Assignment based subjective Questions

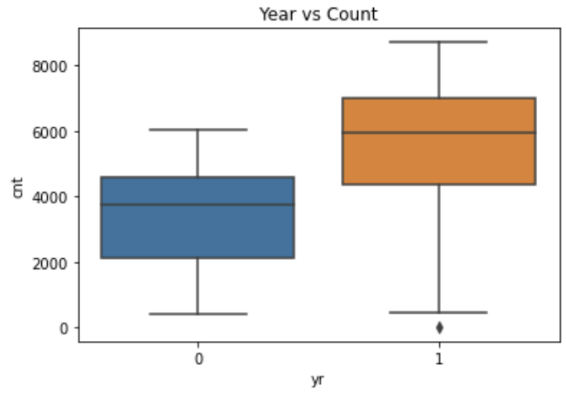
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

1)year

Year has two values in data set 2018 and 2019. We can see a significant increase in the demand from 2018 to 2019.

```
sns.boxplot(data=df,x='yr',y='cnt')
plt.title("Year vs Count")
```

```
Text(0.5, 1.0, 'Year vs Count')
```



2)Weather situation

Weather situation has three values in data set.

Weather_good :Clear, Few clouds, Partly cloudy, Partly cloudy

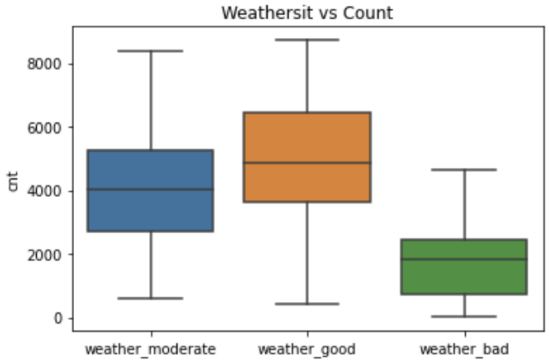
Weather_moderater:Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist

weather_bad:Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds

We can say from plot that when when weather is bad demand is very less and when weather is good we can see increase in demand.

```
sns.boxplot(data=df,x='weathersit',y='cnt')
plt.title("Weathersit vs Count")
```

```
Text(0.5, 1.0, 'Weathersit vs Count')
```

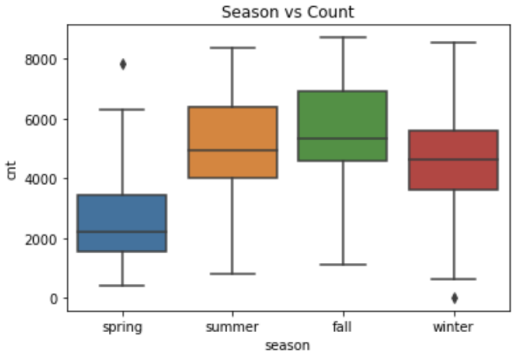


3)Season

Wecan see demand during spring is very less and demand during the seasons summer and fall is increased and slight decrease in winter season.

```
sns.boxplot(data=df,x='season',y='cnt')
plt.title("Season vs Count")
```

```
Text(0.5, 1.0, 'Season vs Count')
```

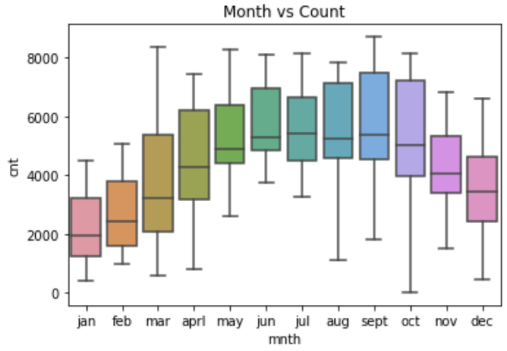


4)Month

Demand from january to may increase and remained same till october and decreased during november and december.

```
sns.boxplot(data=df,x='mnth',y='cnt')
plt.title("Month vs Count")
```

```
Text(0.5, 1.0, 'Month vs Count')
```



