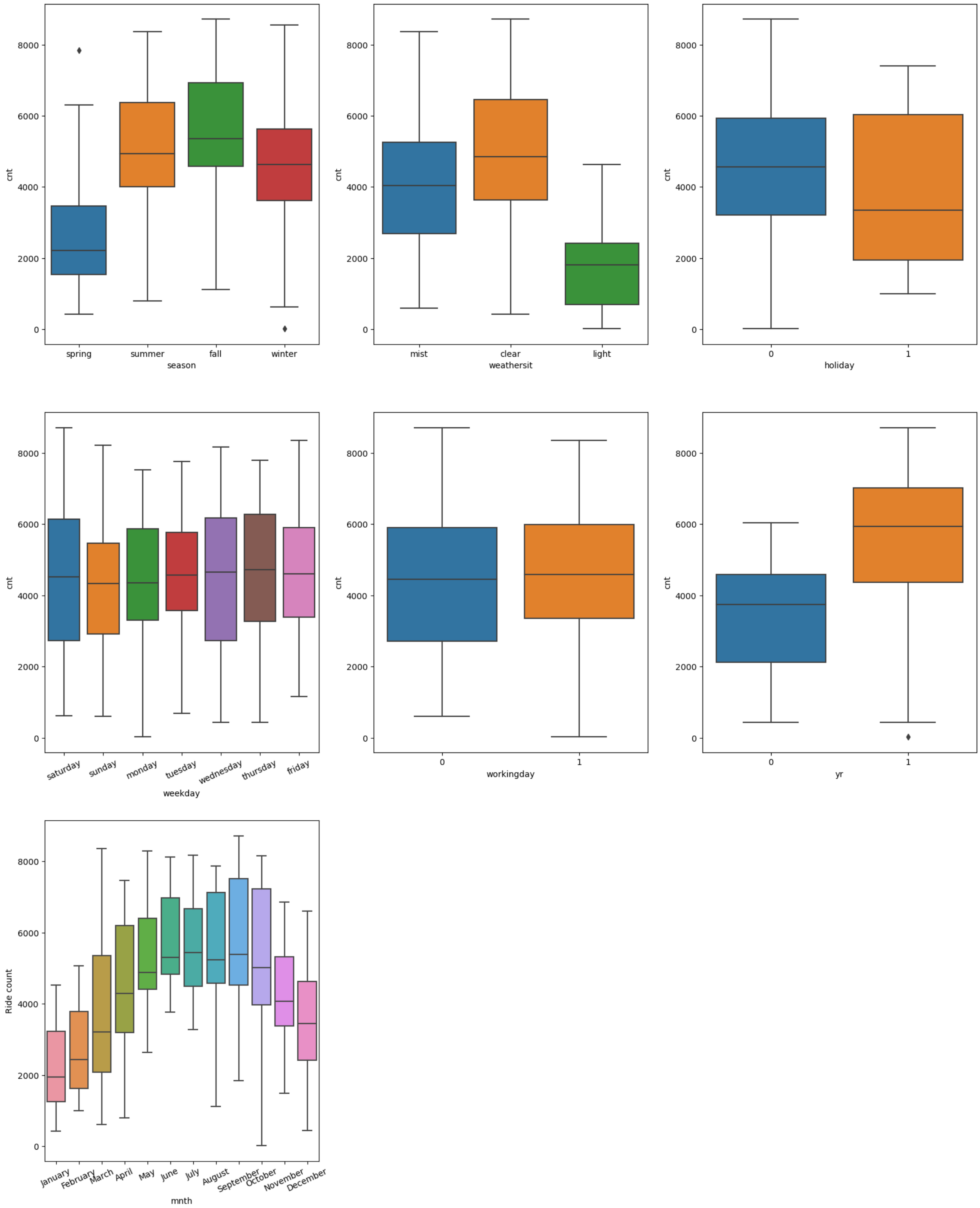
Assignment-based Subjective Questions

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



The above box plots are for categorical variables against the dependent variable. Below are the inferences obtained from the above plots.

- Comparing 2018 and 2019, the year 2019 has increased number of bookings with respect to almost all variables.
- Fall seems to be season where highest number of bookings happened.
- Ride count increased between June to October as they correspond to Summer and Fall seasons in USA.
- September month seems to have the highest number of bookings.
- Bookings is less on holidays compared to the other days which is reasonable.
- Working day or non-working day seems similar, not much difference.
- Ride count is higher on clear days and then on mist days other than light.
- Thursday, Friday, Saturday have an increase in the number of bookings compared to the rest of the days in the week.

