## Assignment-based Subjective Questions

Q1 - From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

A –

- Season in fall(3) has median more than 5000 and count of rental bikes is more than 6000
- There is a significance increase in total number of rental bikes count in the year 2019 with around 2500
- Most number of bikes has been rented during May to October
- Peak at June and September
- If its a working day or if its a holiday then more bikes are rented
- No bikes are rented for weather situation (4) Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog
- When weather is (1) clear, Few clouds, Partly cloudy, Partly cloudy then most number of bikes are rented

Q2. Why is it important to use drop_first=True during dummy variable creation?

A: It reduces the extra column that can be ignored which was created during dummy variable creation.

From the case study for the column season which were having 4 seasons: (spring,summer,fall, winter)

```
In [1372]: season_dummy.head()
```

Out[1372]:

| | spring | summer | winter |
|---|---|---|---|
| **0** | 1 | 0 | 0 |
| **1** | 1 | 0 | 0 |
| **2** | 1 | 0 | 0 |
| **3** | 1 | 0 | 0 |
| **4** | 1 | 0 | 0 |

It created three only, which mean if spring, summer and winter all are 0 then it is fall.

Same has been highlighted in the notebook case study as well:

```
In [1372]: season_dummy.head()
```

Out[1372]:

| | spring | summer | winter |
|---|---|---|---|
| **0** | 1 | 0 | 0 |
| **1** | 1 | 0 | 0 |
| **2** | 1 | 0 | 0 |
| **3** | 1 | 0 | 0 |
| **4** | 1 | 0 | 0 |

```
In [1373]: weathersit_dummy.head()
```

Out[1373]:

| | lightRain | mistCloud |
|---|---|---|
| **0** | 0 | 1 |
| **1** | 0 | 1 |
| **2** | 0 | 0 |
| **3** | 0 | 0 |
| **4** | 0 | 0 |

```
In [1374]: weekday_dummy.head()
```

Out[1374]:

| | Mon | Sat | Sun | Thu | Tue | Wed |
|---|---|---|---|---|---|---|
| **0** | 0 | 1 | 0 | 0 | 0 | 0 |
| **1** | 0 | 0 | 1 | 0 | 0 | 0 |
| **2** | 1 | 0 | 0 | 0 | 0 | 0 |
| **3** | 0 | 0 | 0 | 0 | 1 | 0 |
| **4** | 0 | 0 | 0 | 0 | 0 | 1 |

```
In [1375]: month_dummy.head()
```

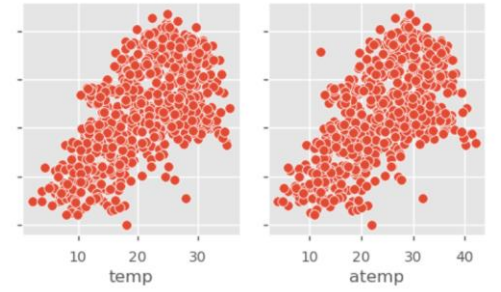
Out[1375]:

| | Aug | Dec | Feb | Jan | Jul | Jun | Mar | May | Nov | Oct | Sep |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **1** | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **2** | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **3** | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **4** | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

- so now for the row **"spring"**, **"summer"**, **"winter"** all as **0** will be having **"fall"** as season
- And **"lightRain"** and **"mistClound"** as 0 then its **"clear"** as weathersit, and as concluded above its never heavy rain
- And for weekdays if all are 0 then its "**Fri**"
- And for months if all are 0 then its "**Apr**"

