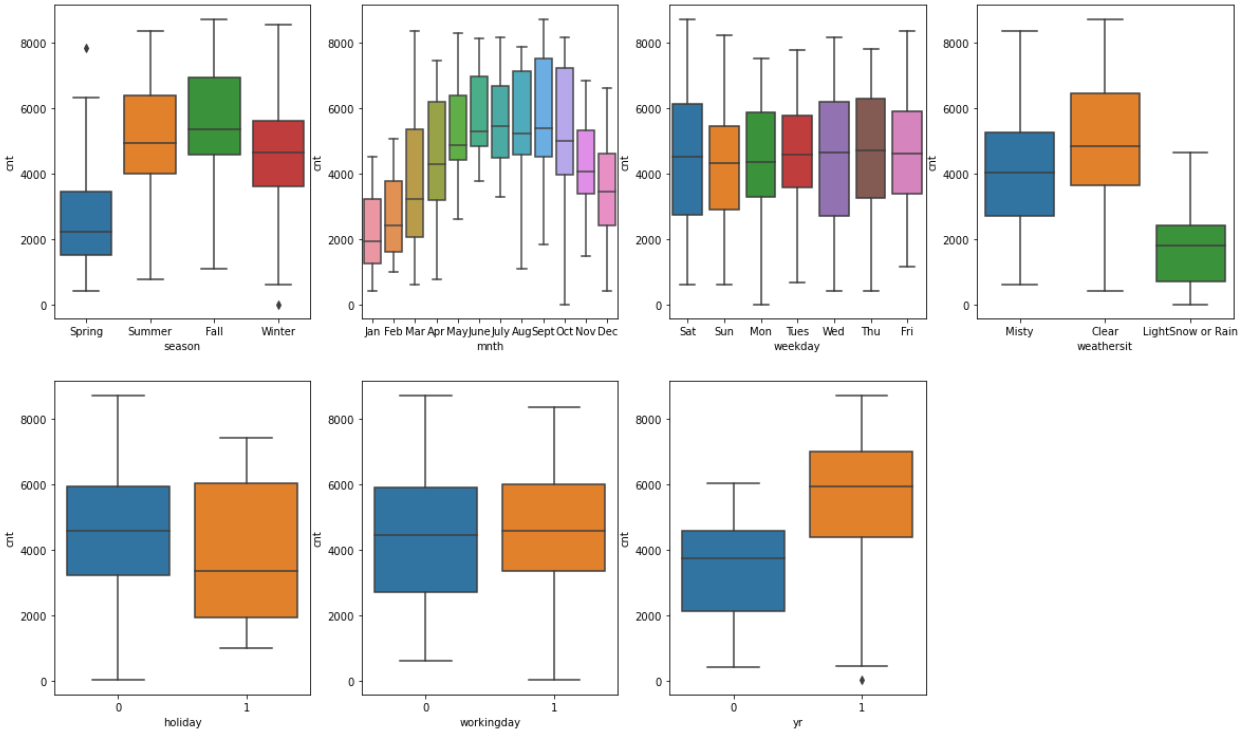
Assignment-based Subjective Questions

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

   Answer: -

   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?

