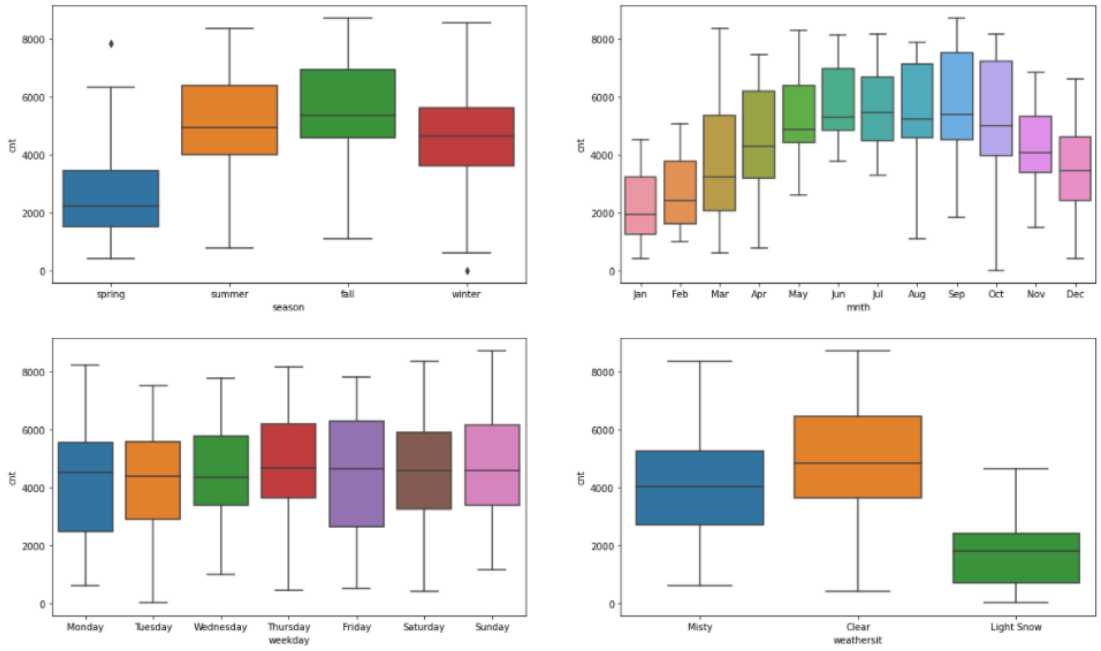
Assignment-based Subjective Questions

Q1) From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Ans: Plotted box plot for categorical variables to understand relationship with target 'cnt' variable



- Interestingly Spring has the least demand of bike rentals whereas Fall has highest demand with median matching demand in summer
- March, April, Sep and Oct have most wide range of bike variables
- We observe that range of demand is highest on Friday and Mondays
- Interestingly the median of bike rentals and is same on all weekdays
- Expectedly the least demand of bikes is in Light Snow weather and the highest on clear days

Thus we infer that weather and season has high impact on target 'cnt' variable. When the weather is clear or it is misty and in Fall and Summer season bike rentals tend to increase.