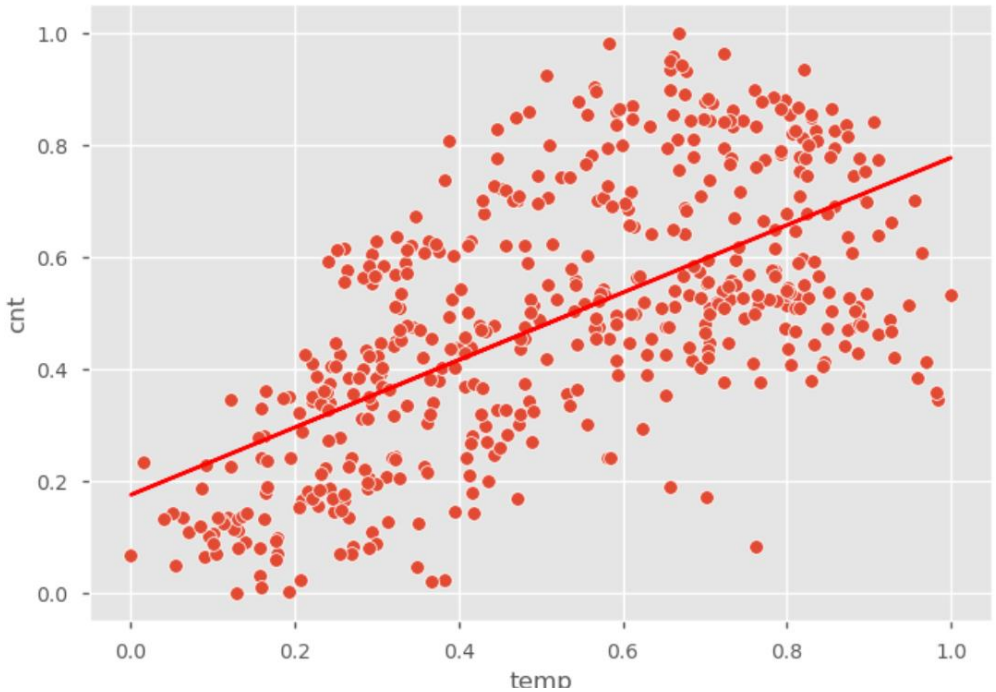
Q3: Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

A: temp and atemp is having the highest coorelation with the target variable



Q4: How did you validate the assumptions of Linear Regression after building the model on the training set?

A: To determine if the assumption is met or not we create a scatter plot for X vs Y graph. The data points should have a straight line in the graph to confirm there is a linear relationship between dependent and independent variable. We did a plot for temp vs cnt




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

A –

- Yr with coeff of 0.2483 which means the demand for the bikes increases in upcoming yr
- If weathersit is (3) Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds with coeff (-0.2788) would cause in decrease in demand for bikes
- In Spring season with coeff(-0.2192) demand for bike will decrease.
- Windspeed coeff (-0.15) and VIF-3.86 which means increase in windspeed will decrease the demand of bikes
- In season Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog no bikes were rented.

## General Subjective Questions

# Q1: Explain the linear regression algorithm in detail.

A: Linear regression is a machine learning algorithm used for predicting numerical values. It is a form of predictive modelling technique which tells us the relationship between the target variable and the predicted variable.

The output variable to be predicted is a continuous variable, for example score of a particular student in predictive analysis, number of bedrooms, etc.

Linear regression makes the key assumption that the relationship between the independent variables and the dependent variable is linear. This means that the change in the dependent variable is directly proportional to changes in the independent variables.

**Simple Linear Regression**: In simple linear regression, there is only one independent variable (X) and one dependent variable (Y). The relationship can be represented as:

$$Y = \beta_0 + \beta_1 * X + \varepsilon$$

$\beta_0$ is intercept

$\beta_1$ is coefficient

**Model Training:** To train the linear regression model, a dataset with observed values of both the independent and dependent variables is passed. The model uses this data for the estimation of coefficients.

