bikes when the weather is clear
- 2019 observed increase of bike rentals when compared to 2018 which indictaes that the business is growing
- Holidays observe more bike rentals

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

A: We use drop_first=True, when creating dummy variables as a good practice to avoid issues related to multicollinearity, simplify interpretation, reduce dimensionality, and improve the performance and stability of the statistical model.

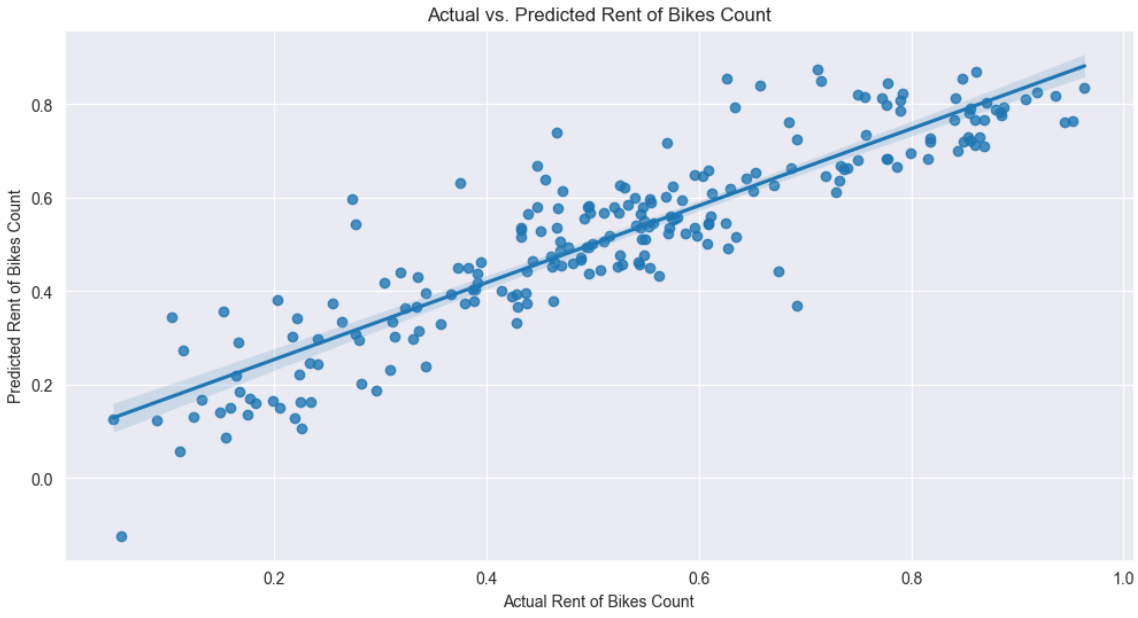
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

A: "temp" is the variable which has the highest correlation with target variable i.e. 0.63.

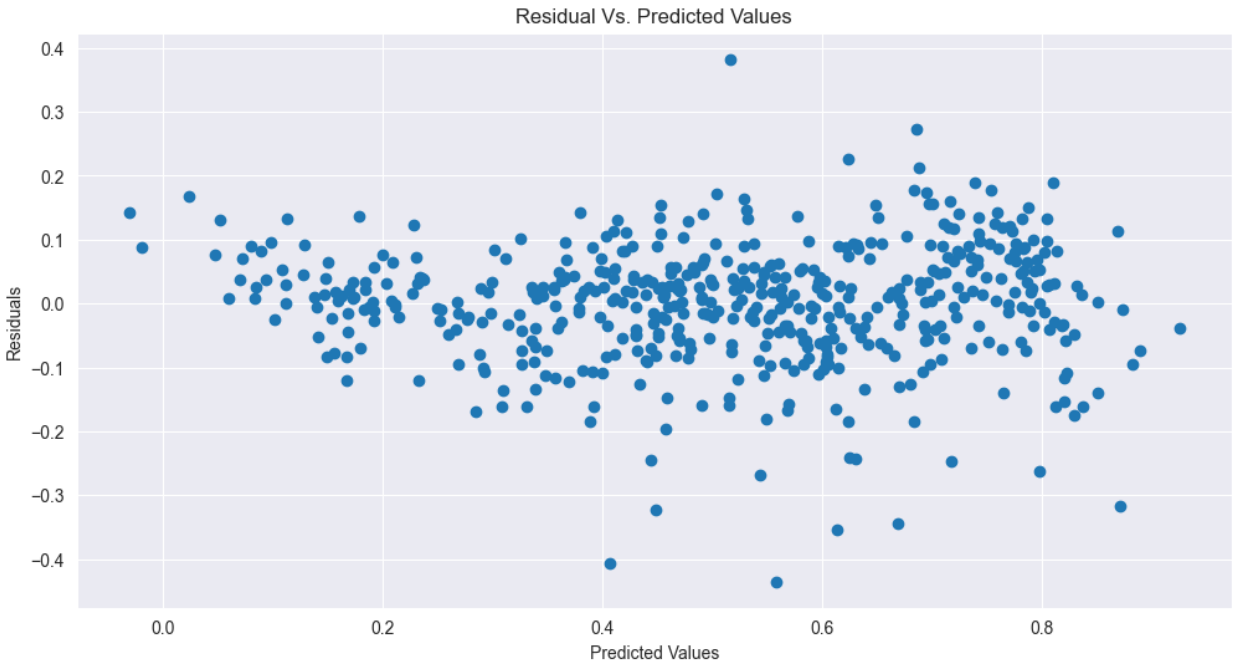
4. How did you validate the assumptions of Linear Regression after building the model on the training set?

