Submitted by : Abs Varkey (C46)

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?***

   Response: The following are the key inferences –

   - **season:** season3 had the max demand among all seasons at approx 32% of the demand. The median was approx 5000 booking (across 2 years). This was followed by season2 & season4 with 27% & 25% of total booking. This indicates, season can be a good predictor for the dependent variable.

   - **years:** Almost 62% of the bike booking were happened in the second year. In comparison to the previous year there is an approx 24% increase in demand. This would imply there is growing trend in demand for shared bikes over the last two years.

   - **mnth:** Almost 10% of the bike booking were happening in the months 5,6,7,8 & 9 with a median of over 4000 booking per month. This indicates, mnth has some trend for bookings and can be a good predictor for the dependent variable.

   - **weathersit:** Almost 69% of the bike booking were happening during 'weathersit1 with a median of close to 5000 booking (for the period of 2 years). This was followed by weathersit2 with 30% of total booking. This indicates, weathersit does show some trend towards the bike bookings and can be a good predictor for the dependent variable.

   - **holiday:** Almost 98% of the bike booking were happening when it is not a holiday which means this data is clearly biased. This indicates, holiday may not be a good predictor for the dependent variable.

   - **weekday:** weekday variable shows very close trend (between 13.5%-14.8% of total booking on all days of the week) having their independent medians between 4000 to 5000 bookings. This variable can have some or no influence towards the predictor.

- **workingday:** Almost 70% of the bike booking were happening in ‘workingday’ with a median of close to 5000 booking (for the period of 2 years). This indicates, workingday can be a good predictor for the dependent variable

