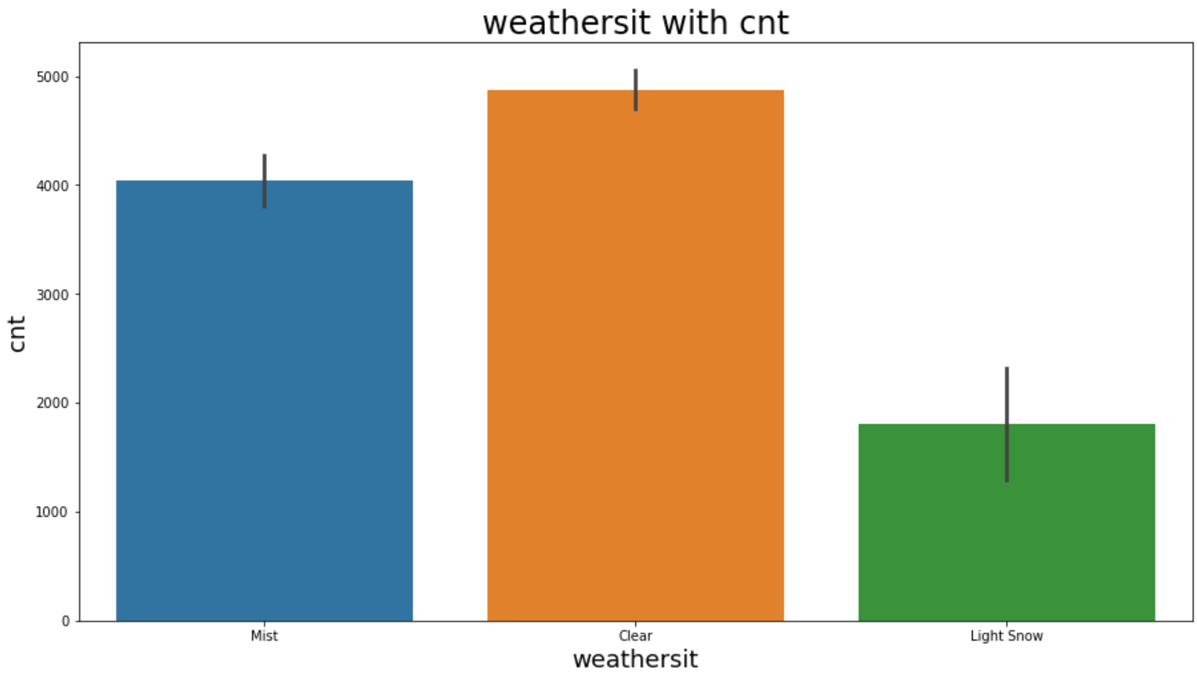
Linear Regression

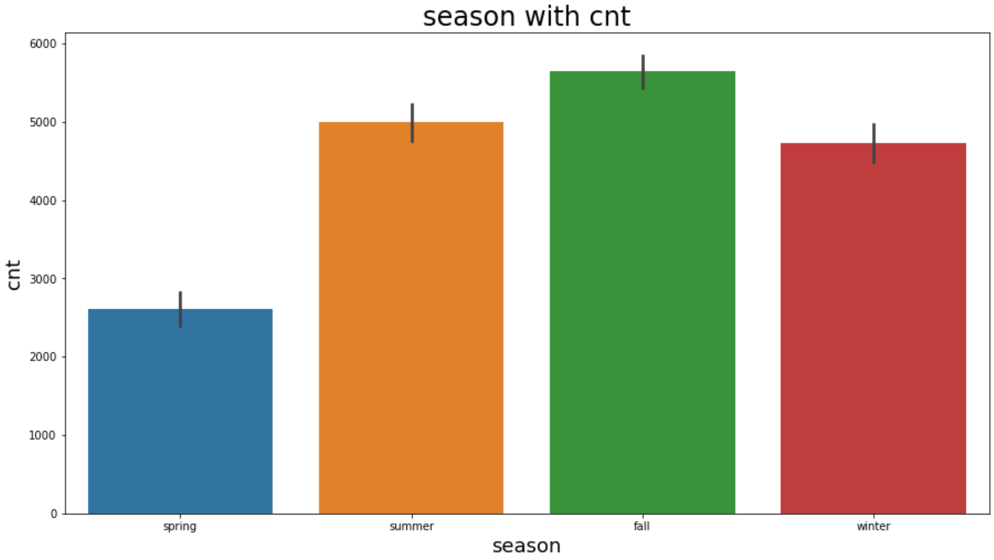
Assignment-based Subjective Questions

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

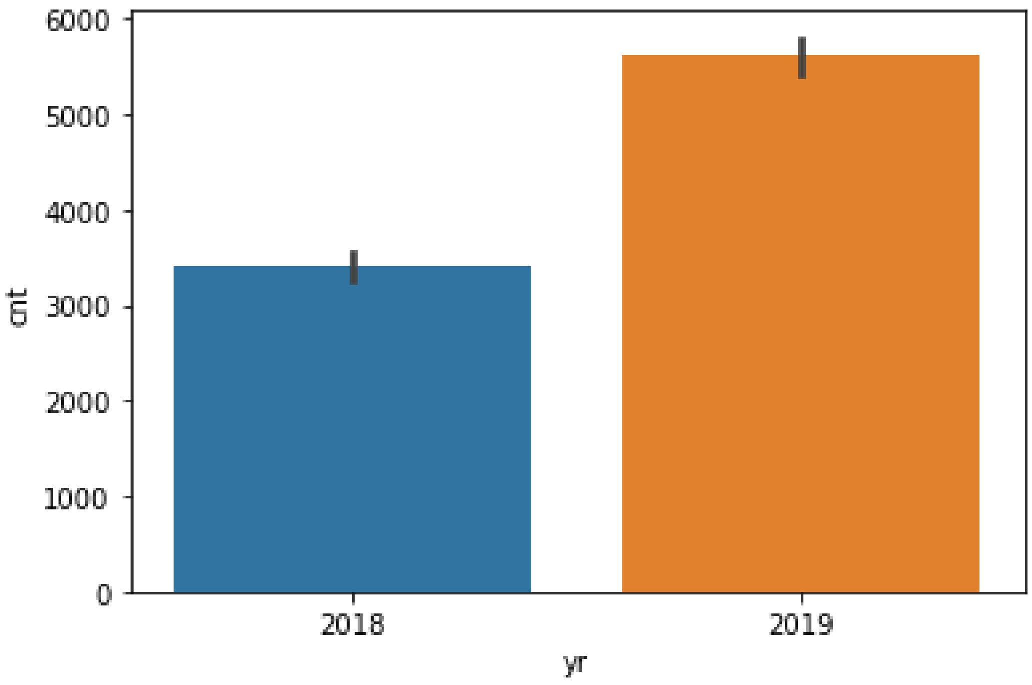
- a. In clear weather people rent more bikes



- In fall season People rent more bike then summer and winter



In 2019 People Rented more bikes then 2018



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

   Ans- drop_first=True is important to use, as it helps in reducing the extra column created during dummy variable creation

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**
   Ans. 'temp' and 'atemp' variable has the highest correlation with the cnt target variable