A: The assumptions of linear regression are validated to be true:

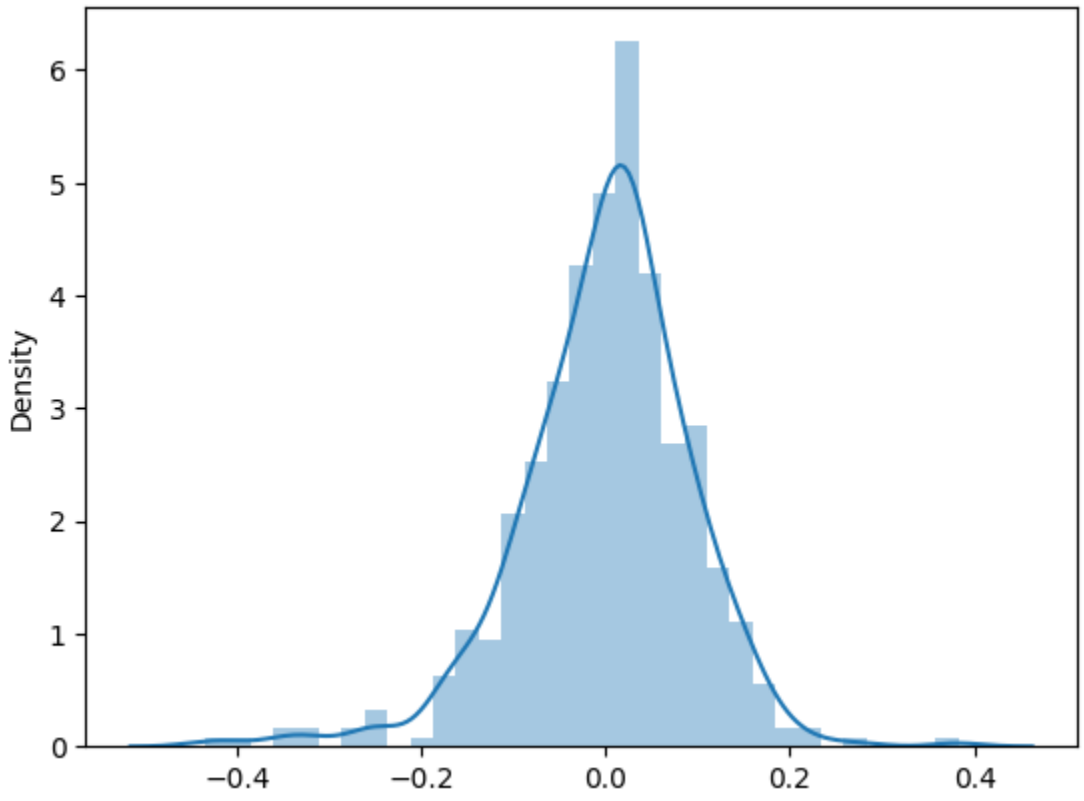
1. Linear relationship between independent and dependent variables:



