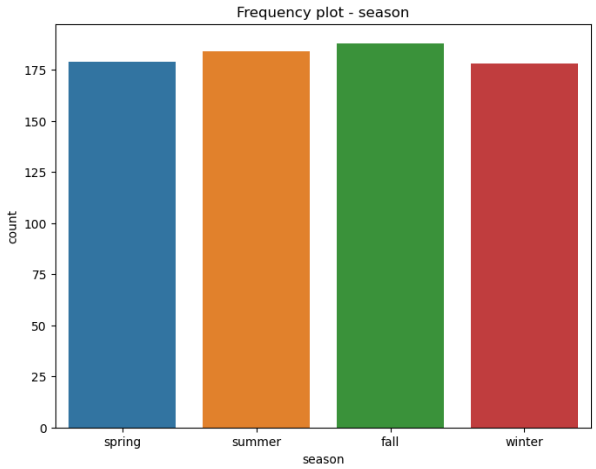
Assignment- based Subjective Questions

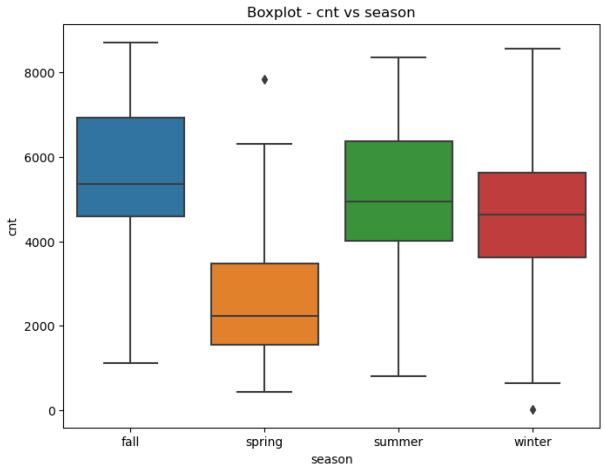
Question1: From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

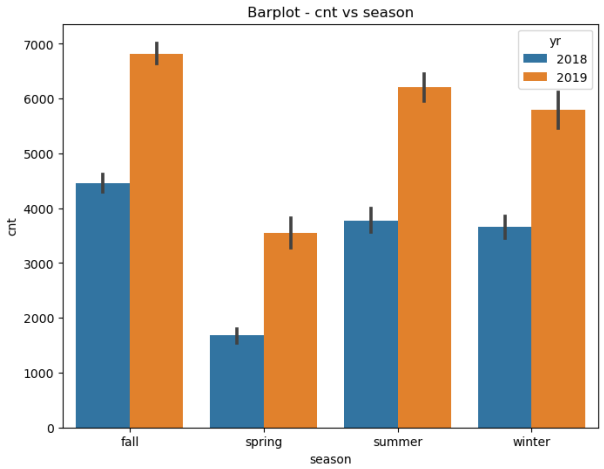
Answer: In the final linear regression model, 8 out of 9 predictors are categorical variables which have significant trend with dependent variable (demand of the bike). Their impact on the target variable is as follows-

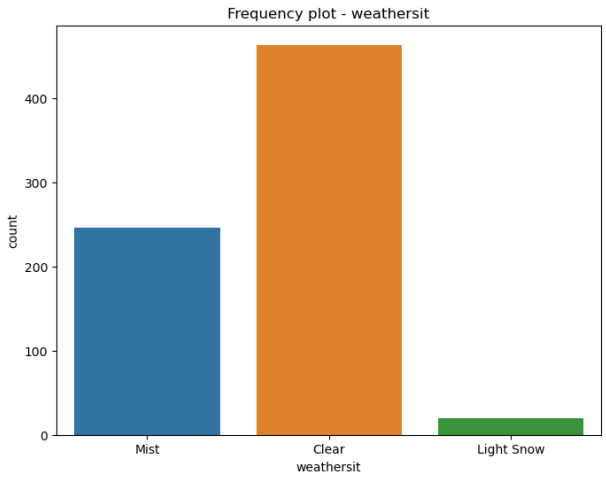
1. **workingday** – This variable has shown **positive weak trend** observed.
2. **season_spring** – This dummy variable represents Spring season which has shown **negative strong trend** with bike demand. It means that people have less preferred to use bike in spring season
3. **yr_2019** – This dummy variable represents year 2019 which has shown **positive strong trend** with bike demand. Bike demand is higher in year 2019 w.r.t year 2018.
4. **mnth_09 Sep** and **mnth_10 Oct** – These two dummy variables represent September and October month. These months have shown **weak positive trend** with bike demand.
5. **weekday_06 Sat** – This dummy variable represents Saturday which has shown **weak positive trend** with bike demand.
6. **weathersit_Light Snow** and **weathersit_Mist** – These dummy variables represent weather situation like Mist and Light snow which have shown **negative strong trend** with bike demand.

