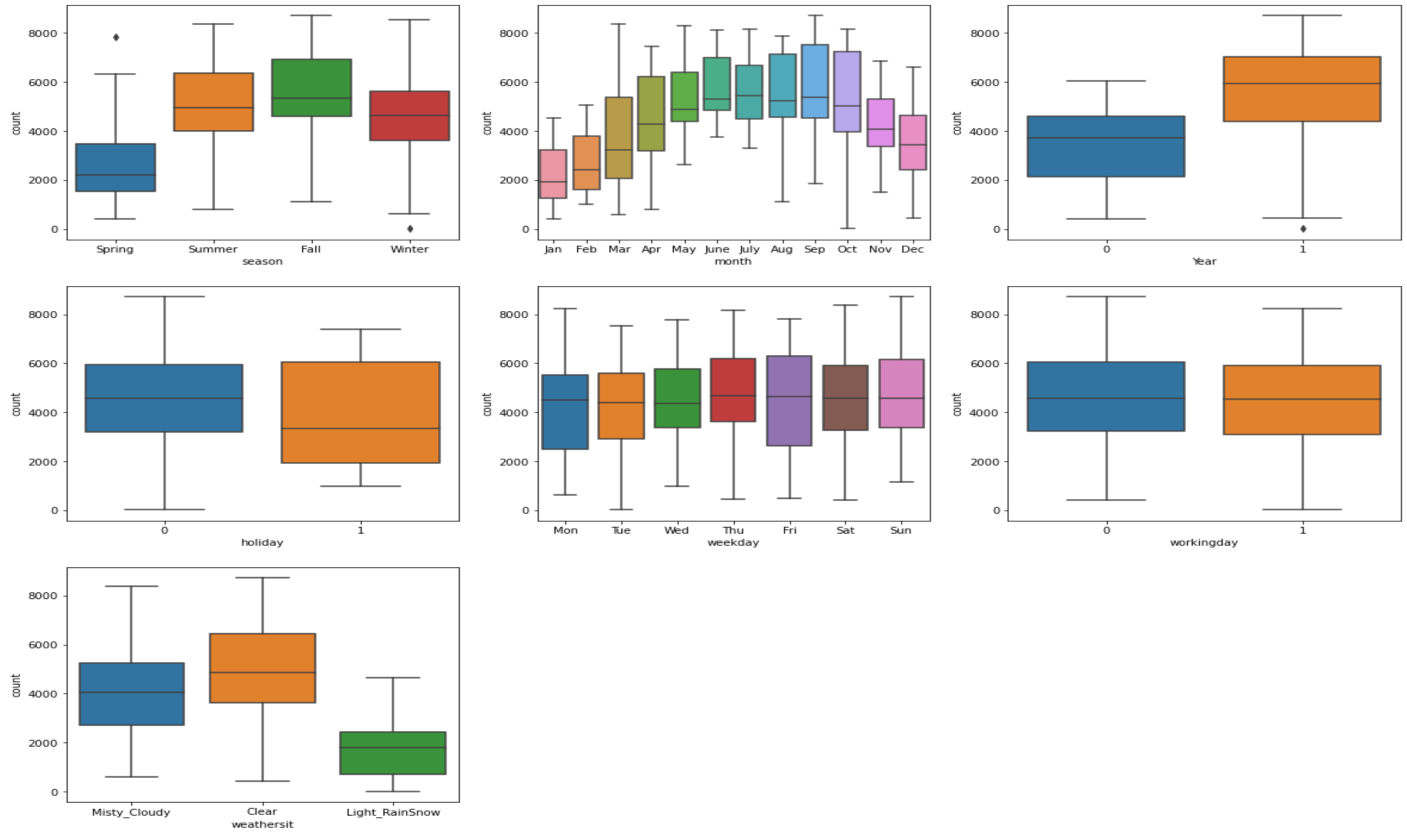
# SUBJECTIVE ASSIGNMENT

## (SUBMITTED BY: GURPREET KAUR, DSC43)

## Assignment-based Subjective Questions

### Q1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

In the Boom-bike assignment dataset, the categorical variables are **Season, Weather Situation, Months, Weekday, Workingday, Holiday, Year**. These were visualized using a boxplot. These variables had the following effect on dependent variables:-



