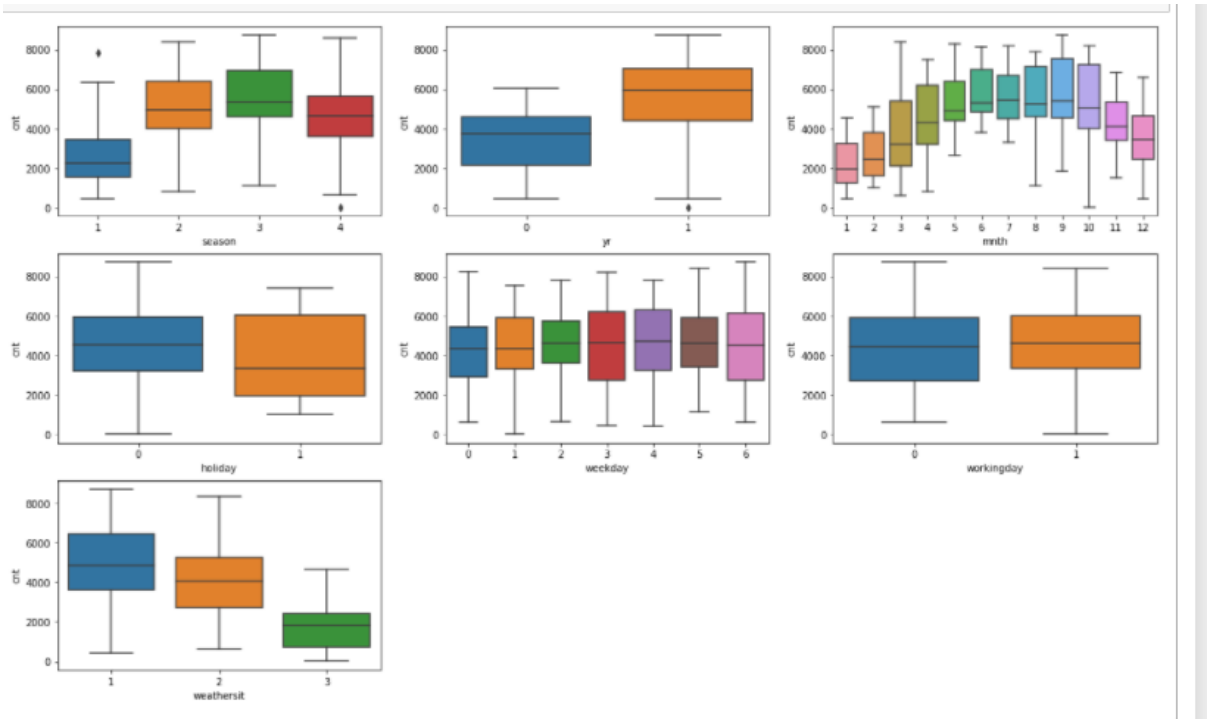
Assignment-based Subjective Questions

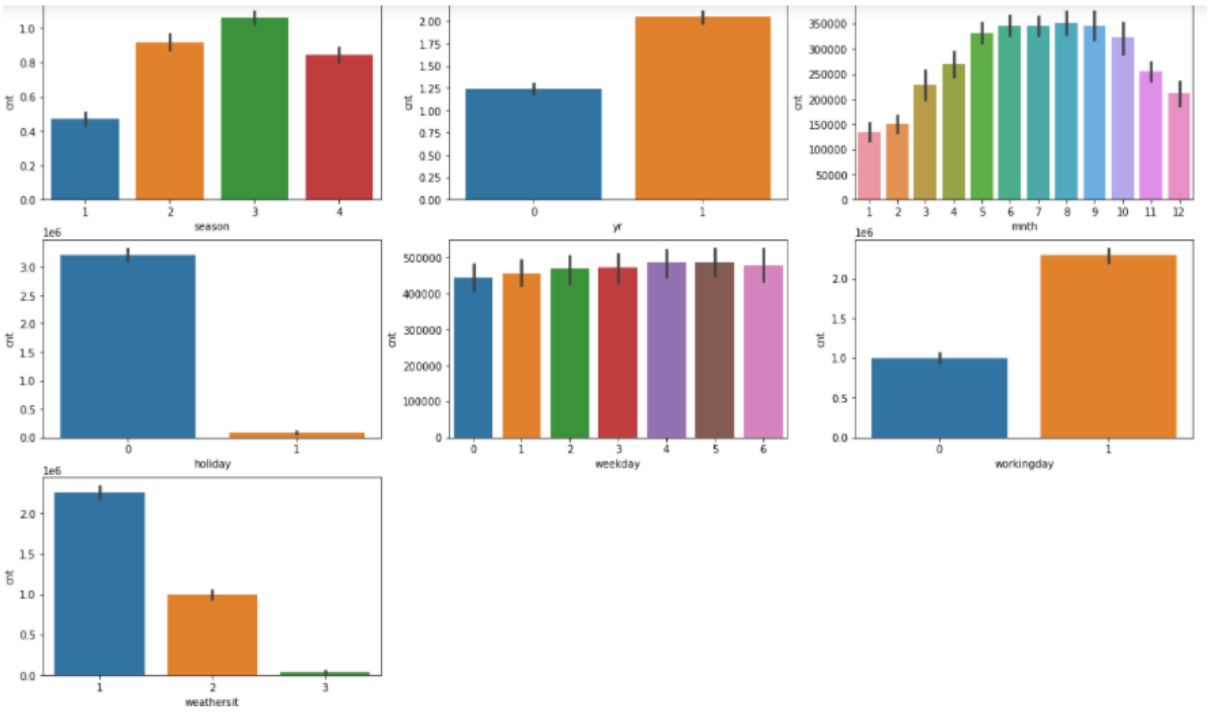
1. ***From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?*** ***(3 Marks)***

We can see the boxplot as well as the bar plot on how the different categorical variables are affecting the dependant variable 'cnt'

Box Plot:



Bar Plot:



From the above diagrams we can see that in the categorical variables section the driver variables are:

a. **Weathersit** - That is for value 1 or 2 i.e., when it is clear or mist then the number of users are high which is correct because in light or heavy rains nobody will prefer bikes, they will mostly prefer four wheelers
b. **Holiday and Working Day** – Here it seems that the persons who are working daily , they are utilizing this service/app most probably for their work purposes as in when there is no holiday and it seems to be a working day , the number of users is high