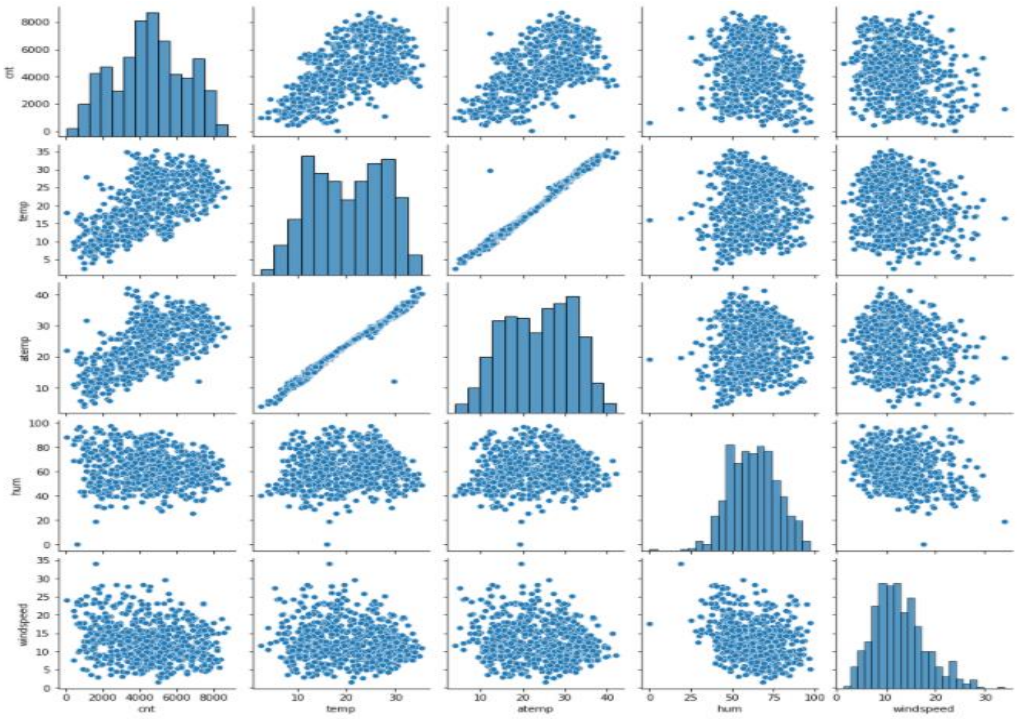
Q2) Why is it important to use *drop_first=True* during dummy variable creation?

Ans: Dummy variables are used to create variables again N number of values of categorical variable. The first variable can be predicted by the remaining variable thus if we do not use ***drop_first=True*** the first variable would be multicollinear with other dummy variables created

from that categorical variable. Hence it is important to use ***drop_first=true*** to address this problem of including a multicollinear variable

Q3) Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

Ans: Below is the pair-plot of numerical variables:

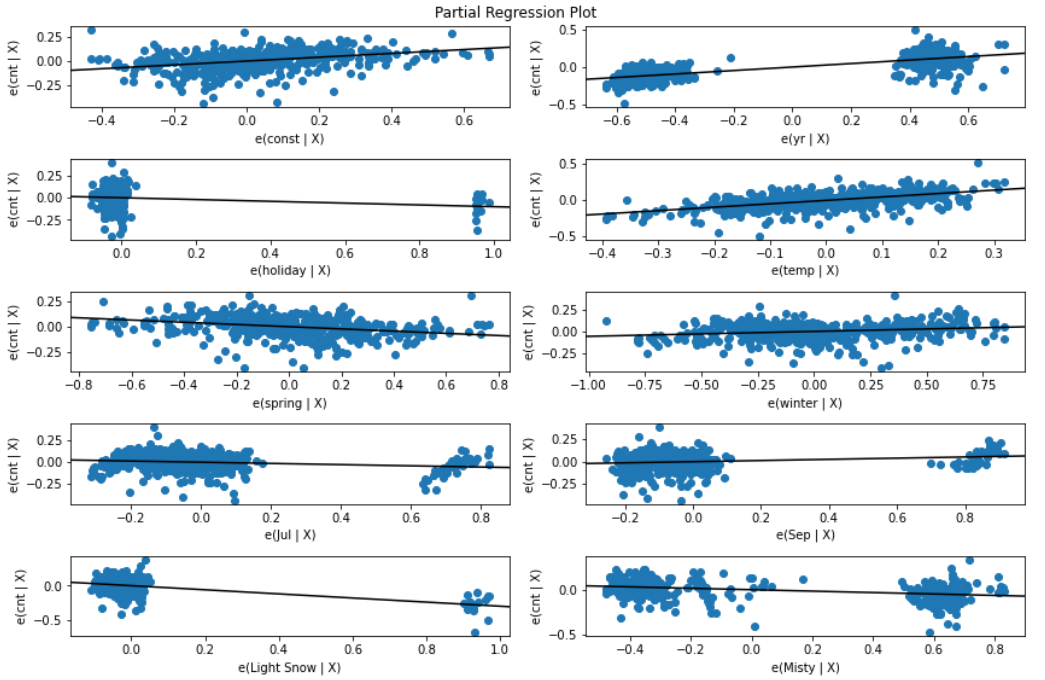


Pair plot above clearly shows that temp is highly collinear with target 'cnt' variable

Q4) How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans: There are three assumptions that I test the final model on training set.
Used Residual plots to check for ***linear relationship*** for numeric variable temp as others are binary.

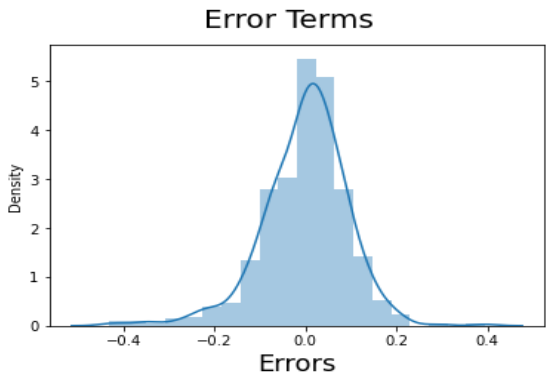
```
fig = plt.figure(figsize=(12,8))
#fig = sm.graphics.plot_regress_exog(LM7, 'temp', fig=fig)
fig = sm.graphics.plot_partregress_grid(LM7, fig=fig)
```



Checked for ***normal distribution of residuals*** by plotting a distplot.

```
fig = plt.figure()
sns.distplot(residual_errors, bins = 20)
fig.suptitle('Error Terms', fontsize = 20)
plt.xlabel('Errors', fontsize = 18)
```

```
Text(0.5, 0, 'Errors')
```

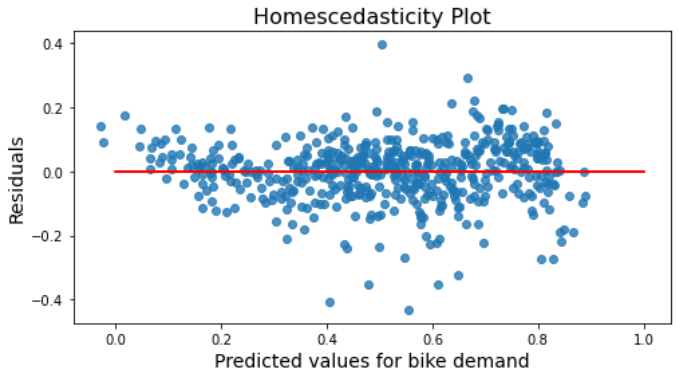


Checked for ***Homoscedasticity*** by using scatter plot and line plot to check if error terms have constant variance

```
plt.subplots(figsize=(8, 4))
plt.scatter(x=y_train_count, y=residual_errors, alpha=0.8)

plt.plot(y_train,(y_train - y_train), '-r')

plt.ylabel('Residuals', fontsize=14)
plt.xlabel('Predicted values for bike demand', fontsize=14)
plt.title('Homescedasticity Plot', fontsize=16)
plt.show()
```



Q5) Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Ans: The top 3 predictors/features that the final model has provided us, which are significant in predicting bike rentals are:

- Temperature that has coefficient of 0.4712 which indicates that it is most important in predicting bike rentals
- Year has positive coefficient which illustrates that with each year bike rentals would grow
- Negative coefficient of Weather condition of Light Snow and Light Rain with Scattered clouds of -0.3002 shows that the bike rentals decline during this weather

General Subjective Questions

Q1) Explain the linear regression algorithm in detail.