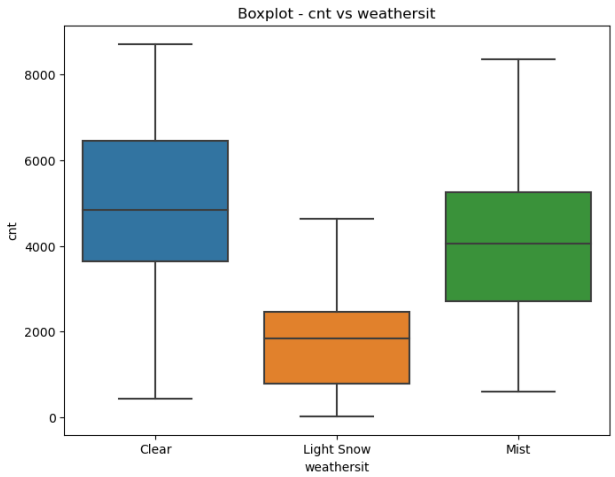
These observations are also supported by Bi-variate and multi-variate analysis with target variable (**cnt**)- (important charts attached below)

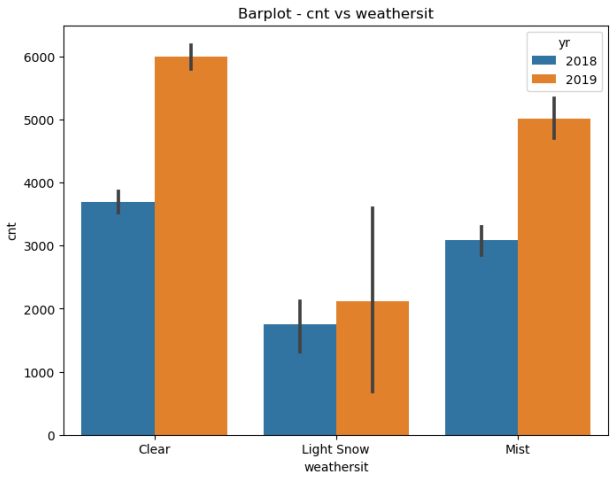


