Assignment-based Subjective Answer

1. We used Box plot to study their effect on the dependent variable (‘cnt’) .

The inference that We could derive were:

- **season**: Almost 32% of the bike booking were happening in season3 with a median of over 5000 booking (for the period of 2 years). This was followed by season2 & season4 with 27% & 25% of total booking. This indicates, season can be a good predictor for the dependent variable.
- **mnth**: Almost 10% of the bike booking were happening in the months 5,6,7,8 & 9 with a median of over 4000 booking per month. This indicates, mnth has some trend for bookings and can be a good predictor for the dependent variable.
- **weathersit**: Almost 67% of the bike booking were happening during ‘weathersit1 with a median of close to 5000 booking (for the period of 2 years). This was followed by weathersit2 with 30% of total booking. This indicates, weathersit does show some trend towards the bike bookings can be a good predictor for the dependent variable.
- **holiday**: Almost 97.6% of the bike booking were happening when it is not a holiday which means this data is clearly biased. This indicates, holiday CANNOT be a good predictor for the dependent variable.
- **weekday**: weekday variable shows very close trend (between 13.5%-14.8% of total booking on all days of the week) having their independent medians between 4000 to 5000 bookings. This variable can have some or no influence towards the predictor. I will let the model decide if this needs to be added or not.
- **workingday**: Almost 69% of the bike booking were happening in ‘workingday’ with a median of close to 5000 booking (for the period of 2 years). This indicates, workingday can be a good predictor for the dependent variable

2. To drop first dummy variable for each set of dummies created.
3. The above Pair-Plot tells us that there is a LINEAR RELATION between 'temp','atemp' and 'cnt'
4. By performing the below tasks.
   - Applying the scaling on the test sets.
   - Dividing into X_test and y_test
5. There are 3 variables below:
   - Temperature (temp)
   - Weather Situation 3 (weathersit_3)
   - Year (yr)

General Subjective Questions

Answer 1

What is Linear Regression?

Linear regression is a **supervised machine learning** algorithm that models the relationship between a **dependent variable** (also known as the target variable) and one or more **independent features** (predictor variables). It assumes a **linear relationship** between these variables and uses a linear equation to represent this relationship.

Here are some key points about linear regression:

1. **Types of Linear Regression**:
   - **Univariate Linear Regression**: When there's only one independent feature.
   - **Multivariate Linear Regression**: When there are multiple independent features.
2. **Equation of Linear Regression**:
   - The basic equation for linear regression is: [Y = X \beta + \epsilon]
     - (Y) represents the dependent variable (what we're trying to predict).
     - (X) is the matrix of independent variables (features).
     - (\beta) is the matrix of coefficients (weights).
     - (\epsilon) represents the error term.
3. **Objective**:
   - Linear regression aims to find the best-fitting linear relationship between the features and the target variable.
   - The goal is to minimize the difference between the predicted values and the actual values.
4. **Assumptions of Linear Regression**:
   - Linearity: Assumes a linear relationship between variables.
   - Independence: Assumes that errors are independent.
   - Homoscedasticity: Assumes constant variance of errors.
   - Normality: Assumes that errors follow a normal distribution.
5. **Cost Function**:
   - The algorithm minimizes the **mean squared error (MSE)** or the **sum of squared differences** between predicted and actual values.
6. **Interpretability and Simplicity**:

    - Linear regression provides interpretable coefficients, helping us understand the impact of each feature on the target variable.
    - Its simplicity makes it a foundational concept for more complex algorithms.
7. **Use Cases**:
    - Linear regression is commonly used for tasks like predicting house prices, sales, salary, and more.

Why is Linear Regression Important?

1. **Interpretability**: The model's equation gives clear coefficients, aiding our understanding of feature impacts.
2. **Foundational Concept**: Linear regression serves as a basis for more advanced models.
3. **Assumption Testing**: It helps validate key assumptions about the data.