1. **Season** – The boxplot showed that spring season had least value of "count" whereas FALL had maximum value of "count". Summer and winter had intermediate value of "count". This indicates, season can be a good predictor for the dependent variable.
2. **Weathersit** -There are no users when there is heavy rain/ now indicating that this weather is extremely unfavourable. Highest count was seen when the weathersit was 'Clear, Partly cloudy'. This indicates, weather can be a good predictor for the dependent variable.
3. **Holiday** - rentals reduced during holiday.
4. **Month** - September saw highest no of rentals while December saw least. The weather situation in December is usually heavy snow. December and January is the beginning of Spring and still winter effect

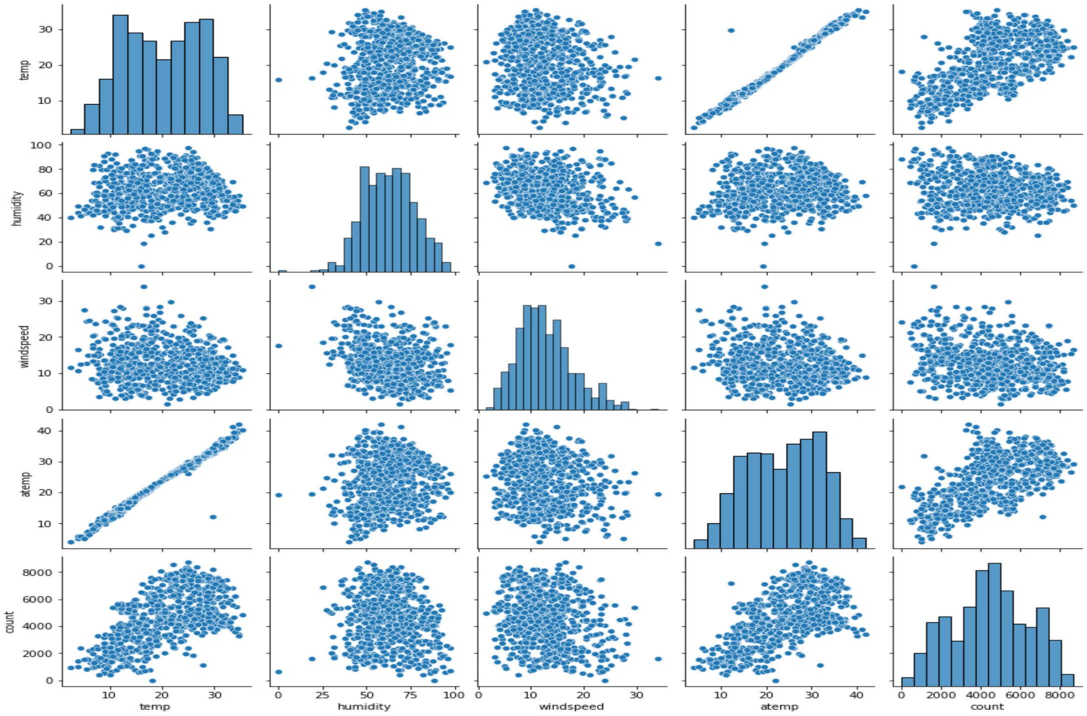
will be there, snow fall will be there. Therefore, the demand for shared bikes is less. From Feb onwards again it pickups.

5. **Year** - The number of rentals in 2019 was more than 2018
6. **Working day**: Almost 69% 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
7. **Weekdays**: This predictor is not affecting more on the target variable.