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

Let's take an example and understand

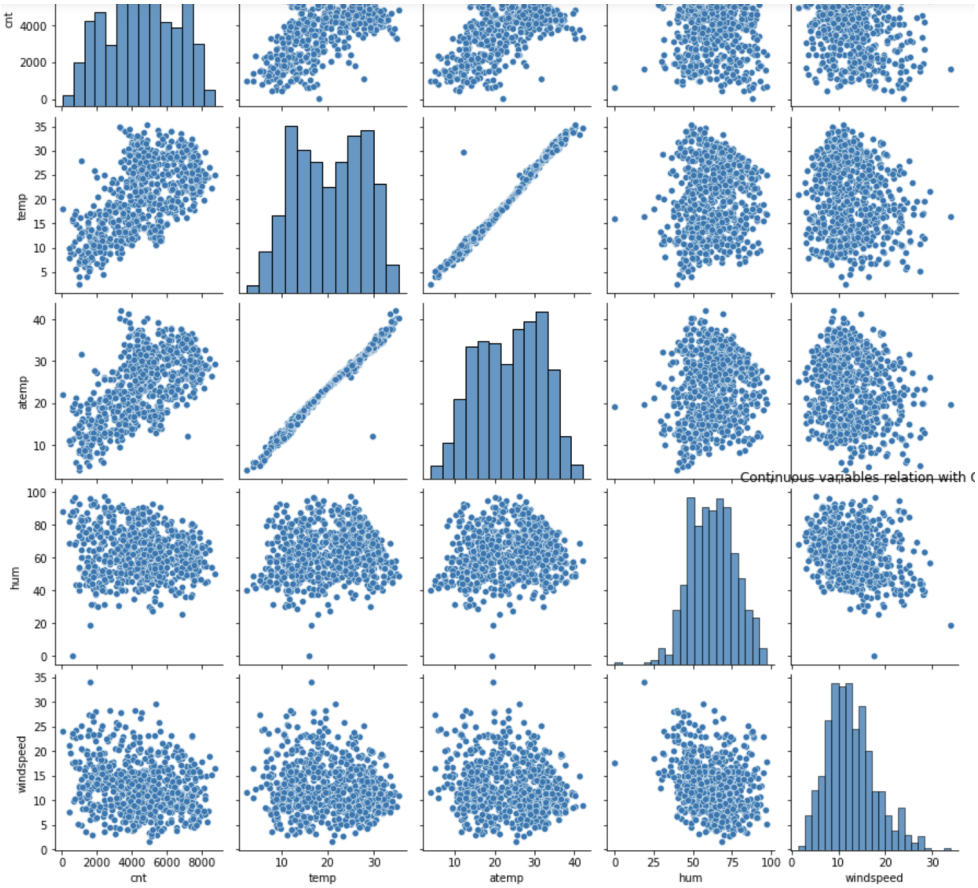
df1=pd.get_dummies(df['season'])

Fall	Spring	Summer	winter
0	1	0	0
0	0	1	0
1	0	0	0
0	0	0	1

Here winter can be explained by 0001. But it can be interpreted with the help of three variables only. If fall, spring and summer are 000 then it is obviously winter. Hence we don't need fourth variable and reduces the correlation among the dummy variables.
Hence we use f1=pd.get_dummies(df['season'],drop_first=True).

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

From looking at the pair plots ‘temp’ and ‘atemp’ has highest correlation among the continuous variables with the target variable.