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

The reason behind creating a dummy variable for each categorical variable with 'n' levels to map/create 'n-1' columns. **drop_first=True** is to reduce the extra columns which can be obtained from the other variables. To avoid redundancy, we are dropping the column. Hence that variables becomes linearly independent. It reduces the correlation created among dummy variables.

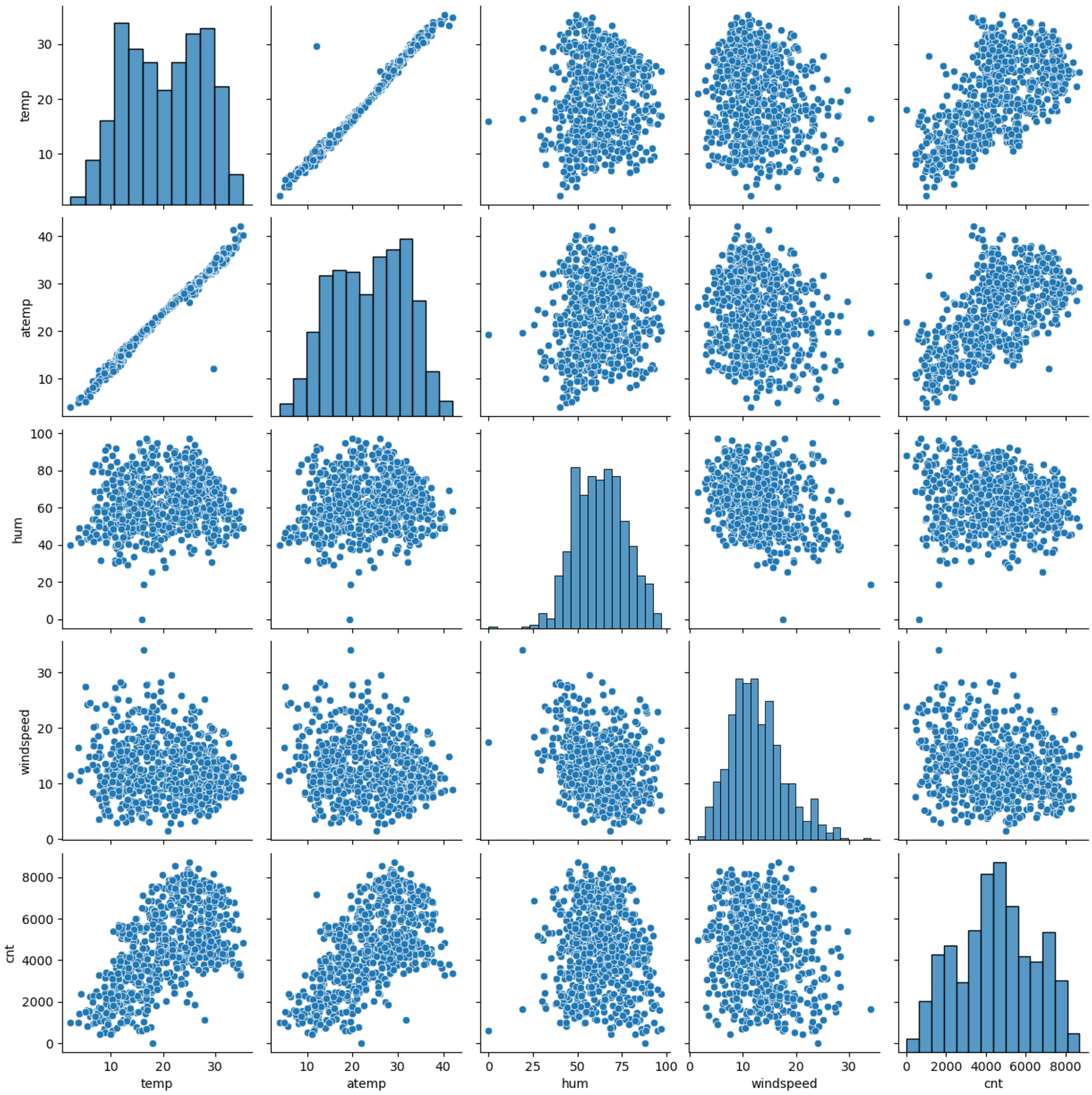
For example, consider a categorical column which has 3 values – status of the application form ('not started', 'in progress', 'completed'). If we just map 2 of them, then the 3rd one is well known if it doesn't fall into the other 2 values. That's why we **create 'n-1' columns only for 'n' levels**.

