Assignment submission by – Venkata Sreeleela Sappa

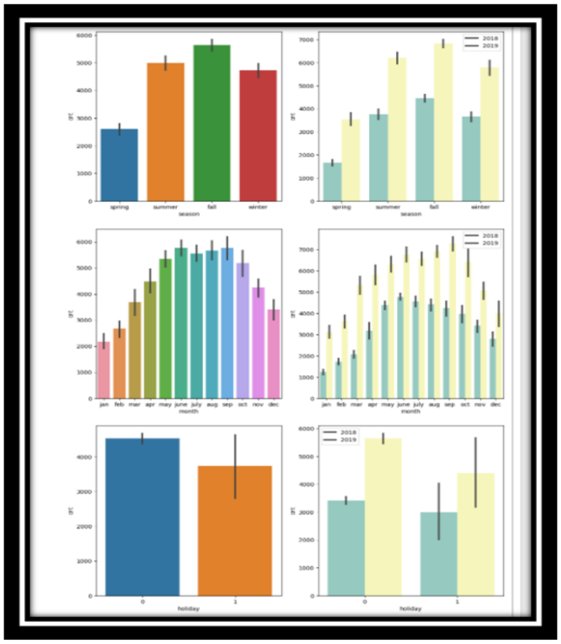
Assignment-based Subjective Questions

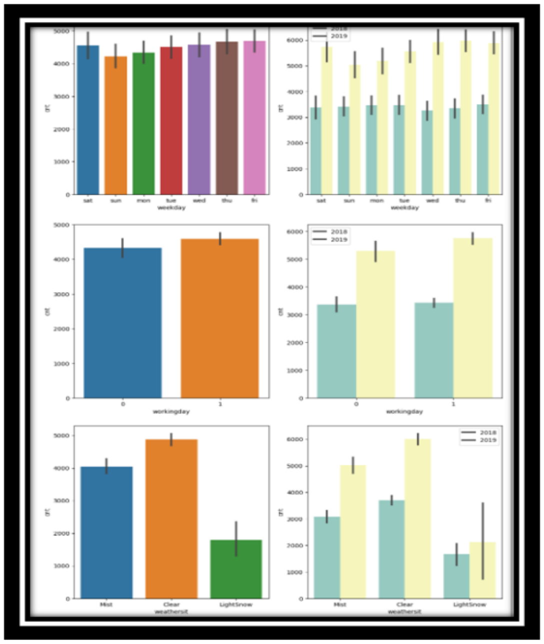
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

A. The following categorical Variables are identified from the data :

- season :spring,summer,fall,winter
- year : 2018,2019
- month: jan,feb,mar,apr,may,june,july,aug,sep,oct,nov,dec
- holiday : 0,1
- weekday : sun,mon,tue,wed,thu,fri,sat
- workingday : 0,1
- weathersit : Clear,Mist,LightSnow,HeavyRain

We can see the Count Plots of each of these variables with hue as cnt(the output variable and the influence of year on them.





Observations:

- Fall season has highest demand.
- Across all seasons booking increased in 2019.
- Demand increase till mid-year and again started to fall .
- Number of bookings per month also increased from year 2018 to 2019 .
- Demand is high when weather conditions are good.
- Irrespective of weather conditions, bookings increased in 2019.
- The influence of holiday/ working day on the number of bookings is minimal.
- Number of bookings is high in 2019 when compared to 2018.

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

A. The use of drop_first=True in dummy variable creation helps to avoid redundant information in the dataset being fed to the model and helps to avoid multicollinearity. This is a very important step when we are using categorical variables as predictors in Linear regression .

