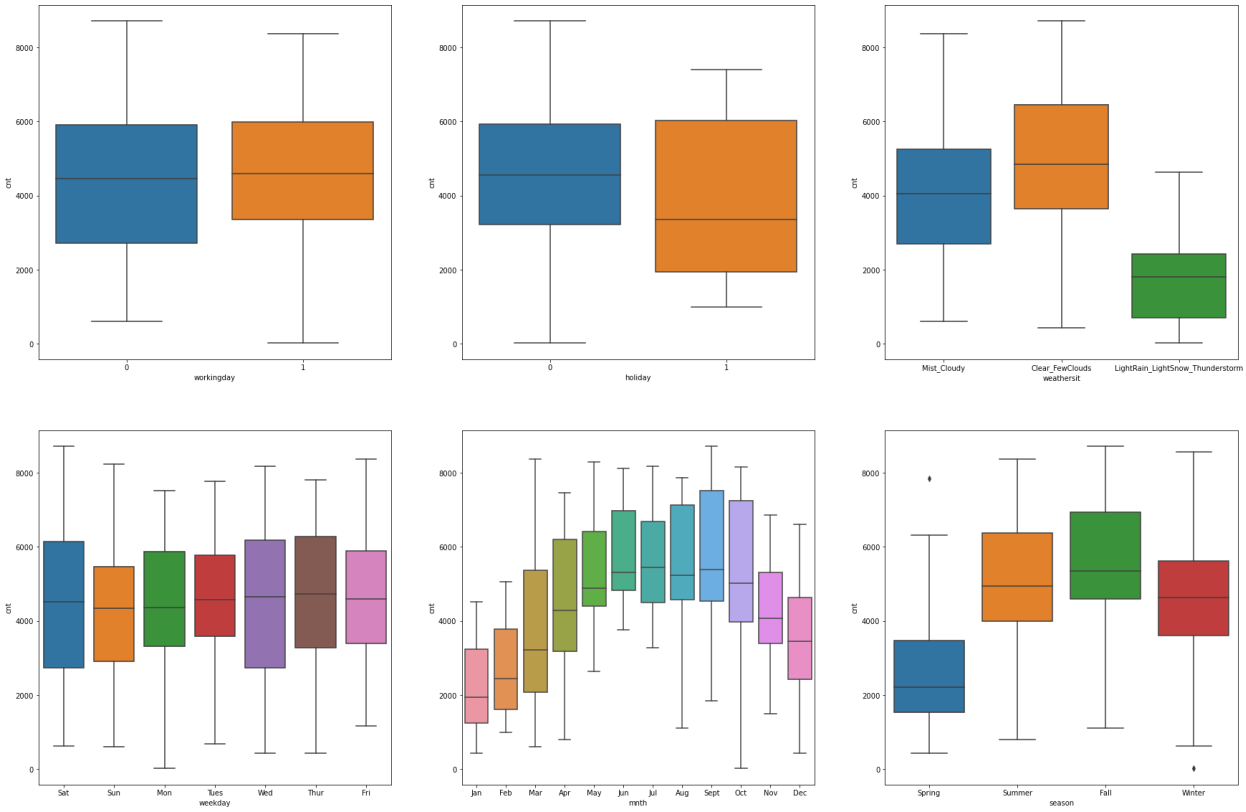
LINEAR REGRESSION ASSIGNMENT

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?**

Answers:



By analysing categorical variables from the dataset it is inferred as follows:

- From the first and fourth plots that are for the variable 'workingday' and 'weathersit' we don't see any significant insights from the graph.
- From the second plot for the variable 'holiday,' the demand has decreased as there is a holiday.
- From the third plot that is for the variable 'weathersit', it is acknowledged that the demand has increased during 'Clear_Few clouds and 'Partly cloudy' weather.

- From the fifth plot that is for the variable 'mnth', it is acknowledged that the demand is continuously increasing till the month of June and after September the demand is decreasing.
- This also shows us that the maximum demand or booking happens during the month of September.
- From the sixth plot that is for the variable 'season' it is acknowledged that maximum booking happens during the 'fall season'.