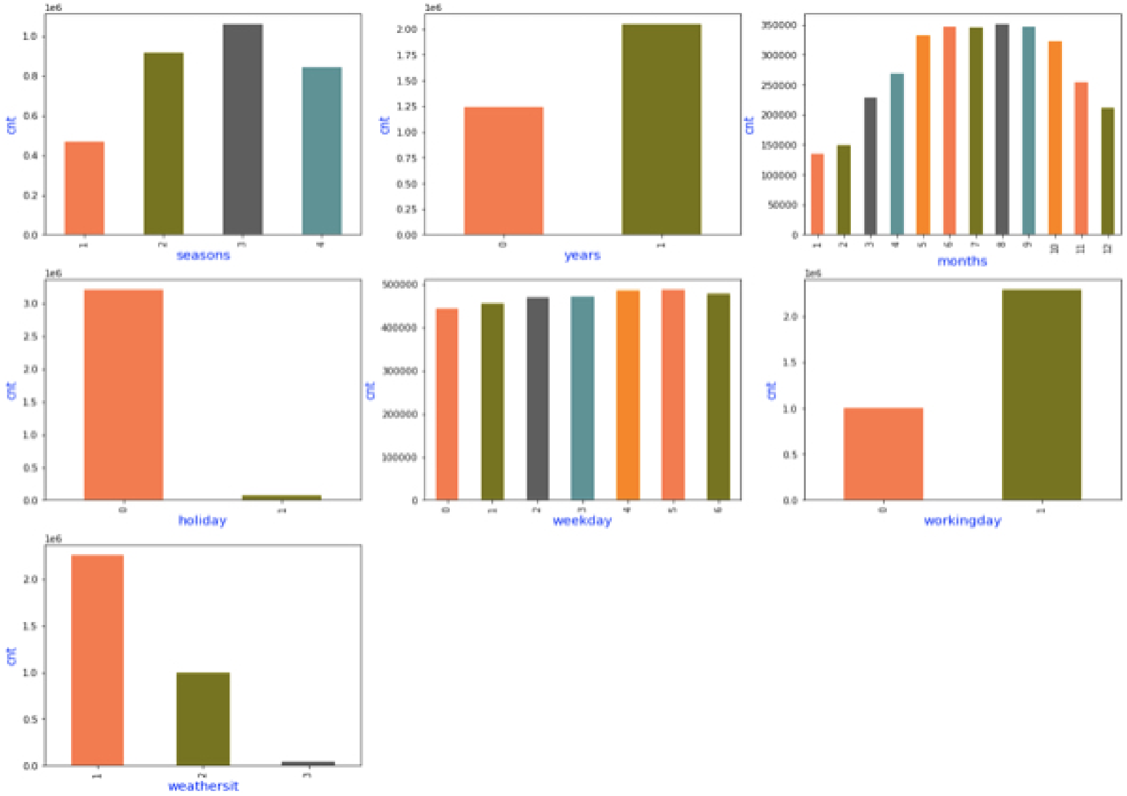


Fig: Bar Chart of Demand (cnt) Vs Categorical Variables

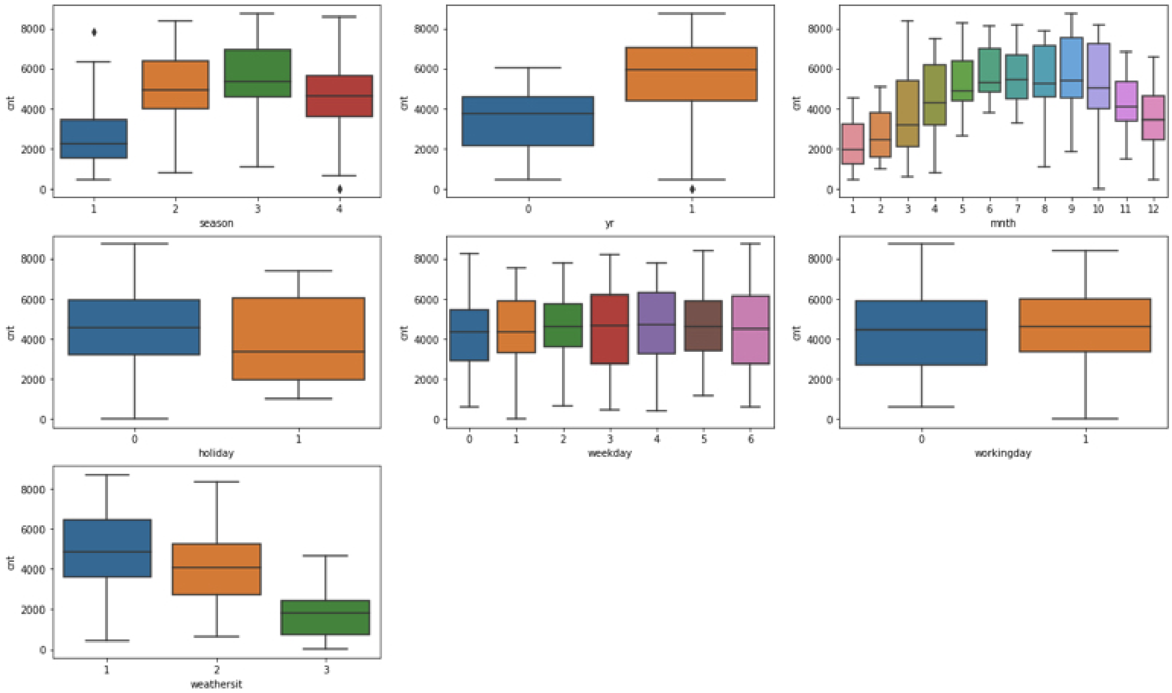


Fig: Box plot of Demand (cnt) Vs Numerical Variables

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

   Response: During the time of dummy variable creation, the pandas get_dummies function creates as many variables as the number of levels the original categorical variable has. The key objective of the regression model is to have the variables independent of each other .The drop_first argument will drop this extra column created , thereby reducing the correlation created among the dummy variables created for the category. Dropping a dummy variable can still explain all the levels by the remaining dummy variable associated with the category thereby also eliminating redundant variables from the final model.

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

   Response: Among numerical variables , temp and atemp appears to have the highest co-relation with the target variable cnt.

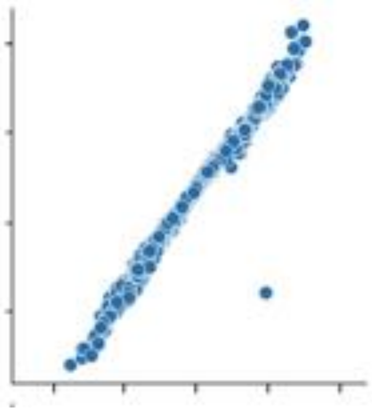


Fig: Pair plot of temp Vs atemp

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

   Response: Some of the key assumptions of Linear Regression that need to be validated post training the model are –

1. **There should be a Linear relationship between independent and dependent variables (linear line or a linear plane) –** The adjusted R-squared value of around 0.835 in our model is indicative of a good linear relationship i.e. Almost 83.5% of the variations is explained by the model. Also on observing the Residual Vs Fitted plot ,the residuals shows no visible pattern. The error terms just appear to be evenly distributed noise around zero (horizontal Red Line) which further helps in clarifying the assumption.
2. **Error terms are normally distributed -** This was done by plotting the histogram of the error terms (residuals) and observing it's distribution.