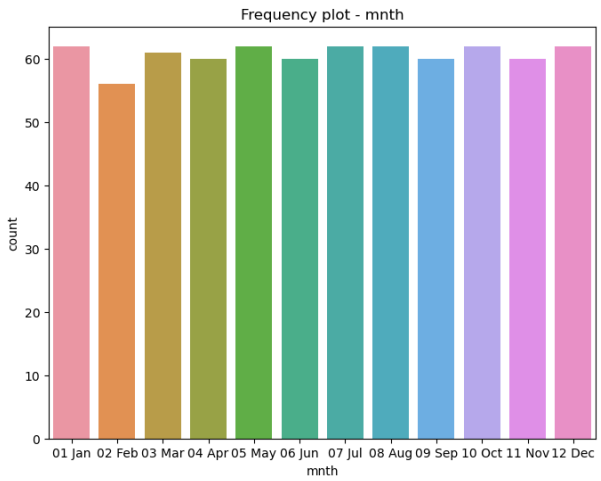


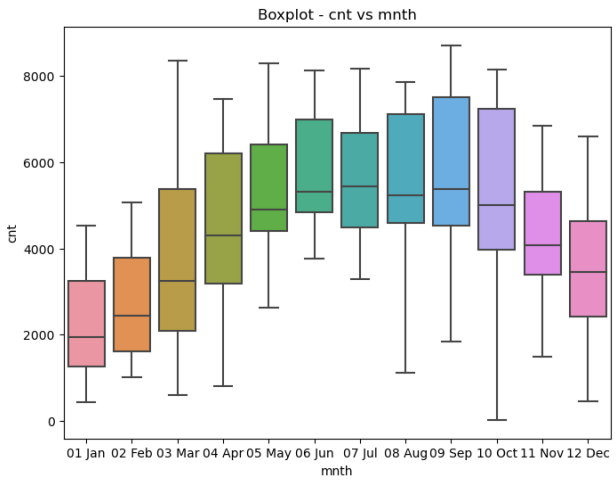


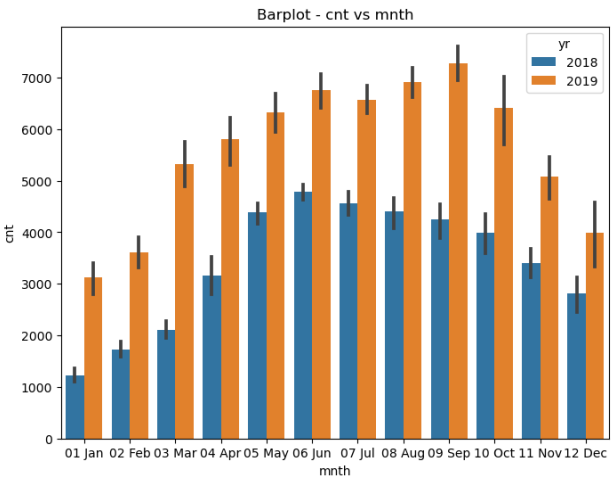










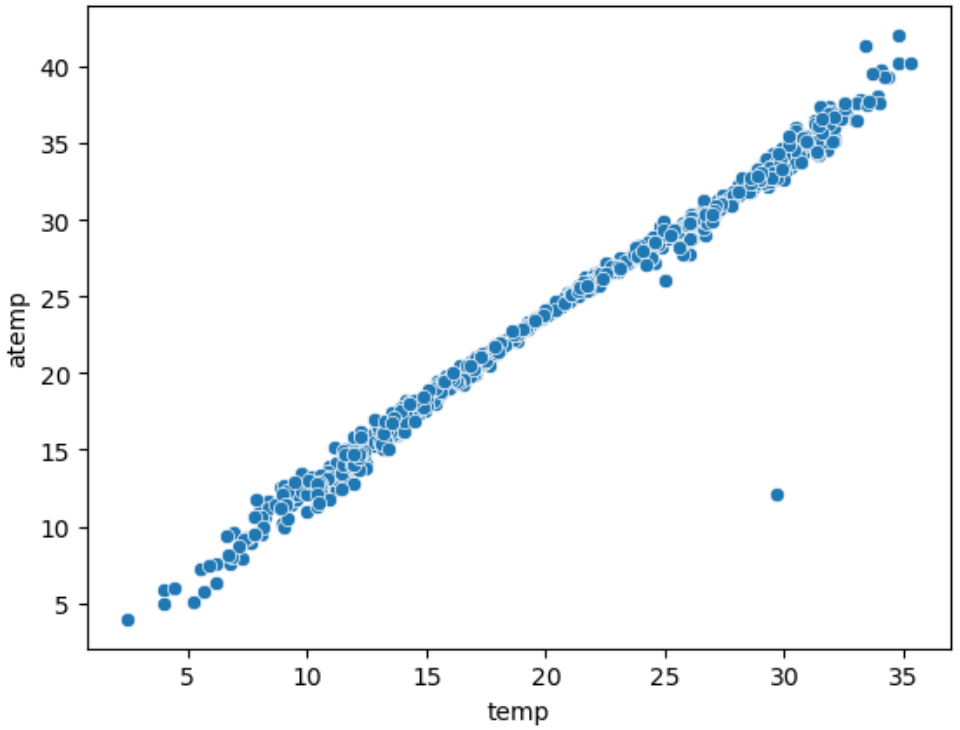


Question 2: Why is it important to use drop_first=True during dummy variable creation?