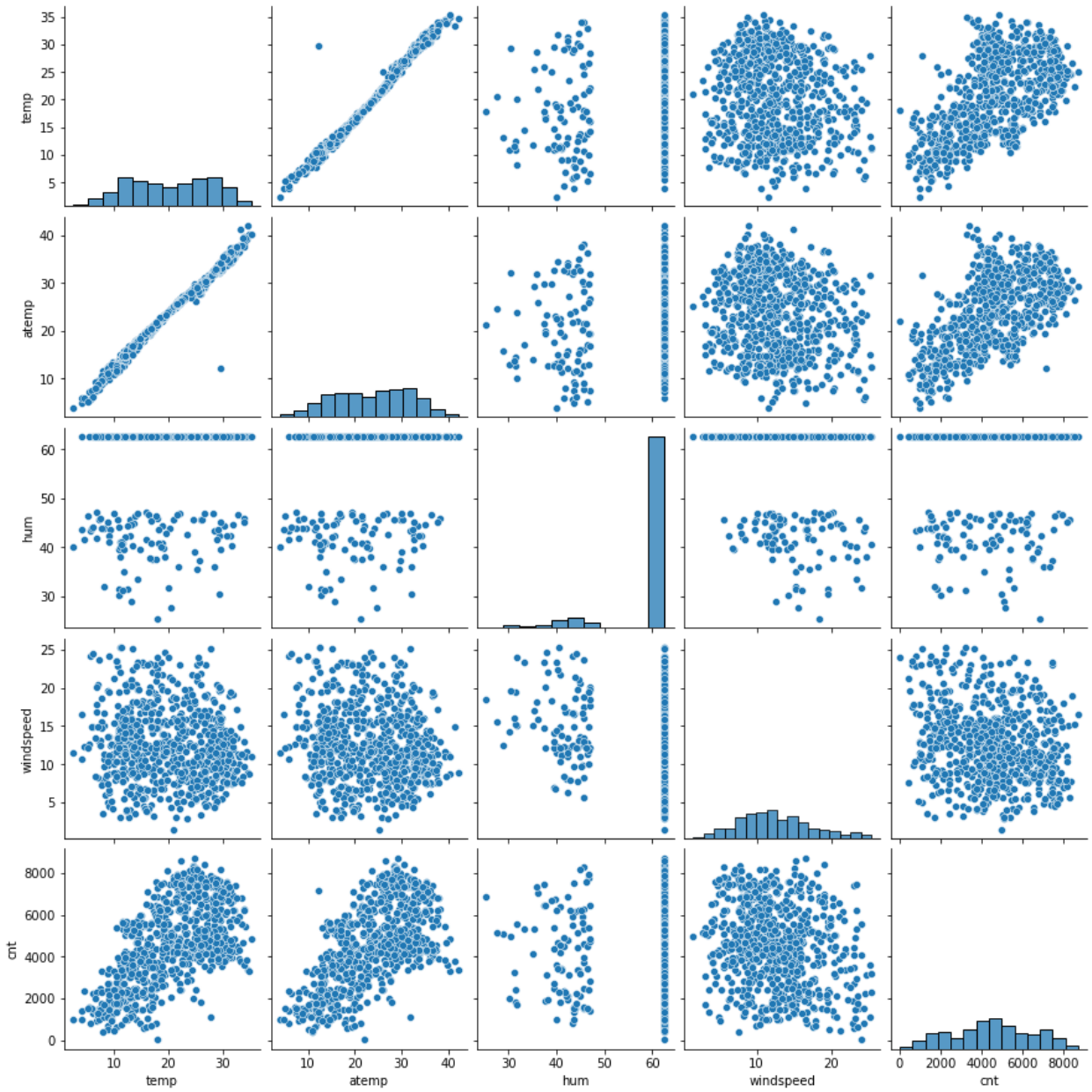
2. **Why is it important to use drop_first=True during dummy variable creation?**

Answers:

It will create one binary 1/0 variable for each categorical level when we use 'get dummies' on the categorical variable. In this example, let's assume that an attribute has m levels. 'm' binary variables are generated by get_dummies by default. The absence(0) of a tier in the 'm-1' binary columns presupposes the presence(1) of the 'm'th binary variable, so it is important to keep this in mind. There is a linear relationship between the binary variables if we have 'm' of them. Since the first level is dropped, we can avoid linear dependence between the independent variables by using the 'drop first=True' condition. 'drop first=True' is being used to prevent multicollinearity in simple terms.

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

Answers:



By looking at the pairplot described above it is acknowledged that 'temp' and 'atemp' variables are highly correlated with the target variable 'cnt'.

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

Answers:

The presumptions of linear regression on a model based on training data can be verified in a number of ways:

- There must be a normal distribution for each independent variable. Otherwise, we'll have to make adjustments.
- By creating a distribution plot, we can calculate the residuals. The distribution of the residuals must be normal in this calculation.
- Using a combination of correlations and a VIF calculation, we iteratively select model variables to make sure that there is no any multicollinear variables utilized in the model.
- If the dependent variable and the independent variable have a linear relationship, we can visualize this using a pairplot/scatter plot.