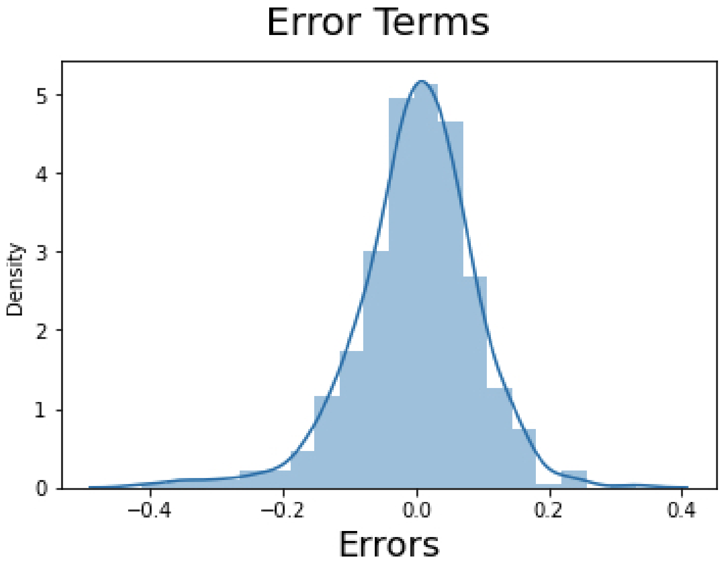


Fig: Distribution plot of Error terms (residuals)

3. **Error terms are independent of each other –** The Durban-Watson value is an indicator of independence of the variables with respect to each other. Our observed value for our final model (lm_5) was 2.044 which is within the reasonable range.
4. **Error terms have constant variance (homoscedasticity) -** The scatter plot for residuals shows no visible pattern. The error terms just appear to be evenly distributed noise around zero which is ideal.

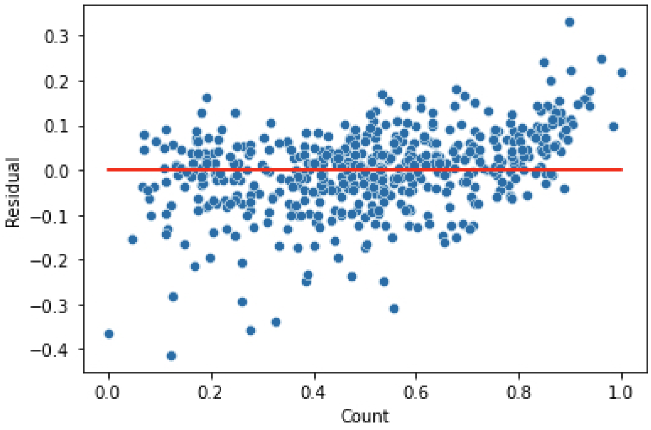


Fig: Scatter plot of Error terms

5. **Multicollinearity should be absent** – The VIF (Variance Inflation Factor) checks was used to check multicollinearity between predictor variables. VIF helps explaining the relationship of one independent variable with all the other independent variables

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

Response: The following are the top three features identified by the model -

- **Temperature (temp)** - unit increase in temp variable could increase bike bookings by 0.5209 units assuming all other variables remain constant.
- **Weather Situation 3 (weathersit_3) -** a unit increase in Weathersit3 variable could decrease bike bookings by 0.2869 units assuming all other variables remain constant.
- **Year (yr) -** a unit increase in year variable could increase bike bookings by 0.2328 units assuming all other variables remain constant.

General Subjective Questions

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

   Response: Regression is used where the output variable to be predicted is a continuous variable e.g. demand for bikes. It falls under the supervised learning methods where previous labelled data is used as the basis to build the model.

   Linear regression is a form of predictive modelling technique which tells us the linear relationship between the dependent (target variable) and independent variables (predictors). It describes the straight line relationship between two variables.

   They can be represented by the following equation

   - y = i + c*X (Simple Linear Regression - Explains the relationship between a dependent variable and one independent variable using a straight line.)
   - y = i + c1*X1 + c2*X2 +…..cn*Xn (Multi Linear Regression - Explains the relationship between a dependent variable and multiple independent variables.)

