## 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?

From my analysis performed on categorical variables of the data set I'm able to infer some of the following effects of categorical variables on the dependent variable:

1. there is a little hike in the number of bikes rent from Jun to September. (dependent variable 'cnt' or counts of bike rent)
2. maximum month counts of bike rent are higher on weekdays rather than weekends.
3. There is a small hike on bike rental in the fall.
4. Clear weather is more suitable in terms of count on bike rental.
5. There is a considerable hike on the count of bike rental in July, especially on holidays.

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

A dummy variable is a numerical variable used in regression analysis to represent subgroups of the sample. Dummy variables are useful because they enable us to use a single regression equation to represent multiple groups. In the simplest case, we could use 0,1 dummy variable. Suppose we have one variable as a season which has only three values 'summer' 'fall' and 'winter'.In that case, we could use 0,1 dummy variable. If combination '0 1' represent summer and '1 0' represent fall than we don't need to create another column for winter as we can easily say that '0 0' represent 'winter'
So for a variable with n levels, we can create n-1 columns of dummy variables to explain that variable properly. For this reason, it's important to use 'drop_first=True' to create n-1 columns for n levels variable. This means that we don't need to write out separate equation models for each subgroup.

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

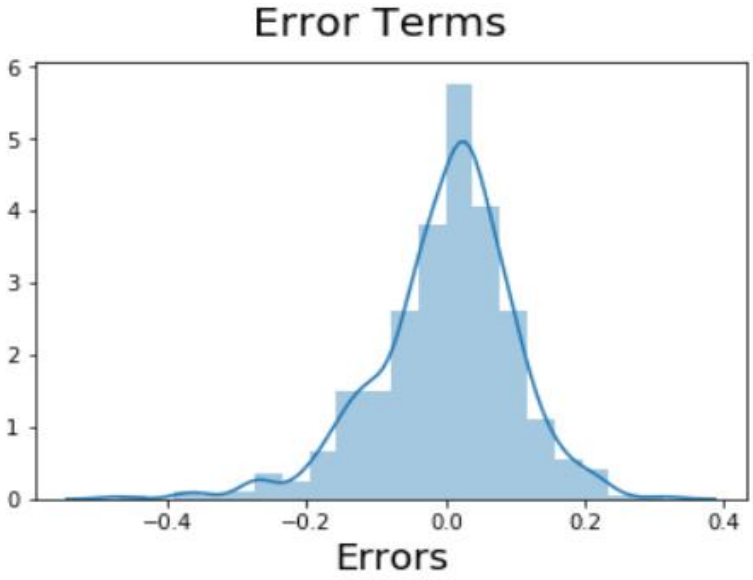
Among numerical variables 'temp' has the highest correlation with the target variable 'cnt'.

Q4. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

The assumption that we make after building the linear regression model on the training data set is that error terms are distributed normally. In support of that,

we have done the residual analysis. Residual represents the error or the difference between the actual target variable or y value and predicted y value by the model.

From the below histogram, we could see that the Residuals are normally distributed. Hence our assumption for Linear Regression is valid.




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

As per our final Model, the top predictor variables that influences the cnt are:

1. temp: A coefficient value of '0.4758' indicated that a unit increase in temp variable, increases the cnt numbers by 0.4758 units.
2. Light Snow: A coefficient value of '0.2562' indicated that a unit increase in Light Snow variable, decreases the cnt numbers by 0.2562 units.
3. yr : A coefficient value of '0.2350' indicated that a unit increase in yr variable, increases the cnt numbers by 0.2350 units.

1. ***Explain the linear regression algorithm in detail***

Linear regression is a **linear model**, e.g. a model that assumes a linear relationship between the input variables (x) and the single output variable (y). More specifically, that y can be calculated from a linear combination of the input variables (x).

In some cases, the value will be linearly upward that means whenever X is increasing Y is also increasing or vice versa that means they have a correlation or there will be a linear downward or a linear upward relationship.