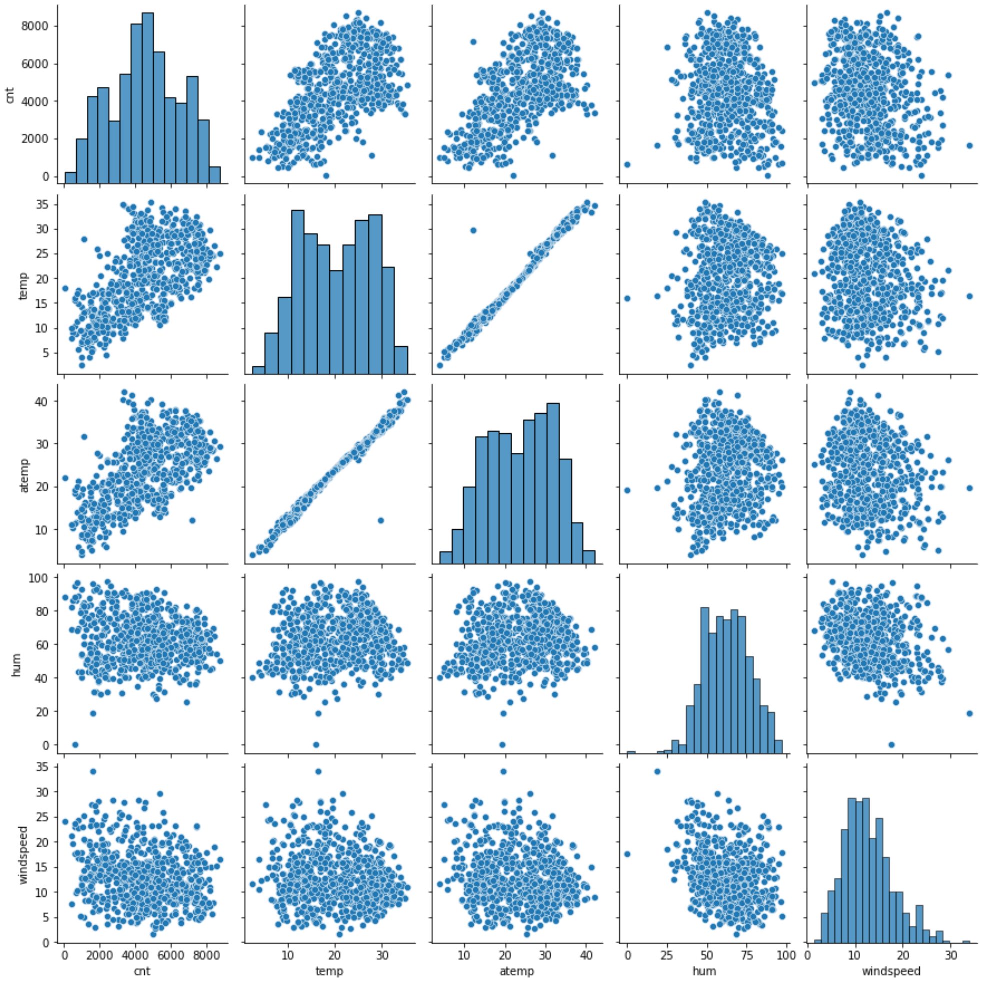
Answer:-

- “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?

Answer:-

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?**
   Answer:-
   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?**
   Answer:-
   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.**

   Answer:-

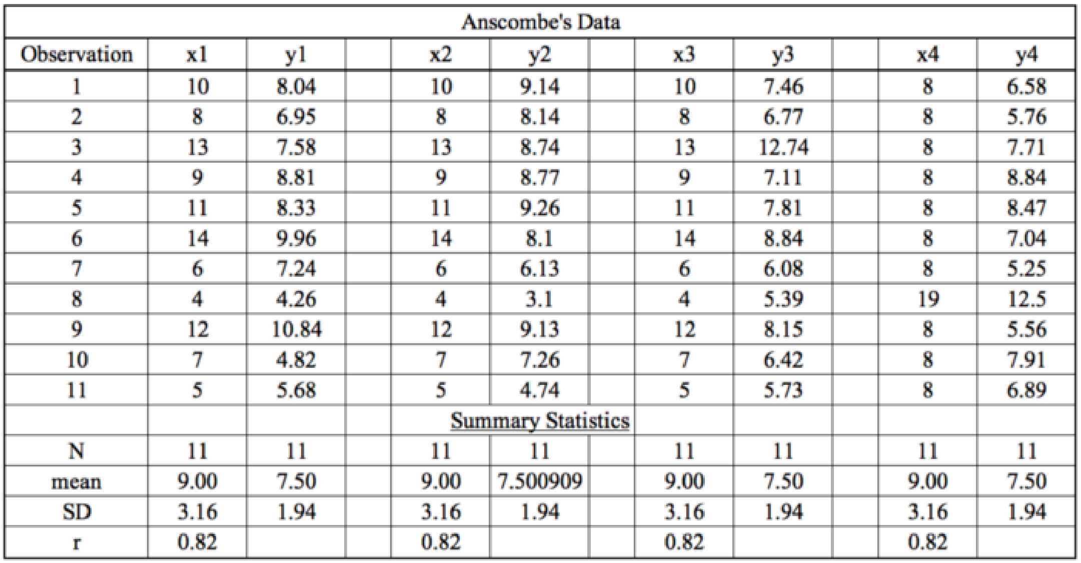
- Linear regression – Linear regression is the supervised machine learning model which explains the relationship between dependent or output variable and independent variable/s that is predictor variable. There are two types of Linear Regression: they are as follows..
  1) Simple linear regression.
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