## Q2. Why is it important to use drop_first=True during dummy variable creation?

- **drop_first=True** is important to use, as it helps in **reducing the extra column** created during dummy variable creation. Hence it **Reduces** the **CORRELATIONS** created among dummy variables.
- If we don't drop the first column then dummy variables will be correlated **(Redundant).** This may affect some models adversely and the effect is stronger when the cardinality is smaller. For example, iterative models may have trouble converging and lists of variable importance may be distorted.
- Another reason is, if we have all dummy variables it **leads to Multi-collinearity** between the dummy variables. To keep this under control, we lose one column.
- **For Example:** Let’s say we have 3 types of values in Categorical column and we want to create dummy variable for that column. If one variable is not furnished and semi-furnished, then it is obvious it will be unfurnished. So, we do not need 3rd variable to identify the unfurnished feature.
- Hence if we have **categorical variable with n-levels, then we need to use n-1 columns to represent the dummy variables.**

## Q3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?



From above pairplots we can observe that **“temp” and “atemp”** are the two numerical variables, which are **Highly Correlated (0.63)** with the target variable **“count”.** There is **Linear Relationship** between temp and atemp. Both of the parameters simultaneously cannot be used in the model due to Multicollinearity.

## Q4. How did you validate the assumptions of Linear Regression after building the model on the training set?

The ***Assumptions Of Linear Regression*** are :

1. **Linearity**: There is a linear relationship between independent and dependent variable. Plot a Scatter plot for X_train_index vs residuals. If there is a linearity in the plots then the assumption holds good. The equally spread residuals around a horizontal line without distinct patterns are a good indication of having the linear relationships.
2. Assumption about the residuals:
    a. **Normality**: Plot distribution plot, Histogram for residuals (sns.distplot((y_train - y_train_pred), bins = 20)). If the distribution of residuals is normally distributed then the assumption of Normality holds good.
    b. **Zero mean**: In the distribution plot if the error terms are normally distributed with mean equal to 0, the assumption holds good.
    c. **Homoscedasticity** (Constant variance) : Check this assumption by examining the scatterplot of “residuals versus X_train_index”; the variance of the residuals should be the same across all values of the x-axis. If the plot shows a pattern then variances are not consistent, and this assumption has not been met.
    d. **Independent error**: By examining the scatterplot of “residuals versus X_train_index”; There should not be any patterns. The pairwise correlation is zero.
3. Assumptions about the estimator:
    a. **Independence/No Multicollinearity**: stating that the variables should be independent of each other i.e no correlation should be there between the independent variables. **(VIF score less than 5 and Heatmap** of **Correlation Matrix** plotted.
    b. Independent variables are measured without errors: If the **p-value of each predictors** is equal to zero or less than 0.05 then we can say the variable is significant.
4. **No Autocorrelation** stating that the error terms(yact – ypred) should be independent of each other. (Durbin-Watson Test)
5. Statistical Analysis Of Final Model (R2, Adj R2, MSE, RMSE, VarianceScore, AIC, BIC, F-Statistics)

## Q5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

The top 3 features contributing significantly towards explaining the demand of the shared bikes are **Temperature(Temp), Year (Yr) & Weather** (**weathersit_Light_RainSnow, Misty_Cloudy** )

1. **Temperature (temp)** - A coefficient value of ‘0.4515’ indicated that a unit increase in temp variable increases the bike hire numbers by **0.4515** units.
2. **Year** - A coefficient value of ‘0.2341’ indicated that a unit increase in year variable increases the bike hire numbers by **0.2341** units. There is a **High demand for the shared bikes in coming years**. We can see demand for bikes will be more in upcoming year 2019 than 2018, so just focus as there is increase in

2019 and might be facing dips in their revenues due to the ongoing Corona pandemic and when the conditions become normal the 'Boom-Bikes' will get very good profit in shared bike business.

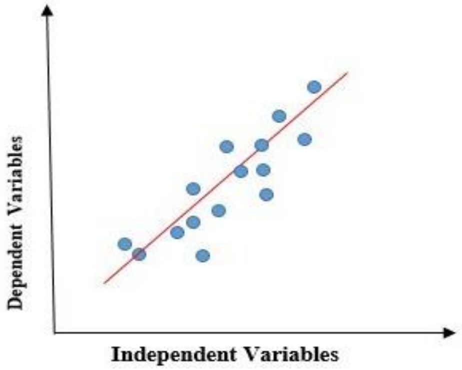
3. **Weather**:
   **a) weathersit_Light_RainSnow** - A negative coefficient value of '-0.2864' indicated that, w.r.t weathersit_Light_RainSnow, a unit increase in weathersit_Light_RainSnow) variable decreases the bike hire numbers by **0.2864 units**. where, weathersit_3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
   **b) weathersit_Misty_Cloudy** - A negative coefficient value of **'-0.0811'** indicated that, w.r.t weathersit_Misty_Cloudy, a unit increase in weathersit_Misty_Cloudy variable decreases the bike hire numbers by 0.0811 units.

