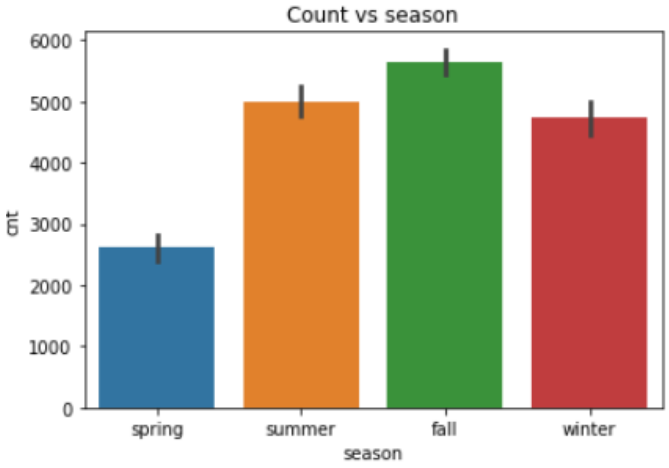
Assignment-based Subjective Questions

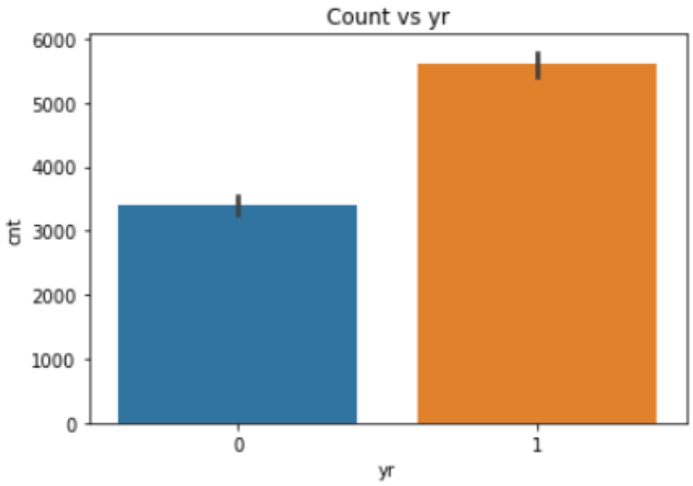
Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Ans1: We are using bar-graph to identify the relationship b/w categorical values and total users:

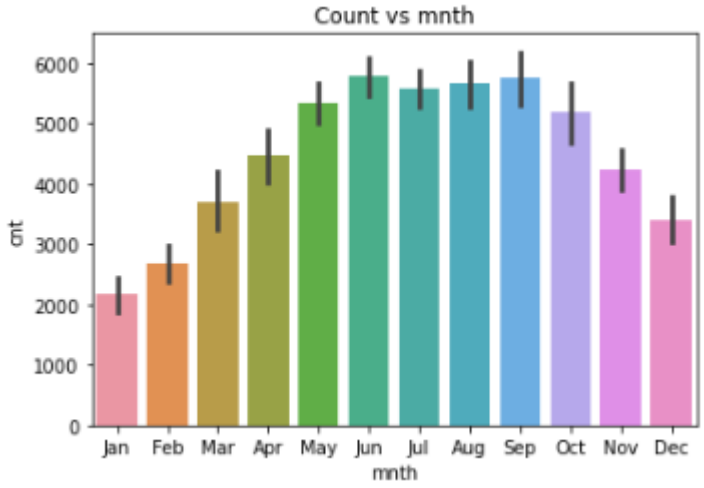
1. Count vs Season: Here we can see, user increased in summer and fall season and low in winter and spring season comparatively.