Answer- Nominal categorical variable with n labels/categories, n-1 Dummy variables with values represented by 0 (non-exist) and 1 (exist) are sufficient to capture all the variable information. Because n-1 dummy variables with all value of 0 (non-exist) indirectly represents the n^{th} 1 (exist) value. Therefore, keeping the n^{th} variable in the data will only add noise. In python, '**drop_first = True**' in pandas library make sure that first dummy variable or first level of categorical variable excluded. This will also help in avoiding multicollinearity.

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

Answer - Both **temp** and **atemp** variables have shown highest correlation (0.63) with target variable (**cnt**). Both temp and atemp variables are **highly correlated** to each other, as observed in the below chart.



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

Answer- There **four** assumptions to be validated in order to perform linear regression-

1. **Linearity** ➔ There should be a linear and additive relationship between dependent (response) variable and independent (predictor) variable(s).

   Validation: Final model has Adjusted R square value of more than 0.8 which is strong linear fit. Additionally, Bi-variate chart also confirmed the linear relationship between bike demand (cnt) and independent predictors.

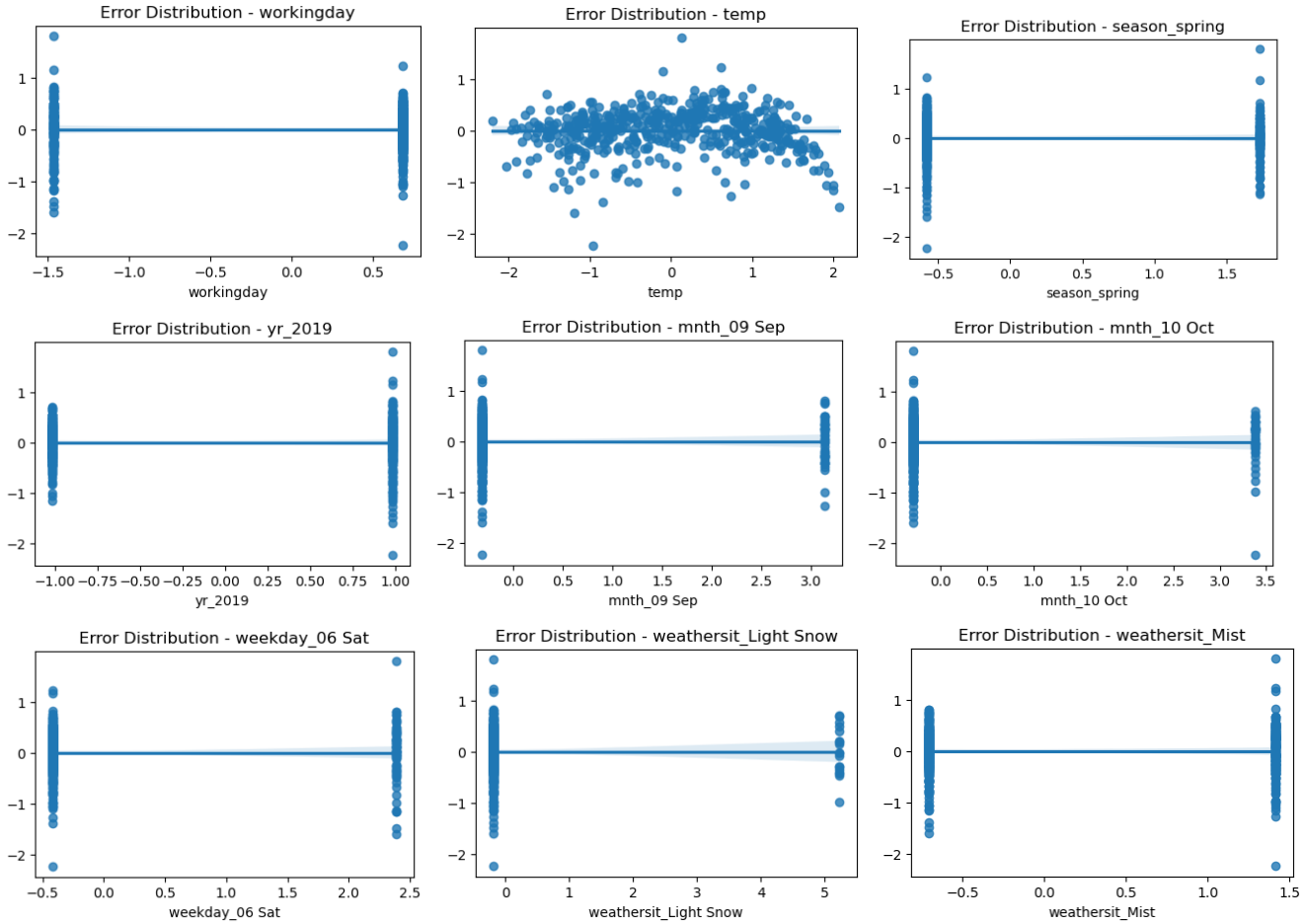
2. **No multicollinearity** ➔ The independent variables should not be correlated. Absence of this phenomenon is known as multicollinearity.

   Validation: Final model does not have VIF greater than 2 for any predictor thus multicollinearity does not exist in the final model.

Model Features	VIF
temp	1.72
season_spring	1.71
workingday	1.62
weekday_06 Sat	1.61
mnth_09 Sep	1.09
mnth_10 Oct	1.08
weathersit_Light Snow	1.07
weathersit_Mist	1.04
yr_2019	1.02
const	1

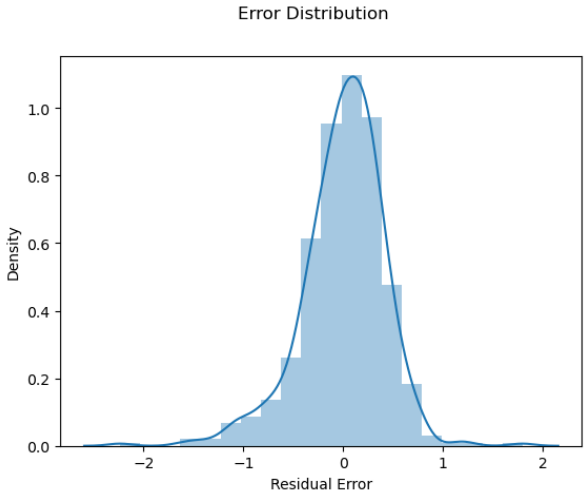
3. **Homoscedasticity** ➔ The error terms must have constant variance. This phenomenon is known as homoskedasticity.

   Validation: From the below charts we can confirm that errors are following Homoscedasticity which means there is no trend observed in the error variance across all the significant predictors



4. **Normality ➔** The error terms must be normally distributed. Validation on the assumptions of Linear Regression was done on basis of below: Residual analysis:

   Validation: From the residual error frequency chart we can confirm that residual errors are normally distributed.



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

Answer- Rescaling is performed in order to make the beta coefficient **comparable**. Without rescaling, strong and weak predictors cannot be distinguished, based on the naked eye.

Based on the final model below these are top 3 features-

1. **season_spring** (beta -0.3093) – This dummy variable represents Spring season which has shown **negative strong trend** with bike demand. It means that people have less preferred to use bike in spring season
2. **yr_2019** (beta +0.5463) – This dummy variable represents year 2019 which has shown **positive strong trend** with bike demand. Bike demand is higher in year 2019 w.r.t year 2018.
3. **temp** (beta +0.3657) -- positive strong trend observed. Higher the temperature, higher is the bike demand

General Subjective Questions

Question 1: Explain the linear regression algorithm in detail.

Answer- Linear regression is a method of finding the best straight line fitting the given data. It tries to find the best linear relationship between the independent variable and the dependent variable. The simplest form of linear regression is single linear regression in which we use one dependent variable and one independent variable. It is represented as below:

$$Y = bx + C$$

Y is the estimated dependent variable
b is the regression coefficient
C is constant
x is the independent variable

The above equation tries to predict the value of Y depending upon the value of x. Linear regression is used to determine the value of b and C in this case, how does a unit change in X effect Y and what will be the value of Y if X is zero.

For Multiple linear regression, the equation is updated as:

$$Y = b1x1 + b2x2 + b3x3\ldots\ldots + C$$ Where b1, b2 are coefficient of x1,x2… and so on.

In multiple linear regression, we have multiple independent variables and we try to find how change in one independent variable effect the dependent variable.

