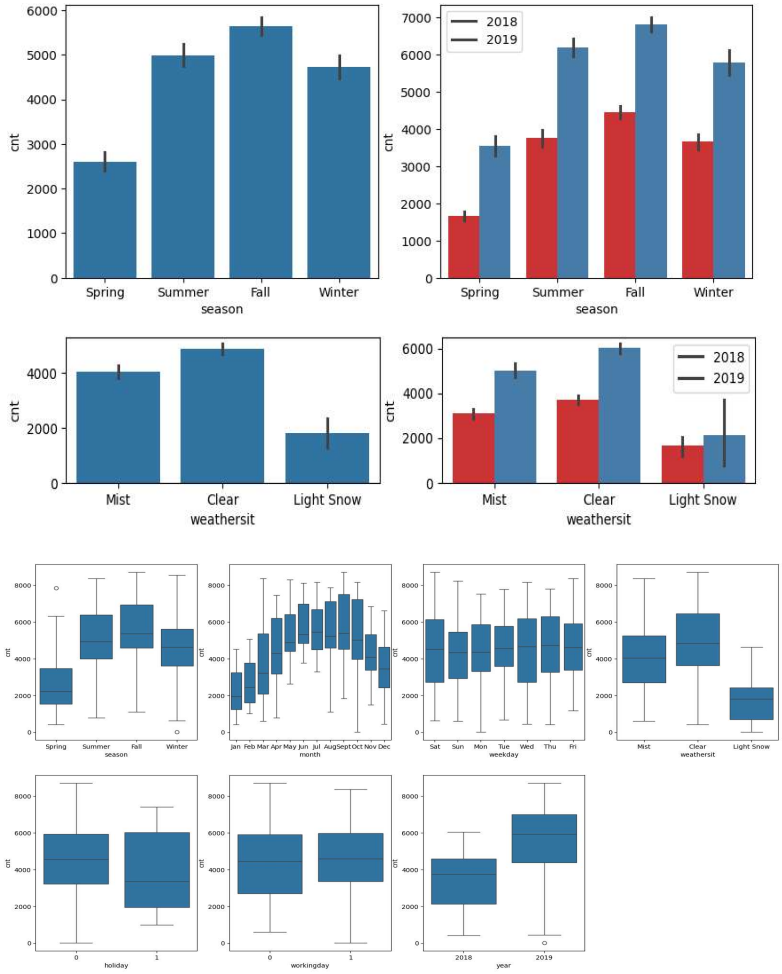
Assignment-based Subjective Questions

1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)**

   I have done analysis on categorical columns using the boxplot and bar plot.

   Below are the few points we can infer from the visualization –

   - Fall Season seems to have attracted more booking. i. e "Fall" Season has more number of Count. And, in each season the booking count has increased drastically from 2018 to 2019.
   - Count of users always increased in all the months of 2018 when compared to the same months of 2019.
   - The mean count of 2018 is much lower than mean count of 2019.
   - The increase in the users trend increased from Jan till Sep just other than July. And after Sep the count decreased.
   - On holidays the count was less when compared to Work day.

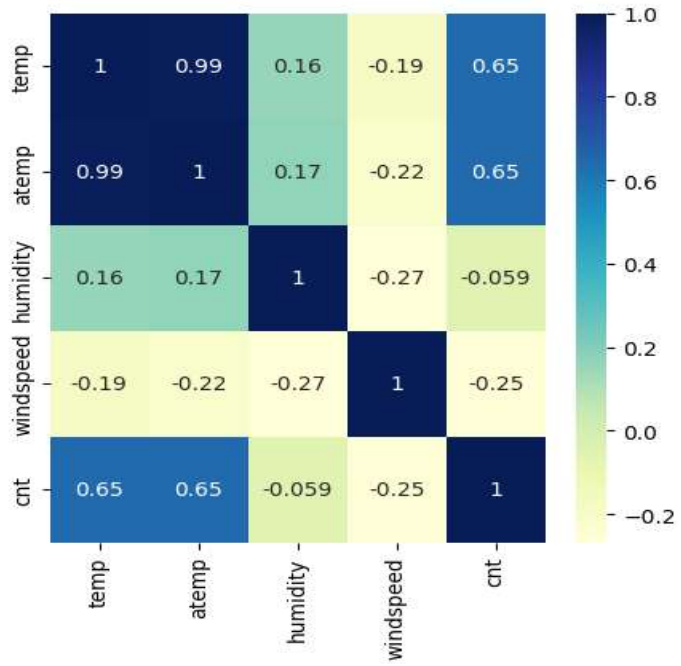


2. **Why is it important to use drop_first=True during dummy variable creation? (2 mark)**