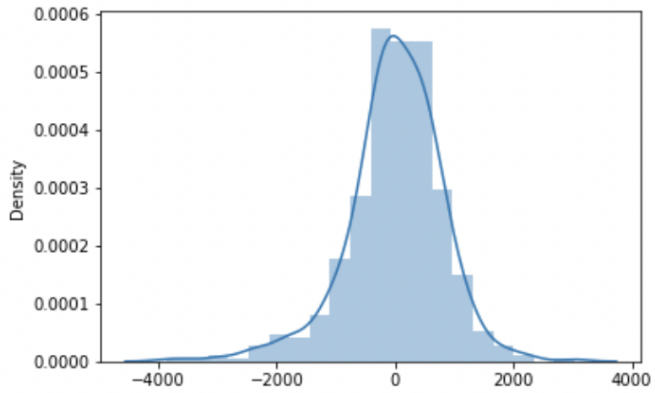


4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Normality of errors:
The residuals must be normally distributed on the training set
Distribution of residuals form my code

```
sns.distplot(y_train - y_train_pred,bins=20)
fig=plt.figure()
fig.suptitle('error terms')
```

```
Text(0.5, 0.98, 'error terms')
```



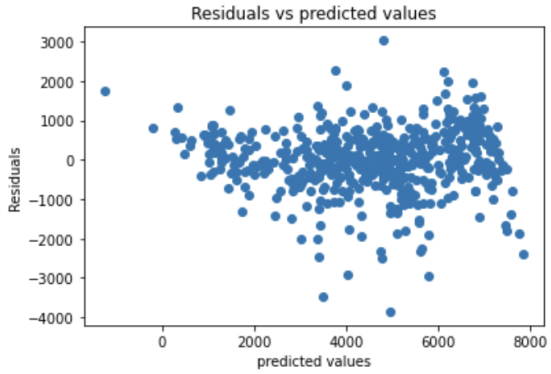

```
<Figure size 432x288 with 0 Axes>
```

Independence of errors:
There should not be relationship between residuals(errors) and y variable

Homoscedasticity:
There should be constant variance of residuals(errors)

```
plt.scatter(y_train_pred,y_train-y_train_pred)
plt.xlabel("predicted values")
plt.ylabel("Residuals")
plt.title("Residuals vs predicted values")
```

```
Text(0.5, 1.0, 'Residuals vs predicted values')
```



We can see there is no clear pattern

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

My model
Y =
1922.91+yr*2046.12+workingday*386.10+atemp*4083.59-windspeed*1005.40-spring*1135.28+winter*430.62-jul*507.48-nov*329.24+sat*546.24-weather_bad*2527.40-weather_moderate*678.09

The three feature influence demand of shared bikes are following
1)atemp(coefficient +4083.59)
2)year(coefficient 2046.12)
3)weathersit_bad(coefficient -2527.40)(Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds)

Subjective Questions

1. Explain the linear regression algorithm in detail.

Ans)It shows the relationship between the continuous variables i.e between the dependent variable (Y) and one or more the dependent variable (X). Linear regression one of the supervised machine learning algorithm. It performs regression task. Regression is a line or curve that passes through all data points such that vertical distance between data points and line is minimum.
Linear regression makes predictions for continuous/real or numeric variables such as sales, salary, age, product price, etc.

y=mx+c
y=a0+a1x
a0=intercept of a line
a1=coefficient of a line(slope)

Simple Linear regression:
If a single independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Simple Linear Regression.
Multiple Linear regression:
If more than one independent variable is used to predict the value of a numerical dependent variable, then such a Linear Regression algorithm is called Multiple Linear Regression.

Positive relationship
If dependent variable increases as independent variable increases
Negative relationship

If dependent variable decreases as independent variable increases

When working with linear regression, our main goal is to find the best fit line. This means the distance between predicted values and actual values should be minimized.
The different values for the coefficient of lines (a0, a1) gives a different line of regression, so we need to calculate the best values for a0 and a1 to find the best fit line, so to calculate this we use cost function.

Cost function
Cost function optimizes the regression coefficients.
we use the Mean Squared Error (MSE) cost function
It is the average of squared error occurred between the predicted values and actual values.