Value	**Indicators**	
Application Status	**not started**	**in progress**
not started	1	0
in progress	0	1
completed	0	0

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

'temp' and 'atemp' are the numerical variables which has the highest correlation with the target variable.

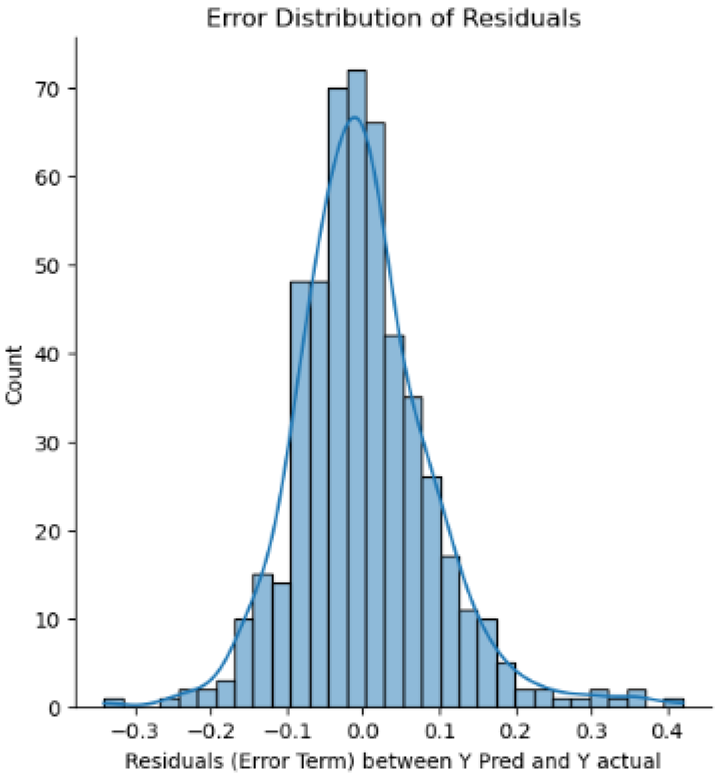
Below is the pair-plot for numerical variables.



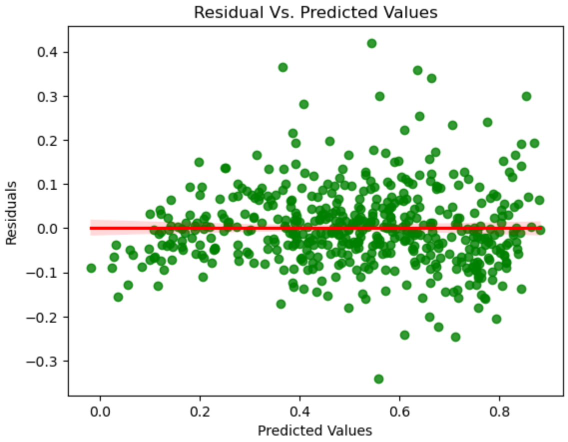
4. **How did you validate the assumptions of Linear Regression after building the model on the training set?**

   → **Linearity** - There is a linear relationship between dependent and independent variables. It is analyzed and visualized based on the pair plot above.

   → **Normality** - Error terms follow a normal distribution. When plotting the residual distribution, it is centered around 0, which is mean=0. It is given in the below plot.



→ **Error terms being independent** – There is no pattern or structure in the plot.



→ **Homoscedasticity** – There is equal variance and there is nothing like concentration of data points in certain specific regions.

→ **Independence of residuals** - Durbin-Watson value of our model is 2.051 which signifies there is no autocorrelation.

→ **Multicollinearity** – There must be insignificant multicollinearity among variables.

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

Based on the final model, below are the top 3 features which contributes significantly towards explaining the demand of the shared bikes.