   Drop_first=True is used to drop the base column on which the dummy variables are created. This helps in reducing the extra column for analysis. This also helps in reducing the correlations created among the dummy variables. When you have a categorical variable with say 'n' levels, the idea of dummy variable creation is to build 'n-1' variables, indicating the levels. If we do not use drop_first=True, then 'n' dummy variables will be created (instead of

n-1 variables), and the predictors (n dummy variables) are themselves correlated which is known as multicollinearity and it, in turn, leads to **Dummy Variable Trap**.

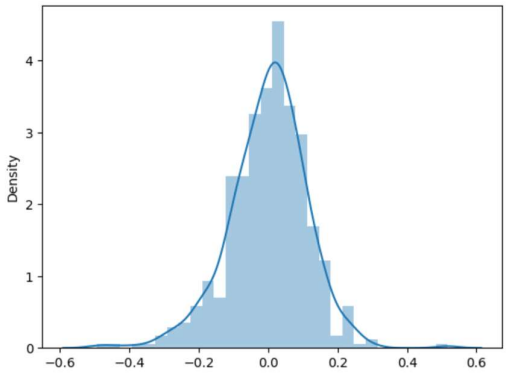
3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)**
   "**temp**" & "**atemp**" is the numerical variable which has the highest correlation with the target variable.



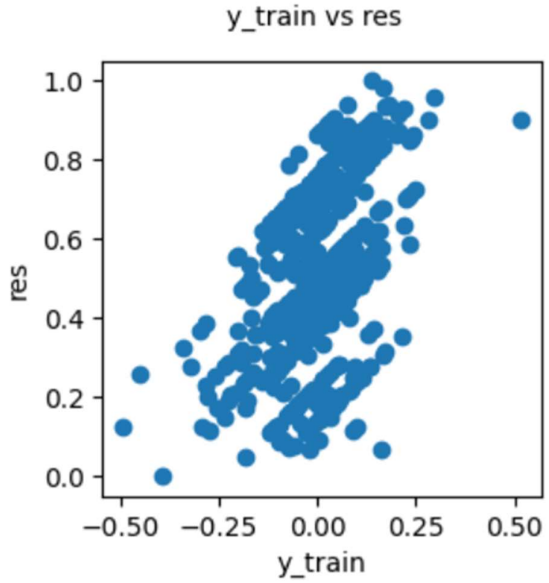
4. **How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)**
   There are four assumptions associated with a linear regression model:

   Linearity: The relationship between X and the mean of Y is linear.
   Since we are in the process of building a model in the format of a straight line which is linear (y=mx + c), this assumption is taken care of.

   Normality: For any fixed value of X, Y is normally distributed.
   After building the model the residuals/errors are plotted in found that they are normally distributed.



Homoscedasticity: Error Terms have constant variance.



Independence: Observations are independent of each other.

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)**
   Based on the final model, below are the 3 top features contributing significantly towards the demand of shared bikes –
   - ✓ Year
   - ✓ Spring
   - ✓ Light Snow

General Subjective Questions

1. **Explain the linear regression algorithm in detail. (4 marks)**
   Linear regression is a type of supervised machine learning algorithm that computes the linear relationship between a dependent variable and one or more independent features. It plays a pivotal role in predicting continuous outcomes. Linear regression predicts the relationship between two variables by assuming a linear connection between the independent and dependent variables. It seeks the optimal line that minimizes the sum of squared differences between predicted and actual values.
   Applied in various domains like economics and finance, this method analyses and forecasts data trends. When there is only one Independent variable upon which the prediction is made then it is Simple Linear Regression, where as if there is more than one independent variable, it is Multiple Linear Regression. Multiple linear regression can involve several independent variables.

