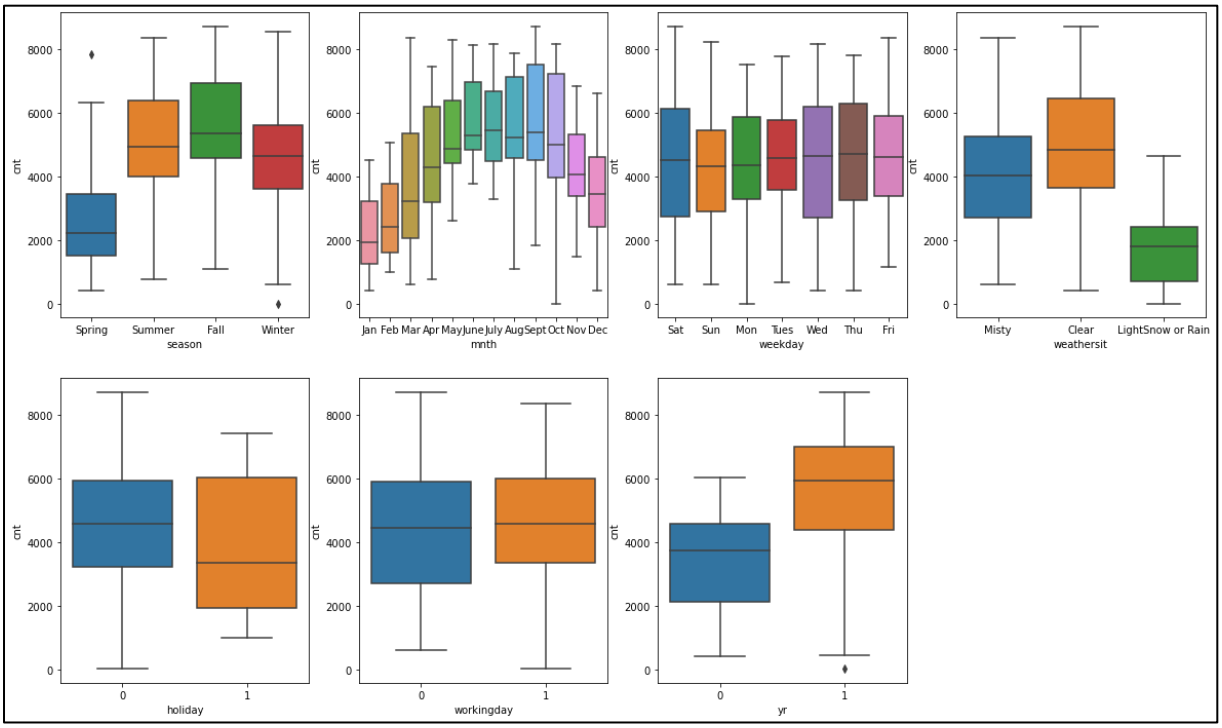
Assignment-based Subjective Questions

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

   ANS - As per the analysis done on the categorical variables using the boxplots and bar plots the following conclusions can be made.



   - Fall season has more demand and has more bookings as compared to any other season.
   - In each season the booking count has increased through 2018 to 2019.
   - Maximum Number of bookings are done in the months of May, june, july, aug, sep and oct. Trend increased starting of the year till September and then it started decreasing till the end of year.
   - Thu, Fir, Sat and Sun have more number of bookings as compared to mon, tues, wed.
   - Clear weather definitely contributes to attracted more booking and light snow or rainy seasons affects the number of bookings as the booking are the lowest in these weather condition.
   - There are fewer bookings on the Holidays.
   - Bookings are to be almost equal either on working days or Holidays.
   - The demand increased in the year 2019 as compared to 2018 hence there is a Growth in the business.