Coefficients:

- → temp: 0.492
- → yr : 0.2339
- → Weathersit_light: -0.2883

General Subjective Questions

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

Linear regression is one type of statistical analysis and a machine learning algorithm which comes under supervised learning. This is predictive modelling method which helps us to identify the relationship between the dependent (target) and independent variables (predictors). It shows the linear relationship, to find how the value of the dependent variable changes based on the corresponding change in the independent variables. If the independent variable value increases or decreases, then the dependent variable value also gets increased or decreased accordingly. The **best fit straight line** against the given data points needs to be identified in order to get the accurate predictions minimizing the errors. It can be both **positive correlation or negative correlation.**

There are 2 types.

1. If there is only 1 predictor, then it is **Simple Linear Regression (SLR).**
2. If there are multiple predictors, then it is called **Multiple Linear Regression (MLR).**

SLR – It is based on the slope and intercept and the linear equation is given as below, which is used to predict the most accurate predictions.

$$Y=mx+b$$

Where, y – dependent variable, x – independent variable, m – slope, b-intercept

MLR – When there are multiple predictors, then the linear equation has multiple coefficients associated with the variables.

$$Y=a+b_1 X_1 + b_2X_2 \quad + \ldots + b_nX_n$$

Where b_1, b_2, b_n are coefficients of the respective variables.

Assumptions of MLR

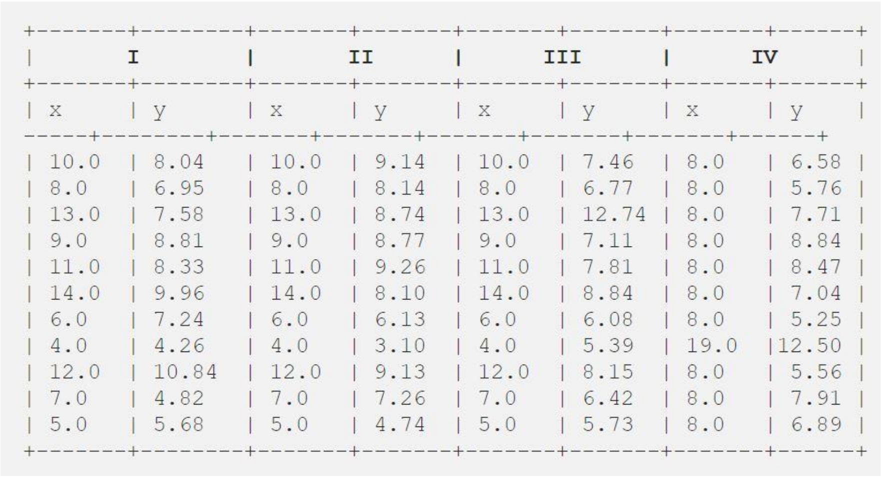
1. **Linearity** - Linear relationship between X and Y
2. **Normality** – Y is normally distributed for any fixed value of X. Error terms must follow normal distribution.
3. **Independence** - Observations are independent of each other.
4. **Multicollinearity** – Predictors have no correlation between them, which means none of them should be highly dependent on each other.
5. **Homoscedasticity** – The residuals have constant variance at every data points in the model.

   Limitations:
   Data needs to be cleansed of any outliers and need to have linear relationship which serves as the most basic condition to predict using linear regression.

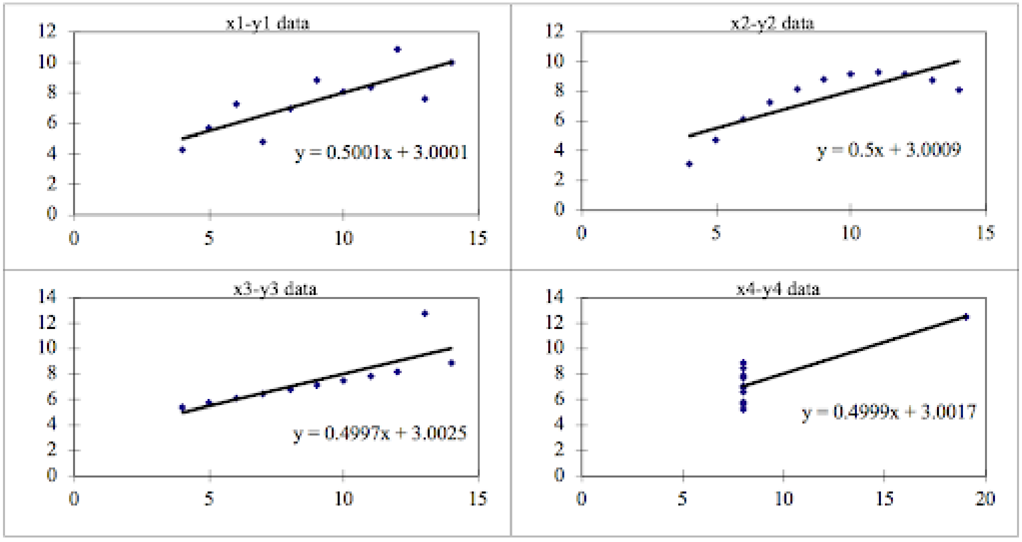
2. Explain the Anscombe's quartet in detail.