MSE=1/N * $\sum i = 1\ i = n$ (yi-(a1xi*a0))^2

Gradient descent:
Gradient descent is used to minimize the MSE by calculating the gradient of the cost function.A regression model uses gradient descent to update the coefficients of the line by reducing the cost function.It is done by a random selection of values of coefficient and then iteratively update the values to reach the minimum cost function.

2. Explain the Anscombe's quartet in detail.

Anscombe's quartet comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties.

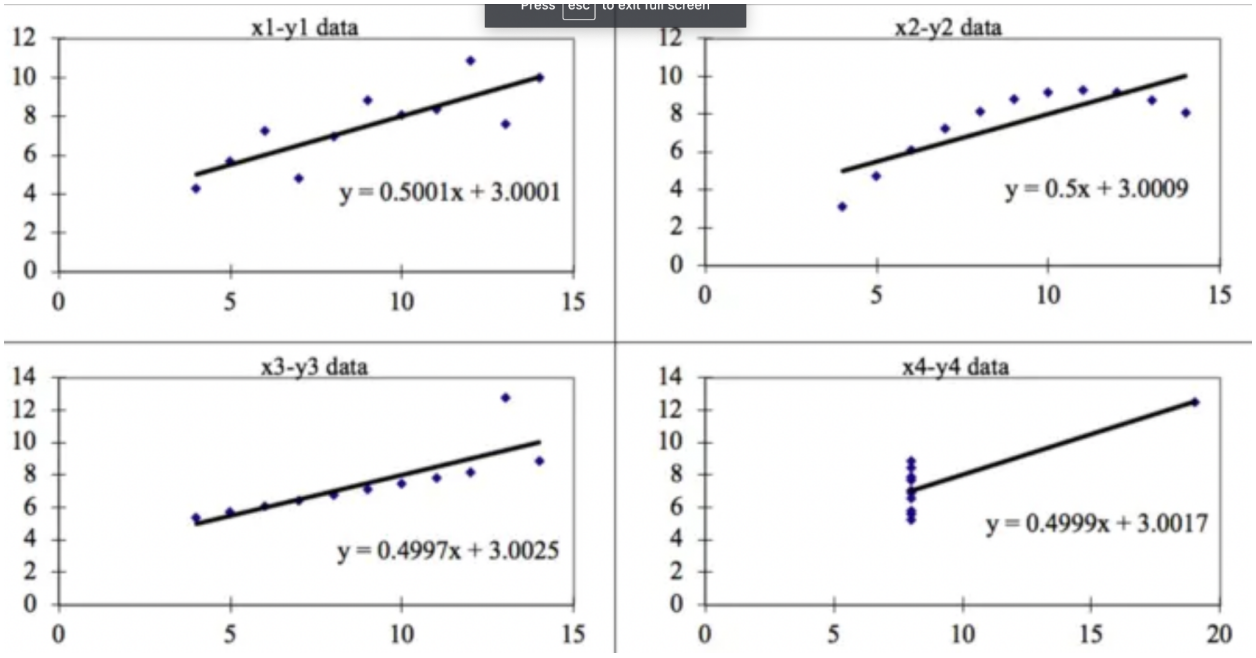
These four data sets have nearly same statistical observation like mean and variance. This tell us the importance of visualizing the data before it can be used for models.

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

The statistical information for all sets can seen 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
				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	