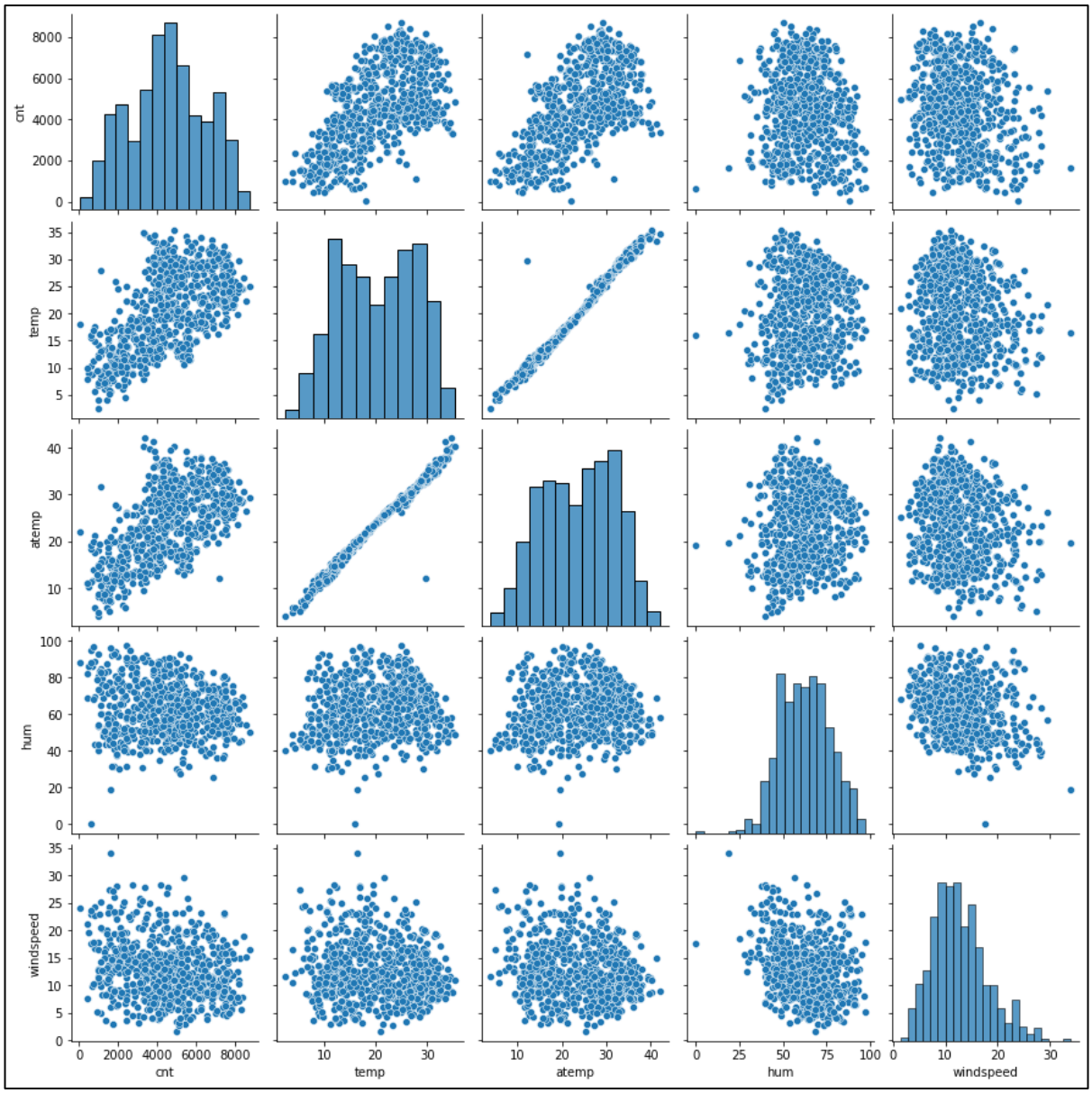
2. Why is it important to use **drop_first=True** during dummy variable creation?(2 mark)

   ANS - "drop_first = True" is important because it helps in reducing the extra column created during creating the dummy variables. Hence it reduces the correlations among dummy

variables.

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

ANS - From thePair Plot, temp & atemp seems to be highly correlated as compared to other variables .



4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

ANS - Validation of the assumptions of Linear Regression Model Are done on the basis of the following points.

- Based on its linearity.
- The Variables should show linearity.
- Errors should be normally distributed.
- There should be very less multicollinearity between variables.
- There should be Homoscedasticity.

- There should be no auto-correlation.

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

ANS - On the basis of the final model the top 3 features contributing significantly towards explaining the demand of the shared bikes are.

- Windspeed
- Temp
- Season_winter.

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

ANS –

- **Linear regression**: - Linear regression is the supervised machine learning model which explains the relationship between dependent or output variable and independent variable/s that is predictor variable. There are two types of Linear Regression: they are as follows.

  1) Simple linear regression.

  2) Multiple linear regression.

- **Algorithm:-** Below are the steps for linear regression –

  1) Import Necessary libraries like pandas, seaborn, matplotlib, sklearn and statsmodels.