   Where, X and y are two variables on the regression line.

   c = Slope of the line

   i = y-intercept of the line

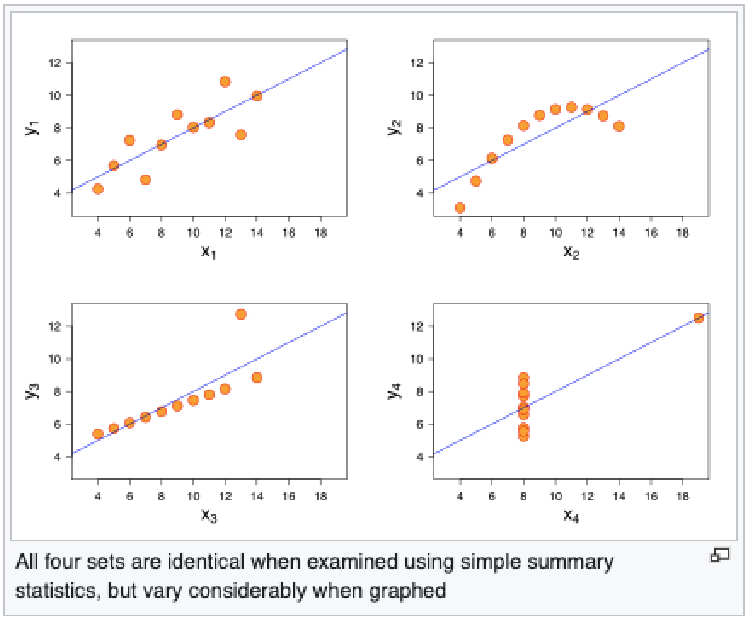
   X = Independent variable from dataset that are used as predictors

   y = Dependent or target variable from dataset that we want to predict

   Thus the objective of a Linear regression model is to generate the best fit line represented by a linear equation as above. This is done by deriving the expression that has the minimal RSS (Residual sum of squares). The Residuals for a data point are computed by subtracting predicted value of dependent variable (y-predicted) from actual value of dependent variable (y-actual) provided in the data set.

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

Response: The Anscombe quartet was constructed by statistician Francis Anscombe in 1973 to primarily illustrate the importance of visualization and observation before analysing the model building. To emphasise this he built 4 data sets(quartet) with similar statistical information of variance and mean for all the x and y point in the data sets. However on plotting these sets in simple scatter chart (x,y) the observations were quite different in each of the four cases .



All four sets are identical when examined using simple summary statistics, but vary considerably when graphed

Fig: Anscombe’s Quartet – Source Wikipedia (https://en.wikipedia.org/wiki/Anscombe%27s_quartet)

- Dataset 1: Fit the linear regression model well
- Dataset 2: Did not fit the linear regression model well as the plot represented more or a curve (non-linear)
- Dataset 3: Showed outliers that could not be handled by the linear regression model
- Dataset 4: Also showed outliers that could not be handled by the linear regression model

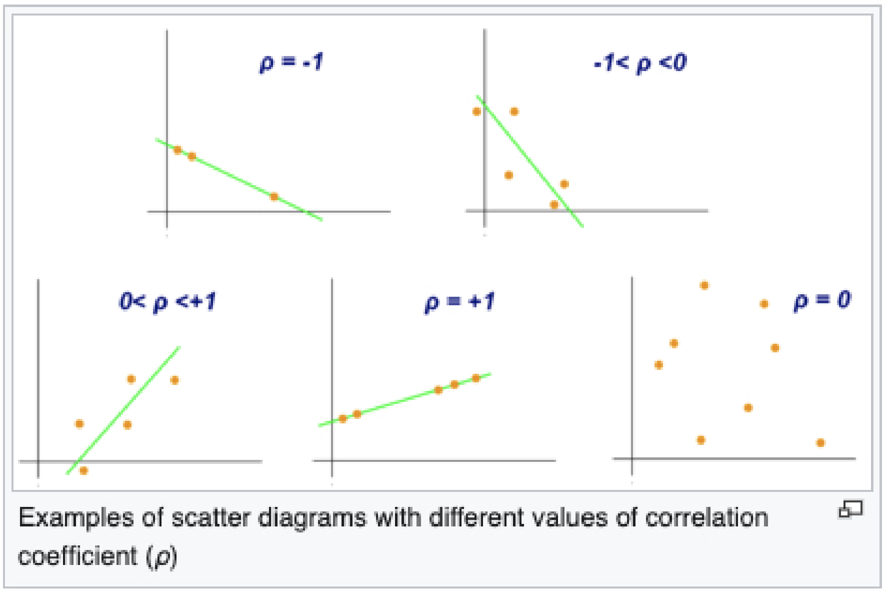
Thus all the anomalies of the data such as outliers, diversity of the data, linear separability etc should be visualised before implementing the Linear regression model determining if the Linear regression model is indeed a good fit for the provided data set.