   Simple Linear Regression
   This is the simplest form of linear regression, and it involves only one independent variable and one dependent variable. The equation for simple linear regression is:
   $Y = \beta 0 + \beta 1X$
   where:
   - Y is the dependent variable
   - X is the independent variable

- β0 is the intercept
- β1 is the slope

Multiple Linear Regression

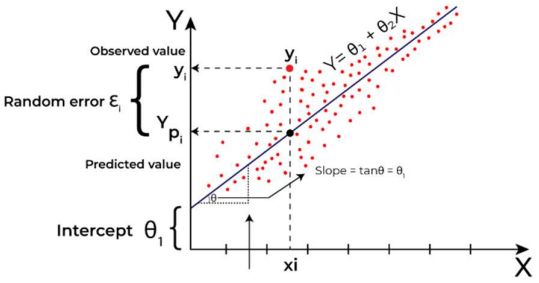
This involves more than one independent variable and one dependent variable. The equation for multiple linear regression is:

Y = β0 + β1X1 + β2X2 + ……. βnXn

where:

- Y is the dependent variable
- X1, X2, …, Xp are the independent variables
- β0 is the intercept
- β1, β2, …, βn are the slopes

The goal of the algorithm is to find the best Fit Line equation that can predict the values based on the independent variables.



2. **Explain the Anscombe's quartet in detail. (3 marks)**

Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (x, y) points.

For all the 4 datasets -

Mean of X – 9

Sample Variance of X – 11

Mean of y – 7.5

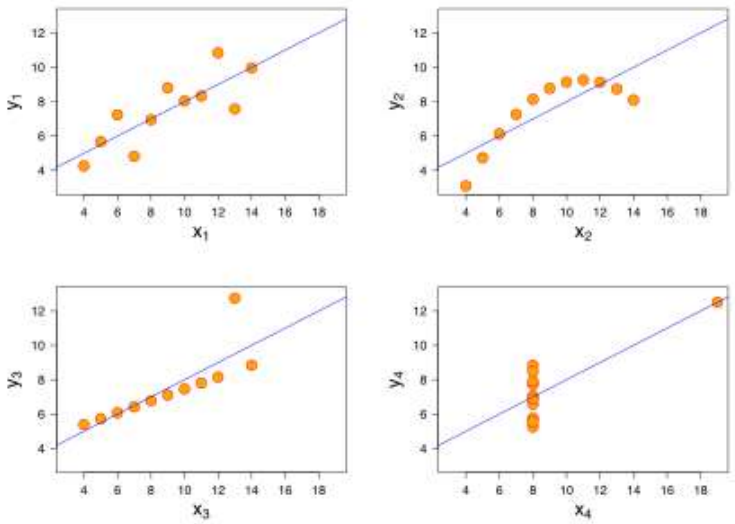
Sample Variance of y – 4.125

Correlation between x and y – 0.816

Linear regression line – y=3.00 + 0.500X

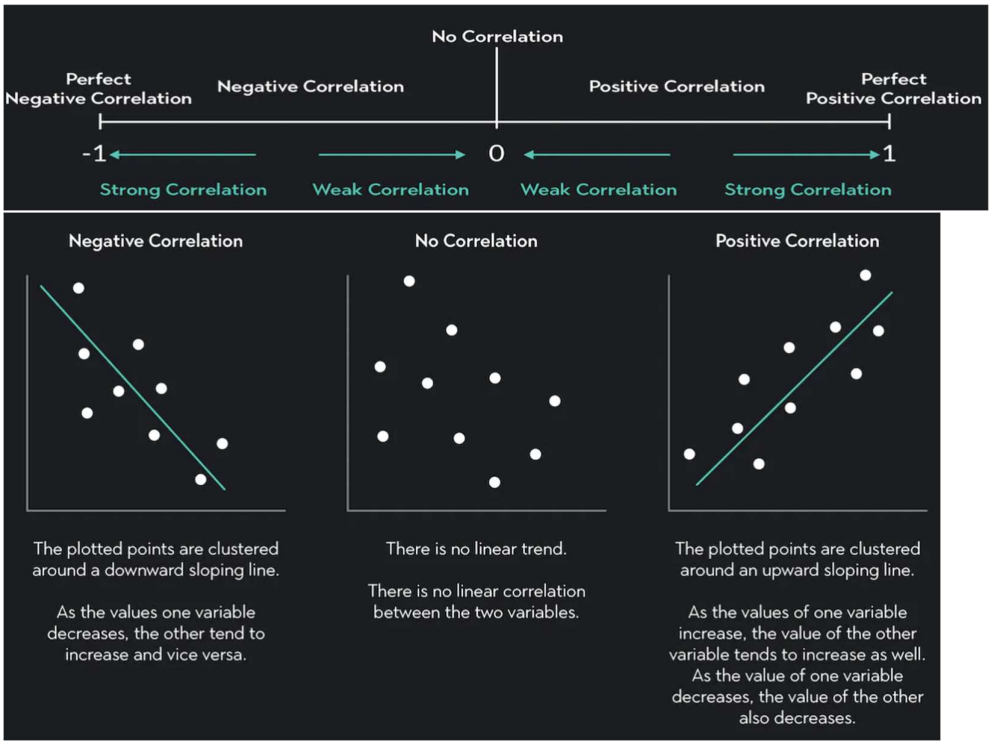
Coefficient of determination of the linear regression: R2 – 0.67