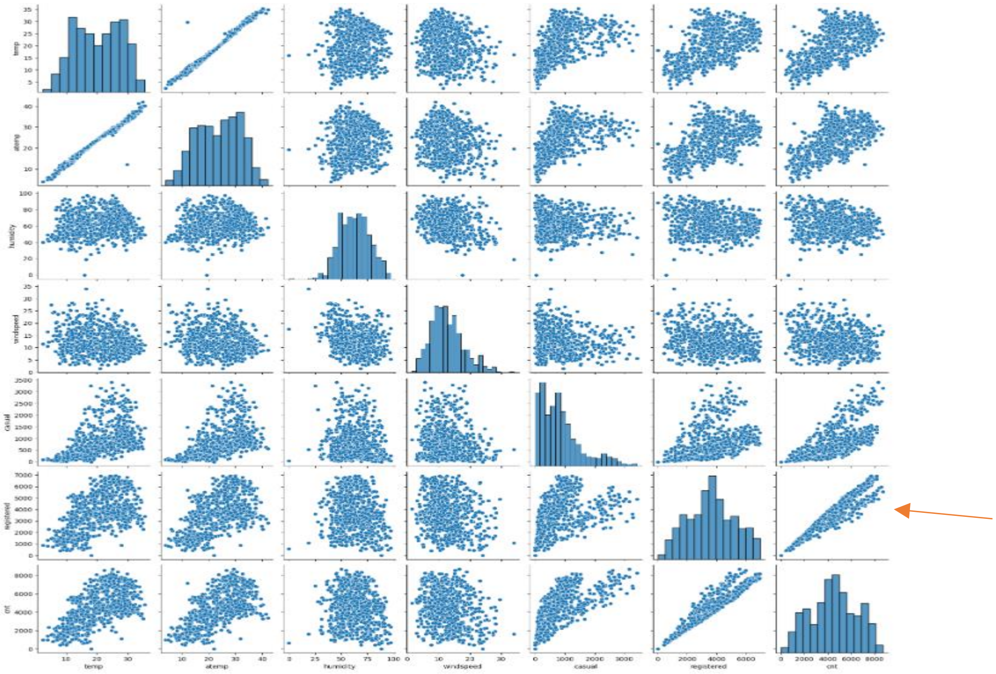
- **Prevent Redundant Information :**
  When creating dummy variables, dropping the first category also helps in avoiding redundant information. When we are converting the varaiables to dummies, it is always possible to derive the value of the nth category using the values of the remaining n-1 categories of the variable. For example, if we have a categorical variable(size) with three categories S,M and L , we can derive the value of L if S and M are known

S	M	Derived L
0	0	1
1	0	0
0	1	0

- **Avoid multicollinearity:**
  When creating dummy variables, if we include all possible categories, it can lead to multicollinearity issues. Multicollinearity occurs when two or more predictor variables are highly correlated with each other. In this case, including all dummy variables can introduce perfect multicollinearity because one category can be perfectly predicted from the others. By dropping one of the dummy variables, ensuring that each category is represented independently and avoiding perfect multicollinearity.

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

A. The pair plot for the numerical variables (temp : temperature in Celsius,atemp: feeling temperature in Celsius,hum: humidity,windspeed: wind speed,casual: count of casual users,registered: count of registered users,cnt: count of total rental bikes including both casual and registered) is below,

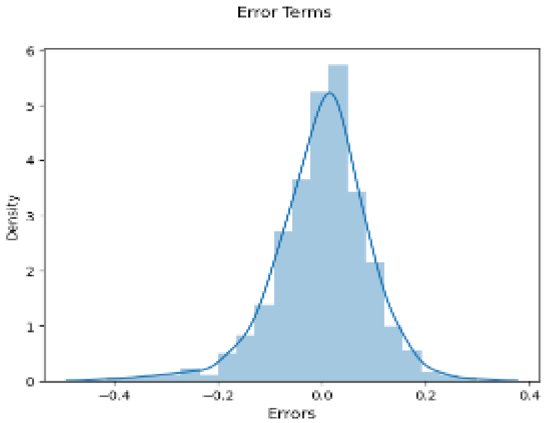


We can see that the variables **registered** has the highest correlation with the target variable cnt.

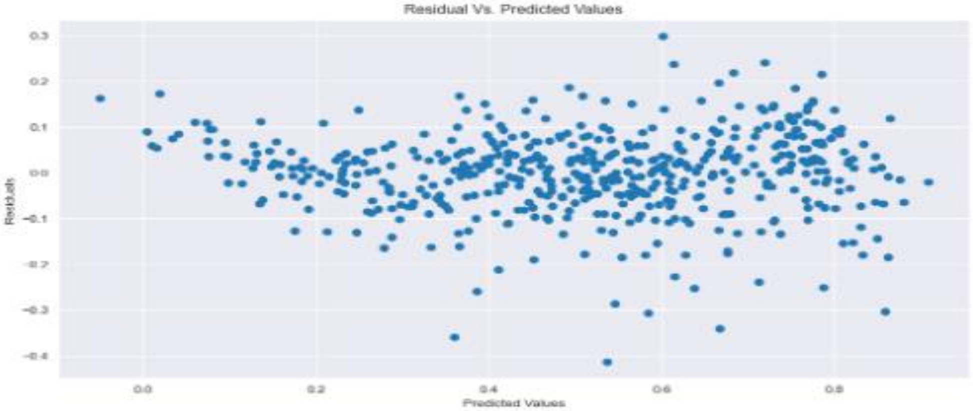
4. **How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)**

A. I have verified the following assumptions of linear regression through residual analysis.

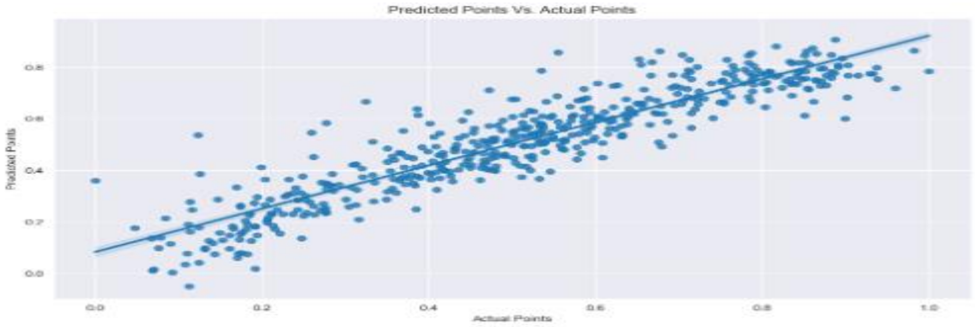
- **Normal distribution of error terms** :
  I have plot the distplot(distribution of a variable against the density distribution) of the error terms or residuals (actual value – predicted value of target variable) and observed that error terms are normally distributed.



- **Error terms are independent of each other** :
  I have plotted a scatter plot of the error terms and checked that there are no evident patterns among the distribution, proving the assumption of linear regression that error terms are independent of each other.



- **Homoscedasticity(Constant Variance)**:
- **Linearity**:
  I have plot a regplot(plot data and draw a linear regression model fit) to verify the above two assumptions. The plot shows the linearity as well as the homoscedasticity assumptions.



5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)**

A. The equation derived from the final model of the multi linear regression is

cnt= 0.27 + 0.47 * temp + 0.23 * year + 0.1 * winter + 0.07 * sep + 0.06 * sat + 0.05 * workingday + 0.04 * summer -0.04 * dec -0.04 * nov -0.05 * jan -0.05 * july -0.06 * Mist - 0.06 * spring -0.15 * humidity -0.19 * windspeed -0.26 * LightSnow

If we closely observe the coefficients in the equation,
We can see that the top 3 features contributing significantly towards explain the demand of shared bikes are

1. **Temp**(temperature in Celsius) , with a positive coefficient of **0.47**. This explains that assuming all other features are constant, an increase in temperature can influence the demand for sharing bikes positively.
2. LightSnow(**weathersit(3)** - Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds) , with a negative coefficient on **-0.26**. This explains that these weather conditions are quite unfavorable for the demand of shared bikes and has a good negative impact on demand.
3. Year (**yr** : year (0: 2018, 1:2019)) : with a positive coefficient of **0.23** ,positively impacts the demand for shared bikes. It implies that there is a significant increase in demand for shared bikes year on year.

General Subjective Questions

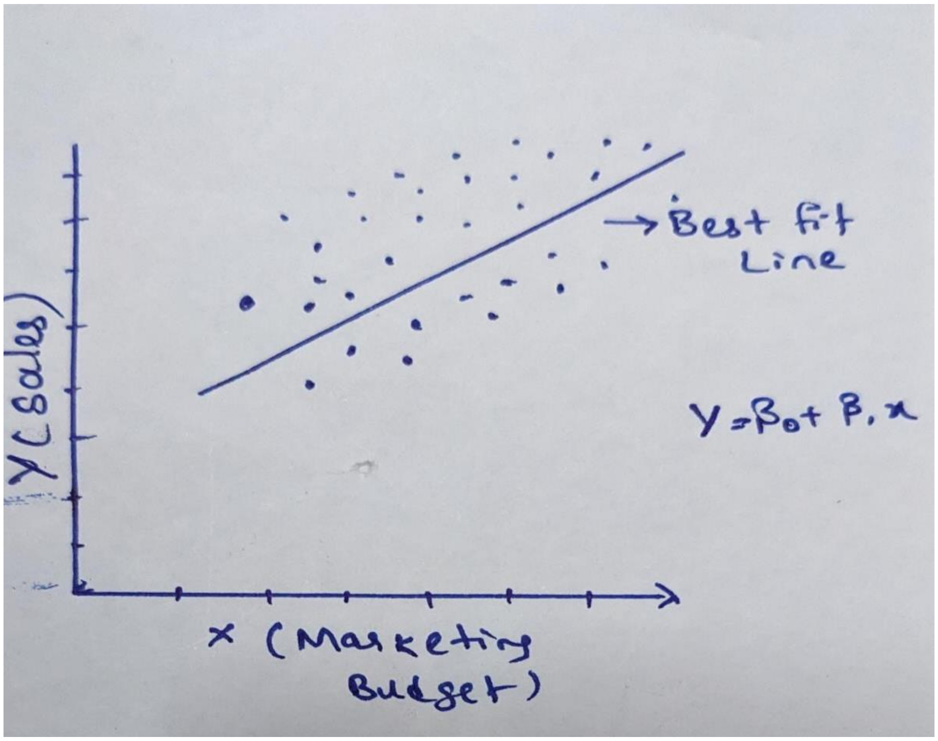
1. **Explain the linear regression algorithm in detail. (4 marks)**

A. Regression is a supervised learning algorithm used for predicting a continuous outcome variable based on one or more predictor variables. Linear regression is the most widely used predictive analysis algorithms in the industry . It assumes a linear relationship between the predictors and the outcome variable. The algorithm aims to find the best-fitting line (or hyperplane in higher dimensions) that minimizes the difference between the predicted and actual values.

Linear regression can be classified into two categories.

1. **Simple Linear Regression:**

The most elementary type of regression model is the simple linear regression which explains the relationship between a dependent variable and one independent variable using a straight line. The standard equation of the regression line in simple linear regression is **$Y = \beta_0 + \beta_1 X$.**
where:
y is the outcome variable to be predicted.
x is the predictor variable.
β_0 is the intercept term (the value of y when x is zero).
β_1 is the coefficient (or slope) that represents the effect of x on y.



The goal of linear regression is to estimate the coefficients b0 and b1 that minimize the difference between the predicted values (y) and the actual values of the outcome variable. This process is often achieved using a method called Ordinary Least Squares (OLS) estimation, which minimizes the sum of squared residuals (the differences between the predicted and actual values).

2. Multiple Linear Regression:

Linear regression algorithms can handle multiple predictor variables by extending the equation to include additional coefficients and predictors. This type of regression is called Multiple Linear Regression (MLR). The equation for multi Linear regression is
$Y=\beta 0+\beta 1X1+\beta 2X2+\ldots+\beta pXp$
where:
$X1, X2, \ldots, Xp$ are the predictor variables.
$\beta 0, \beta 1, \ldots, \beta p$ are the coefficients corresponding to each predictor variable.
The coefficients explain the corresponding variable when all other factors or predictor variables are constant. Once the coefficients are estimated, the linear regression model can be used to make predictions on new data. Given the values of the predictor variables, the model calculates the predicted outcome variable using the learned coefficients.

Assumptions of Linear Regression :

The assumptions of linear regression are :

1. Linear relationship between X and Y
2. Error terms are normally distributed (not X, Y)
3. Error terms are independent of each other
4. Error terms have constant variance (homoscedasticity)

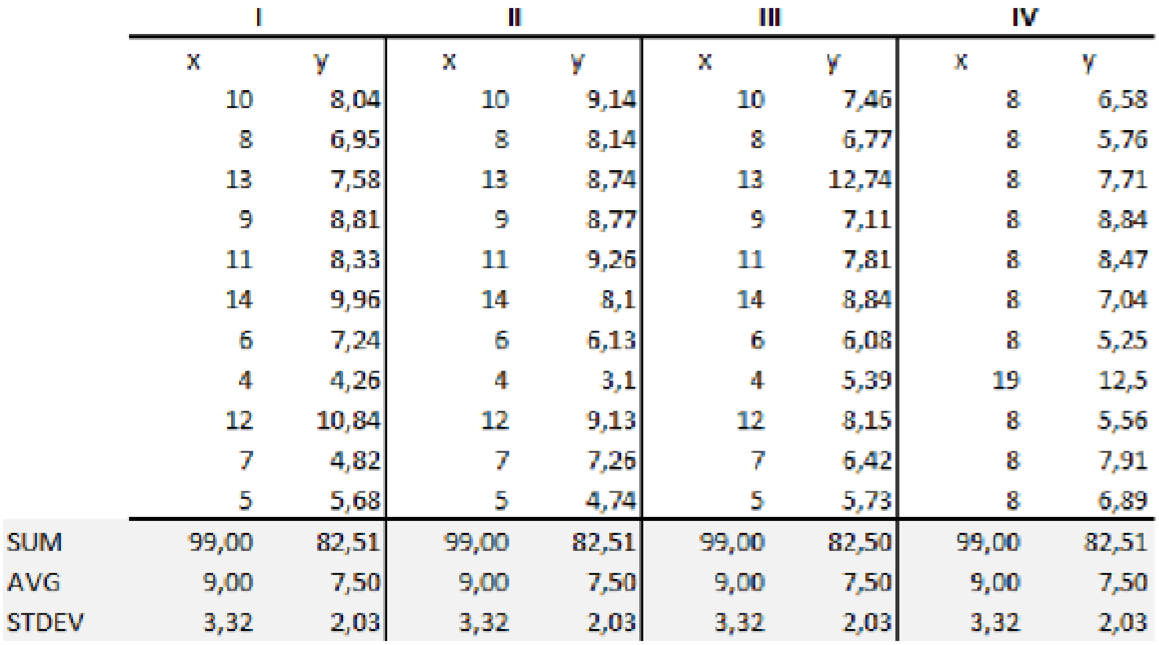
With these assumptions we can go ahead and make inferences about the model which, otherwise, we wouldn't have been able to. There is no assumption on the distribution of X and Y, just that the error terms must have a normal distribution.

Steps in linear Regression :

1. Reading and understanding the data **.**
2. Visualizing the data for initial analysis and data cleaning(EDA)
3. Data preparation: Converting categorical variables with varying degrees of levels into dummy variables (numerical in nature) so that these variables can be represented during model building to contribute to the best fitted line for the purpose of better prediction.
4. Splitting the data into train and test sets and rescaling the predictors.
5. Building a linear model using :
   - ➢ Forward Selection : Start will null and then keep adding predictors one by one to see the model behavior.
   - ➢ Backward Selection: We add all the variables at once and then eliminate variables based on high multicollinearity (VIF>5) or insignificance (high p-values >0.05).
   - ➢ RFE(Recursive Feature Elimination): This is an automated version of feature selection technique where we select that we need "m" variables out of "n" variables and then machine provides a list of features with importance level given in terms of rankings.
6. Residual analysis of train data: It tells us how much the errors (y_actual — y_predicted) are distributed across the model. A good residual analysis will signify that the mean is centred around 0 when we plot a distribution of error terms.
7. Making predictions using the final model and evaluation of the model on the test set: A difference of 2–3% between r2_score of train and test score is acceptable as per the standards.

2. Explain the Anscombe's quartet in detail. (3 marks)

A. Anscombe's quartet was constructed in 1973 by statistician Francis Anscombe to illustrate the importance of plotting data before you analyze it and build your model. These four data sets have nearly the same statistical observations, which provide the same information (involving variance and mean) for each x and y point in all four data sets. However, when you plot these data sets, they look very different from one another. This helps us understand the importance of data visualization, rather than just relying on statistical inferences.
We can see the data sets in the image below.



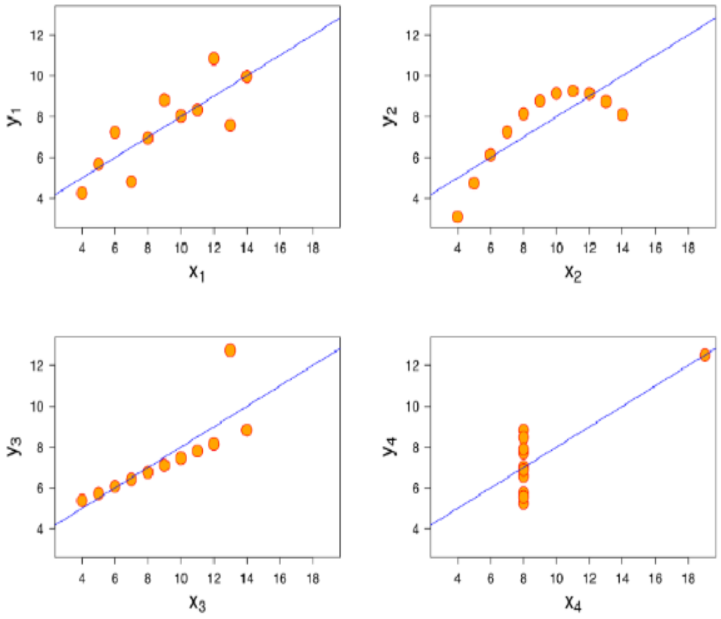
	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	8	5,56
	7	4,82	7	7,26	7	6,42	8	7,91
	5	5,68	5	4,74	5	5,73	8	6,89
SUM	99,00	82,51	99,00	82,51	99,00	82,50	99,00	82,51
AVG	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
STDEV	3,32	2,03	3,32	2,03	3,32	2,03	3,32	2,03

As we can see in the representation above,

The four datasets have exactly same statistical inferences

1. All the data sets have the same Sums of X and Y
2. All the data sets have same mean values
3. All the data sets have exactly same Standard deviation

But when we plot their distribution,



We can see that

- Data Set 1: fits the linear regression model pretty well.
- Data Set 2: cannot fit the linear regression model because the data is non-linear.
- Data Set 3: shows the outliers involved in the data set, which cannot be handled by the linear regression model.
- Data Set 4: shows the outliers involved in the data set, which also cannot be handled by the linear regression model.

As we can see, Anscombe’s quartet helps us to understand the importance of data visualization and how easy it is to fool a regression algorithm. So, before attempting to interpret and model the data or implement any machine learning algorithm, we first need to visualize the data set to help build a well-fit model.

3. **What is Pearson’s R? (3 marks)**

A. Pearson's r, also known as the Pearson correlation coefficient or Pearson's correlation, is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. It assesses the degree to which the variables are linearly related to each other. The Pearson correlation coefficient (r) ranges between -1 and 1. The sign of r indicates the direction of the relationship:

- A positive value ($r > 0$) indicates a positive linear relationship, meaning that as one variable increases, the other variable tends to increase as well.
- A negative value ($r < 0$) indicates a negative linear relationship, meaning that as one variable increases, the other variable tends to decrease.