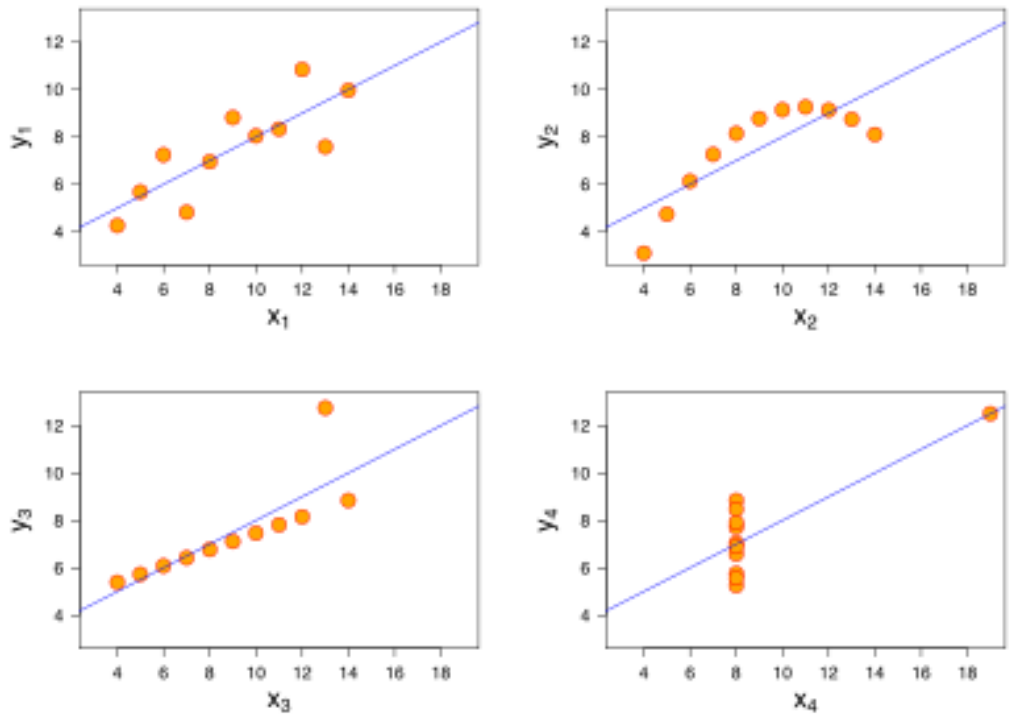
in a simple regression problem (a single x and a single y), the form of the model would be:

$y = B0 + B1*x$

The representation is a linear equation (same as the equation of a straight line) that combines a specific set of input values (x) the solution to which is the predicted output for that set of input values (y). As such, both the input values (x) and the output value are numeric. The linear equation has one factor to each input value or column, called a coefficient and represented by beta (Bi) i can be 1,2…n. B0 is an additional coefficient termed as y intercept when Bi = 0 then y = B0

We use regression to predict the change in price of stock or product.

Q2. ***Explain the Anscombe's quartet in detail***




Above graphs tells us four different story but they are statistically same

The first scatter plot (top left) appeared to be a simple linear relationship. clean and well-fitting linear models

second graph (top right) was not distributed normally; while a relationship between the two variables is obvious.it is not linear and also pearson's co efficient is not relevant.

In the third graph (bottom left), the distribution is linear, but should have a different regression line , the calculated regression is thrown off by an outlier.

Fourth graph shows that one outlier is enough to produce a high correlation coefficient.

