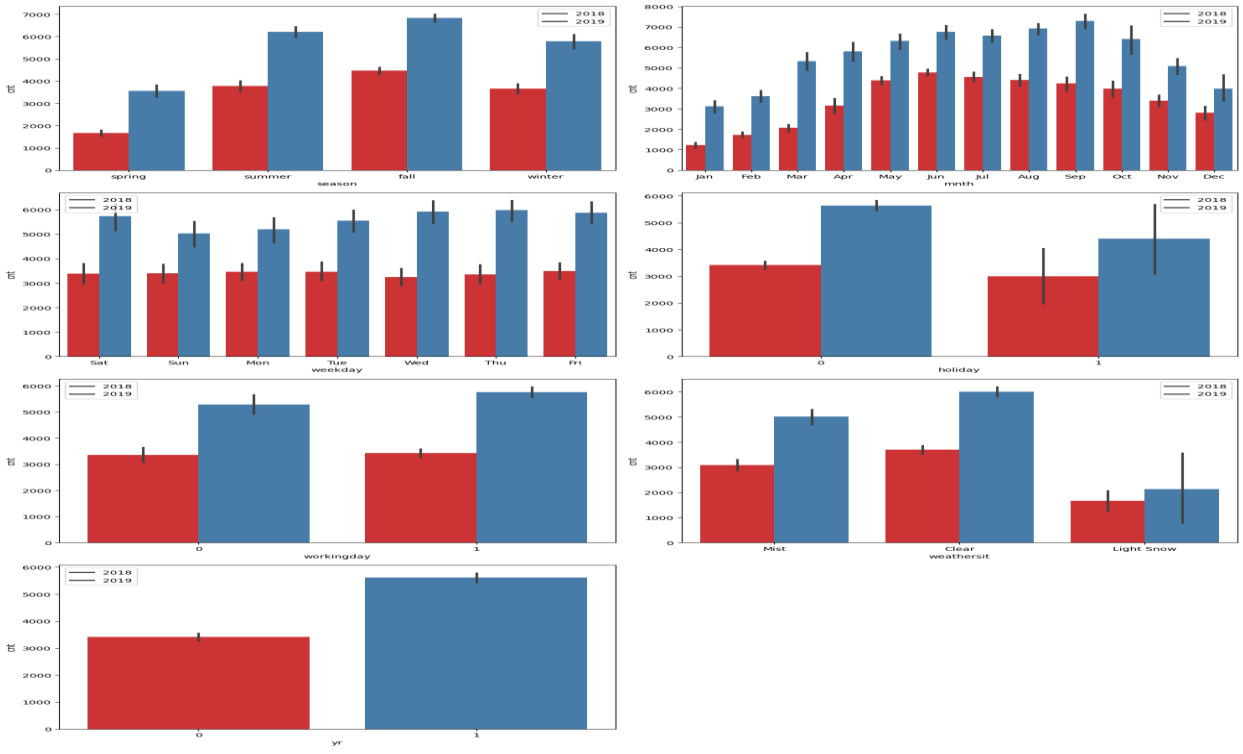
Assignment-based Subjective Questions

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

Answer: Based on the analysis done on the categorical variables, the following inferences were made:

- The demand for bikes is high in the fall in both the years.
- The demand for bikes is high in the months of May, June, July, August, September, and October in both the years.
- Clear weather has the highest demand for bikes in both the years.
- Thu, Fri and Sat have the highest demand for bikes in both the years.
- The demand for bikes is high on holidays in both years.
- Working days and non-working days have the same demand for bikes in both the years.
- In all the plots, 2019 has a higher demand for bikes than 2018.



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

Answer: The drop_first=True is used to avoid the dummy variable trap scenario where multicollinearity occurs due to one variable being a perfect predictor of another variable. When we create dummy variables, we are creating new binary variables (0 or 1) for each category in the original variable. If we have n categories and create n dummy variables without dropping one, we will end up with a perfect multicollinearity situation. This is because one dummy variable can be perfectly predicted from all the other variables.

This can cause problems in models like linear regression where it can make it impossible to calculate the coefficients of the model or lead to unstable and unreliable estimates.

From the bike sharing assignment, 'bikeshare_season', 'bikeshare_mnth', 'bikeshare_weekday', 'bikeshare_weathersit' will have n-1 columns where n is the number of unique categories in the original column. Each row in these dataframes represents a row in the original bikeshare dataframe and the columns represent the categories of the original column.