The quartet is often used to illustrate the importance of looking at a set of data graphically before starting to analyse according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets.



3. What is Pearson's R? (3 marks)

- Pearson correlation coefficient (PCC) is a correlation coefficient that measures linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations.
- The correlation coefficient ranges from −1 to 1.
- An absolute value of exactly 1 implies that a linear equation describes the relationship between X and Y perfectly, with all data points lying on a line.
- The correlation sign is determined by the regression slope: a value of +1 implies that all data points lie on a line for which Y increases as X increases, and vice versa for −1.
- A value of 0 implies that there is no linear dependency between the variables.



4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)**

Feature scaling is a method used to normalize the range of independent variables or features of data. Since the dataset contains different columns & the values of the columns will vary in a wide range, if we build a model on top of the column data as such, ML algorithm will tend to weigh values based on the higher values. Hence to avoid them, we bring all the values in common specific range, so that weightage of each feature is calculated approximately proportionately.

There are two major methods to scale the variables, i.e. Standardisation and MinMax scaling.

Standardisation basically brings all of the data into a standard normal distribution with mean zero and standard deviation one.

MinMax Scaling, on the other hand, brings all of the data in the range of 0 and 1. The formulae in the background used for each of these methods are as given below:

- Standardisation: $x = \frac{x - mean(x)}{sd(x)}$
- MinMax Scaling: $x = \frac{x - min(x)}{max(x) - min(x)}$

SNO	Normalized Scaling	Standardized Scaling
1	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2	Brings all data in the range of 0 & 1	Brings all data into Standard Normal Distribution, with mean=0 & SD=1
3	It is really affected by outliers.	It is much less affected by outliers.
4	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? (3 marks)**

VIF – Is the Variable Inflation Factor. VIF calculates how well one independent variable is explained by all the other independent variables combined.

VIF is calculated as below –

$$VIF_i = \frac{1}{1 - {R_i}^2}$$

So if we analyse why VIF becomes infinite, it becomes infinite when denominator is 0, i. e, when R2 =1. R2=1 for that variable with it has perfect correlation with the other independent variables. Hence for this if we drop one of the independent variable, the VIF drops.

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)**

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data possibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. It is used to compare the shapes of distributions, providing a graphical view of how properties such as location, scale, and skewness are similar or different in the two distributions.

Q–Q plots can be used to compare collections of data, or theoretical distributions.
For example, if you are testing if the distribution of age of employees in your team is normally distributed, you are comparing the quantiles of your team members' age vs quantile from a normally distributed curve. If two quantiles are sampled from the same distribution, they should roughly fall in a straight line.

Importance :-
Q-Q plot I commonly used to test distribution amongst 2 different datasets.
For example, if dataset 1, the age variable has 200 records and dataset 2, the age variable has 20 records, it is possible to compare the distributions of these datasets to see if they are indeed the same. This can be particularly helpful in machine learning, where we split data into train-validation-test to see if the distribution is indeed the same. It is also used in the post-deployment scenarios to identify covariate shift/dataset shift/concept shift visually.