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

Answers:

According to the final model the top 3 features contributing significantly towards explaining the demand of the shared bikes are as follows:

- temp
- yr
- holiday
- season
- weekday Sunday
- Month

Here,

- yr: coefficient of yr depicts that yr will increase bike hiring by 2019.944545 value if the unit increase in yr variable
- season_Spring: coefficient of season_Spring depicts that season_Spring will decrease bike hiring by 1026.694652 value if the unit increase in season_Spring variable.
- season_Winter: coefficient of season_Winter depicts that season_Winter will increase bike hiring by 534.701970 value if the unit increase in season_Winter variable.
- mnth_Sept: coefficient of mnth_Sept depicts that mnth_Sept will increase bike hiring by 626.386774 value if the unit increase in mnth_Sept variable.
- weathersit_LightRain_LightSnow_Thunderstorm: coefficient of weathersit_LightRain_LightSnow_Thunderstorm depicts that weathersit_LightRain_LightSnow_Thunderstorm will decrease bike hiring by 2685.592063 value if the unit increase in weathersit_LightRain_Ligthe htSnow_Thunderstorm variable.

- weathersit_Mist_Cloudy: coefficient of weathersit_Mist_Cloudy depicts that weathersit_Mist_Cloudy will decrease bike hiring by 676.125770 value if the unit increase in weathersit_Mist_Cloudy variable.
- temp: coefficient of temp depicts that temp will increase bike hiring by 676.125770 value if the unit increase in a temp variable.
- weekday_Sun: coefficient of weekday_Sun depicts that weekday_Sun will decrease bike hiring by 414.921335 value if the unit increase in weekday_Sun variable.

General Subjective Questions:

1. Explain the linear regression algorithm in detail.

Answers:

- When it comes to machine learning, linear regression is the simplest and most common type of model training. In linear regression, as the name suggests, the two variables on the x-axis and y-axis should be linearly correlated.
- When running a sales promotion, for example, you can look back at previous promotions and plot them on a chart to see if there is an increase in the number of customers when you rate them. With the help of the previous historical data, you can then try to figure out or estimate what will be the count or what will be the count. For example: In this case, the goal is to make an accurate prediction about what the future will hold based on past performance.
- Mathematically, the equation of linear regression is described as,

$$y = a + bx$$

Here a and b are given and x and y are the two variables on the regression line.

$$a = \frac{(\sum y)(\sum x^2) - (\sum x)(\sum xy)}{n(\sum x^2) - (\sum x)^2}$$

$$b = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$$

b = Slope of the line
a = y-intercept of the line
x = Independent variable from dataset
y = Dependent variable from dataset

2. Explain the Anscombe's quartet in detail.

Answers:

In Anscombe's quartet example, the underlying distributions can be vastly different even if their descriptive statistics are the same. All four examples of distributions are included in this quartet. Each x has a mean of 9, a standard deviation of 11, and a standard deviation of 4.125 for each y. We can see from the example below how important it is to visualise data in order to gain a better understanding of it. Outliers and other extreme values can also be explained by looking at this data set.

Datasets:

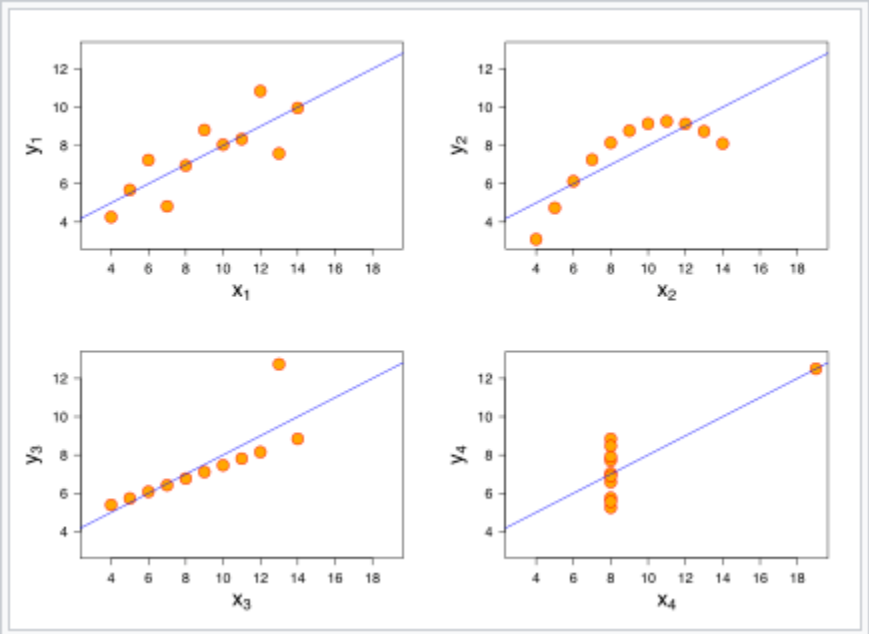
Anscombe's quartet

I		II		III		IV	
x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Descriptive stats:

Property	Value
Mean of x	9
Sample variance of x : s_x^2	11
Mean of y	7.50
Sample variance of y : s_y^2	4.125
Correlation between x and y	0.816
Linear regression line	$y = 3.00 + 0.500x$
Coefficient of determination of the linear regression : R^2	0.67

Graphs:



3. What is Pearson's R?

Answers:

In statistical terms, Pearson's correlation coefficient is the test statistic that measures the statistical relationship between two continuous variables. Because it is based on the method of covariance, it is widely regarded as the best method for determining the degree to which two variables are related. An association's magnitude, or correlation, as well as its direction, can be found here.

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

Answers:

It's possible to standardise the data's independent features by using Feature Scaling. It's used in data pre-processing to deal with wide ranges of magnitudes, values, and unit conversions. A machine learning algorithm tends to weight higher values higher and lower values lower regardless of the unit of the values if feature scaling is not done.

- Min-Max Normalization: Features or observations can be scaled between 0 and 1 by using this technique.
- Standardization: Using this technique, a feature value is rescaled so that it has a distribution with zero mean value and one variance, which is extremely effective.

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

Answers:

To measure how much variance an estimated regression coefficient increases due to collinearity, VIF is the index. When determining VIF, we fit a regression model between the variables. To get an estimate of the coefficient of determination R1, we would fit models like the following:

$$X\_1=C+ \alpha\_2 X\_2+\alpha\_3 X\_3+\cdots$$

$$VIF\_1 =1/(1-R\_12)$$

In the next step, we fit the model between X2 and the other independent variables to estimate R2:

$$X\_2=C+ \alpha\_1 X\_1+\alpha\_3 X\_3+\cdots$$

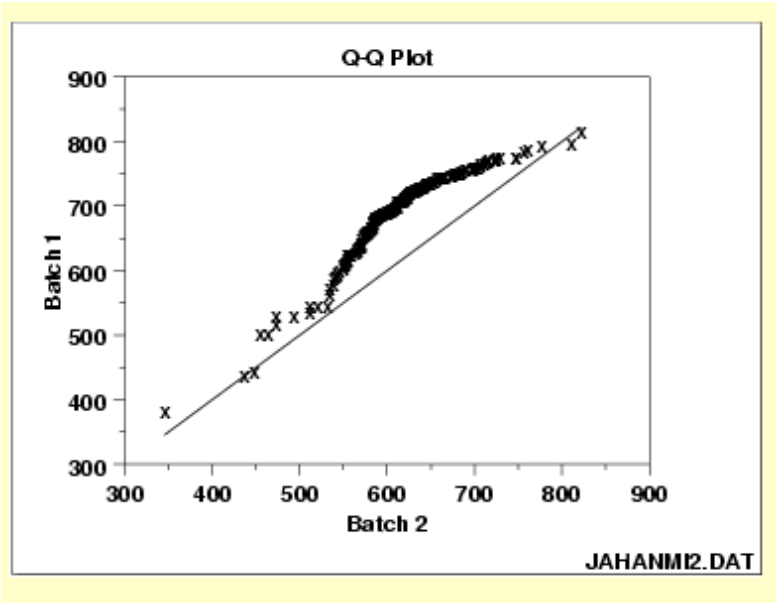
$$VIF\_2=1/(1-R\_22)$$

As long as all the independent variables are orthogonal to each other, the VIF is 1. VIF = infinity if there is a perfect correlation. There is a correlation between variables with a large VIF value. Because of multicollinearity, a VIF of 4 indicates that the model coefficient variance is inflated by a factor of four.

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

Answers:

The quantiles of two distributions are graphically plotted in relation to one another in a Q-Q plot. Plotting quantiles against quantiles is another way of saying this. When interpreting a Q-Q plot, we should focus on the line that reads "y = x" whenever possible. In statistics, it's known as the "45-degree line." This implies that the quantiles in each of our distributions are the same. An outlier in either of the distributions could indicate that one is unbalanced in relation to the other.


Q-Q Plot
900
800
700
600
500
400
300
Batch 1
300
400
500
600
700
800
900
Batch 2
JAHANMI2.DAT