For a linear regression following assumptions are made:

1. There is linear relation between dependent and independent variables
2. The error terms are normally distributed
3. It is assumed the residual terms have same variance
4. Residual terms are independent of each other

Question 2: Explain the Anscombe's quartet in detail.

Answer- Anscombe's quartet developed by a statistician **Francis Anscombe** highlights the importance of visually analyzing the data. He created four datasets with nearly identical statistics but appear completely different when graphed. The idea was to stress the importance of visual analysis and counter the impression "numerical calculations are exact, but graphs are rough."

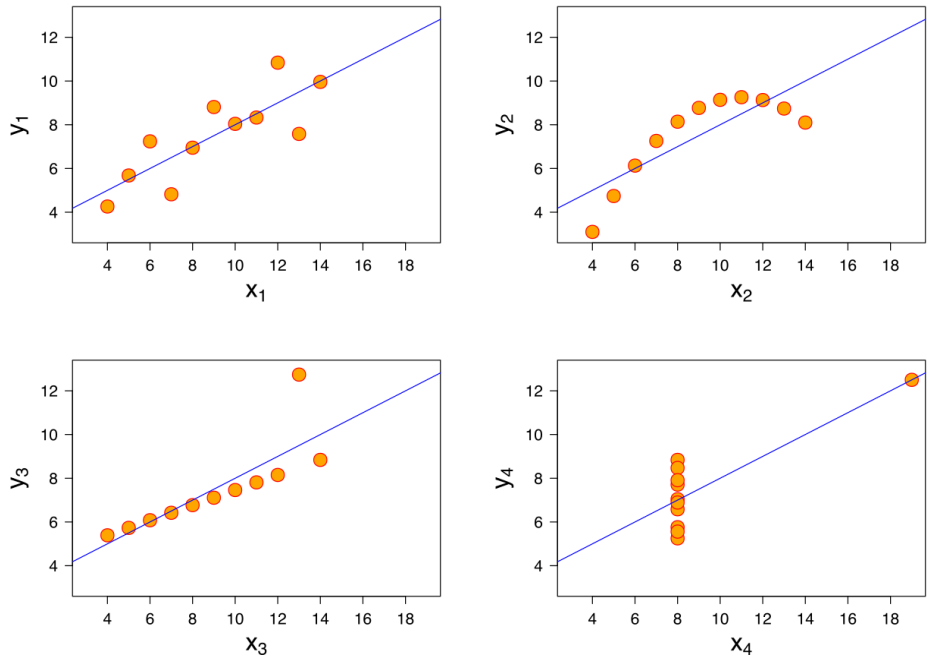


Image source: Wikipedia

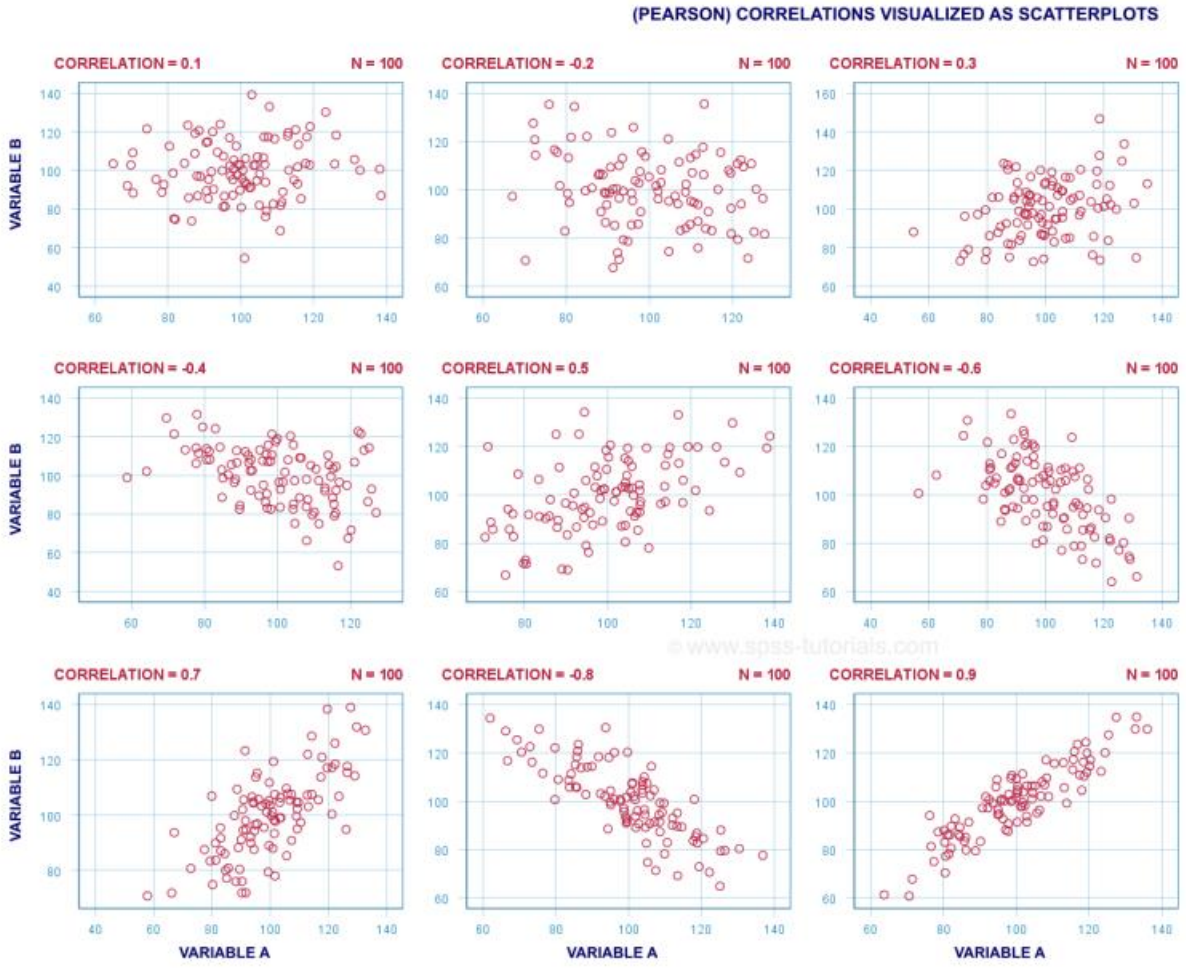
Above graphs tells us four different story but they are statistically same.

1) The first scatter plot (top left) appeared to be a simple linear relationship. Clean and well-fitting linear models
2) second graph (top right) was not distributed normally; while a relationship between the two variables is obvious.it is not linear and also pearson's co efficient is not relevant.
3) In the third graph (bottom left), the distribution is linear, but should have a different regression line, the calculated regression is thrown off by an outlier.
4) Fourth graph shows that one outlier is enough to produce a high correlation coefficient.

This Anscombe's quartet **emphasizes the importance of visualization in Data Analysis**.

Question 3: What is Pearson's R?