The magnitude of r represents the strength of the relationship:

- r = 0 indicates no linear relationship between the variables.
- r = 1 or -1 indicates a perfect linear relationship, where all data points lie on a straight line.
- r = 1 represents a perfect positive linear relationship,
- r = -1 represents a perfect negative linear relationship.
- r between 0 and 1 (or between 0 and -1) represent varying degrees of linear relationship, with larger magnitudes indicating stronger associations.

Pearson's r is calculated by dividing the covariance of the two variables by the product of their standard deviations. The formula for calculating Pearson's r between variables X and Y, each with n data points, is as follows:

r = (Σ((Xi - X_mean) * (Yi - Y_mean))) / (n * StdDev(X) * StdDev(Y))

where:

Xi and Yi are individual data points of variables X and Y, respectively.
X_mean and Y_mean are the means of variables X and Y, respectively.
StdDev(X) and StdDev(Y) are the standard deviations of variables X and Y, respectively.

Pearson's r is commonly used to assess the strength and direction of linear relationships between variables. It provides valuable insights into the degree of association between two continuous variables, enabling researchers to understand the nature and extent of their relationship.

4. **What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)**

A. Scaling : Scaling is a preprocessing step in data analysis and machine learning that involves transforming the numerical values of variables to a standardized range. It ensures that all variables have a similar scale and distribution, which can be beneficial for certain algorithms and analyses.

Scaling is performed for the following reasons:

- Equalize variable influence: Variables with larger magnitudes or wider ranges can dominate the analysis or algorithm's outcome. Scaling helps prevent this by giving equal weight to all variables, making their influence comparable.

- Facilitate convergence: Many optimization algorithms used in machine learning are sensitive to the scale of variables. Scaling can help algorithms converge faster and more reliably by making the optimization landscape more balanced.

- Improve interpretability: Scaling variables to a similar range can make it easier to interpret the model coefficients. It allows for fair comparisons of the variable contributions to the outcome.

There are two common types of scaling: normalized scaling and standardized scaling.

- **Normalized Scaling**: Normalization scales the values of a variable to fit within a specified range, typically between 0 and 1. It is achieved by subtracting the minimum value from each data point and then dividing by the range (maximum value minus minimum value). Normalization preserves the shape and distribution of the variable but compresses it within the defined range.
  X= (X-min(X)) / (max(X) – min(X))

- **Standardized Scaling**: Standardization, also known as z-score scaling, transforms the values of a variable to have a mean of 0 and a standard deviation of 1. It is achieved by subtracting the mean from each data point and then dividing by the standard deviation. Standardization shifts the variable distribution to have a mean of 0 and equalizes the spread.
  X=(X – mean(X))/ StdDev(X)

key differences :

- Normalization restricts the variable values within a specific range, while standardization centers the variable around the mean and scales it by the standard deviation.
- Normalized scaling is useful when the original range of the variable is not relevant, and you want to ensure all values fall within a specific range. Standardized scaling is beneficial when you want to compare variables that have different scales and distributions, or when the mean and standard deviation are meaningful for interpretation.

The choice between normalized scaling and standardized scaling depends on the specific requirements of the analysis or algorithm being used, as well as the characteristics of the data and the variables involved.

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)**

A. VIF (Variance Inflation Factor) : It is a statistical measure used to assess the severity of multicollinearity in a regression analysis. Multicollinearity occurs when predictor variables in a regression model are highly correlated with each other. The VIF quantifies how much the variance of the estimated regression coefficient is inflated due to multicollinearity. It measures the extent to which the variance of a coefficient estimate is increased compared to the scenario where there is no correlation between the predictor variables.