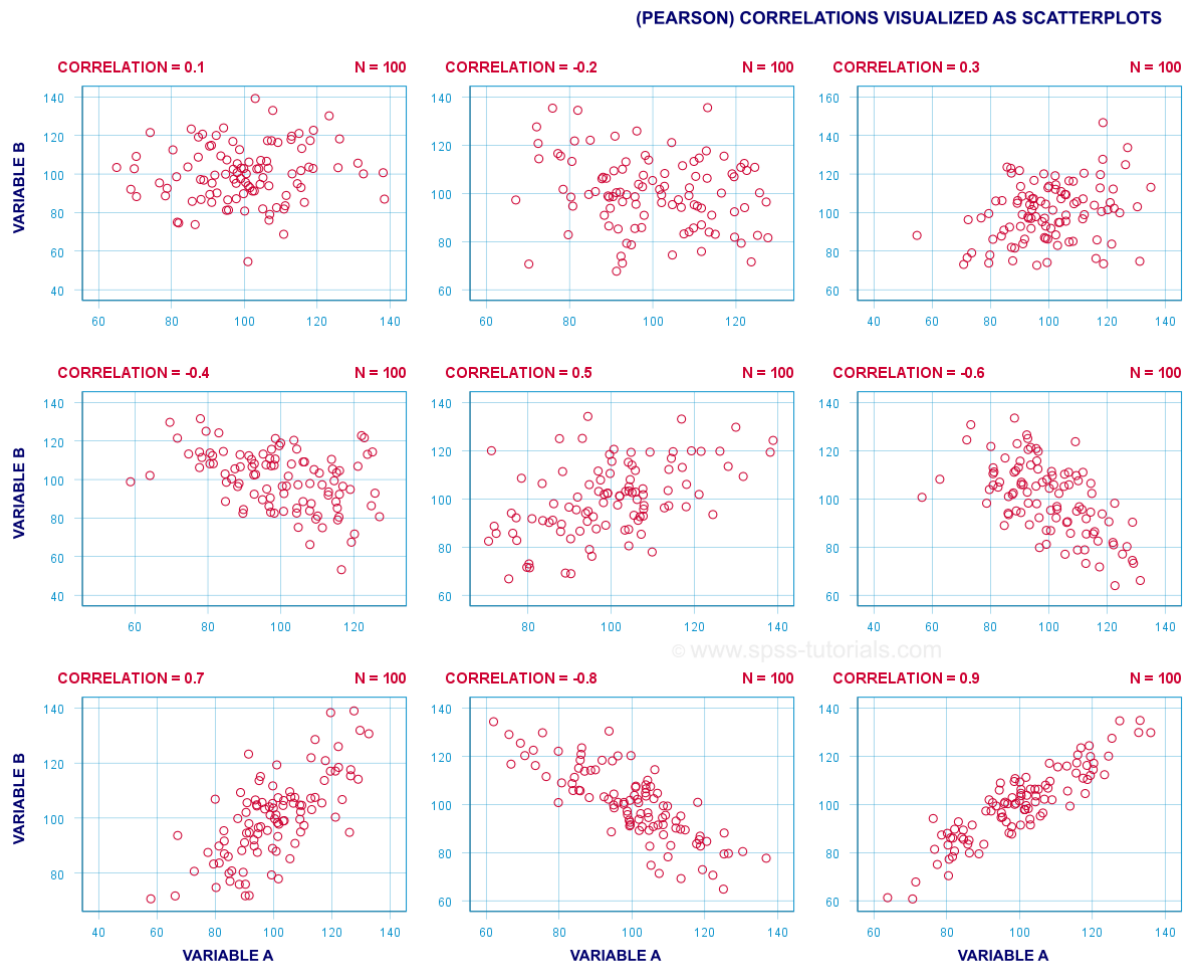
This Anscombe's quartet emphasizes the importance of visualization in Data Analysis.

3. ***What is Pearson's R?***

Correlation between sets of data is a measure of how well they are related. The most common measure of correlation in stats is the Pearson Correlation.

**Pearson's correlation coefficient** is the test statistics that measures the statistical relationship, or association, between two continuous variables.

Following graphs showing **Pearson's correlation coefficient** between two variables




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

.

Scaling is a technique to standardised the values of a feature so that algorithm can predict properly. If an algorithm is not using scaling method then it can consider the value 500 gram to be greater than 5 kg but that's actually not true and hence, the algorithm will give wrong predictions.

There are two types of scaling .

One of the reasons that it's easy to get confused between scaling and normalization is because the terms are sometimes used interchangeably and, to make it even more confusing, they are very similar! In both cases, you're transforming the values of numeric variables so that the transformed data points have specific helpful properties. The difference is that, in scaling, you're changing the range of your data and in normalization you are changing the shape of your data.

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

Since the formula for VIF = 1 / (1 – R2) therefore if R2 is equal to 1 then the denominator will become 0. Since the denominator becomes 0 therefore it will be infinity.

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

Quantile-Quantile (Q-Q) plot, is a graphical tool

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