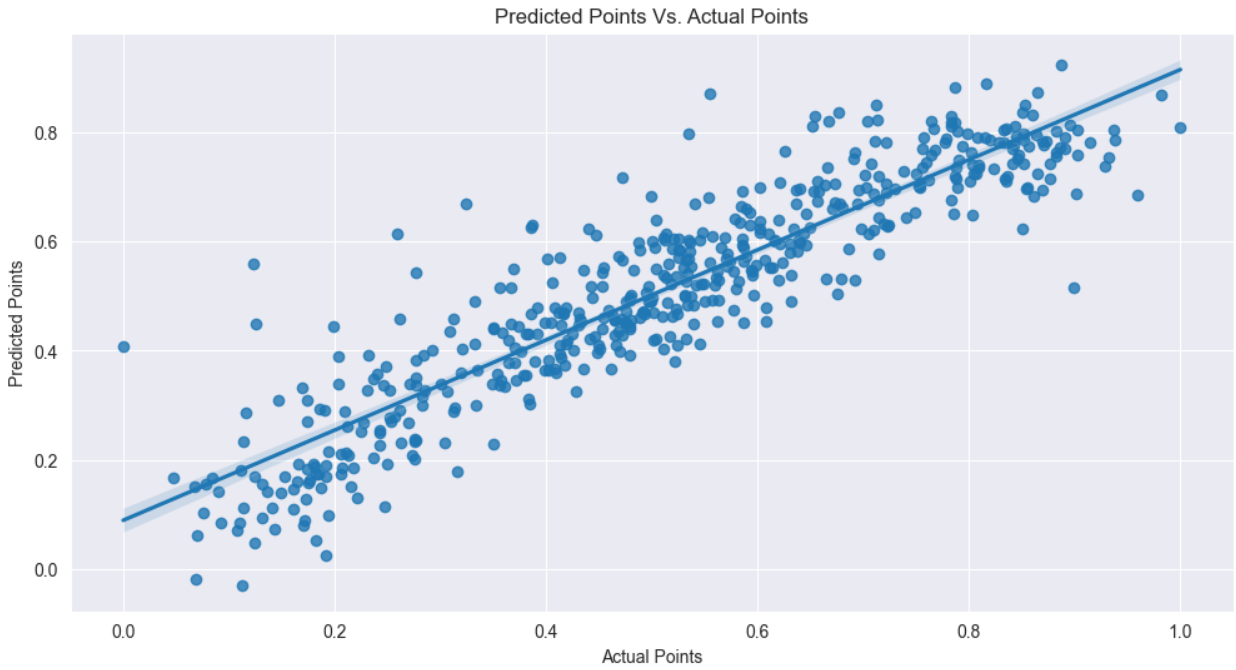
2. . Error terms are independent of each other:



3. Error terms are normally distributed:



4. Error terms have constant variance (homoscedasticity):



5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
A: After looking at the results, I believe that ‘yr’, ‘weathersit’ and ‘season’ contribute significantly towards explaining the demand of the shared bikes.

General Subjective Questions

1. Explain the linear regression algorithm in detail.
A: Linear regression is a fundamental supervised machine learning algorithm used for modeling the relationship between a dependent variable (target) and one or more independent variables (features or predictors). It assumes a linear relationship between the features and the target. Here's a detailed explanation of linear regression:

1. Objective of Linear Regression:
Linear regression aims to find a linear equation that best describes the relationship between the independent variables (features) and the dependent variable (target).
The linear equation allows us to make predictions about the target variable based on the values of the independent variables.

2. Assumptions of Linear Regression:
Linearity: The relationship between the features and the target is assumed to be linear.
Independence: The errors (residuals) should be independent of each other.

Homoscedasticity: The variance of the errors should be constant across all levels of the independent variables.
Normality: The errors should follow a normal distribution.
No multicollinearity: The independent variables should not be highly correlated with each other.

3. Simple Linear Regression:
In simple linear regression, there is one independent variable (feature) and one dependent variable (target).
The linear equation can be represented as: $Y = b0 + b1 * X + \varepsilon$, where:
Y is the target variable.
X is the independent variable.
b0 is the y-intercept (constant).
b1 is the slope (coefficient) that represents the change in Y for a one-unit change in X.
ε represents the error term, which accounts for the unexplained variation in Y.