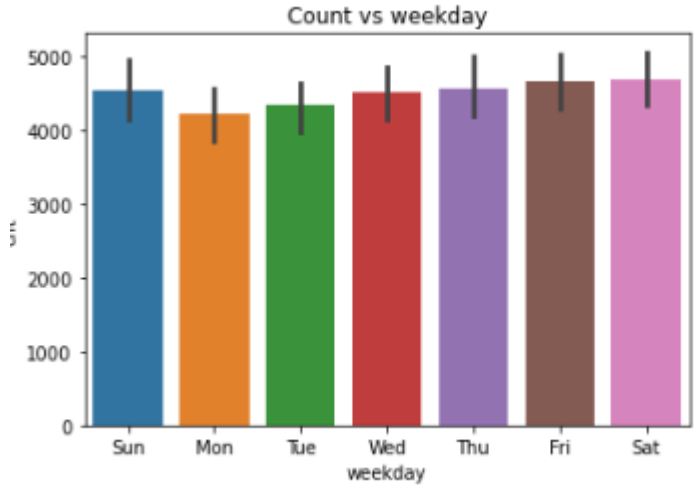
2. Count vs Year: Significantly users grown as compare to last year.



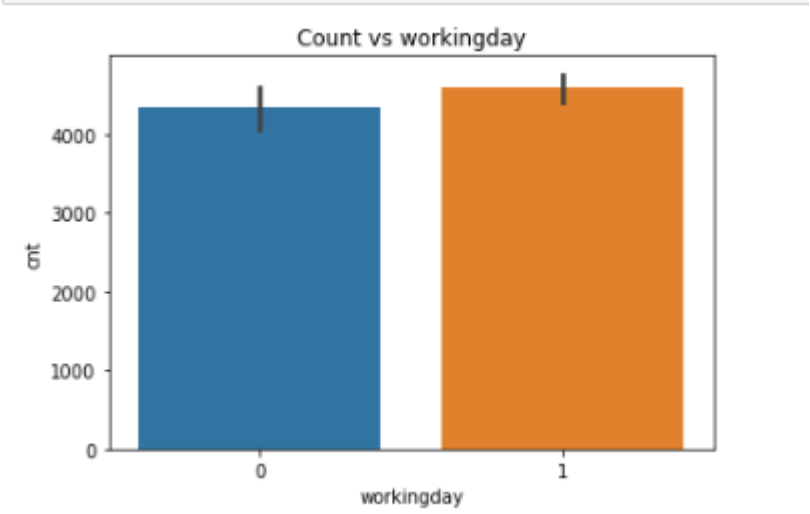
3. Count vs Month: Based on below graph we can see count is changing as per month changed:



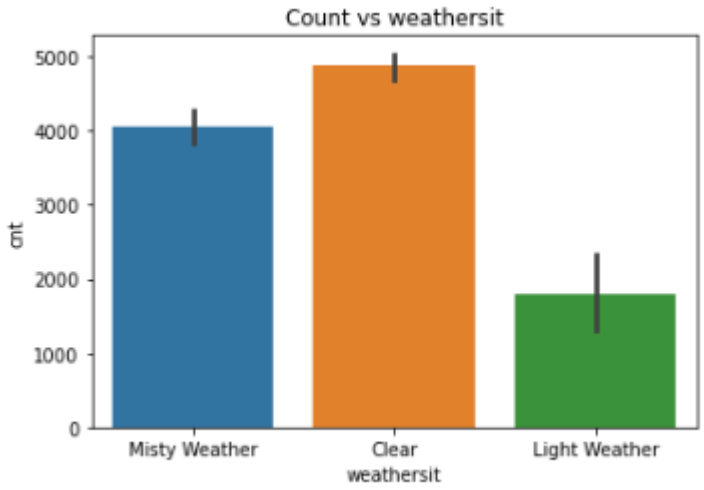
4. Count vs Weekday: We can see the variation is less in weekdays:



5. Count vs Working day: Working /Holiday are not showing high variations with respective of users count



6. Count vs Weather Situation: Most of the person sing bike in Misty and clear weather:



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

Ans2: It is used to reducing the extra column during the creation of dummy values of categorical values. Generally it's using to reduce the correlation among dummy variables.

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

Ans3: The highest correlation was showing by **temp, atemp** with respective of the target variables. We have checked ViF and then removed the **atemp.** Hence, **temp** is become the most correlated variable.