- When there a Misty weather and light-snow the most of the people are not coming out. So there is a less demand for shared bikes. Now seeing to weathersit variable, we have got negative coefficients for Mist +cloudy and Lightsnow weather so we can **give offers to avoid inflation in Bike-Rental business**. Also, they could probably use this time to service the bikes without having business impact.
- **The demand for shared bikes are more when the weather is clear, temp is high and there is least demand when there is snow fall or temp is low.**
- The demand for shared bike is reducing in spring season. The Boom bike company can give some spring season offers to increase the demand in this season. Expand the business in Spring season.

# General Subjective Questions

## Q1. Explain the linear regression algorithm in detail.

- Linear Regression is a **Supervised learning Algorithm**. It is method of finding the **Best Straight Line** fitting to the given data. i.e. finding the best **Linear Relationship** between the **Independent And Dependent Variable.**
- Linear Regression model used to predict the unseen dependent variable by using the independent variables. The Linear Regression Algorithm uses ***Least Sum of Residuals Squares*** to find the best linearly fitted model.
- Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this **Regression Technique** finds out a linear relationship between x (input) and y(output).

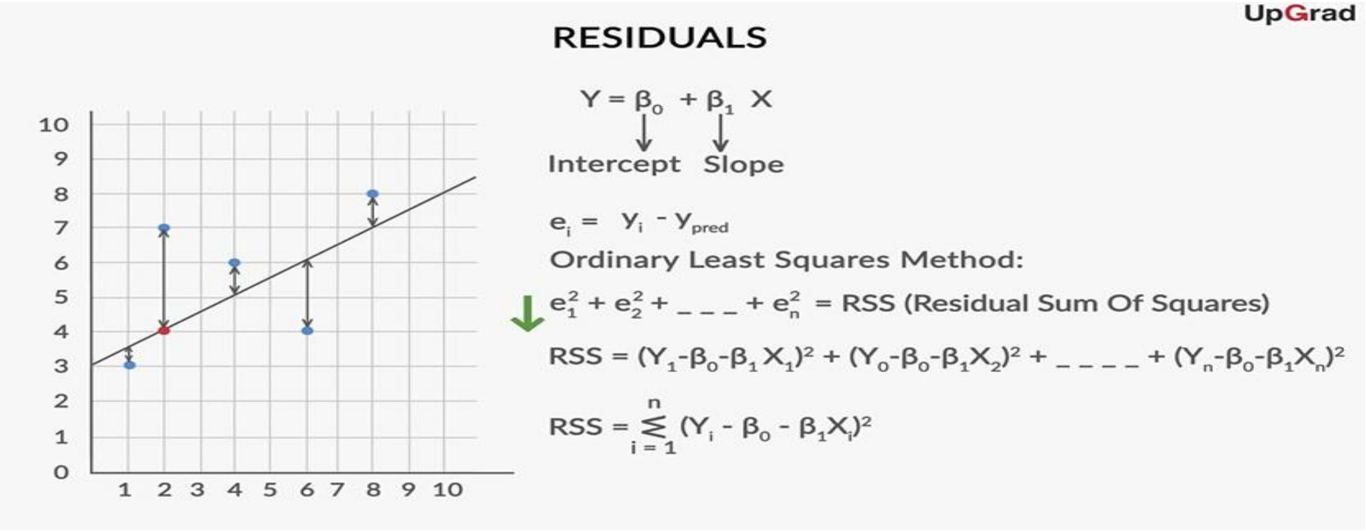


The above graph presents the linear relationship between the dependent variable and independent variables. When the value of **x (independent variable**) increases, the value of **y (dependent variable)** is likewise increasing.

## BEST-FIT LINE:

The best-fit line is found by minimising the expression of **RSS (Residual Sum of Squares)** which is equal to the sum of squares of the residual for each data point in the plot. Residuals for any data point is found by subtracting predicted value of dependent variable from actual value of dependent variable.