Answer 2:

1. **What is Anscombe's Quartet?**
    - Anscombe's quartet consists of **four distinct datasets**, each containing eleven (x, y) points.
    - Surprisingly, all four datasets share nearly identical **simple descriptive statistics** (such as mean, variance, and correlation), yet they exhibit **very different distributions** when graphed.
2. **Why Is It Important?**
    - The quartet was created by the statistician **Francis Anscombe** in 1973.
    - It serves two critical purposes:
        - **Graphing Data**: Anscombe wanted to emphasize the importance of **graphing data** alongside numerical calculations.
        - **Influential Observations**: The quartet demonstrates how **outliers** and other influential observations can impact statistical properties.
3. **The Four Datasets**:
    - Each dataset contains eleven (x, y) pairs. Here they are:

Dataset	x	y
I	10.0	8.04
	8.0	6.95
II	10.0	9.14
	8.0	8.14
III	10.0	7.46
	8.0	6.77
IV	8.0	6.58
	8.0	5.76

4. **Graphical Insights**:
   - Let's explore the quartet visually:
     - **Dataset I (Top Left)**: Appears as a simple linear relationship, suitable for linear regression.
     - **Dataset II (Top Right)**: Shows a non-linear relationship, rendering the Pearson correlation coefficient irrelevant.
     - **Dataset III (Bottom Left)**: Linear but influenced by an outlier, affecting the correlation coefficient.
     - **Dataset IV (Bottom Right)**: High correlation due to a single high-leverage point, despite other data points showing no clear relationship.
5. **Takeaways**:
   - Always **graphically examine data** before diving into specific relationships.
   - Basic statistical properties alone may not capture the nuances of real-world datasets.

Answer 3:

Pearson's R, also known as the **Pearson correlation coefficient**, is a fundamental statistical measure used to quantify the **strength and direction of the linear relationship** between two quantitative variables. Let's explore it in detail:

1. **Definition**:
   - The Pearson correlation coefficient, denoted as (r), ranges between **-1** and **1**.
   - It assesses how closely the data points align with a **linear trend**.
   - Specifically, it is calculated as the **covariance** of the variables divided by the **product of their standard deviations**.
2. **Interpretation**:
   - The sign of (r) indicates the **direction** of the relationship:
     - **Positive correlation** (when ($r > 0$)): As one variable increases, the other tends to increase.
     - **Negative correlation** (when ($r < 0$)): As one variable increases, the other tends to decrease.
     - **No correlation** (when ($r \approx 0$)): The variables are not linearly related.
3. **Strength of Correlation**:
   - The magnitude of (r) reflects the strength:
     - ($|r| > 0.5$): **Strong correlation**
     - ($0.3 < |r| \leq 0.5$): **Moderate correlation**
     - ($0 < |r| \leq 0.3$): **Weak correlation**
4. **Use Cases**:
   - Researchers and analysts employ Pearson's (r) to:
     - Investigate relationships between variables (e.g., height and weight).
     - Assess the impact of one variable on another.
     - Validate hypotheses about associations.
5. **Visual Representation**:
   - Imagine a scatter plot: (r) quantifies how closely the points cluster around a **best-fit line**.
   - If (r) is positive, the line slopes upward; if negative, it slopes downward.
6. **Inferential Aspect**:
   - Pearson's (r) is not only descriptive but also **inferential**.
   - It helps test whether the relationship observed is **statistically significant**.

Answer 4:

Scaling is a crucial step in **preprocessing data** for machine learning models. Let's explore it in detail:

1. **What is Scaling?**
    - **Scaling** refers to transforming input data to a specific range or distribution, ensuring that all features or variables are on a similar scale.
    - It allows machine learning models to handle different variables effectively and make accurate predictions.
2. **Why Is Scaling Performed?**
    - **Distance-based algorithms** (e.g., k-nearest neighbors) are biased toward numerically larger values if data is not scaled.
    - **Tree-based algorithms** are less sensitive to feature scale.
    - Scaling helps machine learning and deep learning algorithms train and converge faster.
3. **Normalization (Min-Max Scaling)**:
    - **Range**: Transforms features to be within [0, 1] or sometimes [-1, 1].
    - **Formula**: ($X_{\text{new}} = \frac{{X - X_{\text{min}}}}{{X_{\text{max}} - X_{\text{min}}}}$)
    - **Use Case**: When there are no outliers.
    - **Geometric Effect**: Squishes data into an n-dimensional unit hypercube.
4. **Standardization (Z-Score Normalization)**:
    - **Transformation**: Subtract mean and divide by standard deviation (Z-score).
    - **Formula**: ($X_{\text{new}} = \frac{{X - \text{mean}}}{{\text{Std}}}$)
    - **Use Case**: When data follows a Gaussian distribution (but not necessarily).
    - **Geometric Effect**: Translates data to the mean vector of original data to the origin.
5. **Differences Between Normalization and Standardization**:

Aspect	**Normalization**	**Standardization**
Scaling Range	[0, 1] or [-1, 1]	Not bounded to a specific range

Aspect	Normalization	Standardization
Handling Outliers	Sensitive	Less affected
Scikit-Learn Transformer	MinMaxScaler	StandardScaler
Geometric Effect	Squishes data into a hypercube	Translates data to origin

Answer 5:

The occurrence of an **infinite value** for the **Variance Inflation Factor (VIF)** is an interesting phenomenon. Let’s explore why this happens:

1. **What is VIF?**
    - The VIF is a measure used to assess **multicollinearity** in regression models.
    - It quantifies how much the **variance of the estimated regression coefficient** is inflated due to the presence of correlated predictor variables.
2. **Reasons for Infinite VIF**:
    - When the VIF is **infinite** for a specific independent variable, it indicates that this variable can be **perfectly predicted** by other variables in the model.
    - Here’s why this might occur:
        - **Perfect Multicollinearity**: One or more variables are **linear combinations** of other variables. For example: [$X_j = X_{\text{other}} \beta + \epsilon$]
            - In this equation, (X_j) can be perfectly predicted using other variables ((X_{other})).
            - As a result, the VIF becomes infinite.
        - **R-squared Equals 1**: When the coefficient of determination ((R^2)) approaches 1, the VIF also becomes 1.
            - This situation arises when you have **more predictors than observations** (i.e., ($k > N$)).
            - All regressions end up having ($R^2 = 1$), leading to infinite VIF values for all variables.

3. **Handling Infinite VIF**:
    - If you encounter infinite VIF values, consider the following steps:
        - **Identify Problematic Variables**: Perform actual regressions for each variable against all others (e.g., (X_j = X_{\text{other}} \beta + \epsilon)).
        - **Check Coefficients**: Examine the coefficients to identify the problematic variables.
        - **Reduce Regressors**: If you have more variables than observations, find ways to use a smaller set of regressors (e.g., forward stepwise regression).
4. **Practical Implications**:
    - When all VIF values are infinite, it's essential to investigate the underlying reasons.
    - Adjust your model by removing problematic variables or using dimensionality reduction techniques.

Answer 6:

What is a Q-Q Plot (Quantile-Quantile Plot)?

A **Q-Q plot** (short for **quantile-quantile plot**) is a graphical tool used to assess whether a set of data plausibly follows a **theoretical distribution**, such as the **normal distribution**, exponential distribution, or uniform distribution. Here's how it works:

1. **Construction**:
    - A Q-Q plot compares the **quantiles** (ordered values) of the **observed data** with the quantiles of a **theoretical distribution**.
    - If the points on the plot roughly form a **straight diagonal line**, it suggests that the data follows the assumed distribution.
2. **Interpretation**:
    - The Q-Q plot helps us:
        - Verify the **normality assumption** (whether the residuals of a model are normally distributed).
        - Detect deviations from the expected distribution.
        - Identify **outliers** or unusual data points.
3. **Key Interpretations**:
    - When comparing two datasets (e.g., residuals vs. theoretical quantiles):
        - **Similar Distribution**: If points lie on or close to a straight line at a **45-degree angle** from the x-axis, the datasets have a similar distribution.

        - **Y-values < X-values**: If y-quantiles are lower than x-quantiles, the data has heavier tails.
        - **X-values < Y-values**: If x-quantiles are lower than y-quantiles, the data has lighter tails.
        - **Different Distribution**: If points deviate significantly from the 45-degree line, the distributions differ.

4. **Importance in Linear Regression**:
    - **Normality Assumption**: Linear regression assumes that the **error terms** (residuals) are **normally distributed**.
    - **Model Validity**: Checking the normality of residuals using Q-Q plots ensures the validity of regression results.
    - **Robustness**: If residuals deviate from normality, consider robust regression techniques.