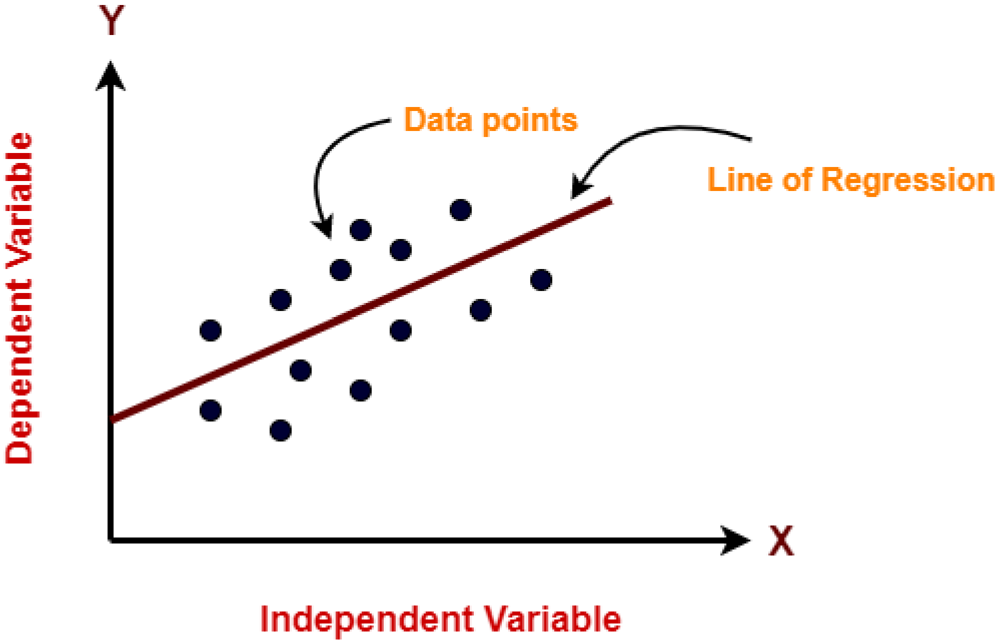
4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**
   Ans - When we build a model, we do residual analysis on the train dataset to see the error distribution, and then we validate our model by using the r2 score on the test dataset.

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**
   Ans – yr, Temp, spring are contributing high

General Subjective Questions

Ans -

Linear regression is a statistical model that examines the linear relationship between two variables: the independent variable (x) and the dependent variable (y). It estimates the relationship by fitting a linear equation to the observed data, which is of the form y = ax + b, where a is the slope of the line and b is the intercept. Linear regression is used to predict the value of a dependent variable based on the value of an independent variable. It can also be used to identify the strength of the relationship between the two variables and to examine which factors may influence the dependent variable.

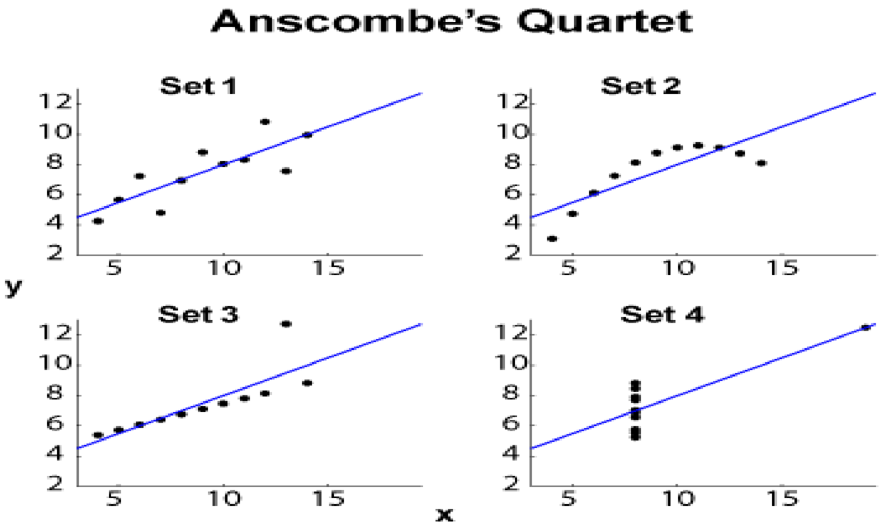


The Line is a bestfit line or it’s a prediction line

The linear regression uses gradient descent algorithm.

2. Explain the Anscombe’s quartet in detail.

Ans - includes four datasets with nearly identical simple descriptive statistics, each dataset having 11 (x,y) points. Anscombe emphasizes the importance of graphing the data before analyzing it, as well as the effects of outliers.