  2) Reading and understanding the dataset.

  3) After reading we Check for missing values, outliers etc.

- Understand potential independent variables based on dataset and business requirements.

- **Prepare data for modeling:-**

  1) Handle categorical and binary variables.

  2) Check if assumptions are met as per type of regression model.

- **Split dataset into train and test set**

- **Train the model:-**

  1) Check for significant variables (using train set) in case of multiple linear regression.

  2) Drop out insignificant variables

  3) Repeat step 5 until best coefficients found

- **Predictions and model evaluation on test set:-**

1) Predict target variable values using a test set.

2) Evaluate the model using the cost function.

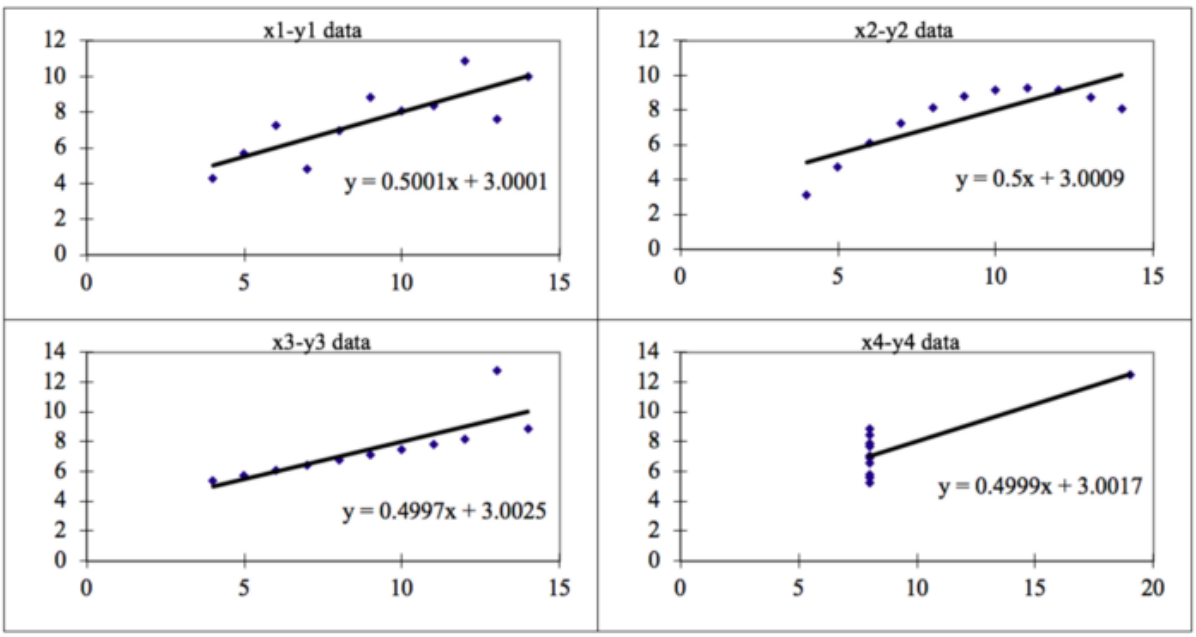
2. Explain the Anscombe's quartet in detail. (3 marks)

ANS –

- Anscombe's quartet can be defined as a group of four datasets which are nearly identical in simple descriptive statistics, but there are some peculiarities in dataset that fools the regression model if it is built.They have very different distributions and appear differently when plotted on scatter plots.
- Example:-

Anscombe's Data									
Observation	x1	y1		x2	y2	x3	y3	x4	y4
1	10	8.04		10	9.14	10	7.46	8	6.58
2	8	6.95		8	8.14	8	6.77	8	5.76
3	13	7.58		13	8.74	13	12.74	8	7.71
4	9	8.81		9	8.77	9	7.11	8	8.84
5	11	8.33		11	9.26	11	7.81	8	8.47
6	14	9.96		14	8.1	14	8.84	8	7.04
7	6	7.24		6	6.13	6	6.08	8	5.25
8	4	4.26		4	3.1	4	5.39	19	12.5
9	12	10.84		12	9.13	12	8.15	8	5.56
10	7	4.82		7	7.26	7	6.42	8	7.91
11	5	5.68		5	4.74	5	5.73	8	6.89
				Summary Statistics					
N	11	11		11	11	11	11	11	11
mean	9.00	7.50		9.00	7.500909	9.00	7.50	9.00	7.50
SD	3.16	1.94		3.16	1.94	3.16	1.94	3.16	1.94
r	0.82			0.82		0.82		0.82	