Graphs plotted below

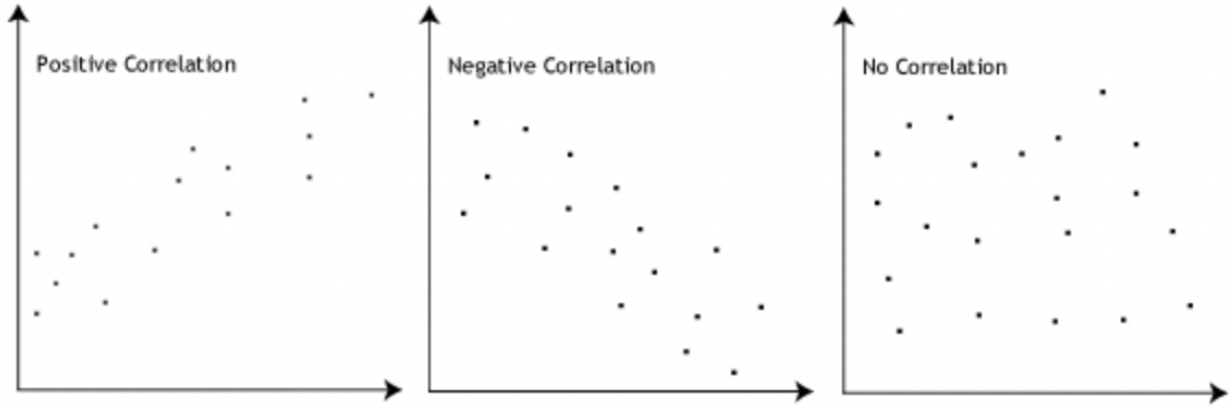


In first data set linear regression fits well. In second data set linear regression does not fit well and data is non linear. Third and fourth data set has outliers and cannot be handled by regression model.

3)What is Pearson's R?

Ans)Correlation coefficients are used to find how strong a relationship is between two continuous data. It is the covariance of two variables, divided by the product of their standard deviations.The Pearson's correlation coefficient formulas return a value between -1 and 1.

1:indicates a strong positive relationship.As one variable increases another one also increases.
-1:indicates a strong negative relationship.As one variable increases another one decreases.
A result of zero indicates no relationship at all.



Pearson's formula as follows

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r= correlation coefficient
xi= values of x variable in sample
yi=values of y variable
xbar=mean of x
ybar=mean of y

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

Scaling:
Scaling is used to normalize the independent variables or features of a data in a fixed range. It is performed during the data preprocessing step to handle highly varying values. If feature scaling is not done, model consider greater values as higher and smaller values as lower, regardless of the units. Some machine learning algorithms based on the mapping distance between data points.

Normalized scaling:
Normalized scaling technique in which values are shifted and rescaled so that they end up between 0 and 1.It is also known as min-max scaling.

x_new=(x-xmin)/(x_max-x_min)

X_min = minimum value of a feature
X_max = maximum value of a feature

Normalization scaling is useful when there are no outliers in the data set.
This transformation squishes the n-dimensional data into an n-dimensional unit hypercube.
It is useful when we don't know about the distribution.
Minimum and maximum values are used for scaling.

Standardized scaling:
Standardization is another scaling technique where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.

x_new= (x - mean)/std

x=data point
mean= mean of the variable
std= standard deviation

Standardized scaling is used when the distribution follows a Gaussian distribution.
It translates the data to the mean vector of original data to the origin and squishes or expands.
Mean and standard deviation used for scaling.
It is less affected by the outliers.

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

VIF(variance influence factor):
VIF is a measures the amount of multicollinearity in a set of multiple regression variables. VIF for a regression model variable is equal to the ratio of the overall model variance to the variance of a model that includes only that single independent variable.This ratio is calculated for each independent variable. A high VIF indicates that the associated independent variable is highly collinear with the other variables in the model.
Detecting multicollinearity is important because while multicollinearity does not reduce the explanatory power of the model, it does reduce the statistical significance of the independent variables.

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

A large variance inflation factor (VIF) on an independent variable indicates a highly collinear relationship to the other variables that should be considered or adjusted for in the structure of the model and selection of independent variables.
When vif is infinity , it means perfect correlation between two independent variables. This means r2=1 ,to solve this problem you have to drop one of the variable which is causing multicollinearity.

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

Q-Q plots are also known as Quantile-Quantile plots. As the name suggests, they plot the quantiles of a sample distribution against quantiles of a theoretical distribution.Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.

This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.

A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight. Here's an example of a Normal Q-Q plot when both sets of quantiles truly come from Normal distributions.