Answer: Pearson's R or Pearson product-moment correlation coefficient measures linear correlation between two variables X and Y. It is a measure of strength of the relation between two variables and their association with each other and explains the effect on one variable when other variable changes. Its value ranges from +1 to -1, +1 being total positive linear correlation, 0 meaning no linear correlation and -1 meaning total negative linear correlation. A positive correlation means that when X increases Y also increases whereas a negative correlation means when X decreases Y also decreases. It is represented as ρ for population and r for sample. Following graphs showing Pearson's correlation coefficient between two variables



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

Answer: Scaling is a method used to normalize/standardize the range or features of the data. Range of data for different variable may vary widely and that is the reason it is suggested to do scaling in the data pre-processing step when using a machine learning algorithm. When applying a machine learning algorithm, say linear regression, the gradient descent will take iterations to fit the line. If we have two variables whose range vary widely, gradient descent will be able to work it out in lesser number of iteration while will need a larger number of iteration for variable with a larger range. Hence, scaling is applied to bring down the cost function gradient descent.

Normalization is a scaling method in which the values are rescaled in such a way that they end up between 0 and 1. This is also known as Min-Max Scaling.

X' = X- Xmin / Xmax – Xmin

Xmax = Maximum value of X
Xmin = Minimum value of X

Normalization is helpful in cases where data follows a Gaussian distribution. It subsidizes the effect of outliers as it has a bounding range.

Standardization is a method in which we rescale the value to be centred around mean with a unit standard deviation. Mean of the attributes become 0 and the distribution has a unit standard deviation.