The Linear Regression Algorithm will find these co-efficient by **Gradient Decent Method (Iterative Process)** using **Sum of Least Square Error to find the best linear fit model.**



RESIDUALS

$Y = \beta_0 + \beta_1 X$

Intercept Slope

$e_i = y_i - y_{pred}$

Ordinary Least Squares Method:

$e_1^2 + e_2^2 + \_\_\_ + e_n^2$ = RSS (Residual Sum Of Squares)

$RSS = (Y_1-\beta_0-\beta_1X_1)^2 + (Y_o-\beta_0-\beta_1X_2)^2 + \_\_\_\_ + (Y_n-\beta_0-\beta_1X_n)^2$

$RSS = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1X_i)^2$

**<u>COST FUNCTION</u>:** Cost function **Optimizes the Regression Coefficients** or weights and measures how a linear regression model is performing. The cost function is used to find the accuracy of the **mapping function** that maps the input variable to the output variable. This mapping function is also known as the **Hypothesis function.**

In Linear Regression, **Mean Squared Error (MSE**) cost function is used, which is the average of squared error that occurred between the predicted values and actual values.

By simple linear equation **y=mx+b** we can calculate MSE as:

$$MSE = \frac{1}{N}\sum_{i=1}^{n}(y_i - (a_0 + a_1x_i))^2$$

Using the MSE function, we will change the values of $a_0$ and $a_1$ such that the MSE value settles at the minima. Model parameters $\boldsymbol{x_i}$, ***(a₀,a₁)*** can be manipulated to minimize the cost function. These parameters can be determined using the gradient descent method so that the cost function value is minimum.

Cost function(J) of Linear Regression is the **Root Mean Squared Error (RMSE)** between predicted y value (pred) and true y value (y).

$$minimize\frac{1}{n}\sum_{i=1}^{n}(pred_i - y_i)^2$$

$$J = \frac{1}{n}\sum_{i=1}^{n}(pred_i - y_i)^2$$

There are **<u>Two Types</u> of Linear regression** based on the number of independent variables.

1. **Simple Linear Regression**: where there is only one independent variable (x) and a target variable (y)
2. **Multiple Linear Regression:** where there can be two or more independent variables (x1, x2, x3, …, xn) and a target variable (y). The independent variables are known as “predictor variables” and the dependent variables are known as “output” or “target” variables.

**<u>Gradient Descent:</u>**

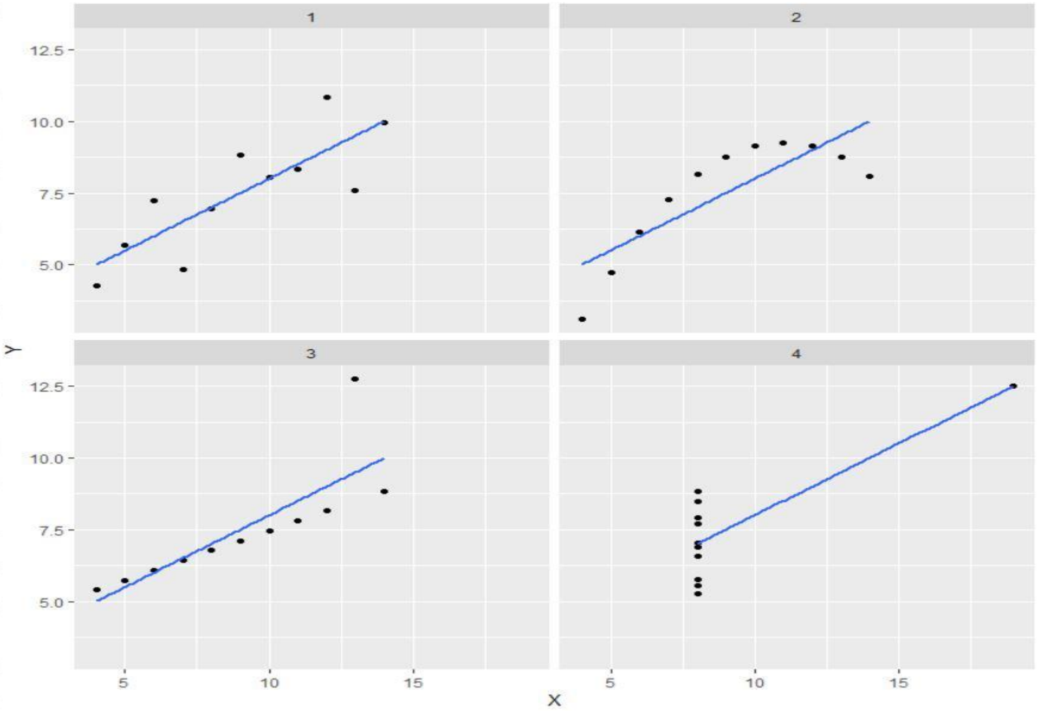
To update θ1 and θ2 values in order to reduce **Cost function (minimizing RMSE value)** and achieving the **best fit line the model uses Gradient Descent**. The idea is to start with random θ1 and θ2 values and then iteratively updating the values, reaching minimum cost.

