# 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?**

   In this dataset, the categorical variables were season, yr, mnth, holiday, weekday, workingday and weathersit. These were visualize using Boxplot. These variables had the following effect on out dependent variable “cnt”:

` **a. Season:** The below boxplot showed that spring season had least value of cnt whereas fall season had maximum value of “cnt”. Summer and winter had intermediate value of “cnt”.

**b. Weathersit:** There we no rentals when there is heavy rain/ snow which means weather is extremely unfavorable. Highest count was seen when weathersit was clear, partly cloud.

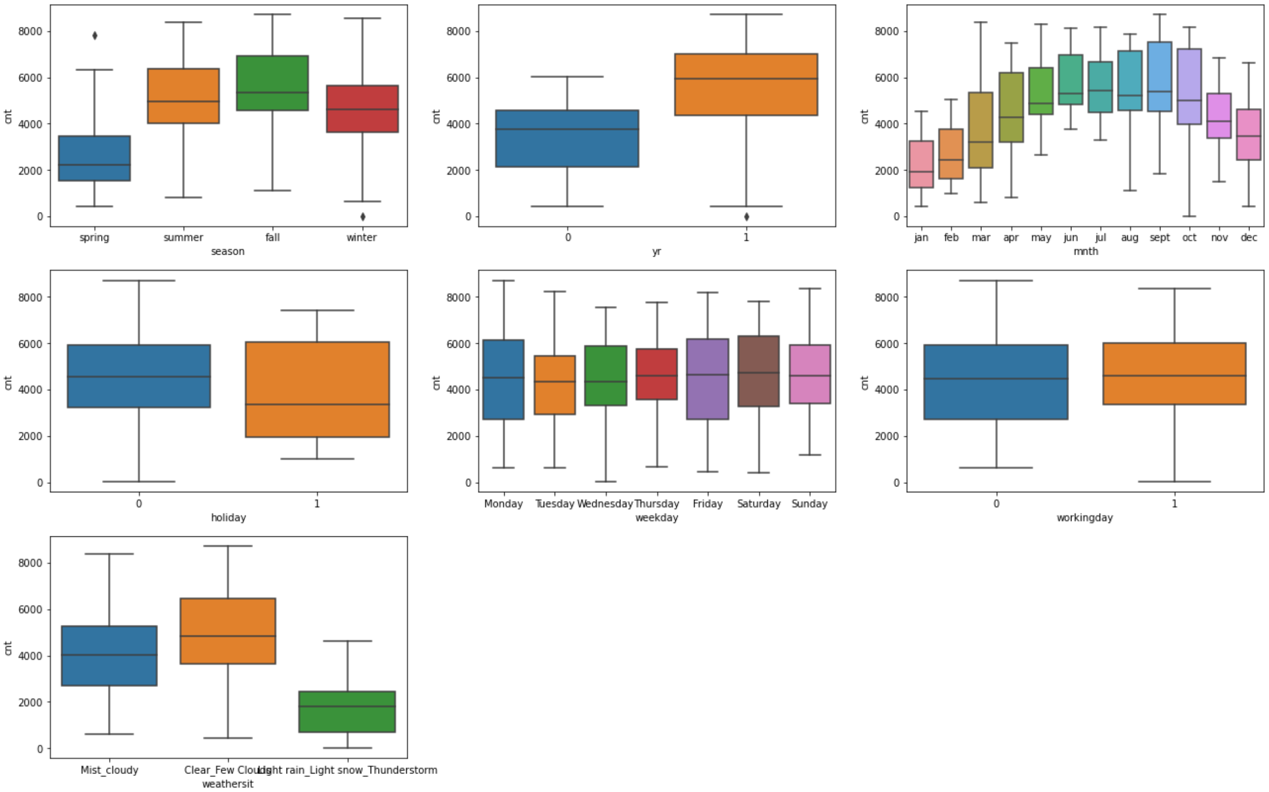
**c. Holiday:** rentals reduces when there are holidays.

**d. Mnth:** September saw highest no. of rentals while January saw least. This observation can also be made as per weathersit as the weather situation in January is usually snowy.