The most common method for finding the coefficients is the Ordinary Least Squares (OLS) method, which minimizes the sum of squared residuals.

$$\beta_1 = \Sigma((X_i - \bar{X}) * (Y_i - \bar{Y})) / \Sigma((X_i - \bar{X})^2)$$

$$\beta_0 = \bar{Y} - \beta_1 * \bar{X}$$

Xi is the value of independent variable for the $i^{th}$ data point.
$\bar{X}$ is the mean
$Y_i$ is the value of the dependent variable for the $i^{th}$ data point.
$\bar{Y}$: Mean of the dependent variable.

**Evaluation:** After training, we assess the model's performance. Common evaluation metrics for linear regression include Mean Square Error, Root Mean Squared Error and R-squared ($R^2$).

$$R^2=1-(RSS/TSS)$$

$$TSS = \Sigma(Y_i - \bar{Y})^2$$

$$RSS = \Sigma (Y_i - \hat{Y})^2$$

**Multiple Linear Regression:**

This is also termed as the statistical technique to understand the relationship between one dependent variable and several independent variables.

$$Y = \beta_0 + (\beta_1 * X_1) + (\beta_2 * X_2) + .... + (\beta_n * X_n)$$

$X_1$ to $X_n$: Multiple independent variables.

$\beta_0$ to $\beta_n$: Coefficients for each independent variable.

**Assumptions of Linear Regression:**

- Linear independence of predictors.
- Homoscedasticity (constant variance of residuals).
- Normally distributed residuals.
- No multicollinearity (correlation between independent variables).

**Overfitting**: To prevent overfitting in cases with many features or multicollinearity, you can use regularization techniques like Ridge or Lasso regression.

**Prediction**: Once the model is trained and evaluated, you can use it to make predictions on new, unseen data by plugging in values for the independent variables.

## Q2. Explain the Anscombe's quartet in detail.

A: Anscombe's quartet is used to illustrate the importance of exploratory data analysis (EDA) and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone.

Anscombe's quartet serves as a reminder that relying solely on summary statistics can be misleading and may not capture the full complexity of a dataset.

Here are the four datasets in Anscombe's quartet, along with their characteristics:

**Dataset I:**

- x-values: [10.0, 8.0, 13.0, 9.0, 11.0, 14.0, 6.0, 4.0, 12.0, 7.0, 5.0]

- y-values: [8.04, 6.95, 7.58, 8.81, 8.33, 9.96, 7.24, 4.26, 10.84, 4.82, 5.68]

Summary statistics:

- Mean of x: 9.0, Mean of y: 7.50
- Variance of x: 10.0, Variance of y: 3.75
- Correlation between x and y: 0.816
- Linear regression equation: y = 3.00 + 0.50 * x

**Dataset II:**

- x-values: [10.0, 8.0, 13.0, 9.0, 11.0, 14.0, 6.0, 4.0, 12.0, 7.0, 5.0]
- y-values: [9.14, 8.14, 8.74, 8.77, 9.26, 8.10, 6.13, 3.10, 9.13, 7.26, 4.74]

Summary statistics:

- Mean of x: 9.0, Mean of y: 7.50
- Variance of x: 10.0, Variance of y: 3.75
- Correlation between x and y: 0.816
- Linear regression equation: y = 3.00 + 0.50 * x

**Dataset III:**

- x-values: [10.0, 8.0, 13.0, 9.0, 11.0, 14.0, 6.0, 4.0, 12.0, 7.0, 5.0]
- y-values: [7.46, 6.77, 12.74, 7.11, 7.81, 8.84, 6.08, 5.39, 8.15, 6.42, 5.73]

Summary statistics:

- Mean of x: 9.0, Mean of y: 7.50
- Variance of x: 10.0, Variance of y: 3.75
- Correlation between x and y: 0.816
- Linear regression equation: y = 3.00 + 0.50 * x