Anscombe's quartet was named person called Francis Anscombe in 1973. It consists of a set of four datasets which has identical descriptive statistical properties in terms of **Correlation**, **Mean**, **Variance**, **R-Squared** and **Linear Regression** lines. They have different representations when we do a scatter plotting on the graph. Below are the representations. This is used to demonstrate the purpose of **Exploratory Data Analysis (EDA),** importance of **visualizing** the data and to explain that summary statistics alone cannot give us good insights. **Identifying trends**, **outliers** and other minor details may be missed out.

The 4 datasets, each include 11 x-y pairs of data. They have a unique and similar connection between X and Y, similar and unique variability patterns but distinctive correlation strengths. Despite this, each dataset is having the same summary statistics, same mean, variance, regression line and correlation coefficient.



I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89



- **Top left plot** – Has a linear relationship and fits the regression line.
- **Top right plot** – There is a relationship but not linear. It cannot fit the regression line as the data points are non-linear.
- **Bottom left plot** – It has a outlier based on the fit regression line, hence assumption of linear regression doesn't hold here as it cannot handle outliers.
- **Bottom right plot** – The data points do not indicate any relationship between the variables.

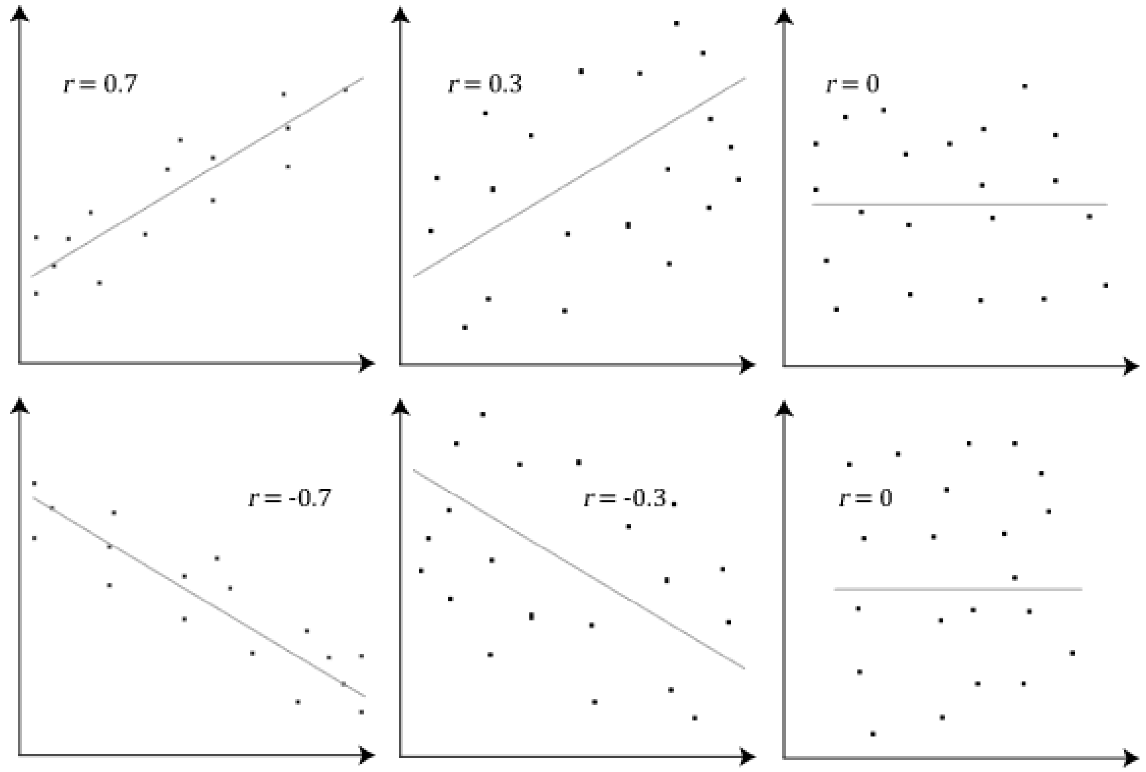
This quartet is often used to visualize the data graphically and how this step is important for building a regression model. It explains the relationship if it is present or not and how misleading the basic statistic properties can be when it comes to real-time big datasets.