The formula for calculating the VIF of a predictor variable is:

$VIF = 1 / (1 - R^2)$
where R^2 represents the coefficient of determination for the regression model that predicts the specific predictor variable using the other predictor variables.

The phenomenon of obtaining infinite values of the Variance Inflation Factor (VIF) typically occurs when there is **perfect multicollinearity among the predictor variables**. Perfect multicollinearity happens when two or more predictor variables are perfectly correlated, meaning one predictor can be expressed as a linear combination of the others. That is if their R2 values is 1 (R= 1 or R= -1) . When this happens the VIF become 1/(1-1) = Infinity.

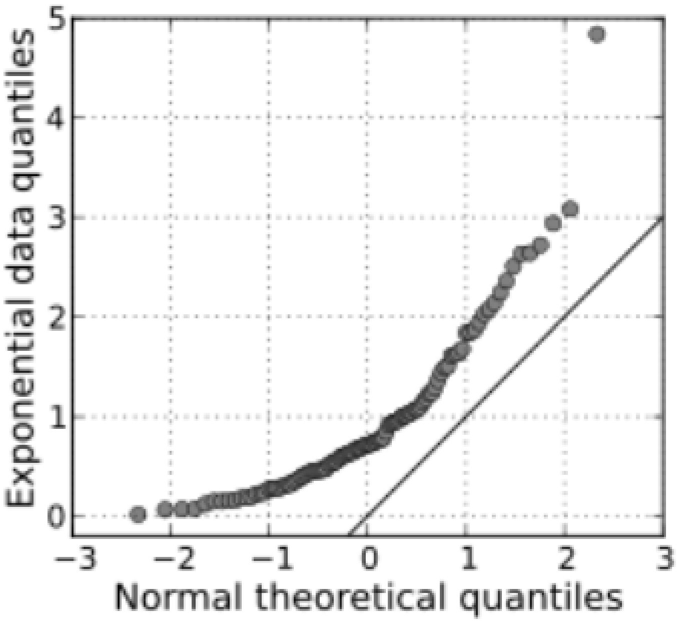
When we encounter a VIF of Infinity, we need to follow these steps .

- Remove one of the Correlated variables : Retain the variable that is more meaningful or theoretically relevant to analysis.
- Consider data transformations: If the variables have nonlinear relationships, we can try applying mathematical transformations, such as logarithmic or power transformations, to mitigate multicollinearity. This can help retain the valuable information without the perfect linear relationship.
- Combine correlated variables: Instead of using individual variables, you can create composite variables or aggregate measures that capture the shared information. This

can be done through techniques like principal component analysis (PCA) or factor analysis.

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)**

A. 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. 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.



A Q Q plot showing the 45 degree reference line.

Importance of Q-Q plots in Linear Regression:

- Assessing Normality: Linear regression assumes that the residuals are normally distributed. A Q-Q plot can reveal departures from this assumption. If the data points on the Q-Q plot deviate significantly from the expected diagonal line, it suggests a departure from normality. Deviations like skewness, heavy tails, or outliers indicate potential violations of the normality assumption.

- Identifying Data Transformations: When the residuals deviate from normality, it may be necessary to transform the data to meet the assumption. Q-Q plots help identify the nature of the deviation, suggesting the appropriate transformation. For example, if the residuals exhibit a skewed distribution, a logarithmic or power transformation might be necessary to achieve normality.

- Model Assessment: Q-Q plots provide a visual assessment of the model's fit to the assumption of normality. They help evaluate the adequacy of the regression model and identify potential issues that might affect the reliability and interpretation of the model.

- Detecting Outliers: Q-Q plots can also be used to detect outliers, which are data points that significantly deviate from the expected distribution. Outliers can affect the model's performance, and their identification through Q-Q plots helps in deciding how to handle them, such as removing or transforming them.

Overall, Q-Q plots are valuable diagnostic tools in linear regression to evaluate the assumption of normality, guide data transformations, detect outliers, and assess the model's fit. By visually examining the observed versus expected quantiles, analysts gain insights into the distributional properties of the residuals and make informed decisions about the regression model.