Ans: Linear regression algorithm falls under the category of supervised learning methods. The primary utility to check/predict linear relationship between target/dependent variable with independent variables. Dependent variable is the predictor that needs to be predicted and independent variables are leveraged to predict dependent variable's value.

As the name 'Linear' suggests that variables depicted on X and Y-axis are correlated linearly. The linear regression equation in terms of Mathematics is written by following equation:

$Y = \beta_0 + \beta_1 X$, where β_0 is the slope and β_1 is the intercept

β_0 the slope is given by following formula:

$$\beta_0 = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

$$\beta_1 = \frac{\sum y - b(\sum x)}{n}$$

Where

x and y are the variables for which we will make the regression line.

- β_0 = Slope of the line.
- β_1 = Y-intercept of the line.
- x = Values of the first data set.
- y = Values of the second data set.

This equation would help to find best fitting line also known as regression line. This is the concept used by linear regression model where we use regression line to show predicted score y for each possible value of variable x.

Q2) Explain the Anscombe's quartet in detail.

Ans: To demonstrate importance of data visualization, statistician Frances Anscombe developed Anscombe's quartet to help understand data by plotting it before analyzing the data. It is composed on 4 datasets where each dataset consists of 11 x and y points. The uniqueness of these datasets is that when we see descriptive characteristics, they appear nearly same however when they are plotted using scatter plot these dataset appear very different.

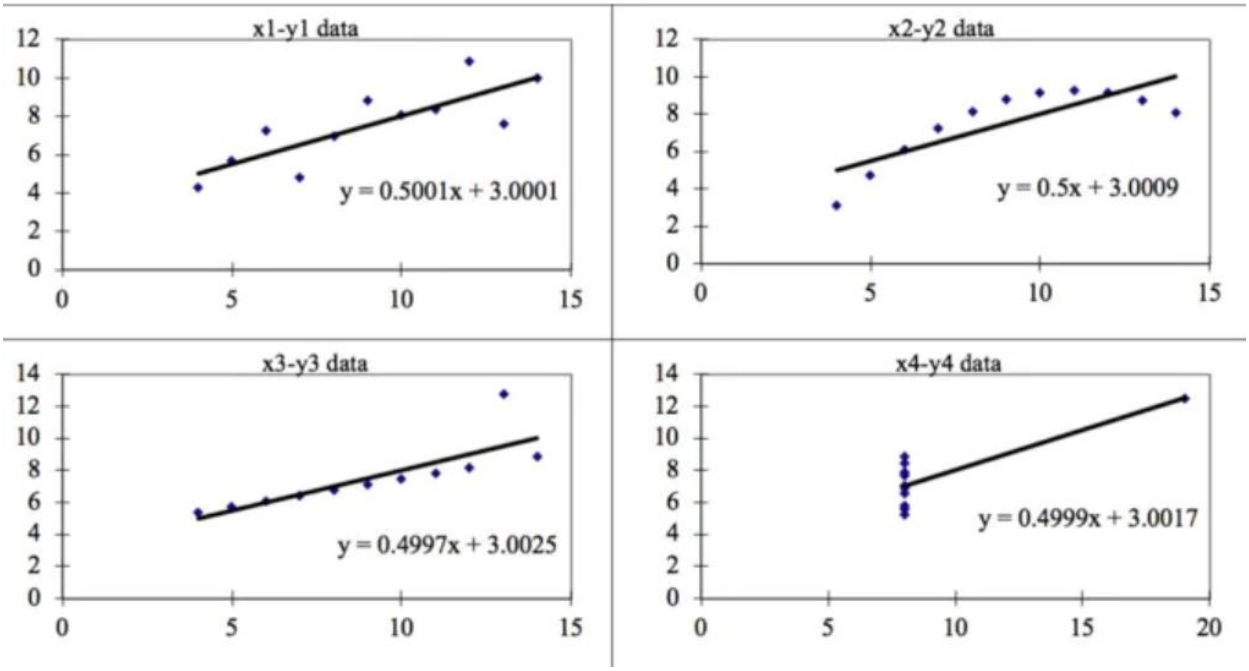
The dataset is given below:

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

When we see descriptive characteristics of above datasets we see they appear similar

				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	

However when we plot these datasets, we get following visuals



They clearly show importance of visualizing data despite checking descriptive information of datasets.

Q3) What is Pearson's R?

Ans: For given any two variables, to measure linear relationship between them Pearson's R is used, as it provides numerical strength summary of linearity between these two variables. Pearson's R is also referred to as Pearson's correlation coefficient or bivariate correlation. Pearson's R cannot represent the data adequately if the relationship between variables is not linear.

Greek letter ρ is used to represent Pearson's R and for any given pair of variable (A,B) the formula to calculate ρ is:

$$\rho A,B = \frac{cov(A,B)}{\sigma A \sigma B}$$

Where:

- cov is the covariance
- σA is the standard deviation of A
- σB is the standard deviation of B

Pearson's R varies between -1 and +1 where R value indicates following:

R = -1 means perfect negative linearity between variables

R = 0 means no linearity between variables

R = 1 means there is perfect positive linearity between variables

They are depicted in scatter plots below:

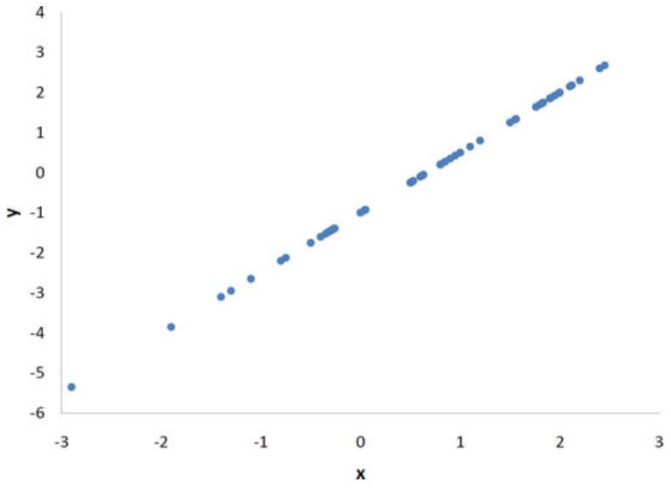


Figure 1. A perfect positive linear relationship, r = 1.

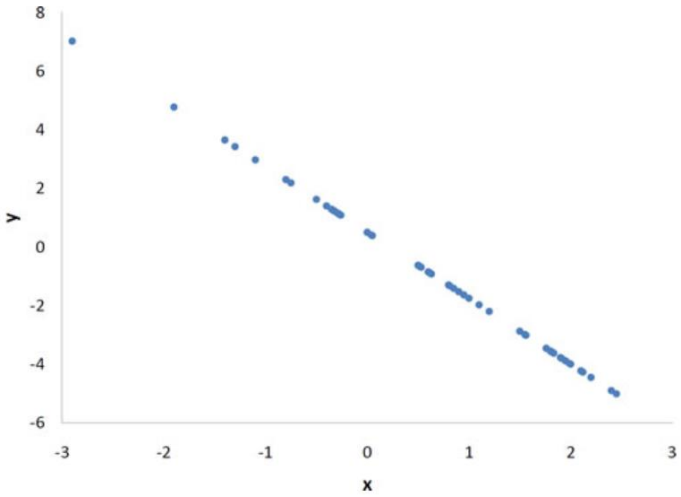
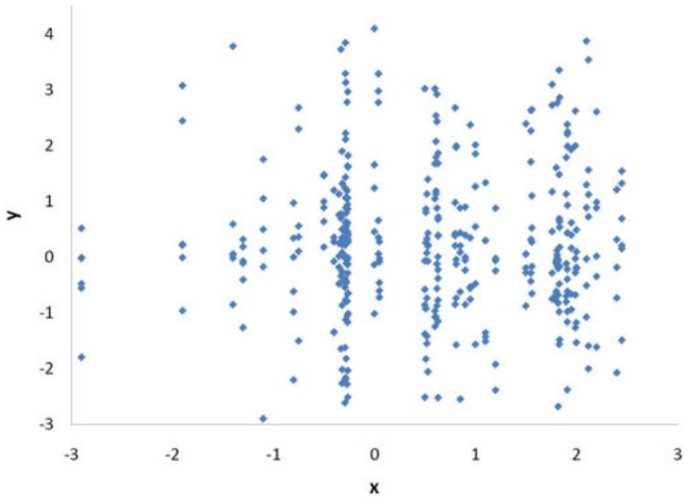