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

Answer: 'temp' has the highest correlation with the target variable

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

Answer: I validated the assumptions of Linear Regression model based on

a. Normality of error terms: error terms are normally distributed with mean 0
b. No multicollinearity: This was checked using the VIF values and correlation matrix
c. Homoscedasticity: There should be no visible pattern in residual values

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

   Answer: The top 3 features contributing significantly towards explaining the demand of the shared bikes are
   - Temp
   - Winter
   - Sep

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Answer: Linear regression is a statistical method that is used for predictive analysis. It allows us to summarize and study relationships between two continuous quantitative variables. The linear regression model assumes a linear relationship between the input variable/independent variable (X) and the single output variable/dependant variable (Y). The output variable (Y) can be calculated from a linear combination of the input variables (X).

Best Fit Line: Linear regression finds the best line that predicts Y from X. The method of least squares is used to find the best-fitting line for the observed data. A residual is the difference between the actual value of Y and the predicted value of Y.

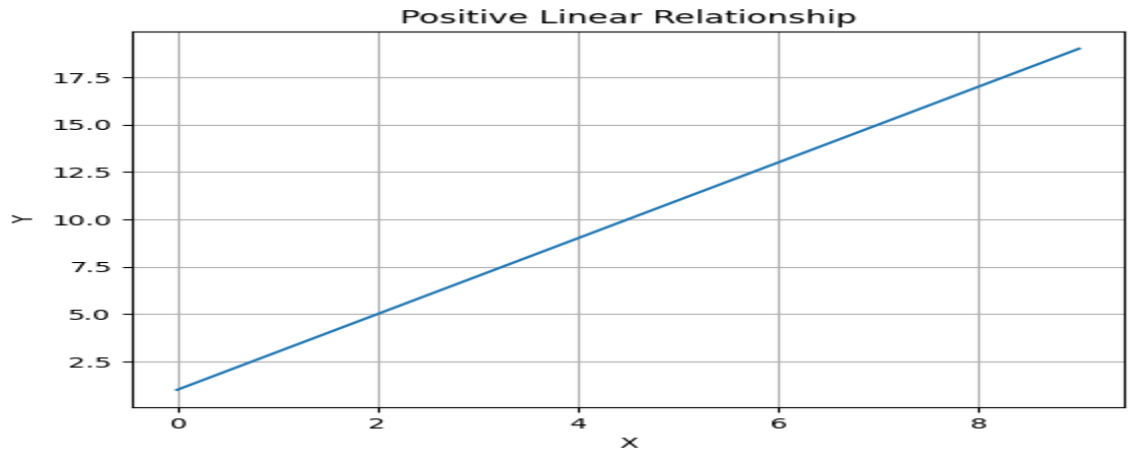
The line is represented by the equation Y = A+ BX + e.

where A and B are coefficients that are estimated by regression, X is the input variable, and e is the error term.

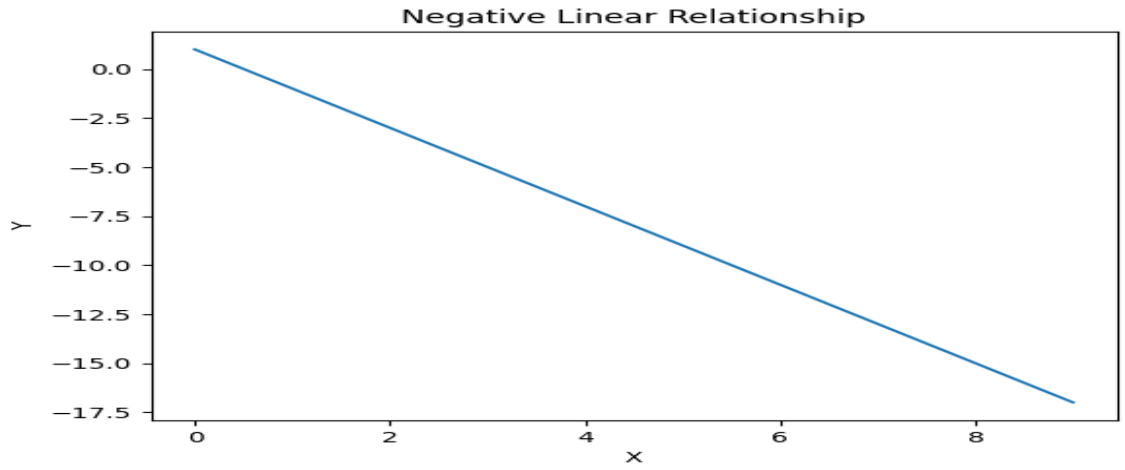
B is the slope of the regression line which infers the effect X has on Y. A is the constant.

Linear relationship can be positive or negative:

1. Positive Linear Relationship: In a positive linear relationship, as one variable increases, the other variable also increases. Similarly, as one variable decreases, the other variable also decreases.