c. **Month -** April,May June,July, August is where the app is having higher number of footfall
d. **Season -** It seems during the fall the number of users are high compared to all the other seasons
e. **Year** - May be because as year on year, there is more advertisement and people come to know about the service hence more people are using the service

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

Here it is important to do as suppose we can avoid an extra variable. Let's for take as example season variable if we don't use drop_first=True and if we use the difference is given as below. Now since it refers to same feature, it can help reducing correlation as well as the VIF factor. Hence important in doing so.

```
pd.get_dummies(df['season'])
```

	1	2	3	4
0	1	0	0	0
1	1	0	0	0
2	1	0	0	0
3	1	0	0	0
4	1	0	0	0
...	...	...	...	...
725	1	0	0	0
726	1	0	0	0
727	1	0	0	0
728	1	0	0	0
729	1	0	0	0

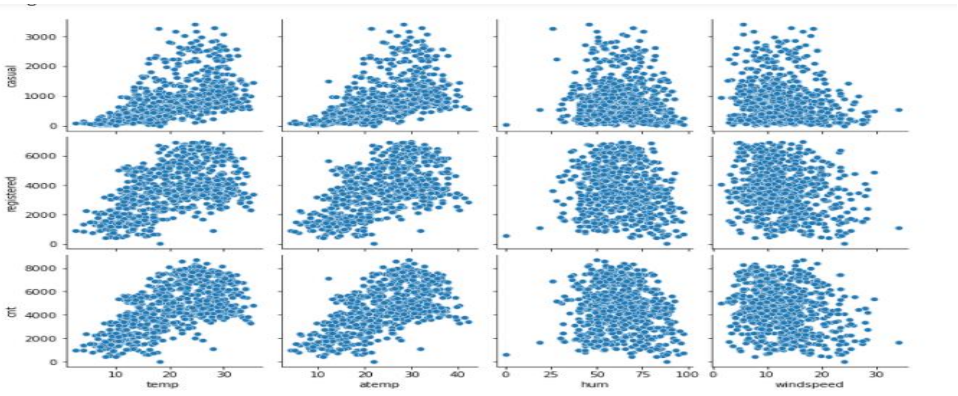
730 rows × 4 columns

```
pd.get_dummies(df['season'], drop_first = True)
```

	2	3	4
0	0	0	0
1	0	0	0
2	0	0	0
3	0	0	0
4	0	0	0
...	...	...	...
725	0	0	0
726	0	0	0
727	0	0	0
728	0	0	0
729	0	0	0

730 rows × 3 columns

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

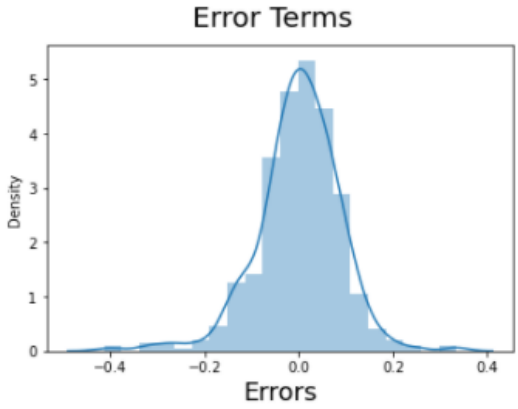


Looking at the above plot, 'temp' and 'atemp' is having the highest correlation, if we plot the correlation then 'atemp' has the highest correlation with value of 0.63

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

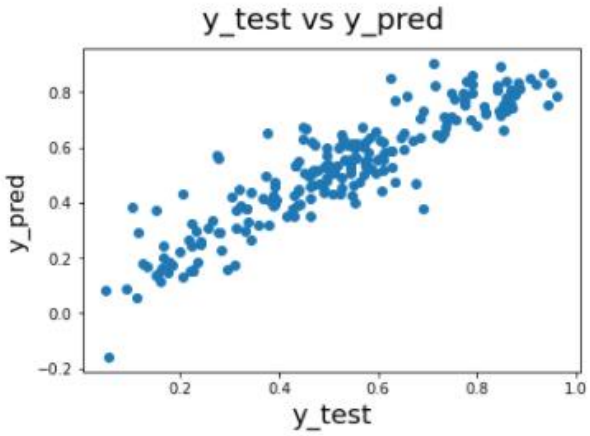
We validated the assumptions of linear regression by

a. Plotting the error terms and then validating whether it is normally distributed or not



Here the error terms seems to be normally distributed thus validating our assumption

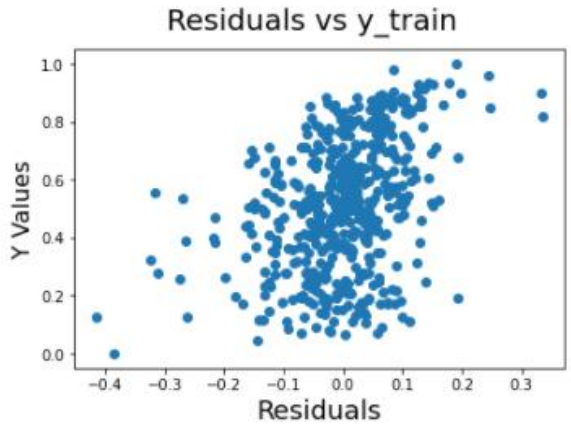
b. Proving the linearity of the model by plotting actual vs predicted values



c. Proving there is no multi collinearity by looking at the VIF

	Features	VIF
0	const	53.47
3	workingday	1.88
5	hum	1.88
12	weekday_6	1.79
8	season_4	1.72
4	temp	1.59
13	weathersit_2	1.57
11	mnth_10	1.49
9	mnth_8	1.46
7	season_2	1.38
14	weathersit_3	1.25
10	mnth_9	1.24
6	windspeed	1.19
2	holiday	1.16
1	yr	1.03

d. Calculating the residuals and plotting it so that there is no pattern among them



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

Based on the final model the top 3 features contributing significantly towards explaining the demand of shared bikes are