4. Multiple Linear Regression:
In multiple linear regression, there are multiple independent variables (features) and one dependent variable (target).
The linear equation can be extended to: $Y = b0 + b1 * X1 + b2 * X2 + ... + bn * Xn + \varepsilon$, where:
Y is the target variable.
X1, X2, ..., Xn are the independent variables.
b0 is the y-intercept (constant).
b1, b2, ..., bn are the slopes (coefficients) for each independent variable.
ε represents the error term.

5. Training the Model:
The goal of training is to find the optimal values for the coefficients (b0, b1, b2, ...) that minimize the sum of squared residuals (the difference between predicted and actual values).

6. Evaluation and Prediction:
Common evaluation metrics for linear regression models include Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared (R^2).
Once the model is trained and evaluated, it can be used to make predictions on new, unseen data by plugging in the values of the independent variables into the linear equation.

2. Explain the Anscombe's quartet in detail.
A: Anscombe's quartet is a famous example in statistics that highlights the importance of data visualization and the limitations of relying solely on summary statistics. It consists of four datasets, each containing 11 data points, with two continuous variables (x and y). On the surface, these datasets have nearly identical summary statistics, but they exhibit very different relationships when visualized. The quartet was created by the statistician Francis Anscombe in 1973 to emphasize the need to visualize data before drawing conclusions. Let's explore each dataset in Anscombe's quartet in detail:

Dataset I:

x-values: [10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5]
y-values: [8.04, 6.95, 7.58, 8.81, 8.33, 9.96, 7.24, 4.26, 10.84, 4.82, 5.68]
Dataset II:

x-values: [10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5]
y-values: [9.14, 8.14, 8.74, 8.77, 9.26, 8.10, 6.13, 3.10, 9.13, 7.26, 4.74]
Dataset III:

x-values: [10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5]
y-values: [7.46, 6.77, 12.74, 7.11, 7.81, 8.84, 6.08, 5.39, 8.15, 6.42, 5.73]
Dataset IV:

x-values: [8, 8, 8, 8, 8, 8, 8, 19, 8, 8, 8]
y-values: [6.58, 5.76, 7.71, 8.84, 8.47, 7.04, 5.25, 12.50, 5.56, 7.91, 6.89]
Key Observations:

Summary Statistics: When you calculate the summary statistics (mean, variance, correlation coefficient, etc.) for each dataset, they appear very similar, if not identical. This can mislead someone into thinking that the datasets are essentially the same.

Visual Differences: The critical insight from Anscombe's quartet is that when you plot these datasets, they exhibit markedly different patterns. Dataset I shows a roughly linear relationship, Dataset II shows a non-linear but well-behaved pattern, Dataset III shows a non-linear pattern with an outlier, and Dataset IV shows a pattern where a single outlier significantly influences the linear fit.

Importance of Data Visualization: Anscombe's quartet underscores the importance of data visualization in understanding data. Summary statistics alone may not reveal the true nature of relationships or uncover unusual observations (outliers).

Statistical Modeling: It highlights that blindly applying statistical models or making decisions based solely on summary statistics can lead to incorrect conclusions.

In summary, Anscombe's quartet serves as a powerful reminder that data visualization is a crucial step in data analysis. It teaches us to be cautious about relying solely on summary statistics and to explore and visualize data thoroughly to gain a deeper understanding of its characteristics and relationships.

3. What is Pearson’s R?
A: Pearson's correlation coefficient, often denoted as "r" or "Pearson's r," is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous

variables. It is widely used in statistics to assess the degree to which two variables are related. Pearson's r falls within the range of -1 to 1, where:

- r = 1: There is a perfect positive linear relationship between the variables.
- r = -1: There is a perfect negative linear relationship between the variables.
- r ≈ 0: There is little to no linear relationship between the variables.

Key characteristics of Pearson's correlation coefficient:

1.Strength of the Relationship : The absolute value of r indicates the strength of the relationship. Larger absolute values (closer to 1) indicate stronger relationships, while values closer to 0 suggest weaker or no linear relationship.

2. Direction of the Relationship: The sign of r (+ or -) indicates the direction of the relationship:
   - Positive (r > 0): As one variable increases, the other tends to increase.
   - Negative (r < 0): As one variable increases, the other tends to decrease.

