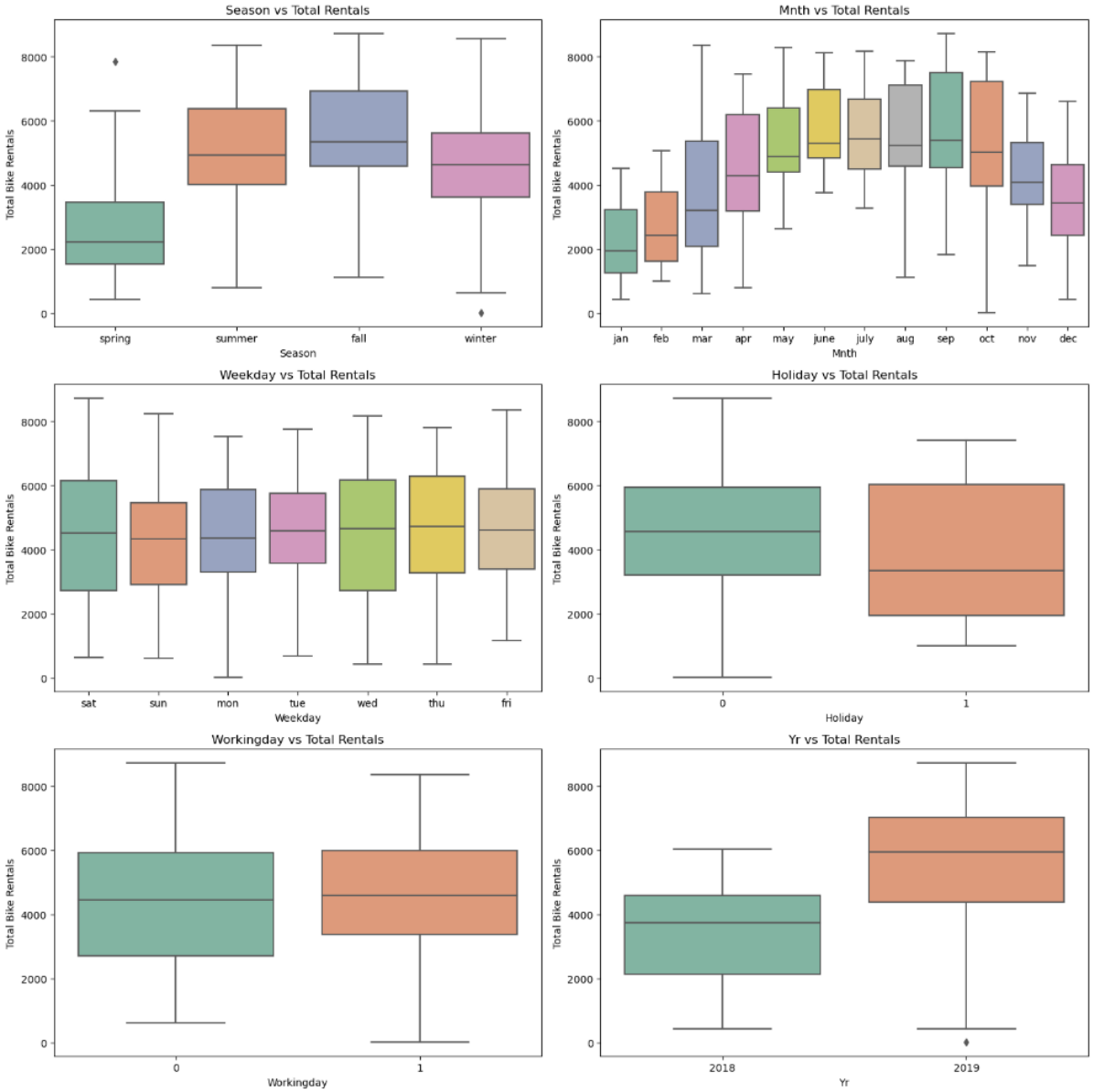
ASSIGNMENT-BASED SUBJECTIVE QUESTIONS:

1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?**



- **Season:** Fall (category 3) had the highest demand; spring (category 1) had the least.
- **Mnth**: Rentals peaked in September, decreased in January due to snowfall.
- **Weekday**: Bike demand remained consistent daily.
- **Holiday**: Rentals decreased on holidays.
- **Workingday**: Bookings were similar on working/non-working days.
- **Weathersit**: No users during heavy rain/snow; clear, partly cloudy had the highest count.
- **Yr**: 2019 had more users than 2018.

