# 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?**

   There are a couple of categorical variables namely season, mnth, yr, weekday, working day and weathersit. These categorical variables have a major effect on the dependent variable 'cnt'. The below fig shows the correlation among them.
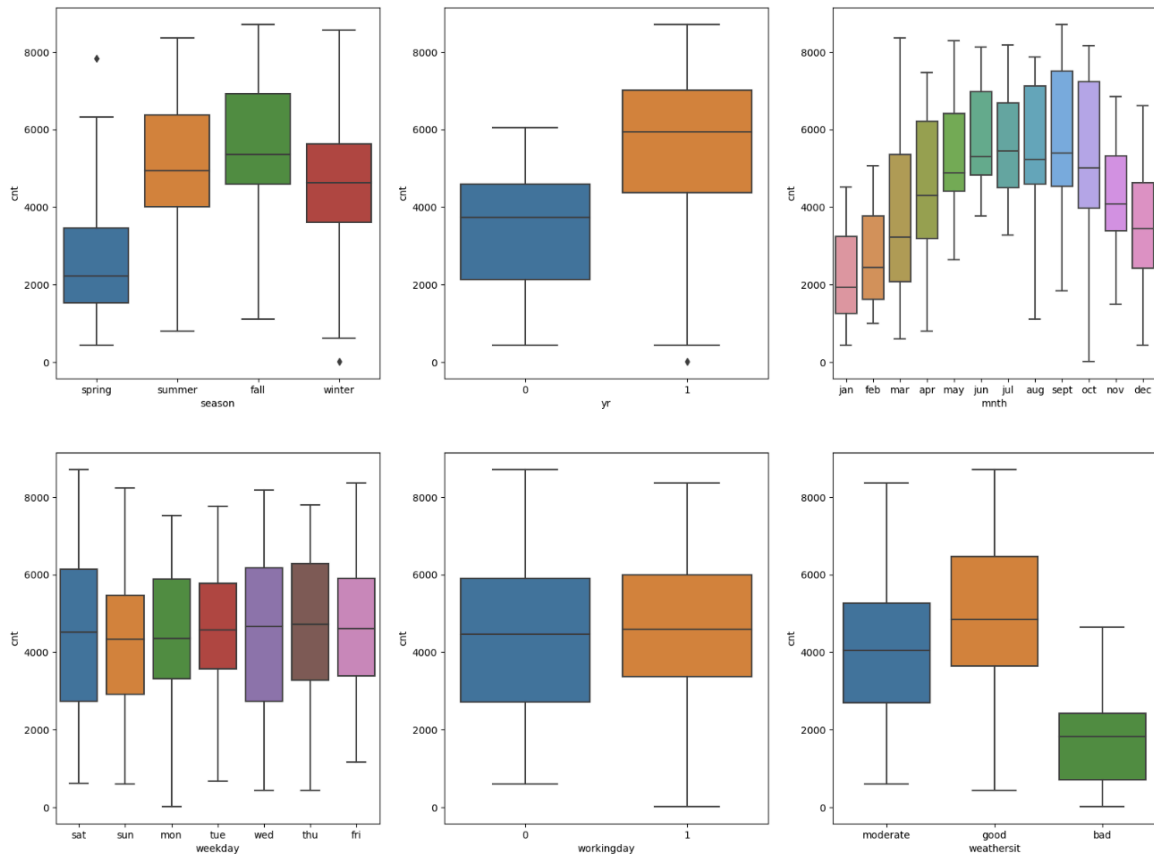

   These variables are visualised using bar plot Box plot both.

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

   The intention behind the dummy variable is that for a categorical variable with 'n' levels, you create 'n-1' new columns each indicating whether that level exists or not using a zero or one. Hence drop_first=True is used so that the resultant can match up n-1 levels. Hence it reduces the correlation among the dummy variables.
   Eg: If there are 3 levels, the drop_first will drop the first column.

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?**
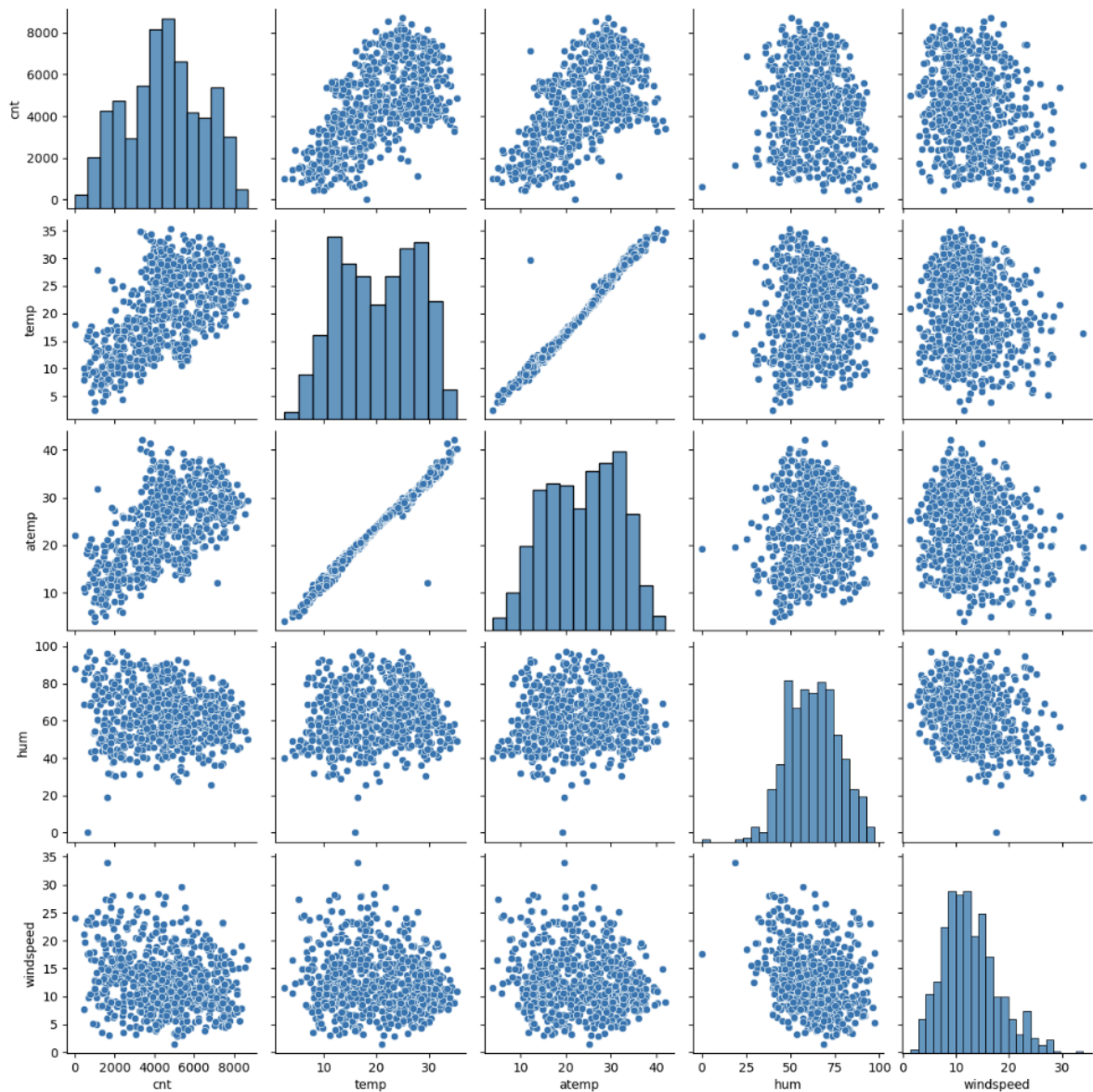
The 'temp' and 'temp' variables have the highest correlation when compared to the rest with target variables as 'cnt'.

4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**
Linear Regression models are validated based on Linearity. No auto-correlation, Normality of error, Homoscedasticity, Multicollinearity.

5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**
Top 3 features that has significant impact towards explaining the demand of the shared bikes are temperature, year and season

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**
    Linear regression is a form of predictive modelling technique which tells us the relationship between the dependent(target variable) and independent variables(predictors). Since linear regression shows the linear relationship, which means it finds how the value of the dependent variable is changing according to the value of the value of the independent variable. If there is a single i/p variable(x), such linear regression is called simple linear regression. And if there is more than one input variable, such linear regression is called simple linear regression. And if there is more than one input variable, such linear regression is called simple linear regression. The linear regression model gives a sloped straight line describing the relationship within the variables.
    A regression line can be a Positive Linear Relationship or a Negative Linear Relationship. The goal of the linear regression algorithm is to get the best values for a0 and a1 to find the best fit line and the best fit line should have the least error.
    In Linear regression, RFE or Mean Squared Error(MSE) or cost function is used, which helps to figure out the best possible values for a0 and a1, which provides the best fit line for the data points.

2. **Explain the Anscombe's quartet in detail.**
    Anscombe's Quartet can be defined as a group of four datasets 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 difficult distributions and appear differently when plotted on scaler plots. It was constructed to illustrate the importance of plotting the graphs before analysing and model building, and the effect of other observations on statistical properties. There are these four data set plots which have nearly the same statistical observations which provide the same statistical information that involves variance, and mean of all x,y points in all four datasets.
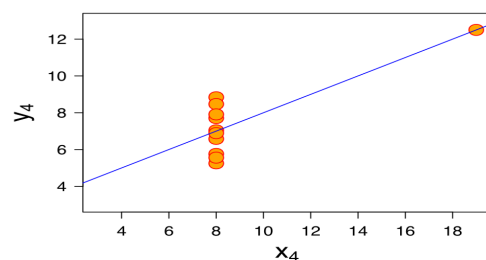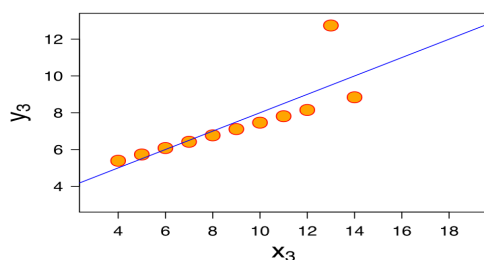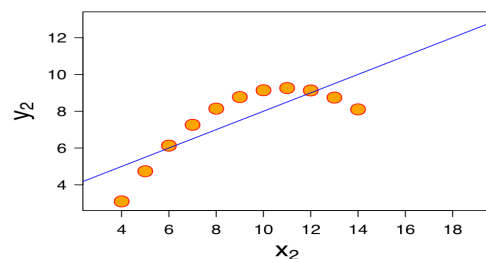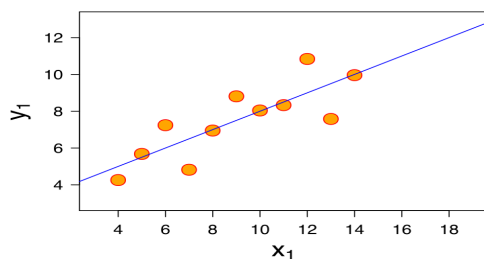
- 1st dataset fits linear regression model as it seems to be linear relationship between X and Y
- 2nd dataset does not show a linear relationship between X and Y, which means it does not fit the linear regression model.
- 3rd dataset shows some outliers present in the dataset which can't be handled by a linear regression model.
- 4th data set has a high leverage point means it produces a high correlation coeff.

Its conclusion is that regression algorithms can be fooled so, it's important to data visualisation before building a machine model.

3. **What is Pearson's R?**
   In Statistics, the Pearson's Correlation Coefficient is also referred to as Person's r, the Pearson product-moment correlation coefficient (PPMCC), or bivariate correlation. It is a statistic that measures the linear correlation between two variables.

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

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**
   Scaling means you're transforming your data so that it fits within a specific scale. It is one type of data pre-processing step where we fit data in a specific scale and speed up the calculations in an algorithm. Collected data contains features varying in magnitudes, units and range. If scaling is not performed than algorithms to weight high values magnitudes and ignore other parameters which will result in incorrect modelling.
   Difference between Normalizing Scaling and Standardize Scaling:
1. In normalized scaling minimum and maximum value of features are used whereas in Standardize scaling mean and standard deviation is used for scaling.
2. Normalized scaling is used when features are of different scales whereas standardized scaling is used to ensure zero mean and unit standard deviation.
3. Normalized scaling scales values between(0,1) or (-1,1) whereas standardized scaling is not having or is not bounded in a certain range.
4. Normalized scaling is used when we don't know about the distribution whereas standardized scaling is used when distribution is normal.
5. Normalized scaling is called scaling normalization whereas standardized scaling is called Z Score normalization.

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen?**
   VIF(VarianceInflationFactor) basically helps explain the relationship of one independent variable with all the other independent variables. The formulation of VIF

is given below: A VIF value of greater than 10 is definitely high, aVIF of greater than 5 should also not be ignored and inspected appropriately.
A very high VIF value shows a perfect correlation between two independent variables. In the case of perfect correlation, we got R2=1, which leads 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.

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.**
Q-Q plot is probability plot, which is a graphical method for comparing two probability distributions by plotting their qualities against each other.
Quantile-Quantile(Q-Q) plot, is a graphical tool to help us assess if a set of data possibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution.
QQ plots can also be used to determine whether or not two distributions are similar or not . If they are quite similar you can expect the QQ plot to be more linear. The linearity assumption can best be tested with scatter plots. Secondly, the linear regression analysis requires all variables to be multivariate normal. This assumption can best be checked with a histogram or a Q-Q Plot.

Importance of QQ Plot in Linear Regression:
In Linear Regression where we have a train and test dataset then we create a Q-Q plot by which we can confirm that both the data train and test data set are from the population with the same distribution or not.
Advantages:
● It can be used with sample size also.
● Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot
Q-Q plot use on two datasets to check.
If both datasets came from population with common distribution
If both datasets have common location and common scale
If both datasets have similar type of distribution shape
If both datasets have tail behaviour