3. Assumes Linearity: Pearson's r specifically measures linear relationships. If the relationship between the variables is nonlinear, Pearson's correlation may not accurately represent the strength of the association.

4. Sensitive to Outliers: Outliers can strongly influence Pearson's correlation. A single extreme data point can artificially inflate or deflate the correlation coefficient.

5. No Causality: Correlation does not imply causation. Even if two variables are strongly correlated, it does not necessarily mean that one variable causes the other. Correlation only measures the degree of association.

Pearson's correlation coefficient is a valuable tool in various fields, including statistics, economics, social sciences, and more. It helps researchers and analysts understand the relationships between variables and can be used for tasks such as feature selection, data exploration, and assessing the effectiveness of models.

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

A: Scaling is a preprocessing technique in data analysis and machine learning that is used to transform the range of variables or features so that they have consistent properties. The main goal of scaling is to ensure that all variables contribute equally to the analysis or modeling process. Scaling is performed for various reasons, including improving model performance, aiding in the interpretation of results, and addressing the sensitivity of certain algorithms to the scale of features. There are two common types of scaling: normalized scaling and standardized scaling, each with its own purpose and methodology.

1. Normalized Scaling:
  - Normalization, also known as min-max scaling, rescales the data to a specific range, typically [0, 1].
  - The formula for min-max scaling is:
    X_normalized = (X - X_min) / (X_max - X_min)

    where:
    - `X` is the original value of a feature.
    - `X_min` is the minimum value of that feature in the dataset.
    - `X_max` is the maximum value of that feature in the dataset.
  - Normalization is particularly useful when you want to bring all variables into the same range, and you know or assume that the data follows a uniform distribution.
  - It preserves the relative relationships between data points.

2. Standardized Scaling (Standardization):
  - Standardization, also known as z-score scaling, transforms the data such that it has a mean (average) of 0 and a standard deviation of 1.
  - The formula for standardization is:
    X_standardized = (X - X_mean) / X_std

    where:
    - `X` is the original value of a feature.
    - `X_mean` is the mean (average) of that feature in the dataset.
    - `X_std` is the standard deviation of that feature in the dataset.
  - Standardization is useful when you want to compare variables that have different units or when you are working with algorithms that are sensitive to the scale of the features (e.g., gradient-based optimization methods).
  - It centers the data around 0 and scales it to have unit variance.

Key Differences:

- Range: Normalization scales the data to a specific range, usually [0, 1], while standardization centers the data around 0 with a standard deviation of 1.

- Preservation of Distribution: Normalization preserves the distribution shape and the relative relationships between data points, while standardization transforms the data to have a standard normal distribution (mean = 0, std = 1).

- Sensitivity: Standardization is more appropriate when dealing with algorithms that are sensitive to feature scales (e.g., support vector machines, k-means clustering), while normalization is often used for algorithms that assume input features have similar scales.

In summary, scaling is performed to ensure that variables or features are on a comparable scale, making it easier to compare, interpret, and analyze them. Normalization and standardization are two common scaling techniques, each suited for different purposes and types of data. The choice between them depends on the specific requirements of your analysis or modeling task.

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

A: The VIF formula clearly signifies when the VIF will be infinite. If the R2 is 1 then the VIF is infinite. The reason for R2 to be 1 is that there is a perfect correlation between 2 independent variables.

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

A: Q-Q plots are the quantile-quantile plots. It is a graphical tool to assess the 2 data sets are from common distribution. The theoretical distributions could be of type normal, exponential or uniform. The Q-Q plots are useful in the linear regression to identify the train data set and test data set are from the populations with same distributions. This is another method to check the normal distribution of the data sets in a straight line with patterns explained below

- **Interpretations:**
- Similar distribution: If all the data points of quantile are lying around the straight line at an angle of 45 degree from x-axis.
- Y values < X values: If y-values quantiles are lower than x-values quantiles.
- X values < Y values: If x-values quantiles are lower than y-values quantiles.
- Different distributions – If all the data points are lying away from the straight line.
- **Advantages:**
- Distribution aspects like loc, scale shifts, symmetry changes and the outliers all can be daintified from the single plot.
- The plot has a provision to mention the sample size as well.