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

Ans5: Assumptions of linear regression after building the model on the raining set:

1. The error terms are normally distributed.
2. There is no Multicollinearity among independent variables.
3. Homoscedasticity is Residual.
4. The R2-Squareded and Adjusted R2-squared of training dataset is ~82.5% and test data set is ~80% which are very closely to each other.
5. P< 0.05 and ViF < 5
6. Predicted value is very nearly too actual values.

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

Ans5: We got below linear regression:

*0.2369 * yr + 0.3800 * temp - 0.164 * windspeed - 0.1308 * season_spring - 0.0517 * mnth_Jan -0.0691 * mnth_Jul + 0.0541 * mnth_Sep + 0.0232 * weekday_Sun - 0.2780 * weathersit_Light Weather - 0.0793 * weathersit_Misty Weather*

Here we can say based on the final model **temperature**, **year** and **weathersit_Light Weather** features contributing significant towards explaining the demand of the shared bikes.

General Subjective Questions

Q1. Explain the linear regression algorithm in detail.

Ans1: Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables they are considering, and the number of independent variables getting used.

Hypothesis function for linear regression:

$Y = \beta 0 + \beta i x i$

When training the model – it fits the best line to predict the value of y for a given value of x. The model gets the best regression fit line by finding the best **β0** and **β1** values.

β0: is the intercept of y
βi: coefficient of x of ith variable
xi: is the i th independent variable in the training input data.

Once we find the best **β0** and **βi** values, we get the best fit line. So when we are finally using our model for prediction, it will predict the value of y for the input value of x.

You will learn primarily about the following two types of linear regression models:

1. Simple linear regression (SLR) -> when we deal with single predictor then we go with SLR.
2. Multiple linear regression (MLR) -> when we deal with multiple predictor then we got with MLR.

MLR are using additional cases which are not using in SLR like Multicollinearity, Homoscedasticity, Residuals etc.

1. Multicollinearity: Two or more than independent variables should not to be highly correlated.
2. Homoscedasticity: A condition in which the variance of the residual, or error term, in a regression model is constant.
3. Residuals: We can identify how well a line fits an individual data point.

Q2: Explain the Anscombe's quartet in detail.

Ans2:

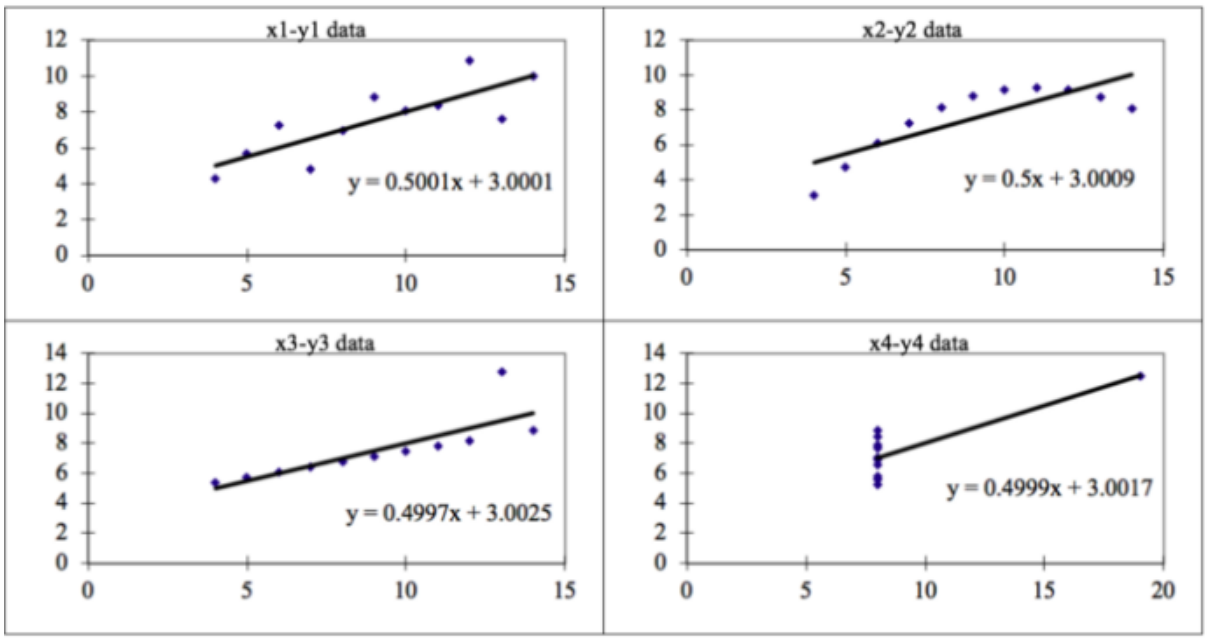
3. **Anscombe's Quartet** can be defined as a group of four data sets which are **nearly identical in simple descriptive statistics**, but there are some peculiarities in the dataset that **fools the regression model** if built. They have very different distributions and **appear differently** when plotted on scatter plots.
4. This tells us about the importance of visualising the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc. Also, the Linear Regression can be only be considered a fit for the **data with linear relationships** and is incapable of handling any other kind of datasets. These four plots can be defined as follows:

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 these four datasets are approximately similar and can be computed as follows:

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:

5. 

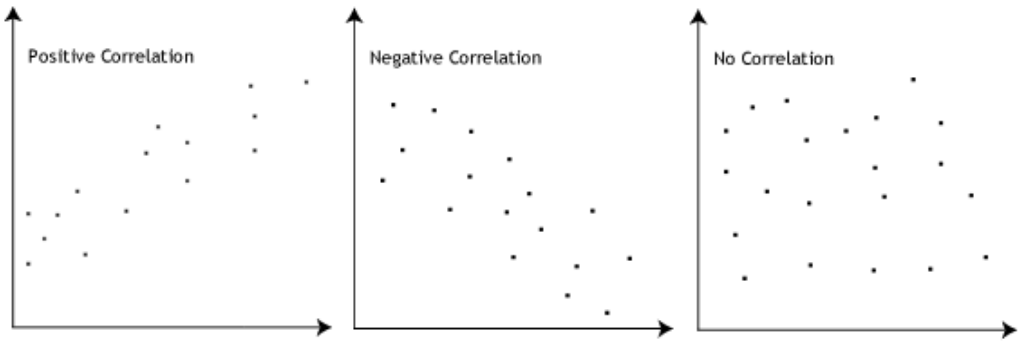
The four datasets can be described as:

6. Dataset 1: this fits the linear regression model pretty well.
7. Dataset 2: this could not fit linear regression model on the data quite well as the data is non-linear.
8. Dataset 3: shows the outliers involved in the dataset which cannot be handled by linear regression model
9. Dataset 4: shows the outliers involved in the dataset which cannot be handled by linear regression model

Q3. What is Pearson's R?

Ans3: Pearson's Correlation Coefficient is also referred to as Pearson's r, the Pearson product-moment correlation coefficient (PPMCC), or bivariate correlation. It is a statistic that measures the linear correlation between two variables. Like all correlations, it also has a numerical value that lies between -1.0 and +1.0.

The Pearson correlation coefficient, *r*, can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases. This is shown in the diagram below:



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

Ans4: **Scaling**: Normalization typically means rescales the values into a range of [0,1]. Standardization typically means rescales data to have a mean of 0 and a standard deviation of 1 (unit variance).
Scaling Performed Reason: It's a data Pre-Processing which is applied to independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.
Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude. It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

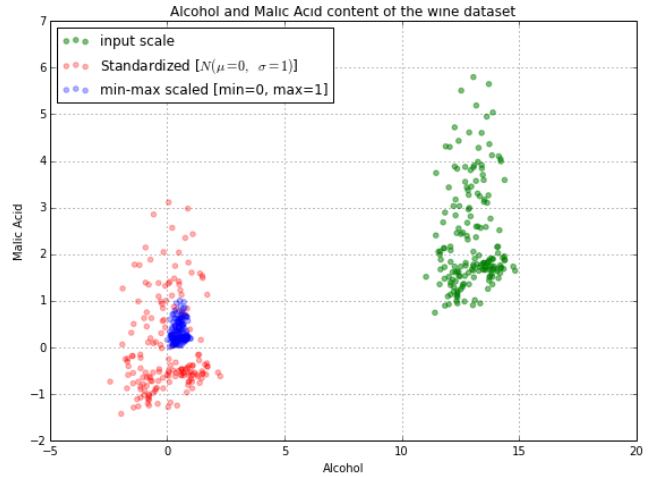
1. **Normalized Scaling:** It is used when features are of different scales. Minimum and maximum value of features are used for scaling. Scales values between [0, 1] or [-1, 1].
Scikit-Learn provides a transformer called MinMaxScaler for Normalization.