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

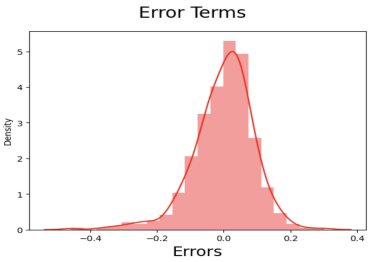
- drop_first=True is used to prevent multicollinearity in regression models.
- It drops one of the dummy variable columns to avoid perfect multicollinearity, where one variable can be linearly predicted from others.
- This ensures the model's stability and prevents issues in coefficient interpretation.

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

Based on the pair-plot and the correlation matrix heatmap, the numerical variable **"atemp" (feeling temperature) has the highest correlation with the target variable "cnt"**. Its correlation coefficient is **approximately 0.631**, indicating a moderately strong positive correlation.

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

- **Linearity of the Model:** Checking whether the relationship between the independent variables and the dependent variable is linear.
- **Independence of Residuals:** Ensuring that the residuals (the differences between the actual and predicted values) are independent of each other.
- **Normality of Residuals:** Verify that the residuals follow a normal distribution by creating a histogram of the residuals and checking if they approximate a normal distribution.



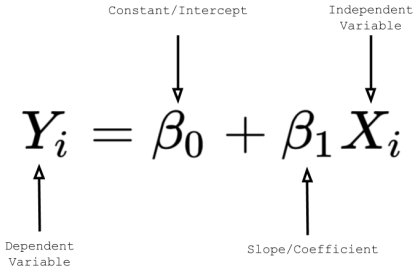
- **No Multicollinearity:** Ensure that there is no high multicollinearity among the independent variables. Calculate the Variance Inflation Factor (VIF) for each independent variable to check for multicollinearity. And checked p-value, if P value is greater than 0.05 and VIF values above 5 are typically a cause for concern so I have dropped the feature.
- **Model Performance Metrics:** Finally, by evaluating the model's performance on the test set using appropriate metrics such as R-squared. Which gave me around 79% on test data and 83% on train data explaining a significant portion of the variance and examining the linearity assumption by plotting observed vs. predicted values.

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

   In the provided example, the top three significant features were mentioned as "**Year (yr)**," "**Season**" and "**Temperature (temp)**."

GENERAL SUBJECTIVE QUESTIONS:

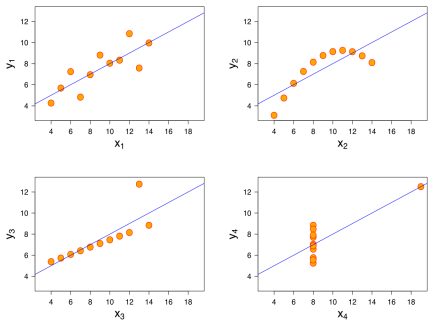
1. **Explain the linear regression algorithm in detail.**



- Linear regression is a supervised learning algorithm used for predicting a continuous target variable based on one or more predictor variables.
- It assumes a linear relationship between predictors and the target.
- The algorithm finds the best-fit line (or hyperplane in multiple dimensions) that minimizes the sum of squared differences between predicted and actual values.
- The goal is to minimize the MSE by adjusting θ0 and θ1. There are different methods to do this, such as gradient descent, normal equaϴon, or using libraries like scikit-learn.
- Coefficients for predictors are estimated, allowing predictions for new data points.

2. **Explain the Anscombe's quartet in detail.**

Anscombe's Quartet, created by statistician Francis Anscombe, consists of four datasets with nearly identical statistical characteristics but drastically different distributions and visual representations.It underscores the importance of graphing data before analysis and the impact of outliers and influential observations on statistical properties.



1. The first scatter plot (top left) suggests a straightforward linear relationship.
2. The second plot (top right) displays a non-normal distribution with a nonlinear connection.
3. In the third graph (bottom left), although the distribution is linear, a single outlier significantly alters the regression line, reducing the correlation coefficient from 1 to 0.816.

4. Lastly, the fourth graph (bottom right) illustrates how a single high-leverage point can yield a high correlation coefficient, emphasizing the need to consider influential data points.

3. What is Pearson's R?

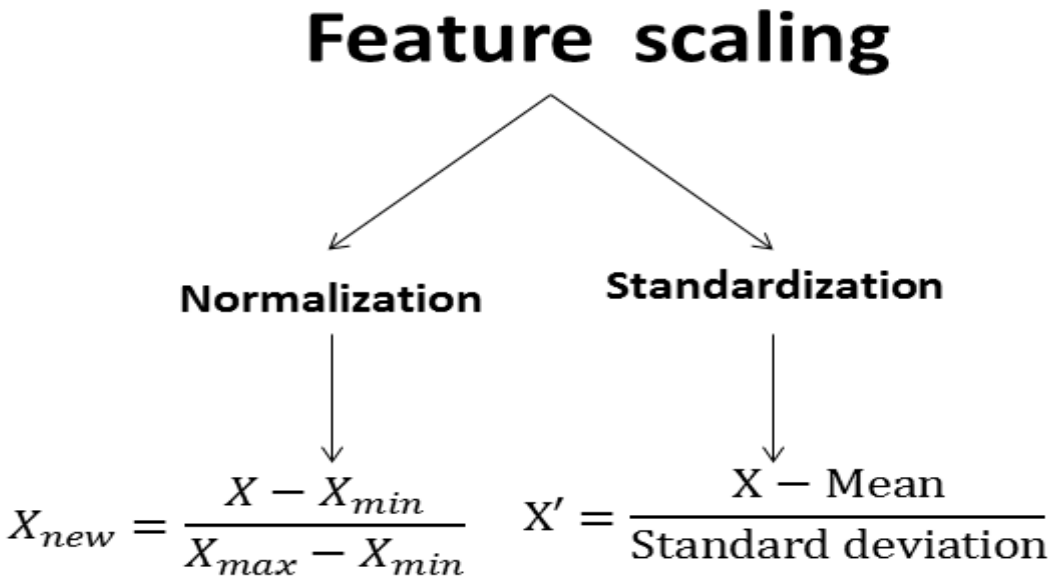
Pearson's R, or the Pearson correlation coefficient, is a statistic that measures the linear correlation between two variables.

It ranges from -1 (perfect negative correlation) to 1 (perfect positive correlation), with 0 indicating no linear correlation.

It quantifies the strength and direction of the linear relationship between two variables.

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

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



- Scaling is the process of transforming variables to a standard scale. It's performed to bring all variables to a common scale, preventing some variables from dominating others during modeling.
- Normalized scaling scales variables to a range between 0 and 1.
- Standardized scaling (z-score scaling) scales variables to have a mean of 0 and a standard deviation of 1.
- Normalization is appropriate when variable distributions are not necessarily Gaussian. Standardization is suitable when variable distributions are approximately Gaussian.

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

VIF (Variance Inflation Factor) measures multicollinearity, which occurs when predictor variables in a regression model are highly correlated.

VIF can be infinite when one predictor is a perfect linear combination of others, making it impossible to estimate its coefficient independently. This is known as perfect multicollinearity.

If that independent variable can be explained perfectly by other independent variables, then it will have perfect correlation and its R-squared value will be equal to 1.

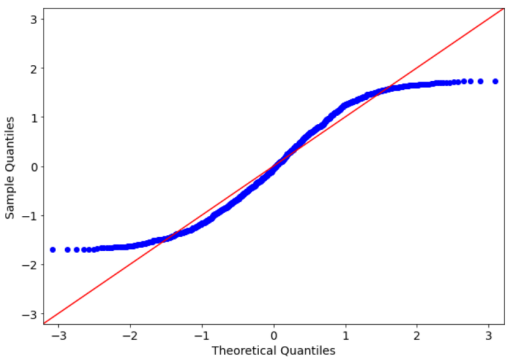
So, VIF = 1/(1-1) which gives VIF = 1/0 which results in "infinity" The numerical value for VIF tells you (in decimal form) what percentage the variance (i.e. the standard error squared) is inflated for each coefficient. For example, a VIF of 1.9 tells you that the variance of a particular coefficient is 90% bigger than what you would expect if there was no multicollinearity — if there was no correlation with other predictors.

A rule of thumb for interpreting the variance inflation factor:

1 = not correlated.
Between 1 and 5 = moderately correlated.
Greater than 5 = highly correlated.

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

A Q-Q (quantile-quantile) plot is a graphical tool to assess if a dataset follows a particular theoretical distribution, typically the normal distribution.It plots quantiles of the observed data against quantiles of the expected theoretical distribution.
In linear regression, Q-Q plots are used to check the normality assumption of residuals. If residuals follow a straight line in the plot, they are approximately normally distributed, which is important for accurate inference and prediction in regression.we should see the points forming a line that's roughly straight. The q-q plot is used to answer the following questions:



- Do two data sets come from populations with a common distribution?
- Do two data sets have common location and scale?
- Do two data sets have similar distribution shapes?
- Do two data sets have similar tail behavior?