- When these models are plotted on a scatter plot, all datasets generates a different kind of plot that is not interpretable by any regression algorithm which is fooled by these peculiarities and can be seen as follows:



- The four datasets can be described as:

A. Dataset 1: this fits the linear regression model pretty well.
B. Dataset 2: this could not fit linear regression model on the data quite well as the data is non-linear.
C. Dataset 3: shows the outliers involved in the dataset which cannot be handled by linear regression mode
D. Dataset 4: shows the outliers involved in the dataset which cannot be handled by linear regression model

- The following four datasets are intentionally created to describe the importance of data visualization and how any regression algorithm can be fooled by the same. Hence, all the important features in the dataset must be visualized before implementing any machine learning algorithm on them which will help to make a good fit model.

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

ANS –

- In Statistics, the Pearson's Correlation Coefficient is referred to as Pearson's R
- Pearson's r is a numerical summary of the strength of the linear correlations between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.
- However, it cannot capture nonlinear relationships between two variables and cannot differentiate between dependent and independent variables. Like all correlations, it also has a numerical value that lies between -1.0 and +1.0.

$$r = \frac{N\Sigma xy-(\Sigma x)(\Sigma y)}{\sqrt{[N\Sigma x^2-(\Sigma x)^2][N\Sigma y^2-(\Sigma y)^2]}}$$

- N = the number of pairs of scores
- Σxy = the sum of the products of paired scores
- Σx = the sum of x scores
- Σy = the sum of y scores
- Σx2 = the sum of squared x scores
- Σy2 = the sum of squared y scores
- r = 1 means the data is perfectly linear with a positive slope (i.e., both variables tend to change in the same direction)
- r = -1 means the data is perfectly linear with a negative slope (i.e., both variables tend to change in different directions)
- r = 0 means there is no linear association
- r > 0 < 5 means there is a weak association
- r > 5 < 8 means there is a moderate association
- r > 8 means there is a strong association.

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

ANS -

- Feature Scaling is a technique to standardize the independent features present in the data in a fixed range.That is you are transforming the data to make it fit within a specific scale
- It is performed during the data pre-processing to handle highly varying magnitudes or values or units. And speedup the calculation in an algorithm. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.
- Difference between normalized scaling and standardized scaling

Normalized scaling	standardized scaling
Use of maximum and minimum values are done for scaling.	Use of mean and standard deviation is done for scaling
It is used when the features have different scale	It is used when we want to ensure zero mean and unit standard deviation
Scales values between [0, 1] or [-1, 1].	It is not bound to a certain range.
Outliers affect scaling	It is less affected by the outliers
Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.
It is useful when we don't know about the distribution	It is useful when the feature distribution is Normal or

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

ANS –

Here are the key reasons for infinite VIF values:

- Perfect Linear Relationship: When two or more independent variables in the model are perfectly correlated, one can be expressed as a linear combination of the others. This results in a division by zero when calculating VIF, leading to an Infinite value.
- Redundant Variables: In cases where one variable can be precisely predicted from a combination of other variables, multicollinearity arises. This redundancy makes it impossible to calculate VIF for the affected variables.
- Insufficient Data: When the dataset is too small relative to the number of independent variables, VIF calculations can become unstable. In such cases, the estimates are unreliable, and VIF values may be infinite.

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

ANS –

Q-Q plot is a plot of quantiles (Quantile -Quantile plot) it is a graphical technique for determining if the two datasets come from the populations with the same common distribution .it can be also explained as it is thecomparison of 2 probability distributions by plotting tier quantiles against each other's.

The use and importance of a Q-Q plot in linear regression includes:

- Normality Assumption: Linear regression often assumes that the residuals (the differences

between observed and predicted values) are normally distributed. Q-Q plots help verify this assumption by assessing whether the residuals
follow a normal distribution.

- Identification of Outliers: Q-Q plots can reveal outliers and extreme values. Outliers can distort regression results and may need special consideration or data cleaning.
- Model Evaluation: Q-Q plots are valuable for model evaluation. If the residuals closely match a straight line, it suggests that the model assumptions are met, and the regression results are more reliable.
- Data Transformation: When the Q-Q plot indicates non-normality, data transformation techniques may be applied to improve the model's performance and the validity of statistical inferences.