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?**

   The categorical variables in our dataset are weathersit and season they can be analysed in various ways by creating box plots, value counts and also with the help of other numerical variables to visualize graphically.

   To reduce the effort of converting year, month and weekday to categorical variables I have kept them as numerical variables which help during the process of scaling.

   Weathersit – number of people riding the bike is mostly on clear weather days when compared to other days and during September the number of bikes ridded are more when compared to other month when weather is clear.

   Season – the count of people riding number of bikes are different across different months like in march has the highest number of bikes during fall, June has highest number of bikes for both summer and fall and September has highest number of bike rides for winter.

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

   As per general rule of thumb when creating dummy variables if there are n level we keep n-1 levels. It also helps in reducing correlation created among dummy variables. This effect some models where cardinality is smaller.

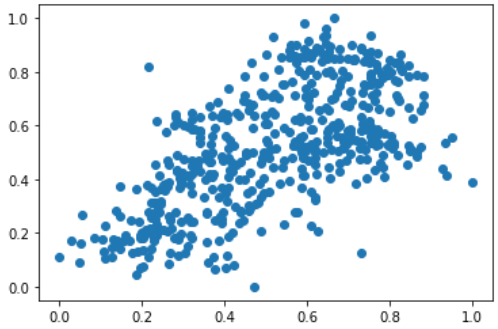
3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable**?

   Atemp and Temp have the highest correlation with Count Variable.

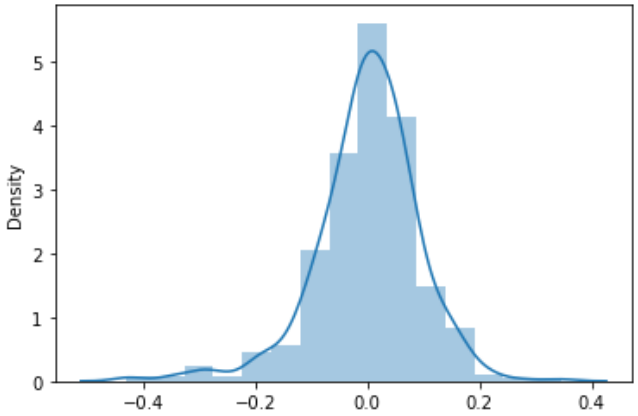
4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**

   There are 4 assumptions of linear regression

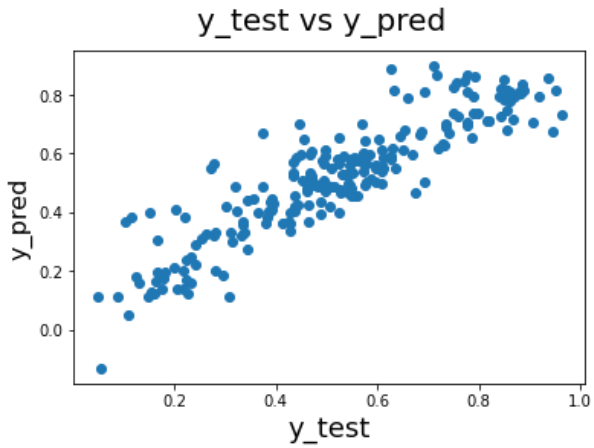
   1. There is a linear relationship between x and y : as per our analysis there is linear relationship between x and y



2. Error Terms are normally distributed with mean, are independent of each other and follow homoscedasticity.



As per our analysis the residual error terms follow above 3 principles as all the residual are normally distributed with the help of VIF, correlation matrix we can check multicollinearity and scatter plot to check homoscedasticity.



5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**
    1. Yr - t value - 28.044
    2. Temp – t value - 14.003
    3. Mist - t value – (-8.867)

General Subjective Questions:

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

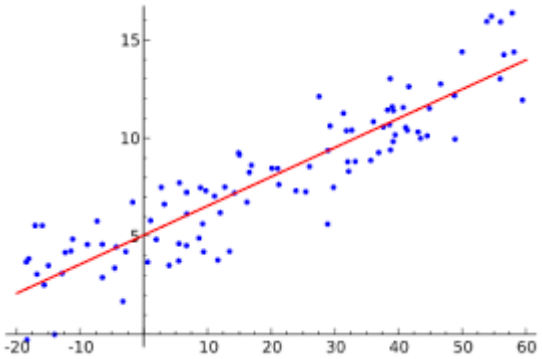
Linear regression is a linear modelling approach to define relationship between dependent and independent variables. If we have one independent variable, then we call it as simple linear regression and when we have more than one independent variable then we call it as multiple linear regression.
It is one of the most fundamental supervised machine learning algorithms.

Equation: $y = \beta 0 + \beta 1Xi1 + \beta 2Xi2 + \cdots . + \beta pXip + \varepsilon i$

where i = 1,2,….n

β0 = intercept when p is 1 then we can say it as regression slope.

Interpretation:



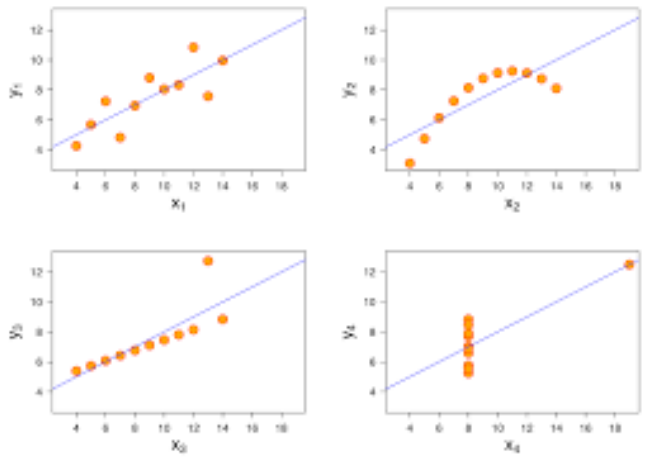
Assumptions of Linear Regression.

1. True relationship is linear.
2. Error are normally distributed.
3. Independence of the observations.
4. Homoscedasticity of Error

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

Anscombe's quartet consist of four simple data sets that have nearly identical simple descriptive statistics but still have different distributions and will appear differently when plotted.

It states the importance of graphing data before analysing and as it will help in understanding effect of outliers, if any influential observations on statistical properties.



3. What is Pearson's R?

Pearson's R measure linear correlation between two variables it has a numerical value lies between -1.0 to 1.0

It cannot capture non-linear relationship between variables and also cannot differentiate between dependent and independent variables.

Formulae:

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\, n\Sigma x^2 - (\Sigma x)^2 \,][\, n\Sigma y^2 - (\Sigma y)^2 \,]}}$$

There are few requirements for Pearson's correlation coefficient:

1. Scale should be interval or ratio.
2. Association of the variables should be linear.
3. No outliers in the data as this will affect the value.
4. Variables should normally distribute approximately.

A positive correlation is which when x increases y also increases. Negative Correlations is which when x decreases y also decreases.

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

Feature scaling is a generally mean putting the feature values into the same range. It is usually performed during data pre-processing stage which is applied for independent variables as it helps in speeding up the calculations in an algorithm.

Scaling effects only the coefficients and none of the other statistical parameters like t - statistics, f – statistics, r- squared etc.

1. Normalization typically rescale it values into a range of 0 to 1 where as Standardization means rescales data to mean of 0 and SD of 1.
2. Normalization is a good practice when the distribution of your data does not follow Gaussian distribution. On the contrary Standardization is used when the data follow Gaussian Distribution.

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

VIF is an index to measure how much variance of an estimated regression coefficient increases due to collinearity. When all the independent variables have perfect correlation then VIF = 1 then VIF = infinity.

Formulae of VIF

VIF = $1/(1-r^2)$

In the case of perfect correlation, the r2 will become 1 which means 1/1-1 which is 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 which helps us to assess if a dataset plausibly comes from theoretical distributions such as Normal exponential or Uniform distribution.

It helps in determining if two data sets come from a population with a common distribution.

Use and Importance of (Q-Q):

1. If datasets come from a population with common distributions.
2. If data sets have common location and scale
3. If data sets have similar distributional shapes
4. If data sets have similar tail behaviour
5. It can be used with small sample sizes also
6. The presence of outliers can be detected from this plot.