3. What is Pearson’s R?

Ans- The **Pearson correlation coefficient (*r*)** is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

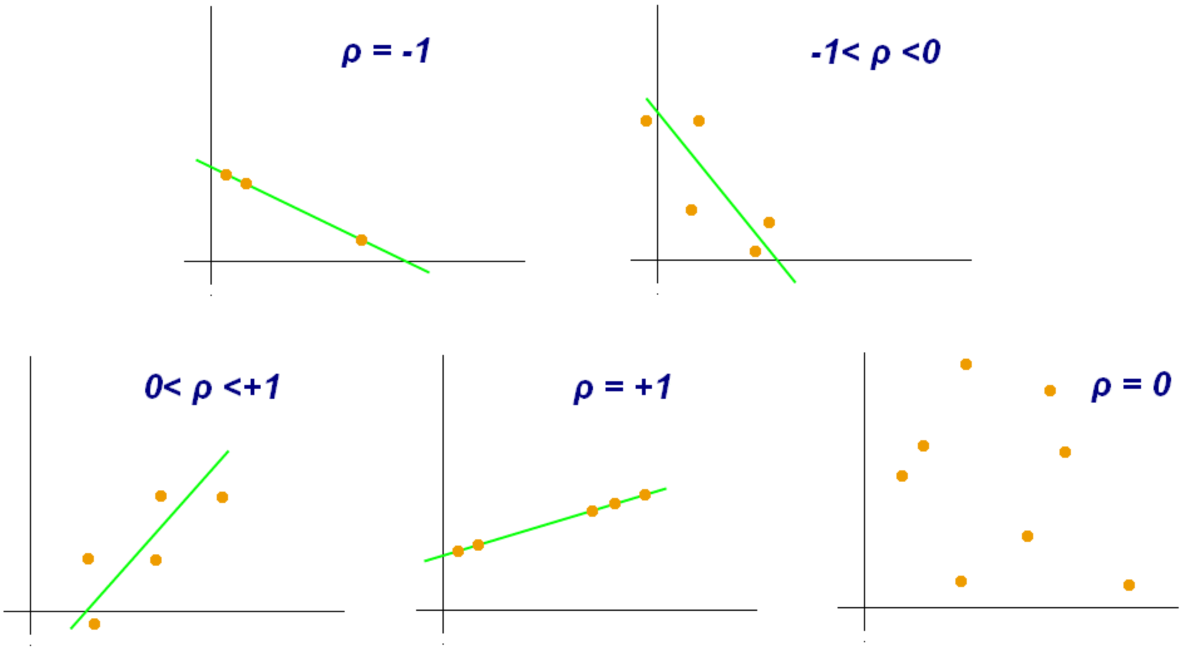
$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

Where,

r = Pearson Correlation Coefficient

x_i = x variable samples y_i = y variable sample

$\bar{x}$ = mean of values in x variable $\bar{y}$ =mean of values in y variable



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

Ans-

Scaling is the process of transforming a set of data into a range of values. This is done to make the data more manageable and easier to analyze.

Scaling is performed to ensure that all the data is on the same level of measurement and to reduce the impact of outliers.

Normalized scaling is a technique used to scale a set of data between two values, typically between 0 and 1. This is done by subtracting the minimum value from each data point and then dividing by the range of the dataset.

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Standardized scaling is a technique used to scale a set of data to have a mean of 0 and a standard deviation of 1. This is done by subtracting the mean from each data point and then dividing by the standard deviation.

$$z = \frac{x - \mu}{\sigma}$$

$\mu =$ Mean

$\sigma =$ Standard Deviation

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

   Ans - This happens when two or more variables in a model are perfectly correlated with each other. This means that the variance of one of the variables can be perfectly predicted from the other variable. In this case, the variance inflation factor is undefined and is represented as infinite.

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

   Ans - A Q-Q plot (Quantile-Quantile Plot) is a graphical method used to assess how closely two sets of data follow a theoretical distribution such as a normal distribution. It is also used to compare two probability distributions. The plot consists of plotting the quantiles of one dataset (usually the dataset of interest) against the quantiles of a second dataset (usually a theoretical distribution). In linear regression, a Q-Q plot is used to compare the residuals of the model to a normal distribution. If the residuals follow a normal distribution, then the points on the Q-Q plot should form a straight line. If the points deviate from the line, then the residuals do not follow a normal distribution and the linear regression model may not be a good fit for the data. Additionally, a Q-Q plot can also be used to detect outliers in the data.