**<u>Assumptions of Linear regression</u>**:

- There is a linear relationship between X and Y
- Error terms are normally distributed with mean zero.
- Error terms are independent of each other.
- Error terms have constant variance (homoscedasticity).

## Q2. Explain the Anscombe’s quartet in detail.

**Anscombe’s Quartet** can be defined as a group of four data sets which are **nearly identical in simple descriptive statistics,** but there are some peculiarities in the dataset that **fools the regression model** if built. They have very **different distributions** and **appear differently** when **plotted on scatter plots.**



**Explanation of this output**:

- In the **First One(Top Left)** if you look at the scatter plot you will see that there seems to be a linear relationship between x and y.
- In the **Second One(Top Right)** if you look at this figure you can conclude that there is a non-linear relationship between x and y.
- In the **Third One(Bottom Left**) you can say when there is a perfect linear relationship for all the data points except one which seems to be an outlier which is indicated be far away from that line.
- Finally, the **Fourth One(Bottom Right)** shows an example when one high-leverage point is enough to produce a high correlation coefficient.

Thus, In the **above plot**, the **first one** seems to be doing a decent job, the **second one** clearly shows that linear regression can only model linear relationships and is incapable of handling any other kind of data. The **third and fourth** images showcase the linear regression models sensitive to outliers. If outliers are not there, we could have got the great line through the data points. So, we should not run a regression without having a good look at our data.

**Anscombe's Quartet illustrate the importance of plotting the graphs, visualizing the data** that can help you **identify the various anomalies** present in the data like outliers, diversity of the data, linear separability of the data, etc. before analysing and model building, and the **effect of other observations on statistical properties**. There are these four data set plots which have nearly same statistical observations, which provides same statistical information that involves variance, and mean of all x,y points in all four datasets.

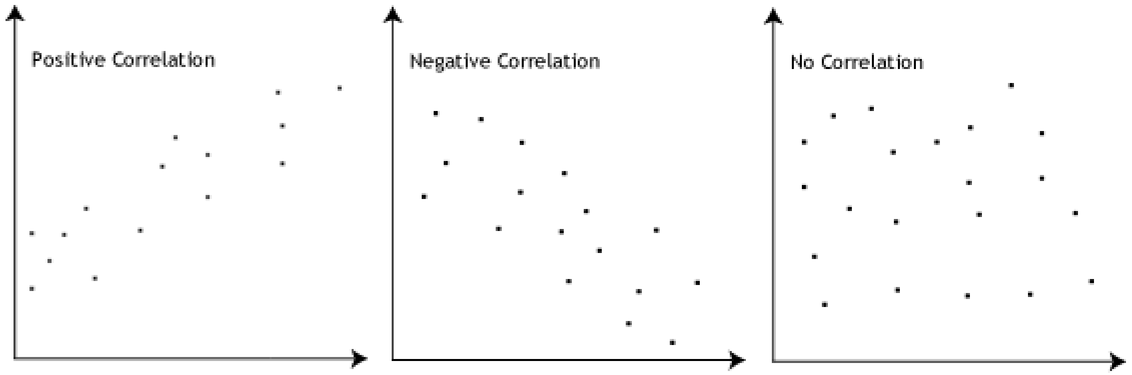
Also, the Linear Regression can be only be considered a fit for the data with linear relationships and is incapable of handling any other kind of datasets.