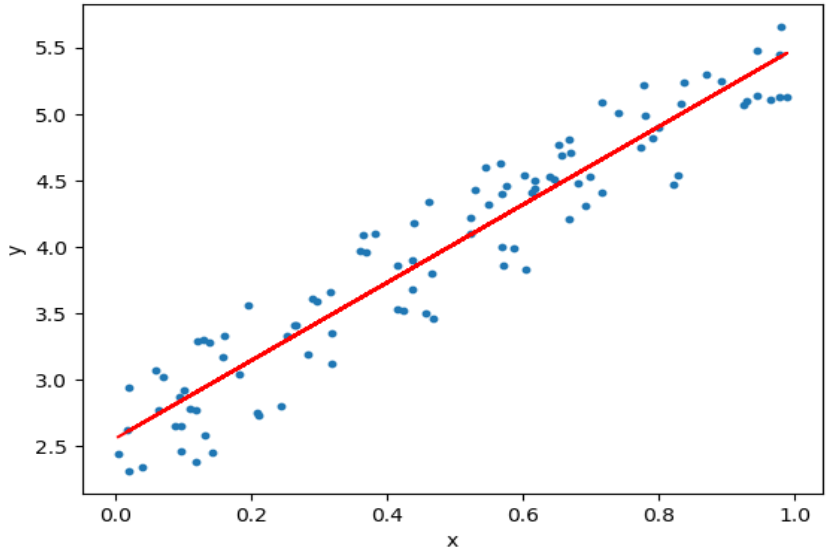
a. 'atemp': The feel temperature greatly affects the demand of bikes and it has a positive correlation, the higher the temp more the demand for shared bikes. Hence it has a higher positive correlation
b. 'year': Based on yearly growth the demand for bikes increases means the penetration of the app increases year on year hence the variable. Has a positive correlation
c. 'weathersit ': This as a negative correlation based on how we see it , it is a categorical variable, so when it is clear or misty we have higher demand , which is also logical as when there is light/heavy rain user prefer four-wheeler
d. 'mnth_9' or fall season : this is a dummy variable with drop_first so month 8 or August has the highest demand in shared bikes or also we can say that in the fall season it has higher demand

The above columns were based on the coefficients of our linear model

General Subjective Questions

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

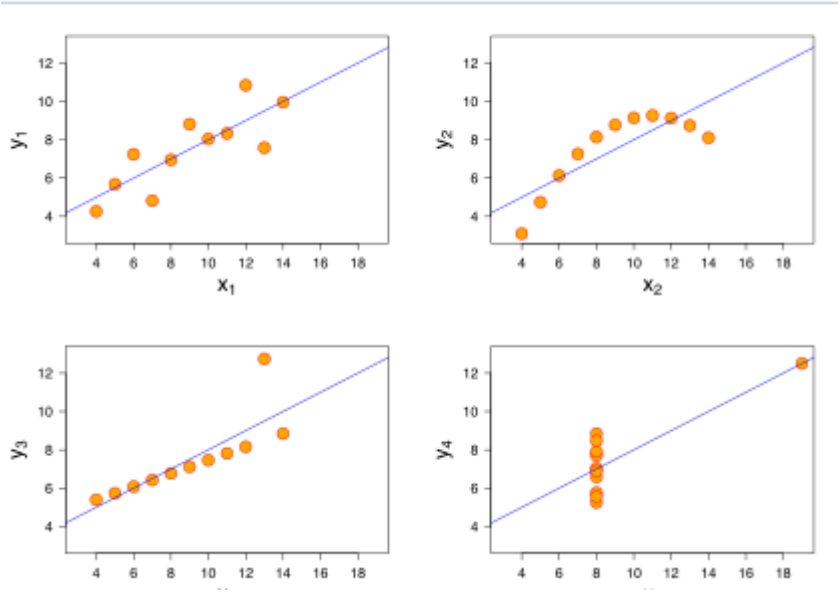
Machine Learning is the process in which we teach machines based on patterns of data. Here the machine can predict like as a classification or a continuous variable. Here if we are predicting a continuous variable it is known as regression. Now in simple terms Linear Regression is a model using which we can predict a continuous variable in which if it is simple linear regression we fit a normal line and if it is multiple linear regression we fit a plane based on the number of variables in that equation.
Linear Regression is a form of supervised learning as we already have the output so we a plot a line between x independent variable and y dependent variable to predict the output. Here x, the independent variable, is also known as the predictor variable and the dependant variable is called the output variable.
For a real-time example we are predicting the prices of house based on a single variable suppose area so we can plot area in x-axis and y as price of the house in the y-axis and plot a line and then predict when we have this much area how much is the cost like:



Now single variables almost is impossible in the real world , so we have multiple predictor variables and then we predict the single output variable. For that we fit a plane in the multiple dimensions. Now for Linear Regression with multiple predictor variables to succeed we are assuming many things that are:

a. There is a linear model or a line or a plane that can be plotted between the predictor and the output variables
b. The Error terms are normally distributed: this means that the error should not have minimum errors
c. No patterns should emerge in Error Terms
d. Error terms should have constant variance as it should not change for different error terms i.e., homoscedasticity

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



Anscombe's quartet is a set of four datasets that have similar mean values .i.e., statistically speaking the four datasets are similar, but then when graphed it gives very different results like the image given above. The main aim of this was to shown how outliers in a graph can heavily distort the

statistical data. So before proceeding with linear regression it is very important to remove outliers in our data. A very common real life example is wealth inequality, if we take the mean of income and if we are having a rich income family and rest all from poor families which represents the Indian society we will be getting a skewed data as the rich family is increasing the mean income , but others/majority is poor. Thus the data is giving us a much distorted view.

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

Pearons' R is a measure of linear correlation between two sets of data. Here the speciality of this is that it shows the relationship as well as nicely explains the variance among its values with respect to each other. The formulae is given by the ratio of the covariance of two variables and the product of their standard deviation. Hence the value will always be between -1 and 1. Hence we got a standardized way on to find the correlation among the output and the predictor variables

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i-1}^{n}(y_i - \bar{y})^2}} \quad \text{(Eq.3)}$$

where:

- n is sample size
- x_i, y_i are the individual sample points indexed with i
- $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (the sample mean); and analogously for $\bar{y}$

Pearson's R is very useful in linear regression to find the correlations among variables and establishing a plane between predictor variables and output variable

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

Scaling means to scale a variable to particular limits. For example we have a predictor variable of x from 1-1000 and output variable 1-10 and suppose we have other predictor variables around the range of 0-10 then at this instant the coefficients of variable x might be less so that it impacts less as a simple mathematical equation of reducing the value. So scaling helps to establish a common scale for all predictor variables to be put against and know the best value.

Scaling is performed for the following reasons:

a. Ease of Understanding – That is to put common predictor variables against a common scale so that it will be easier for us to compare the correlations between variables
b. Faster Convergence of Gradient descent Method – Here we can arrive at the coefficients faster when it is scaled so we get a better performance

Difference between normalized scaling and Standardized scaling is:

Normalized Scaling	**Standardized Scaling**
Scaling occurs with minimum and maximum value of the column of the dataset	Scaling occurs with mean and the standard deviation of the column of the dataset
Range is between 0 to 1	Since this is centred around mean and standard deviation, no specific range

Formulae: $X' = \frac{X - X_{min}}{X_{max} - X_{min}}$	Formulae: $X' = \frac{X - \mu}{\sigma}$

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

VIF or Variance Inflation Factor is defined as the amount collinearity among the predictor variables of the liner regression module. It is given by

The VIF is given by:

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

where 'i' refers to the i-th variable which is being represented as a linear combination of rest of the independent variables. You'll see VIF in action during the Python demonstration on multiple linear regression.

Here sometimes it can become infinity if R is 1 that means that there exists a variable that is exactly represented by other variables hence rendering the current variable useless. Which means the impact of this particular variable on the dependant variable is already explained by other variables. A simple example would be like suppose if season is taken and we are also deriving quadrant from the date, it might be highly correlated since it is similar value when we create dummy and normalize these two variables. Thus the impact is already explained by season factor so we don't need quarter variable which will have infinite VIF value.

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

Q-Q (Quantile - Quantile) Plot is a graphical plot that is used to plot a statistical distribution. We can come to know if it has a normal/exponential or uniform distribution. If suppose we have collected different samples from different populations and then we can compare it by using the Q-Q plot and understand whether the samples have similar distribution or not. So we are equipped to tell whether the sampling is good or not.

It has great importance in linear regression in the sense that if we have many samples we can determine whether they are both coming from the same population or not, like if we have taken bike sharing linear regression from one location , how is its distribution different from another location , whether the same linear model can be used or not. We can determine this using the Q-Q plot.