Answer:-

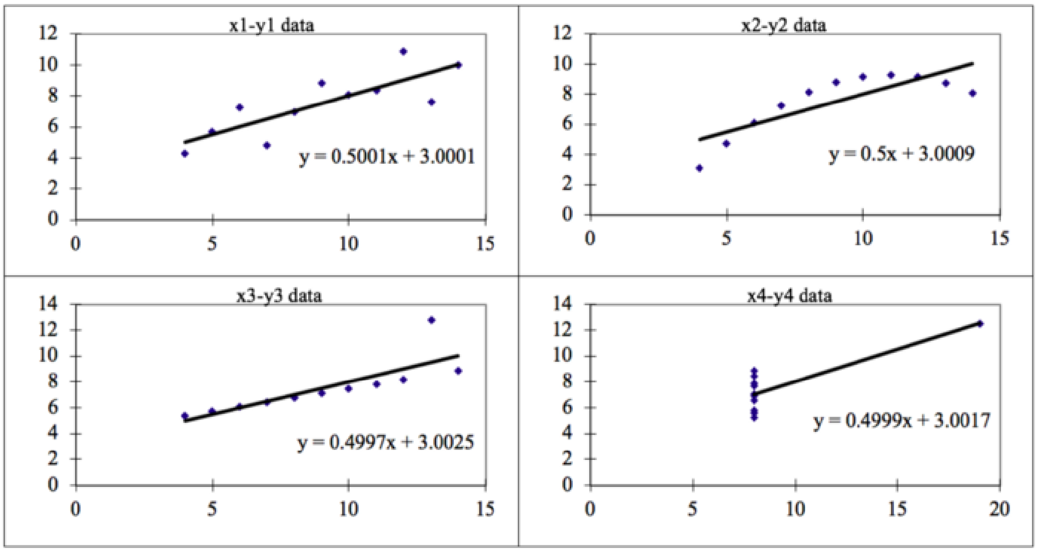
- Anscombe's quartet can be defined as a group of four datasets which are nearly identical in simple descriptive statistics, but there are some peculiarities in dataset that fools the regression model if it is built.They have very different distributions and appear differently when plotted on scatter 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:
- **Dataset 1:** this fits the linear regression model pretty well.
- **Dataset 2:** this could not fit linear regression model on the data quite well as the data is non-linear.
- **Dataset 3:** shows the outliers involved in the dataset which cannot be handled by linear regression mode
- **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?**
Answer:-

- 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 scaling and standardized scaling?

Answer:-

- 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?**

Answer:-

- If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables
- In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To resolve this we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.
- An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables which also show an infinite VIF .

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

Answer:-

- 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 the comparison of 2 probability distributions by plotting tier quantiles against each others.

- Use of Q-Q plot are as follows.
  - A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second dataset. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value. A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.
- Advantages are as followed:-
- It can be used with sample sizes also
- Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.
- It is used to check following scenarios:
  - In two data sets —
    - If there are twe populations with a common distribution
    - They have common location and scale
    - They have similar distributional shapes
    - They have similar tail behavior.