$X' = X-\mu/\sigma$

μ is the mean of the feature variable
σ is the standard deviation of the feature variable

Standardization can be helpful for cases where data doesn't follow Gaussian distribution. It doesn't take care of outliers as it doesn't have a bounding range

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

Answer- In the context of linear regression analysis, the Variance Inflation Factor (VIF) is a measure that quantifies multicollinearity, which is the correlation or high interdependency among predictor variables. VIF is used to assess the extent to which the variance of the estimated regression coefficients is inflated due to multicollinearity.

The formula for calculating the VIF of a predictor variable is as follows: VIF = 1 / (1 - R^2)

where R^2 represents the coefficient of determination obtained by regressing the predictor variable against all other predictor variables.

The VIF value provides insight into how much the variance of a particular predictor's coefficient is inflated due to multicollinearity.
In some cases, the VIF value can be infinite. This occurs when the coefficient of determination (R^2) for a particular predictor variable is equal to 1. A perfect correlation between a predictor variable and other predictor variables can lead to an R^2 of 1, resulting in an infinite VIF.

There are a few scenarios that can cause a predictor variable to have an R^2 of 1 and an infinite VIF:

1. Perfect Linear Relationship: The predictor variable is perfectly linearly related to one or more other predictor variables in the model. In this case, the VIF becomes infinite
   because the variance of the coefficient estimate cannot be determined separately from the other correlated variables.
2. Redundant Predictor: The predictor variable is a linear combination or a duplicate of another predictor variable(s) in the model. When two or more predictor variables provide the same information, it results in perfect multicollinearity and leads to an
   infinite VIF.

Having an infinite VIF suggests severe multicollinearity, indicating that the predictor variable is perfectly predictable from the other variables in the model. This can pose challenges in interpreting the model and estimating the effect of individual predictors. In such cases, it is necessary to address multicollinearity by identifying and resolving the high interdependency among the predictor variables. Techniques such as removing redundant variables, transforming variables, or using dimensionality reduction methods can help mitigate multicollinearity issues.

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

Answer- A Q-Q (quantile-quantile) plot, also known as a quantile plot or normal probability plot, is a graphical tool used to assess the distributional similarity between a sample of data and a theoretical distribution, typically the normal distribution. It helps to determine if the data follows a specific distribution or if it deviates from it.

In a Q-Q plot, the observed quantiles of the data are plotted against the quantiles of the theoretical distribution. If the data follows the theoretical distribution, the points in the plot should lie approximately along a straight line. Deviations from the straight line suggest departures from the assumed distribution.

In the context of linear regression, Q-Q plots are used to assess the assumption of normality for the residuals or errors. The residuals represent the differences between the observed values and the predicted values obtained from the linear regression model. The assumption of normality is important as it enables valid inference, hypothesis testing, and confidence interval estimation in linear regression.

The use and importance of Q-Q plots in linear regression are as follows:

Assessing Normality: By examining the Q-Q plot of the residuals, you can visually assess if the residuals follow a normal distribution. If the points on the

plot closely follow the diagonal line, it suggests that the residuals are approximately normally distributed. On the other hand, deviations from the line indicate departures from normality.

Detecting Skewness and Outliers: Q-Q plots can reveal skewness in the distribution of residuals. If the points on the plot deviate from the straight line, it may indicate a skewed distribution. Additionally, extreme deviations from the line may indicate the presence of outliers in the residuals.

Model Validity: Normality of residuals is a crucial assumption for linear regression models. Violations of this assumption can lead to biased coefficient estimates, incorrect standard errors, and invalid hypothesis testing. Q-Q plots provide a graphical tool to assess the validity of the normality assumption and identify potential issues with the model.

Remedial Actions: If the Q-Q plot reveals deviations from normality, it indicates a need for further investigation and potential remedial actions. It may be necessary to consider transformations of variables, identify influential observations, or explore alternative modeling techniques that can handle non-normal residuals.

In summary, Q-Q plots provide a visual assessment of the distributional assumptions in linear regression, particularly the assumption of normality for residuals. They help to identify departures from normality, skewness, and outliers, which are important considerations in ensuring the validity and reliability of the regression model.