**e. Yr:** The numer of rentals in 2019 is greater than 2018**.**

**f. Workingday:** Rentals are  more when there is workingday.

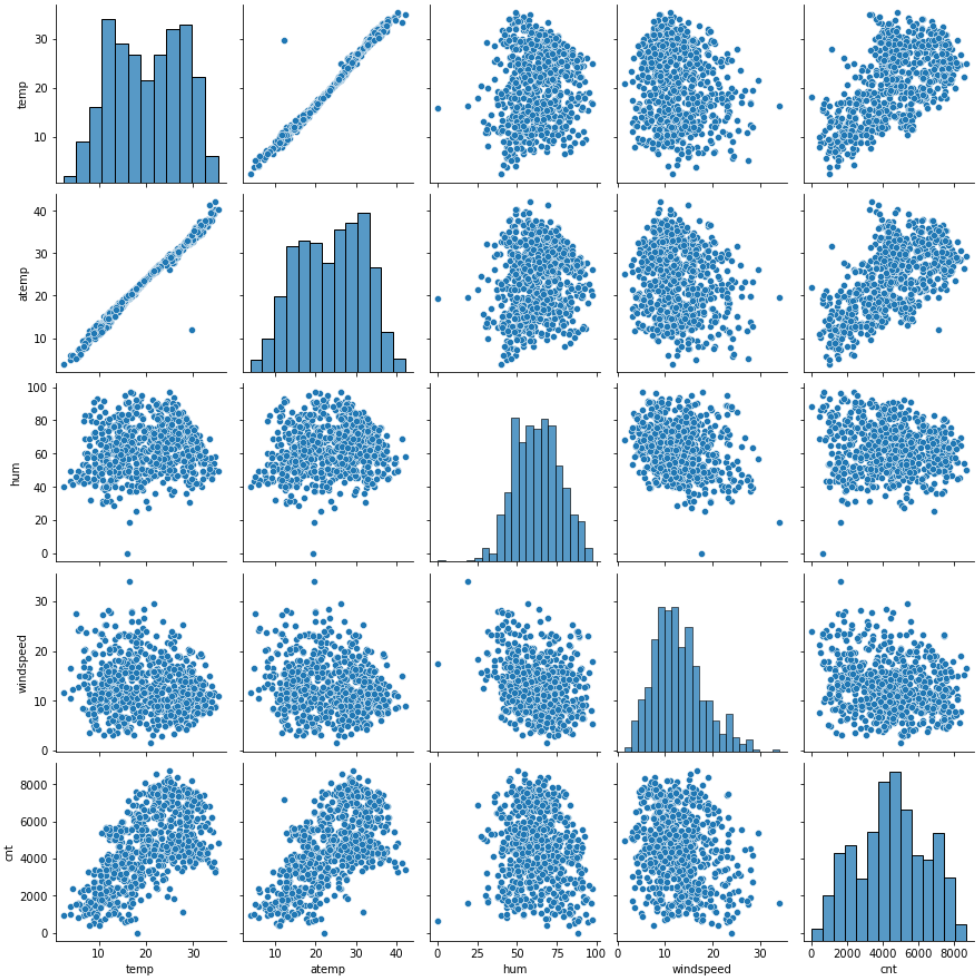
**g. Weekday:** The highest rental are on Monday and the least is on Wednesday.



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

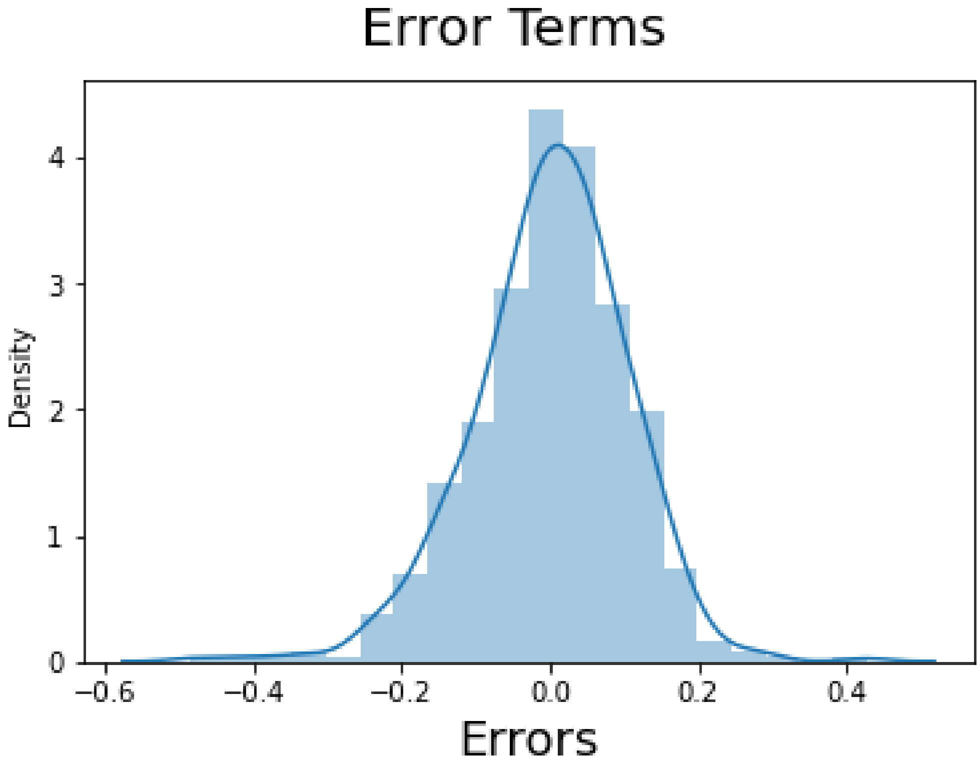
- If we don’t drop the first column then those dummy variables will be correlated and this may affect some model adversely.
- Also, if we have all the dummy variables it will lead to multicollinearity between the dummy variables and to keep this under control, we use drop_first=True.

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



Looking into above pairplot, ‘temp’ and ‘atemp’ are two variables which are highly correlated to target variable ‘cnt’.

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



- Residual distribution should follow normal distribution and centered around 0 (mean=0).
- We validate this assumption about residual by plotting a distplot of residual and see if residual is following normal distribution or not.
- This above diagram shows that the residuals are distributed about mean=0.

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

The top three features contributing significantly towards explaining the demand of the shared bikes are:

a. **temp**: coefficient is + 0.491754

b. **yr**: coefficient is +0.234103

c. **weathersit_Light rain_Light snow_Thunderstorm:** coefficient is -0.234103

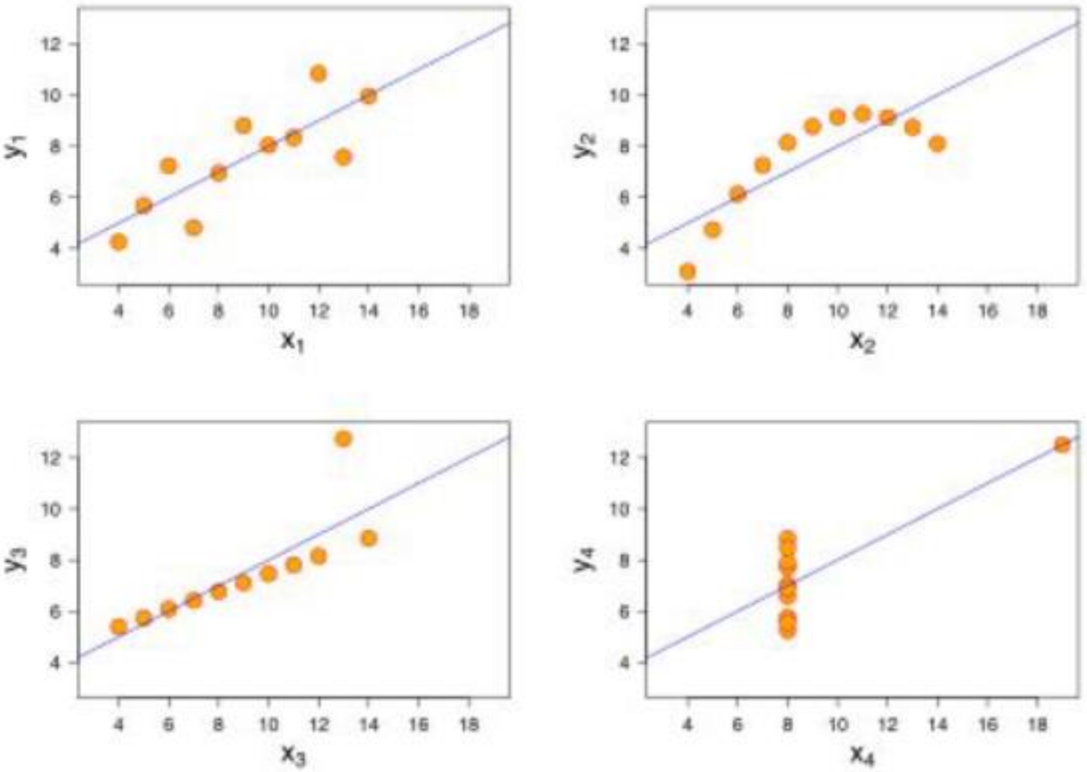
# General Subjective Questions

1. **Explain the linear regression algorithm in detail.**
    - Linear regression is a type of supervised Machine Learning algorithm that is used for the prediction of numeric values. Linear Regression is the most basic form of regression analysis and regression is most commonly used predictive analysis model.
    - Linear regression is based on the popular equation **“y=mx+c”.**
    - This assumes that there is a linear relationship between dependent variable(y) and the independent variable(x) and in regression we calculate the best fit line which describes the relationship between y and x with least error.
    - Regression is performed when the dependent variable is continuous data type and independent variable could be of any data type like continuous, categorical, etc.
    - Regression is broadly divided into two types:
        1. **Simple Linear Regression (SLR):** when dependent variable is predicted using only one independent variable.
        2. **Multiple Linear Regression (MLR):** when dependent variable is predicted using multiple independent variables.
           Equation of MLR will be:
           Y=β0+ β1X1+ β2X2+ β3x3+……+ βnXn
           where: β1-> coefficient of X1 variable
           β2-> coefficient of X2 variable
           β3-> coefficient of X3 variable
           βn-> coefficient of Xn variable
           β0-> constant/intercept

2. **Explain the Anscombe’s quartet in detail.**

    - Anscombe’s quartet is a group of datasets that have same mean, standard deviation and regression line, but which is quantitatively different**.**
    - It is often used to illustrate the importance of looking at a set of data graphically and not only relying on basic statistics.
    - It was developed to emphasize both the importance of graphing data before analyzing

it and the effect of outlier and other influential observation on statistical properties.



1. First scatter plot (top left) appears to be a simple linear relationship.
2. The second plot (top right) is not distributed normally, while there is relation between then which is non-linear.
3. In the third plot (bottom left) the distribution is linear but it should have different regression line.
4. Finally, the last plot (bottom right) shows an example when one-high leverage point is enough to produce high correlation coefficient, even though the older data points do not indicate any relationship between the variables.

3. **What is Pearson's R?**
   - Pearson's r is a numerical summary of the strength of the linear association between the variables.
   - Its value ranges between -1 to 1.
   - It shows the relationship between two sets of data**.**
     - r = 1 means the data is perfectly linear with a positive slope.
     - r = -1 means the data is perfectly linear with negative slope.
     - r= 0 means there is no linear association

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?**
   - Feature scaling is a method used to normalize or standardize the range of independent variables or features of data.
   - It is performed during the data preprocessing stage to deal with various values in the dataset.

- If feature scaling in not done, then a machine learning algorithm needs to weigh greater values, higher and consider smaller values as the lower values, irrespective of the units of the values.
    - **Normalization** is generally used when we know that the distribution of our data will not follow gaussian distribution. It can be useful in algorithms that do not assume any distribution of the data like K-nearest neighbor and Neural Networks.
    - **Standardization can** be helpful in cases where the data follows Gaussian distribution, However, this does not have to be necessarily true and also it does not have bounding range. So, even if we have outliers in data, it will not be affected by standardization.

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen?**
    - VIF (Variance inflation factor) gives how much the variance of the coefficient estimate is being inflated by collinearity (VIF) = 1/(1-R^2) and if there is perfect correlation then VIF= infinite.
    - Here R^2 (R-square) value is value of the independent variable which we want to check how well it is explained by other independent variables.
    - If that independent variable can be explained perfectly, then it will have perfect correlation and its R-square value will be equal to 1. So, VIF = 1/1-1 which gives VIF equals to “infinite”

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.**
    - A q-q plot is a plot of the quantiles of the first dataset against the quantiles of the second dataset.
    - It is used to compare the shapes of distribution.
    - A Q-Q plot is a scatterplot created by plotting both 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.
    - The q-q plot is used to answer the following questions:
        - Do two datasets come from population with a common distribution?
        - Do two datasets have common location and shape?
        - Do two datasets have similar distributional shapes?
        - Do two datasets have similar tail behavior?