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

2. **Standardized Scaing :** It is used when we want to ensure zero mean and unit standard deviation. Mean and standard deviation is used for scaling. It is not bounded to a certain range.

$$x_{\text{stand}} = \frac{x - \text{mean}(x)}{\text{standard deviation } (x)}$$

We can see diff via graph as well.



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

Ans5: VIF = infinity, indicates a perfect correlation between two independent variables. In the case of perfect correlation, we get

R2 =1, which lead to **1/(1-R2)** infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

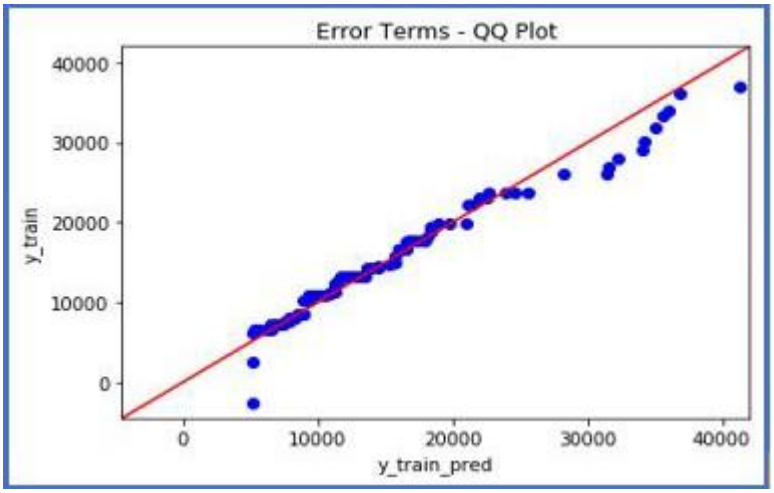
An infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables (which show an infinite VIF as well).

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

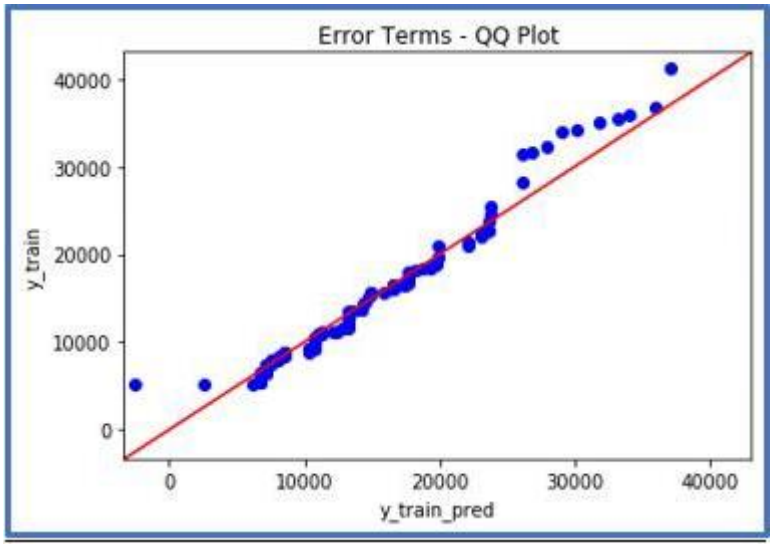
Ans: *Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.*

Interpretation:
A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

Below are the possible interpretations for two data sets.
a) Similar distribution: If all point of quantiles lies on or close to straight line at an angle of 45 degree from x –axis
b) Y-values < X-values: If y-quantiles are lower than the x-quantiles.



c) X-values < Y-values: If x-quantiles are lower than the y-quantiles.



d) Different distribution: If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis