**Assignment**

***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?**

**Ans :** In our dataset, we have 7 categorical columns (i.e. **season, yr, mnth, holiday, weekday,**

**workingday, weathersit**).

1. We can infer that during the **holiday**, average count of bike rents decreases.
2. The same thing is explained by **working day or weekday**, count increases when it is working day. This might be due to people renting bikes to commute to their workplace.
3. This pattern is also followed in **season**, we can see that ‘spring’ has more holiday compare to other season and the count of bike rents is low in that season.
4. When the **weathersit** is good, the count of bike rents increases which makes sense. As we all know that it is very difficult to ride bikes in rainy or cloudy condition and best to ride bikes in clear sky conditions.
5. With **year** pass by there might be chance that the popularity of boombikes would have increase which will increase the count of bike rents
6. According to me, **mnth** might be correlated to season as the seasons are fixed according to months. Hence, we can delete **mnth**.
7. **Why is it important to use drop\_first = True during dummy variable creation?**

**Ans :** When we try to convert the categorical variable to numeric variable we do one hot encoding.

For one hot encoding we create dummy variables. We use drop\_first = True to help reducing

the extra columns during dummy variable creation. Hence it will help us to reduce the correlation created among the dummy variables.

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

**Ans :** Feeling temperature and air temperature are most correlated to the target variables.

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

**Ans :** a) **Linear relationship** between x and y - pair-wise scatterplots may be helpful in validating the linearity assumption as it is easy to visualize a linear relationship on a plot.

1. No **multicollinearity** in independent variables – We used heatmap as well as VIF to check and remove any multicollinearity in independent variables.
2. Errors are **normally distributed** with mean = 0. This can be validated by plotting **distplot** on error (y\_true - y\_pred) and observing that distribution is normal or not.
3. **Autocorrelation in error** - In our dataset we can see that Durbin-Watson = 2.057 which means that error is not auto correlated. We know that if Durbin Watson value is close to 2 then errors will not be autocorrelated.
4. To check **Homoscedasticity** - Plotting residual plot to observe that as the value in x-axis increases the residual in y is random or not.
5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?**

**Ans :** Top 3 features contributing significantly towards explaining the demand are :

1. **Fall season** – we can say that during fall season, the demand increases.
2. **Summer season** – we can say that during summer season, the demand increases.
3. **Light Rain / Light Snow** – we can say that when there is less or no rain and snow, the demand increases as it is best suited for bike riding.

***General Subjective Questions***

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

**Ans :** **Regression analysis** predicts the value of a dependent variable based on the known value of the independent variable. There are two types of Linear Regression:

**Simple Linear Regression**: It is a statistical method that allows to summarize and study relationships between 2 continuous variables.

**Multiple Linear Regression:** Similar to Simple Linear Regression, this technique uses several explanatory variables to predict response of the outcome variable.

**Equation of Simple Linear Regression**:

Random error component

Linear component

– Dependent variable

– Population Y intercept

– Population Slope coefficient

– Independent variable

– Random Error term

**Assumptions of Linear Regression:**

a) There should be a linear relationship between dependent variable and independent variable.

b) Errors should be normally distributed.

c) There should not be any multicolinearity between the errors.

d) Error should not be auto-corelated to each other.

e) Error terms should show homoscedasticity i.e. Errors should have constant variance.

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

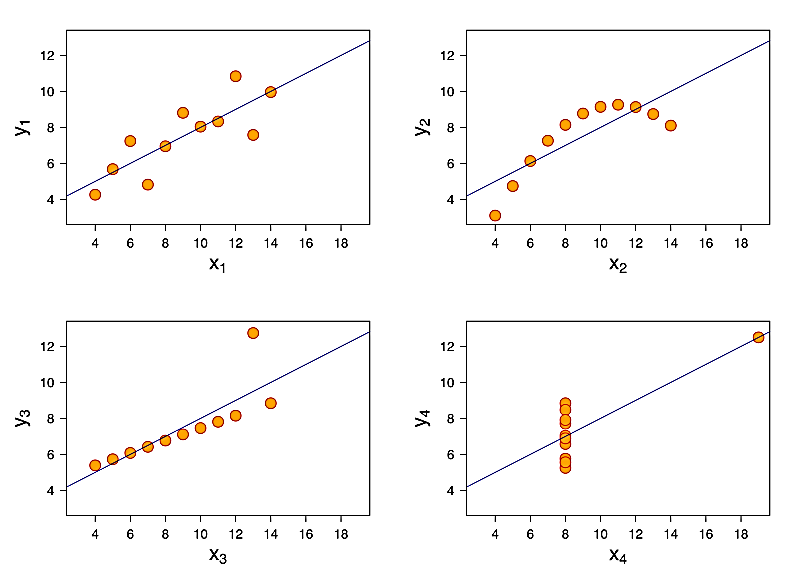
**Ans :** Linear regression has quite a few shortcomings such as :

a) It is sensitive to outliers.

b) It models the linear relationships only.

c) A few assumptions are required to make the inference.

To explain this phenomena, **Francis Anscombe** constructed **Anscombe’s Quartet** to illustrate the **importance**of **plotting the graphs**before analyzing and model building, and the effect of other **observations on statistical properties**.

 The four graphs can be described as:

a) **Graph 1 :**this **fits**the linear regression model pretty well.

b) **Graph 2 :** this **could not fit**linear regression model on the data quite well as the data is non-linear.

c) **Graph 3 :**shows the **outliers**involved in the dataset which **cannot be handled**by linear regression model.

d) **Graph 4 :**shows the **outliers**involved in the dataset which **cannot be handled**by linear regression model.

**3. What is Pearson’s *r* ?**

**Ans :** Pearson’s *r*  is a measure of the strength of a linear association between two variables.

– Pearson correlation coefficient

– values of the x-variable in a sample

– mean of the values of the x - variable

– values of the y-variable in a sample

– mean of the values of the y - variable

The Pearson correlation coefficient, r, can take a range of values from **+ 1 to - 1**.

a) **Value 0** indicates that there is no association between the two variables.

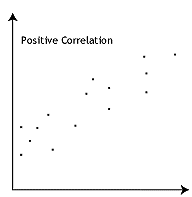
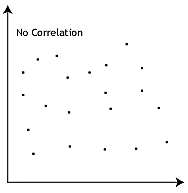
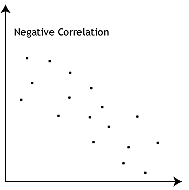
b) **Value greater than 0** indicates a positive association i.e. as the value of one variable increases,

the value of the other variable also increases.

c) **Value less than 0** indicates a negative association i.e. as the value of one variable increases,

the value of the other variable decreases.

This is shown in the diagram below:



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

**scaling and standardized scaling?**

**Ans :** Feature scaling is a method used to normalize the range of independent variables or features of data.

We need to scale features because of two reasons:

a) Ease of interpretation

b) Faster convergence for gradient descent methods

Features can be scaled using two very popular method :

a) **Standardizing :** The variables are scaled in such a way that their mean () is 0 and

standard deviation () is 1.

b) **MinMax Scaling (also Normalized scaling)** : The variables are scaled in such a way that all the values lie between 0 and 1 using the maximum and the minimum values in the data.

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

**Ans :** As we know that,

According to the above formula, we can say that if the denominator = 0 then VIF = infinity.

For denominator to be 0, **R2 should be equal to 1** which means that the variable must be perfectly correlated to other variables.

We can say that if R2 =1 or variable is perfectly correlated then only the value of VIF will be equal

to infinity.

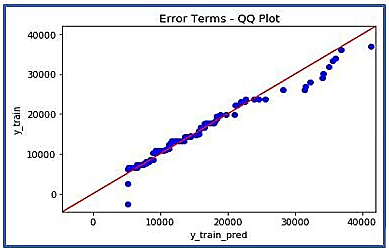
**6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.**

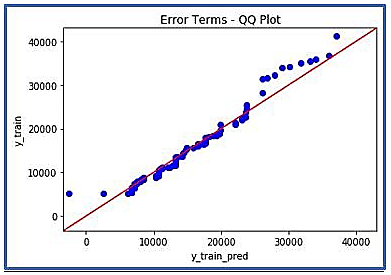
**Ans :** A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second data set.

Below are the possible interpretations for two data sets.

a) **Similar distribution :** If all point of quantiles lies on or close to straight line at an angle of 45°

from X – axis

 b) **Y - values < X - values :** If y-quantiles are lower than the x-quantiles.

 c) **X - values < Y - values :** If x-quantiles are lower than the y-quantiles.

d) **Different distribution :** If all point of quantiles lies away from the straight line at an angle of 45°

from X - axis.