3. *What is Pearson’s R?*

Response: Pearson’s R or Pearson’s correlation coefficient (r or *p*) is a way of measuring the linear correlation between two variables within the data set. It is indicative of the dependency of one variable over the other. It is measured and described on a scale of -1 to 1, where

- **0 :** indicated no correlation between the two variables
- **Between 0 to 1:** Indicates a positive correlation , with 1 indication almost 100% positive correlation.
- **Between 0 to -1:** Indicates a negative correlation, with -1 indication almost 100% negative correlation.

It is used in statistics to test the significant relationship between two variables and measure how close the observations are to a line of best fit. It helps identify if the slope of the best fit line is positive (+ r value) or negative (- r value)



Examples of scatter diagrams with different values of correlation coefficient (ρ)

Fig: Pearson’s correlation coefficient (r) – Source Wikipedia (https://en.wikipedia.org/wiki/Pearson_correlation_coefficient)

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

Response: Generally the dataset provided contains variables/features of varying magnitude of values. The machine learning algorithms tend to associate a higher weightage to larger values in comparison to lower values. This could lead the model to have a higher coefficient bias towards variables exhibiting higher values.

Feature Scaling in machine learning is used to standardize the independent features present in the data to a fixed range.

Thus scaling affects the coefficients and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc.

There are about five major methods to scale the variables in ML -

- **Standardisation scaling -** basically brings all of the data into a standard normal distribution with mean zero and standard deviation one.

  X*new* = X−mean(X)/ sd(X) , sd – Standard Deviation

- **MinMax scaling -** on the other hand, brings all of the data in the range of 0 and 1.

  X*new* = X−min(X)/max(X)−min(X)

- **MaxAbs Scaler -** Works in a very similar fashion and MinMax, but scales in a way that the training data lies within the range [-1, 1] by dividing through the largest maximum value in each feature. It is meant for data that is already centred at zero or sparse data.

- **Normalization -** the process of scaling individual samples to have unit norm.

  This process can be useful if you plan to use a quadratic form such as the dot-product or any other kernel to quantify the similarity of any pair of samples.

$$X_{new} = X - mean(X)/max(X) - min(X)$$

- **Robust Scaling -** In this method, all the data points are subtracted by the median value and then divide it by the Inter Quartile Range(IQR) value. It centres the median value at zero and is robust in manging outliers.

$$X_{new} = X - median(X)/IQR$$

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

Response: It is possible to have a have high collinearity between multiple variables even if no pair of variables have a high corelation. VIF (variance inflation factor) is used to measure the collinearly between multiple predictor variables within the dataset. A variable with a high VIF means it can be largely explained by other independent variables. Thus a variable with an infinite value would mean that is highly correlated with other variables in the dataset.

VIF calculates how well one independent variable is explained by all the other independent variables combined and is computed as below –

$$VIFi = 1 / (1-Ri\^{} 2)$$

where 'i' refers to the i-th variable which is being represented as a linear combination of rest of the independent variables.

In case of a prefect co-relation the R-Squared value would be 1, which would make the denominator 0 and thus the VIF value as infinity.

High VIF scoring variables would need to be removed (after checking the p values) , implying that their impact on the outcome can largely be explained by other variables. However, variables with a high VIF or multicollinearity may be statistically significant $p<0.05$, in which case you will first have to check for other insignificant variables before removing the variables with a higher VIF and lower p-values

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

Response: Q-Q (Quantile-Quantile) plots the quantiles of a sample distribution (Y axis) against the quantiles of a theoretical distribution(e.g. Normal distribution) on the X axis.

If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the y = x line (also called as identity line or 1:1 line). If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line, but not necessarily on the line y = x.