2. Negative Linear Relationship: In a negative linear relationship, as one variable increases, the other variable decreases. The graph of a negative linear relationship slopes downward.



Key assumptions of linear regression:

- Normality: It is assumed that the error terms are normally distributed. We can check if data is normal distributed to avoid bias.
- Multicollinearity: Variable must be independent of each other. We have to check the correlation Matrix
- Homoscedasticity: It is assumed that the residual terms have the same but unknown variance. We have to check for homogeneity of data to variance of output.
- Autocorrelation: Autocorrelation is the correlation between two of the same series

2. Explain the Anscombe's quartet in detail

Answer: Anscombe's Quartet is a group of four datasets that were constructed by statistician Francis Anscombe. Each dataset consists of 11 x-y pairs. Despite having different graphical representations, all four datasets share identical descriptive statistical properties. These properties include:

- The same mean for both x and y
- The same variance for both x and y
- The same correlation coefficient between x and y
- The same linear regression line

When visualized using scatter plots, each dataset exhibits a unique relationship between x and y with different patterns of variability and correlation strengths. A brief description of the four datasets below:

1. Dataset 1: Fits the linear regression model well

2. Dataset 2: Cannot fit the linear regression model because the data is non-linear

3. Dataset 3: Shows the outliers involved in the dataset, which cannot be handled by the linear regression model

4. Dataset 4: Also shows the outliers involved in the dataset, which cannot be handled by the linear regression model

The purpose of Anscombe's Quartet is to emphasize the importance of exploratory data analysis and the potential pitfalls of relying solely on summary statistics. It demonstrates the crucial role of data visualization in spotting trends, outliers, and other vital details that might not be apparent from summary statistics alone. Therefore, before attempting to interpret and model the data or implement any machine learning algorithm, it's important to visualize the dataset.

3. What is Pearson's R?

Answer: Pearson's R is used to measure the linear correlation between two variables. It's a value between -1 and 1, where:

- 1 indicates a strong positive linear relationship.
- -1 indicates a strong negative linear relationship.
- 0 indicates no linear relationship.

The formula for Pearson's R is:

$r = \Sigma[(xi - \bar{x})(yi - \bar{y})] / sqrt[(\Sigma(xi - \bar{x})^2)(\Sigma(yi - \bar{y})^2)]$

Where:

- xi and yi are the individual sample points indexed with i

- $\bar{x}$ and $\bar{y}$ are the means of x and y respectively

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

Answer: Feature scaling is a preprocessing step in machine learning that transforms the range of data values so that they can be compared on common grounds. It is performed to handle the issue of features having different units or scales, which can cause a machine learning algorithm to weigh certain features more heavily than others leading to biased or incorrect results.

For example in healthcare, features like a patient's age, blood pressure, and cholesterol levels are measured on different scales. Scaling allows these features to be used in the same model.

Here is a comparison between Normalization and Standardization scaling:

Criteria	**Normalization**	**Standardization**
Method	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling
Usage	It is used when features are of different scales	It is used when we want to ensure zero mean and unit standard deviation
Range	Scales values between [0, 1] or [-1, 1]	It is not bounded to a certain range
Effect of Outliers	It is really affected by outliers	It is much less affected by outliers
Scikit-Learn Transformer	Scikit-Learn provides a transformer called MinMaxScaler for Normalization	Scikit-Learn provides a transformer called StandardScaler for standardization

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

Answer: If the VIF is infinite which means the denominator of the VIF formula is zero, i.e. R^2 is 1. This happens when a predictor is perfectly correlated with a combination of one or more other variables which leads to perfect multicollinearity.

Perfect multicollinearity is a problem because it means that the affected variables don't provide unique and independent information to the model. This can make the estimates of the regression coefficients unstable and difficult to interpret. When we encounter a situation where VIF is infinite, we might need to remove one or more of the correlated variables from the model.

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

Answer: A Q-Q (Quantile-Quantile) plot is a graphical tool that helps us assess if a dataset follows a particular theoretical distribution. It compares the sorted values of our data (quantiles) against the expected values of the chosen theoretical distribution (quantiles).

In the context of linear regression, a Q-Q plot is used to check the assumption of normality of the residual errors. This is important because many statistical tests rely on the assumption that the residuals are normally distributed. Violations of this assumption can lead to unreliable and misleading results.

Here's how a Q-Q plot can help:

- If the data follows the chosen distribution, the points in the Q-Q plot will approximately lie on the line y = x.
- If the data is skewed, the points will deviate from the line in a certain way.
- If the data has heavy tails, the points will deviate from the line at the ends.