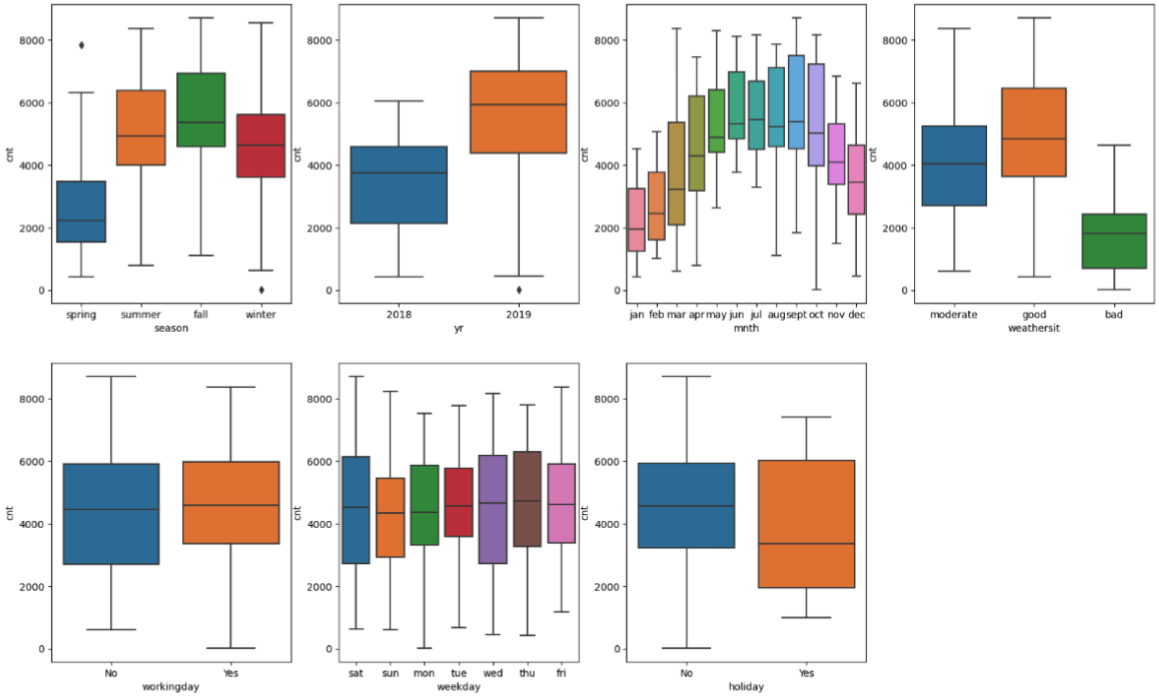
# Assignment-based Subjective Questions

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

Ans: Following are the categorical variables as per the data shared:

- Season
- Year
- Month
- Weathersit
- Workingday
- Weekday
- Holiday



Based on the above boxplot, following can be inferred about the dependency:

- Season – “Fall” has the highest median which shows that the max demand was during this time. Whereas the least was during “spring”.
- Year – “2019” had the highest count of users as compared to year “2018”.
- Month – During “September” the demand for bikes was the highest. It gradually increases from Jan to Sep and then decreases from Sep to December.
- Weathersit – There was no demand of bikes during severe weather condition (Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog). During bad weather (Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds) the demand was least. The max demand was during good weather (Clear, Few clouds, Partly cloudy, Partly cloudy).
- WorkingDay – No much difference on demand of bikes between a working and non-working day.

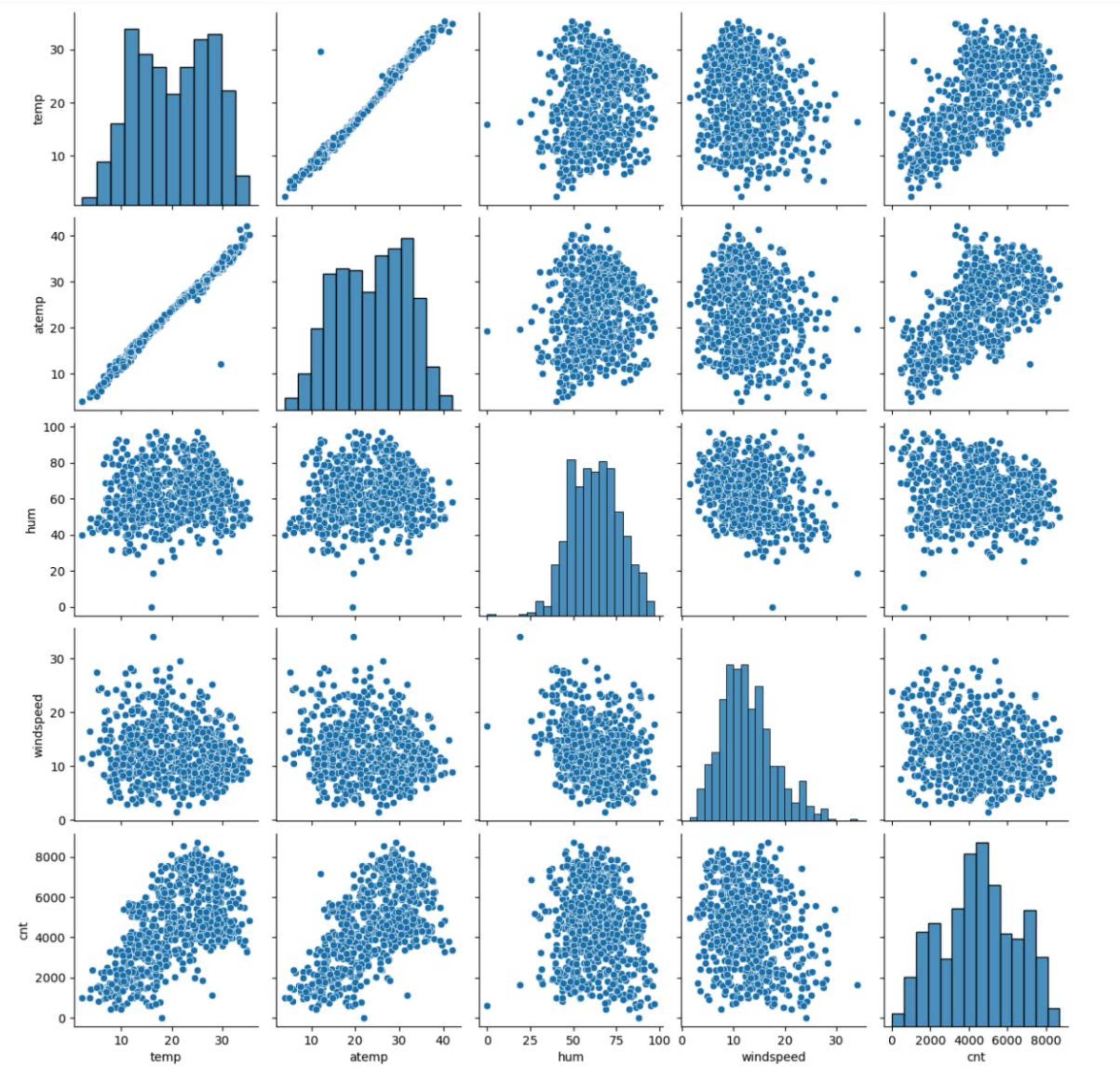
- WeekDay – No much difference on demand of bikes across days of week.
- Holiday – median of the demand of bikes is less during holiday.

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

Ans: For a categorical variable with “n” levels, only (n-1) columns are required to represent the dummy variables. “drop_first=True” helps in reducing the extra columns created during dummy variable creation.

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

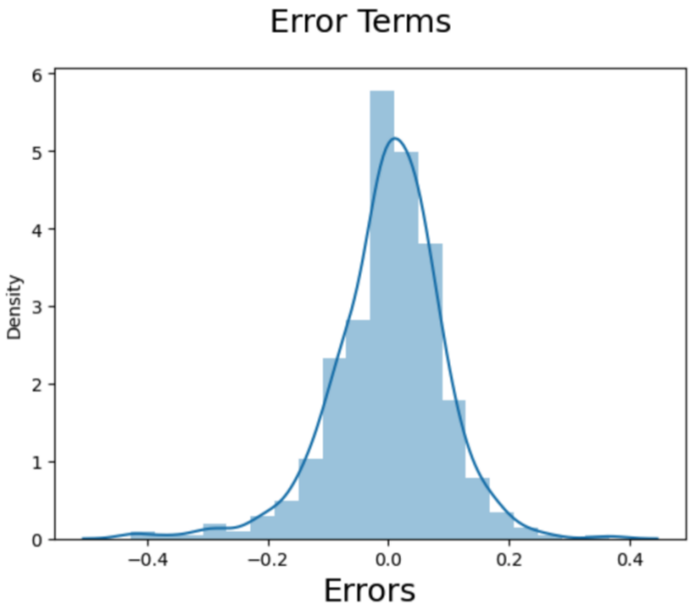
Ans: “temp” and “atemp” have the highest correlation with “cnt”.



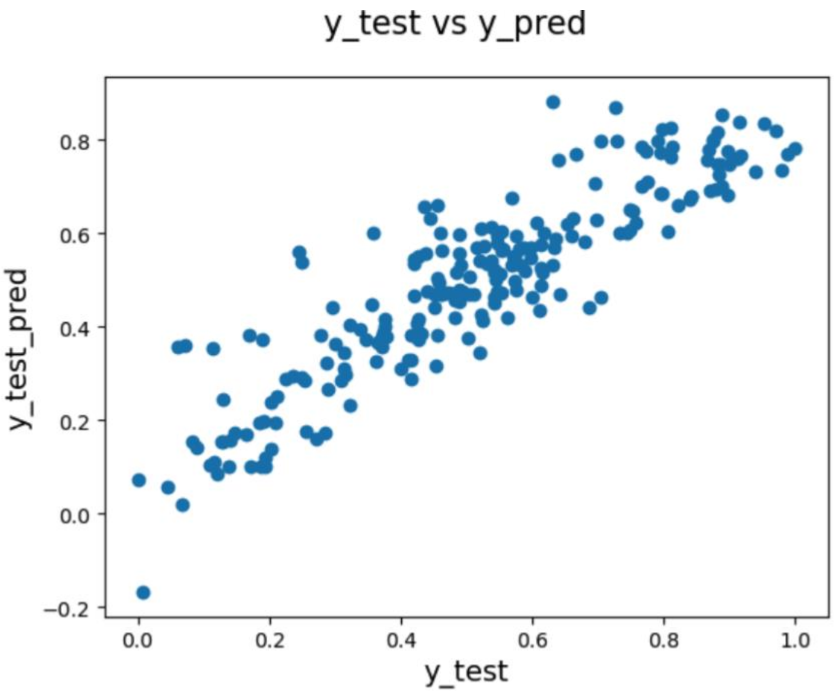
**4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)**

Ans: Validation was done by –

1. Residual distribution is a normal distribution centred around 0(mean). Distplot of residuals is as below:



2. Linear regression assumes there is little or no multicollinearity. VIF was calculated and VIF values for all features was <5.
3. Linear relationship between independent and dependant variables can be visualised by the pairplot.
4. Error terms are independent of each other,
5. Error terms have a constant variance across all values of independent variables.



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

Ans: Top 3 features contributing to the demand of bikes (ignoring year as this :

- Temp  ( Coeff – 0.45)
- Weather ( Bad weather - -0.28)
- Windspeed (-0.14)

# General Subjective Questions

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

Ans:

Linear regression is a machine learning algorithm based on supervised learning. It performs a regression task, which means it predicts a continuous output variable (y) based on one or more input variables (x). It is mostly used for finding out the linear relationship between variables and forecasting.

The basic idea of linear regression is to find a line that best fits the data points, such that the distance between the line and the data points is minimized. The line can be represented by an equation of the form:

$Y= B0 + B1X$

Y -> Dependant variable

X -> independent variable

B -> corelation coefficient

2 types:

a. Simple Linear regression ( One independant variable)

Equation is as mentioned above.

b. Multiple Linear regression ( More than 1 independant variable)

$Y= B0 + B1X1 + B2X2 + B3X3 +\ldots\ldots+ BnXn$

Y -> Dependant variable

Xi -> ith independent variable

B

Assumptions:

1. X and Y have a linear relation
2. 2. Error terms are normally distributed with mean as 0
3. Error terms are independent of each other
4. Error terms have constant variance.

Where, error(ei) = yi – y(pred)

RSS ( cost function) = e1^2+e2^2+e3^2+……+ en^2

Ultimately aim of this model is to minimize the cost function.

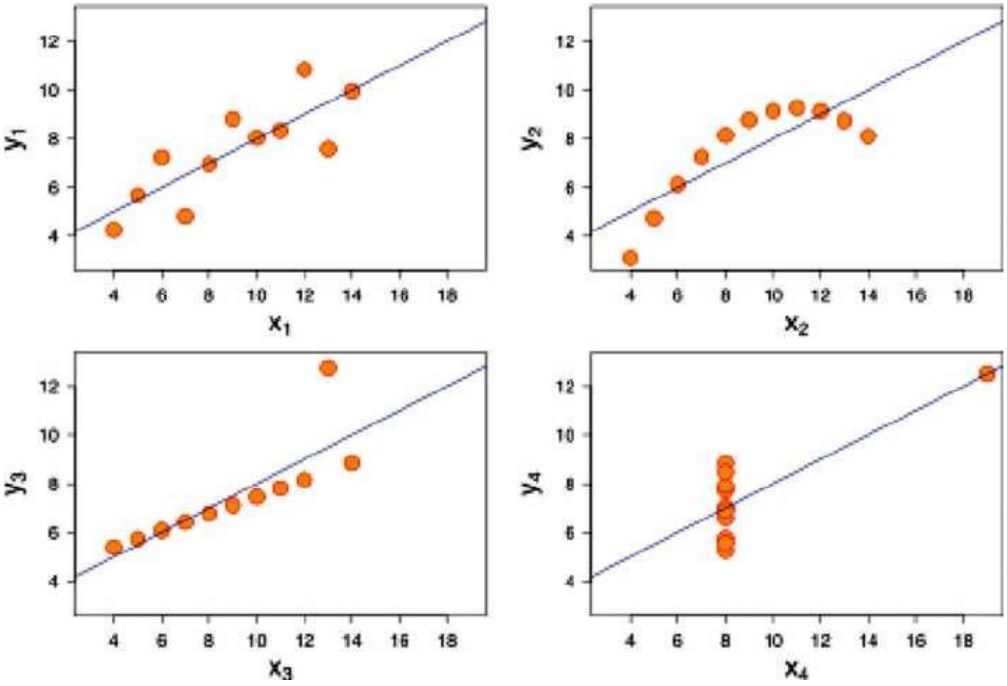
**2. Explain the Anscombe's quartet in detail. (3 marks)**

Ans:

Anscombe's Quartet was developed by statistician Francis Anscombe. It includes four data sets that have almost identical statistical features, but they have a very different distribution and look totally different when plotted on a graph. It was developed to emphasize both the importance of graphing data before analyzing it and the effect of outliers and other influential observations on statistical properties.

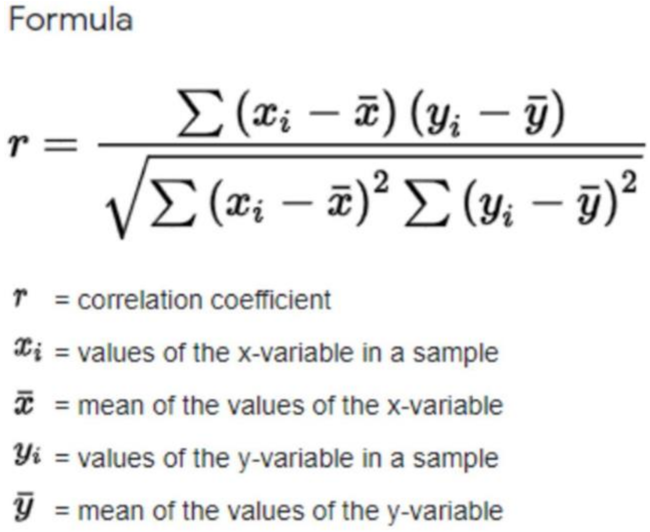
The first scatter plot (top leG) appears to be a simple linear relationship.
The second graph (top right) is not distributed normally; while there is a relation between them is not linear.
In the third graph (bottom leG), the distribution is linear, but should have a different regression line the calculated regression is offset by the one outlier which exerts enough influence to lower the correlation coefficient from 1 to 0.816.

Finally, the fourth graph (bottom right) shows an example when one high-leverage point is enough to produce a high correlation coefficient, even though the other data points do not indicate any relationship between the variables.



## 3. What is Pearson’s R? (3 marks)

Pearson's r is a numerical summary of the strength of the linear association between the variables. It value ranges between -1 to +1. It shows the linear relationship between two sets of data. In simple terms, it tells us “can we draw a line graph to represent the data?”



Formula

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

$r$ = correlation coefficient

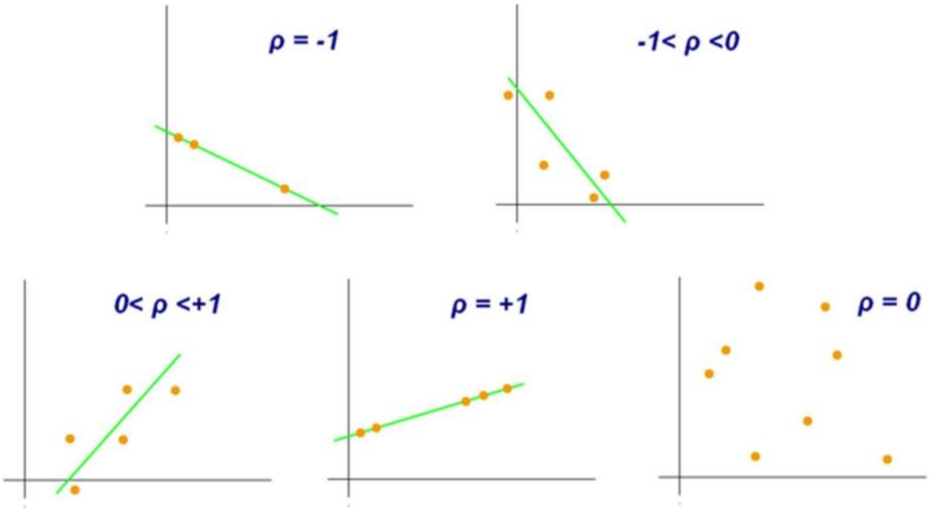
$x_i$ = values of the x-variable in a sample

$\bar{x}$ = mean of the values of the x-variable

$y_i$ = values of the y-variable in a sample

$\bar{y}$ = mean of the values of the y-variable

As can be seen from the graph below, r = 1 means the data is perfectly linear with a positive slope r = -1 means the data is perfectly linear with a 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? (3 marks)**

**Ans:**

**Scaling** is a method used to normalize or standardize the range of independent variables or features of data. Scaling just affects coefficients and not params like t-statistics, f-statistics, p-value, R-squared etc.

It is performed for :

1. Ease of interpretation
2. Faster convergence for gradient descent methods

| S.NO. | Normalized scaling | Standardized scaling |
|---|---|---|
| 1. | Minimum and maximum value of features are used for scaling | Mean and standard deviation is used for scaling. |
| 2. | It is used when features are of different scales. | It is used when we want to ensure zero mean and unit standard deviation. |

| | | |
|---|---|---|
| 3. | Scales values between [0, 1] or [-1, 1]. | It is not bounded to a certain range. |
| 4. | It is really affected by outliers. | It is much less affected by outliers. |
| 5. | Scikit-Learn provides a transformer called MinMaxScaler for Normalization. | Scikit-Learn provides a transformer called StandardScaler for standardization. |

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

Ans:
VIF (Variance Inflation Factor) gives how much the variance of the coefficient estimate is being inflated by collinearity.
$VIF = 1/(1-R^2)$

When the value of VIF is infinite, $(1-R^2) = 0$ which means R-squared (R2) =1.
This shows a perfect correlation between two independent variables. This is the case of perfect correlation.
To solve this 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.**

**Ans:**
The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.

Use of Q-Q plot:
A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second dataset. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the
0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value. A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.

Importance of Q-Q plot:

When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified. If so, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale. If two samples do differ, it is also useful to gain some understanding of the differences. The q-q plot can provide more insight into the nature of the difference than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample tests.