Q-Q can be used to -

- Compare graphically the distributions -
    - of two different data sets e.g. Train and Test.
    - of the features with respect to how much it deviates from a statistical reference model.
    - of the model with respect to how much it deviates from a statistical reference model.
- Plot skewness of the dataset or skewness of the features within the data set

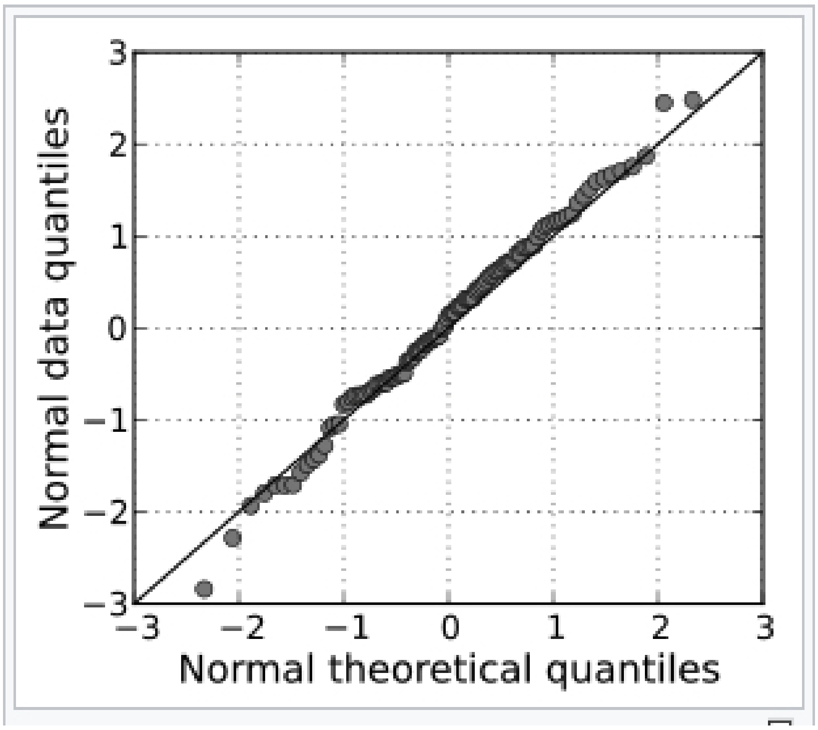


Fig: Q-Q linear plot – Source Wikipedia
https://en.wikipedia.org/wiki/Q%E2%80%93Q_plot

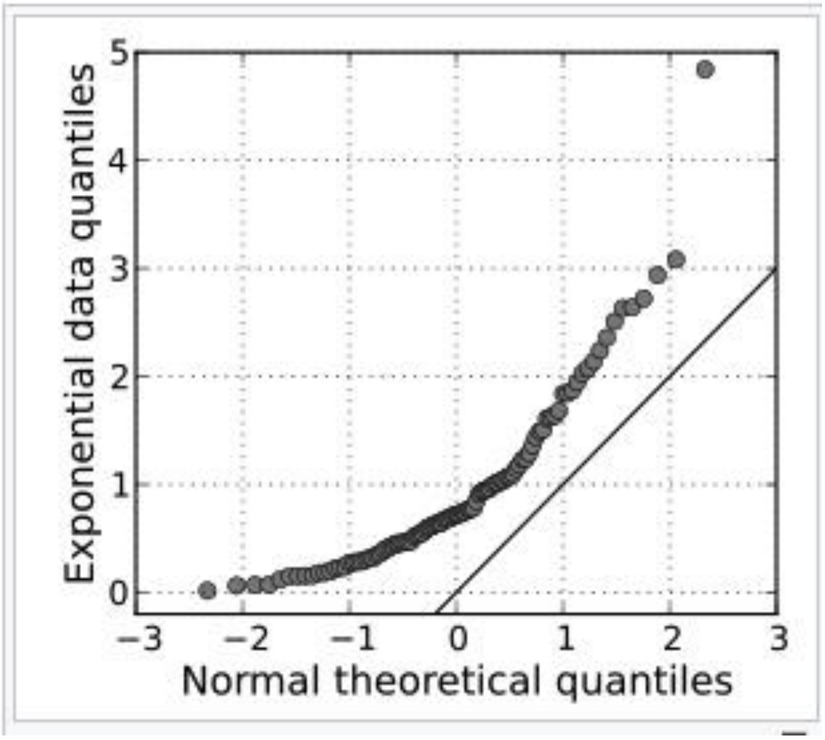


Fig: Q-Q non linear plot – Source Wikipedia
https://en.wikipedia.org/wiki/Q%E2%80%93Q_plot