## Q3. What is Pearson's R?

In statistics, the **Pearson Correlation Coefficient (PCC),** also referred to as **Pearson's r**, the **Pearson Product-Moment Correlation Coefficient (PPMCC), Or the Bivariate Correlation**, is a measure of **Linear Correlation** between two sets of data. It is the **Covariance** of two variables, **divided by the product of their standard deviations**; thus it is essentially a **normalised measurement of the covariance**, such that the result always has a value between **−1 and 1.**

**The Pearson's correlation coefficient varies between -1 and +1** where:

- **r = 1** means the data is perfectly linear with a positive slope ( i.e., both variables tend to change in the same direction)
- **r = -1** means the data is perfectly linear with a negative slope ( i.e., both variables tend to change in different directions)
- **r = 0** means there is no linear association
- **r > 0 < 5** means there is a weak association
- **r > 5 < 8** means there is a moderate association
- **r > 8** means there is a strong association



## Pearson r Formula:

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

Here,

- $r$=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

**Pearson's R or Pearson's correlation coefficient** is defined in statistics as the measurement of the strength of the relationship between two variables and their association with each other. Pearson's R calculates the effect of change in one variable when the other variable changes. The Pearson's R tries to find out two things, the **strength** and the **direction** of the relationship from the given sample sizes.

## Pearson Correlation Coefficient Formula:

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

**Where:**
**N** = the number of pairs of scores
**Σxy** = the sum of the products of paired scores
**Σx** = the sum of x scores
**Σy** = the sum of y scores
**Σx2** = the sum of squared x scores

**Σy2 =** the sum of squared y scores

The Pearson's R  returns the values between -1 and 1.
**Strength:** The stronger the association between the two variables, the Pearson's R value incline towards 1 or -1. Attaining values of 1 or -1 signify that all the data points are plotted on the straight line of 'best fit.' It means that the change in factors of any variable does not weaken the correlation with the other variable. If the Pearson's R lies near 0, the more the variation in the variables.

**Direction:** The negative and positive sign of the Pearson's R tells the direction of the line. The direction of the line indicates a positive linear or negative linear relationship between variables. If the line has an upward slope, the variables have a positive relationship. This means an increase in the value of one variable will lead to an increase in the value of the other variable. A negative correlation depicts a downward slope. This means an increase in the amount of one variable leads to a decrease in the value of another variable.

**Pearson's R Correlation co-efficient** is designed to find the **correlation** between the variables which shows linear relationship and it might not be a measure for if the relationship between the variables is non-linear.

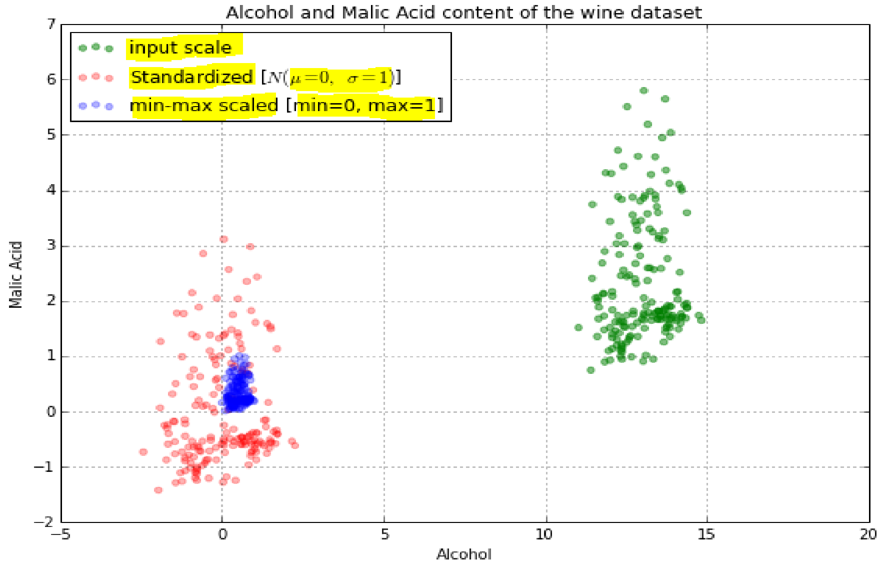
## Q4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