Figure 2. A perfect negative linear relationship, r = -1.



Q4) What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans: In a given datasets variables may be of several types and units. Moreover, range of feature values may differ drastically leading to features with higher values take over features with lower values. It may lead ML models to tend towards features with higher values. To overcome this feature scaling is used.

Feature Scaling allows users to get the features to be on nearly same range of values thus allowing all features to be easily used and process by ML models.

Two widely used types of scaling are Standardized and Normalized scaling. In terms of difference between the two: The standardized scaling minimum and maximum value of the feature is used, where as in normalized scaling mean and standard deviation is used. The scale values of standardized scaling is [0-1] or [-1 to 1] but of normalized scaling it is not bound to any specific range. Normalized scaling is minimally affected by outliers where standardized scaling can get impacted by outliers.

Q5) You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Ans: VIF (Variance Inflation Factor) is used to determine the extent to which two predictors/variables are correlated to each other. Hence VIF is used for understanding multicollinearity between variables. VIF of infinity indicates that there is perfect correlation between two variables that are independent. VIF is given by equation:

VIF = $\frac{1}{1-R^2}$

Where R^2 is R-squared value

If R^2 is 1 that would lead VIF to be infinite.

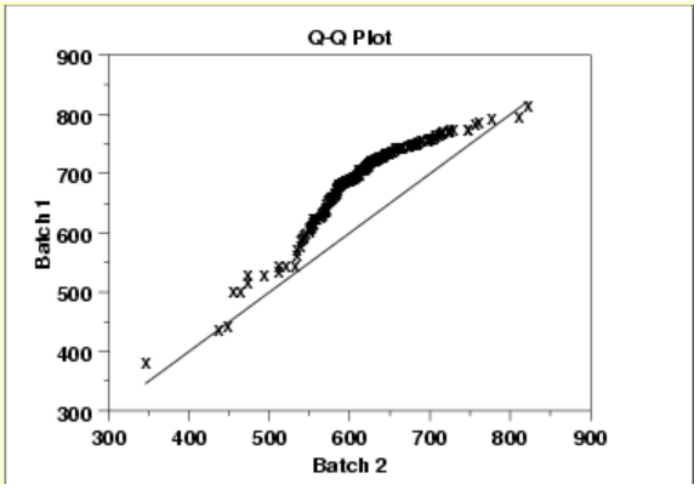
This would mean that there is some predictor/variable that has same correlation with all other variables, hence if we get VIF as infinite we can start with dropping that variable and then check the model

Q6) What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Ans: Q-Q plot also known as Quantile-Quantile plots is graphical way to plot quantiles of two distribution with respect to each other. It is often used to assess the normality in the linear regression model. By quantiles we understand that specified percentage of data below given values. For instance 25 percentile is point where 25 percent of data fall below the value and 75% fall above.

Along with this there is a 45 degree line that is reference line which is also plotted. This is to find out if two variables come from similar distribution they will fall on that line.

Example of a Q-Q plot :



It helps in finding out if two data sets have similar distribution shape or they have common scale. The importance of Q-Q plot in linear regression is that it helps to visualize and check whether the given data meets the assumptions of linear regression such as homoscedasticity and normality.