**Dataset IV:**

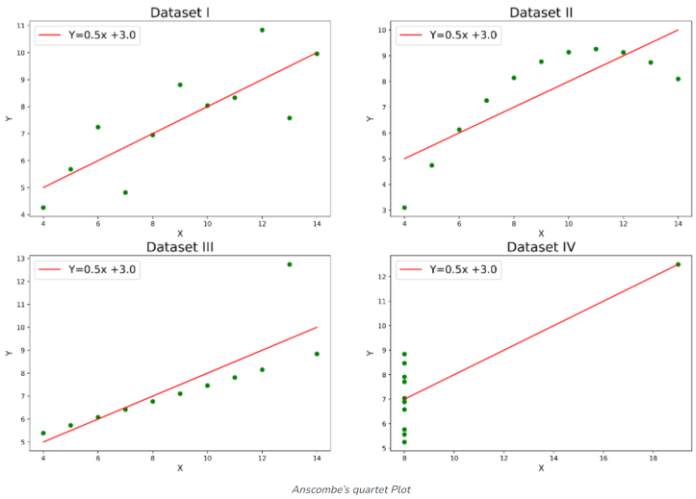
- x-values: [8.0, 8.0, 8.0, 8.0, 8.0, 8.0, 8.0, 19.0, 8.0, 8.0, 8.0]
- y-values: [6.58, 5.76, 7.71, 8.84, 8.47, 7.04, 5.25, 12.50, 5.56, 7.91, 6.89]

Summary statistics:

- Mean of x: 9.0, Mean of y: 7.50
- Variance of x: 10.0, Variance of y: 3.75
- Correlation between x and y: 0.816
- Linear regression equation: y = 3.00 + 0.50 * x

*Key observations from Anscombe's quartet:*

- All four datasets have the same mean, variance, and correlation, and their linear regression equations are nearly identical.
- However, when you graphically plot these datasets, you'll see that they exhibit different patterns, including linear, quadratic, and outlier-heavy relationships.
- This demonstrates the importance of visualizing data to gain a deeper understanding of its underlying structure and relationships, as summary statistics alone may not reveal the true nature of the data.



*Anscombe's quartet Plot*

Anscombe's quartet underscores the need for data exploration, visualization, and graphical analysis to complement traditional statistical methods when working with data. It serves as a cautionary example for researchers and analysts to avoid making assumptions based solely on summary statistics.

## Q: 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 the linear relationship between two continuous variables. It is widely used in statistics to assess the degree of association or correlation between two variables. Pearson's r ranges from -1 to 1, where:

- r = 1 indicates a perfect positive linear relationship.
- r = -1 indicates a perfect negative (inverse) linear relationship.
- r = 0 indicates no linear relationship; the variables are not correlated.

Key characteristics of Pearson's correlation coefficient:

1. *Linear Relationship:* Pearson's r measures only linear relationships. It assumes that the relationship between the two variables can be well approximated by a straight line.

2. *Symmetry:* Pearson's r is symmetric, meaning that swapping the two variables doesn't change the correlation coefficient. In other words, the correlation between X and Y is the same as the correlation between Y and X.

3. *Range:* The range of Pearson's r is between -1 and 1, with -1 indicating a perfect negative linear relationship, 1 indicating a perfect positive linear relationship, and 0 indicating no linear relationship.

4. *Strength:* The magnitude of r indicates the strength of the linear relationship. Values closer to -1 or 1 suggest a stronger linear association, while values closer to 0 suggest a weaker linear association.

5. *Direction:* The sign of r (positive or negative) indicates the direction of the linear relationship. A positive r indicates a positive linear relationship (as one variable increases, the other tends to increase), while a negative r indicates a negative linear relationship (as one variable increases, the other tends to decrease).

6. *Assumption:* Pearson's correlation assumes that the data is normally distributed and that the relationship between the variables is approximately linear. It may not capture non-linear relationships.

The formula for calculating Pearson's correlation coefficient (r) for a sample is as follows:

$$r = \Sigma((X - \bar{X}) * (Y - \bar{Y})) / \sqrt{(\Sigma(X - \bar{X})^2 * \Sigma(Y - \bar{Y})^2)}$$

- X and Y are the individual data points.