3. What is Pearson's R?

Pearson correlation coefficient or Pearson R statistical test measures the strength of linear association between two variables and their relationships and is denoted by r. This indicates how far away all these data points are to the regression line of best fit. It is used to quantify the relationship.

It can take the range of values between **-1 to +1.** Value **0** indicates that there is **no association** between the two variables. **+1 indicates strong positive correlation** whereas **-1 indicates strong negative correlation**. That means, for positive correlation, when value of one variable increases, then the value of other variable also increases. Similarly, for negative correlation, when one value decreases, then the other value also decreases.

The below figure shows, the correlation based on the r value which is the correlation coefficient value.



It is calculated based on the below formula for two variables x and y.

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

Where,

- r = Pearson Coefficient
- n= number of pairs of variables
- ∑xy = sum of products of the paired variables
- ∑x = sum of the x variable
- ∑y= sum of the y variable
- $\sum x^2$ = sum of the squared x variable
- $\sum y^2$ = sum of the squared y variable

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

Feature Scaling is a technique to **normalize** the independent features or variables of data and **standardize** within the **range**. This is performed to bring back all of them in the standardized units so that it doesn't affect our predictions when building the model. It is done as pre-processing step to handle high varying magnitudes or low varying values into a standard single range which will be useful for comparisons. If this is not done, then the model may sometimes include the high weighing features taking only magnitudes into account (ignoring the units) and may ignore the lower values which might be important and hence we may get misleading results.

For example, if we have 100 ml and 1 litre, and if it considers only magnitude, then obviously 100 > 1, then the model might consider 100 as the important feature which may not be so. Hence, we need to scale the features to bring in within the same range in order to handle these issues.

Normalized Scaling	**Standardized Scaling**
Good to use when the data doesn't follow a gaussian distribution	When data follows gaussian distribution. But this may not be necessarily true.
Features are mapped between the range of [0,1].	There is no definite range, hence we transform the data so that it has mean of 0 and standard deviation of 1
Min and max values are used for scaling.	Mean and standard deviation is used.
It is affected by outliers.	Not affected by outliers.
MinMaxScaler can be used from scikit-learn	StandardScaler can be used from scikit-learn

Normalized Scaling

$$z = \frac{x - min(x)}{max(x) - min(x)}$$

Where x is the feature

Standardized Scaling

$$z = \frac{x - \mu}{\sigma}$$

Where x is the feature, μ is the mean, and σ is the standard deviation.

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

Variance Inflation Factor (VIF) is an index measuring how much variance of an estimated regression line coefficient increases because of collinearity. If all of the independent variables are orthogonal to each other, then VIF=1. VIF is estimated based on the below equation.

$$VIF = \frac{1}{1 - R^2}$$

When **VIF is infinity**, which means there is a **perfect correlation**. In this case, then there are variables which are highly correlated to each other, keeping this will affect our model, hence we need to drop one of them.

This means that one variable can be explained very well by other. In this case **R^2 will be 1. As per the equation, it will become 1/(1-1) = 1/0 = infinity.**

Anything for VIF > 10, definitely has high multicollinearity and may need to drop this.

Between 1 to 5 is moderately correlated, greater than 5 is highly correlated and we must take a look to drop one of the variables accordingly.

The real understanding of this, for example, if VIF is 6, then the variance of the model coefficient is inflated by a factor of 6 due to the presence of multicollinearity.

Below actions can be taken when VIF is very high.

- Take a review on variable which has high VIF, eliminate one of them re-build the model and check again. Repeat this process until VIF falls within acceptable range. Decide based on business knowledge as well before dropping.
- Can use PCA – Principal Component Analysis
- Can increase the sample size so that the confidence intervals become narrower.
- Can transform the data to a different space.