**Scaling is a step of data Pre-Processing of Machine Learning**. Scaling is applied to independent variables to **normalize** the data within a particular range. It also helps in **speeding up the calculations in an algorithm**. **Scaling reduced the iterative steps of Gradient Decent Algorithm to converge towards the best-fit Model.**

Most of the times, the collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units it leads to incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

It is important to note that **scaling just affects the coefficients** and none of the other parameters like t-statistic, F-statistic, p-values, R-squared, etc. are affected by scaling.

| 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.** |
| **6. This transformation squishes the n-dimensional data into an n-dimensional unit hypercube.** | **It translates the data to the mean vector of original data to the origin and squishes or expands.** |
| **7. It is useful when we don't know about the distribution** | **It is useful when the feature distribution is Normal or Gaussian.** |
| **8. It is a often called as Scaling Normalization** | **It is a often called as Z-Score Normalization.** |
| **9. Formula Used:** $x$ = x – min(x) / max(x) – min(x) | **Formula Used:** $x$ = x -mean(x) / sd(x) |



## Q5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

If there is **Perfect Correlation**, then **VIF (Variance Inflation Factor) = infinity**. This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get **R2 =1**, which leads to **1/(1-R2) infinity**. To **solve this problem** we need to **drop one of the variables** from the dataset which is causing this perfect **multicollinearity.**

The VIF gives how much the variance of the coefficient estimate is being inflated by collinearity VIF=1/1-R^2 , Where 1-R^2 is the R-square value of that independent variable which we want to check how well this independent variable is explained well by other independent variables. If that independent variable can be explained perfectly by other independent variables, then it will have perfect correlation and it's R-squared value will be equal to 1. So, **VIF = 1/(1-1) which gives VIF = 1/0 which results in "infinity.**

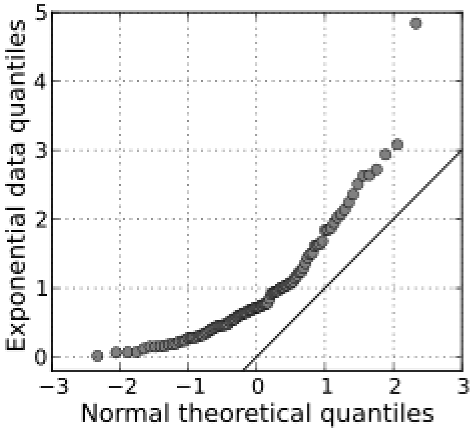
An **infinite VIF value indicates** that the corresponding variable may be expressed exactly by a **Linear** combination of other variables (which show an infinite VIF as well).

## Q6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Q-Q Plots **(Quantile-Quantile plots)** are plots of two quantiles against each other. A quantile is a fraction where certain values fall below that quantile. For example, the median is a quantile where 50% of the data fall below that point and 50% lie above it. The purpose of Q -Q plots is to find out if two sets of data come from the same distribution.

**<u>The use and importance of Q-Q Plot are:</u>**

1. Q-Q plots helps in a scenario of **linear regression** when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.
2. **Many distributional aspects like shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.**
3. A Q–Q plot is used to **compare the shapes of distributions,** providing a graphical view of how **properties such as location, scale, and skewness are similar or different in the two distributions.**



**A 45 degree angle is plotted on the Q-Q plot:**

- If the two data sets come from a common distribution, the points will fall on that reference line.
- If all point of quantiles lies away from the straight line at an angle of 45 degree from x -axis indicates the data sets come from different distribution

**The Q-Q Plot is used to answer the following questions:**

- Do two data sets come from populations with a common distribution?
- Do two data sets have common location and scale?
- Do two data sets have similar distributional shapes?
- Do two data sets have similar tail behavior?

If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line y = x. 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 plots can also be used as a graphical means of estimating parameters in a location-scale family of distributions.

# THANK YOU

## (SUBMITTED BY: GURPREET KAUR, DSC43)