- $\bar{X}$ and $\bar{Y}$ are the means of X and Y, respectively.

Pearson's correlation is a valuable tool for assessing relationships between variables, such as determining the strength and direction of the association between temperature and ice cream sales, or between study time and exam scores. However, it's essential to remember that correlation does not imply causation. A high correlation between two variables does not necessarily mean that one causes the other; it simply indicates a statistical relationship. Causation typically requires further study and experimentation to establish.

## Q: **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**

A: Scaling is a data pre-processing procedure that normalizes data within a specific range. It's used to ensure that all values in a dataset are within a certain range. It accelerate algorithmic calculations.

The primary purpose of scaling is to ensure that all variables have equal influence on the analysis, as many statistical and machine learning algorithms are sensitive to the scale of the input data. Scaling helps in making comparisons and calculations more meaningful and can improve the performance of certain algorithms.

Normalization and standardization are two scaling methods. Normalization typically rescales values into a range in which the data doesn't have Gaussian distribution and is highly affected by outliers

Standardization typically rescales data to have a mean of 0 and a standard deviation of 1 and is being used on data having Gaussian distribution and this is not bounded by range.

*Key differences between normalization and standardization:*

- Range: Normalization scales values to a specific range (usually 0 to 1), while standardization rescales values to have a mean of 0 and a standard deviation of 1.
- Impact on Distribution: Normalization maintains the distribution's shape and only changes the scale, while standardization centers the data around 0 and changes the spread.
- Outliers: Standardization is less affected by outliers compared to normalization. Outliers can disproportionately impact the range-based scaling used in normalization.
- Interpretability: Normalization retains the original units of measurement, while standardization transforms data into z-scores, making it unit less.

The choice between normalization and standardization depends on the specific requirements of your analysis and the characteristics of your data. For instance, if you're working with data where the variable values are on different scales, normalization may be

more appropriate. If your data is expected to follow a normal distribution and you want to mitigate the influence of outliers, standardization may be preferred. In practice, it's common to experiment with both methods and choose the one that results in better model performance or more meaningful insights.

## Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

A: VIF is a measure used to determine multicollinearity in multiple linear regression models. A high VIF indicates a strong correlation between one independent variable and a linear combination of others.

Infinite VIF values occur when there is perfect or near-perfect multicollinearity between independent variables or when the sample size is very small.

## Q6: What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

A: A Q-Q plot, or quantile-quantile plot, is a graphical method for comparing two sets of data. It plots the quantiles of one data set against the quantiles of another data set. A quantile is a fraction of points below a given value.

For example, the median is a quantile, where 50% of the data fall below that point and 50% lie above it.

*A Q-Q plot is created by:*

1. Plotting two sets of quantiles against each other.
2. If both sets of quantiles came from the same distribution, the points should form a line that's roughly straight.

*Use and Importance of a Q-Q Plot in Linear Regression:*

1. Normality Assumption: One of the key assumptions in linear regression is that the residuals should be normally distributed.
2. Model Evaluation: Q-Q plots can be used to assess the normality assumption for the residuals of a linear regression model. By plotting the residuals against the quantiles of a theoretical normal distribution, you can visually inspect whether the residuals deviate significantly from normality.
3. Detecting Outliers: Q-Q plots can help identify outliers and extreme values in the data. Outliers often appear as points that deviate substantially from the expected normal distribution line.

4. Model Improvement: If the Q-Q plot reveals departures from normality, you may need to consider data transformation (e.g., logarithmic transformation) or employ robust regression techniques to account for non-normal residuals.

In summary, a Q-Q plot is a valuable diagnostic tool for assessing the normality assumption in linear regression and other statistical analyses. It provides a visual means to check whether the residuals or other data follow a specified theoretical distribution, helping researchers make informed decisions about model validity and potential data transformations or adjustments.