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

Quantile-Quantile plot (Q-Q plot) is a **graphical** method used for determining if the two samples of data came from the **same population or not**. It is a plot of quantiles of first dataset against the second dataset. It is a **scatterplot**, when both sets of quantiles come from same distribution, we must be seeing the data points forming a **line** which is roughly **straight**.

Use

Quantiles are often referred to as **percentiles**. For example, if 0.4 (or 40%) quantile means that 40% of data fall below and 60% fall above that value. The greater the departure from the fit line, then we would be having strong evidence that the two datasets comes from populations with different distributions. This is useful to compare shapes of distributions, provides graphical view of properties like location, scale, and skewness, to understand if it is similar or different in the two distributions. It helps us to summarize the distribution visually.

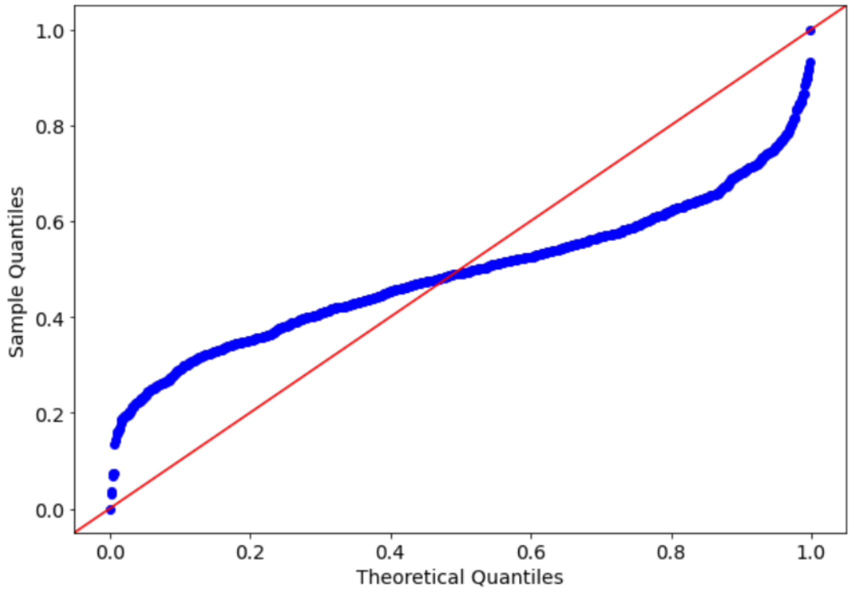
Importance

To determine,

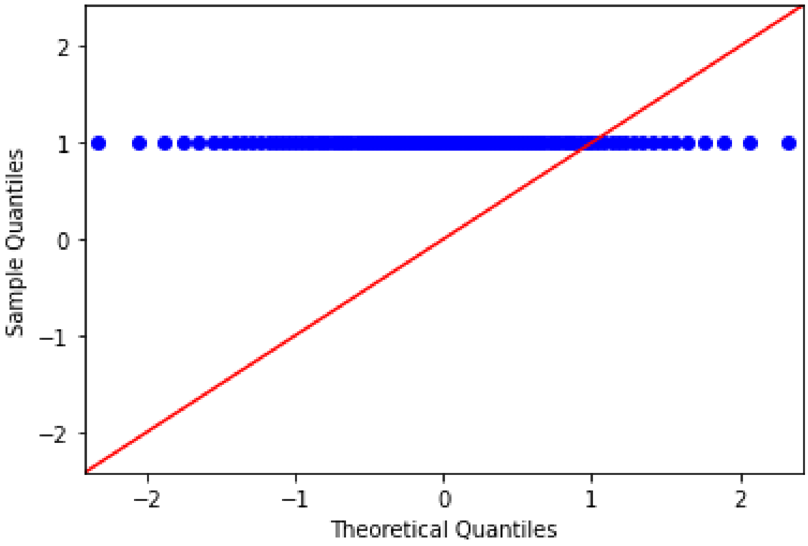
- Skewness of distributions
- If two populations follows same distribution.
- Nature of difference is well known compared to other analytical methods like chi-square.
- If they have common scale and location
- If they have same distributional shapes
- If they have similar tail behaviour.

Types of Q-Q plots

- Left tailed distribution graph



- Uniform distribution graph


2
1
0
−1
−2
Sample Quantiles
−2
−1
0
1
2
Theoretical Quantiles