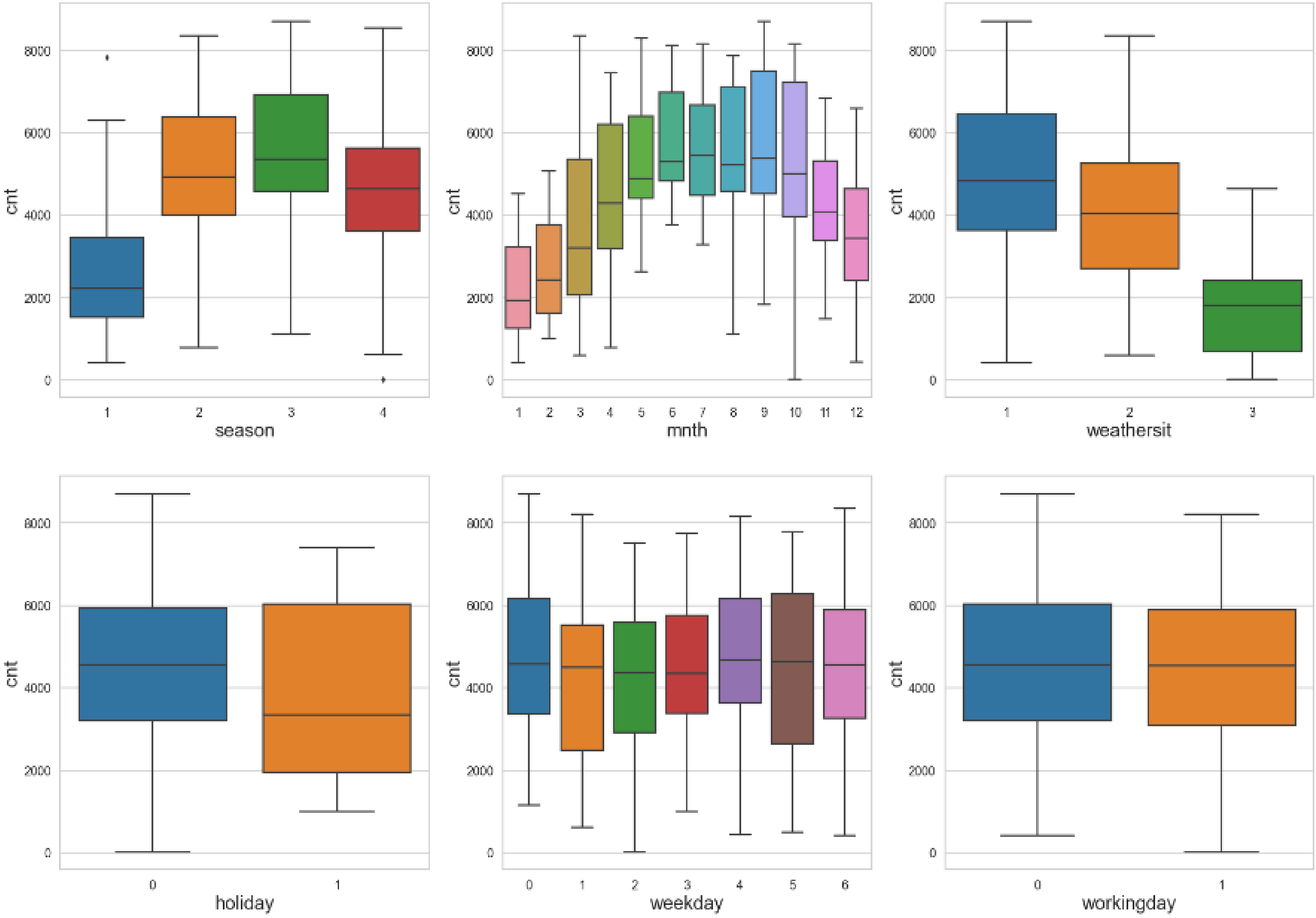
Assignment-based Subjective Questions

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

Visualizing Categorical Variables



❖ Q1 . Continued..

categorical variable details

season : season (1:spring, 2:summer, 3:fall, 4:winter)

weathersit :

1: Clear, Few clouds, Partly cloudy, Partly cloudy -
2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist -
3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog -

categorical variable	Inference	effect on the dependent variable
Season	**seson3** i.e fall is the top season to rent bike. with median above 5000 i.e 5000 bookings. This is followed by season2 i.e summer & season4 i.e winter with median i.e bookings between 4000 to 5000.	season can be a good predictor for the dependent variable.
Month	more than 4000 bike booking per month were happening in the **months 5,6,7,8 & 9**	Month can be a good predictor for dependenat variable.
weathersit	around 5000 bike bookings is happening during '**weathersit1**' i.e Clear, Few clouds, Partly cloudy, Partly cloudy weather. This is followed by **weathersit2** i.e Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist with around 4000 bike bookings This indicates, weathersit does show some trend towards the bike bookings	it can be a good predictor for the dependent variable.
Weekday:	it shows very close trend.	weekday is not a good predictor for the dependent variable it also carries same information as of holiday so we can drop this variable.
Holiday:	most of the bookings happend on working day	can be consider as good predictor

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

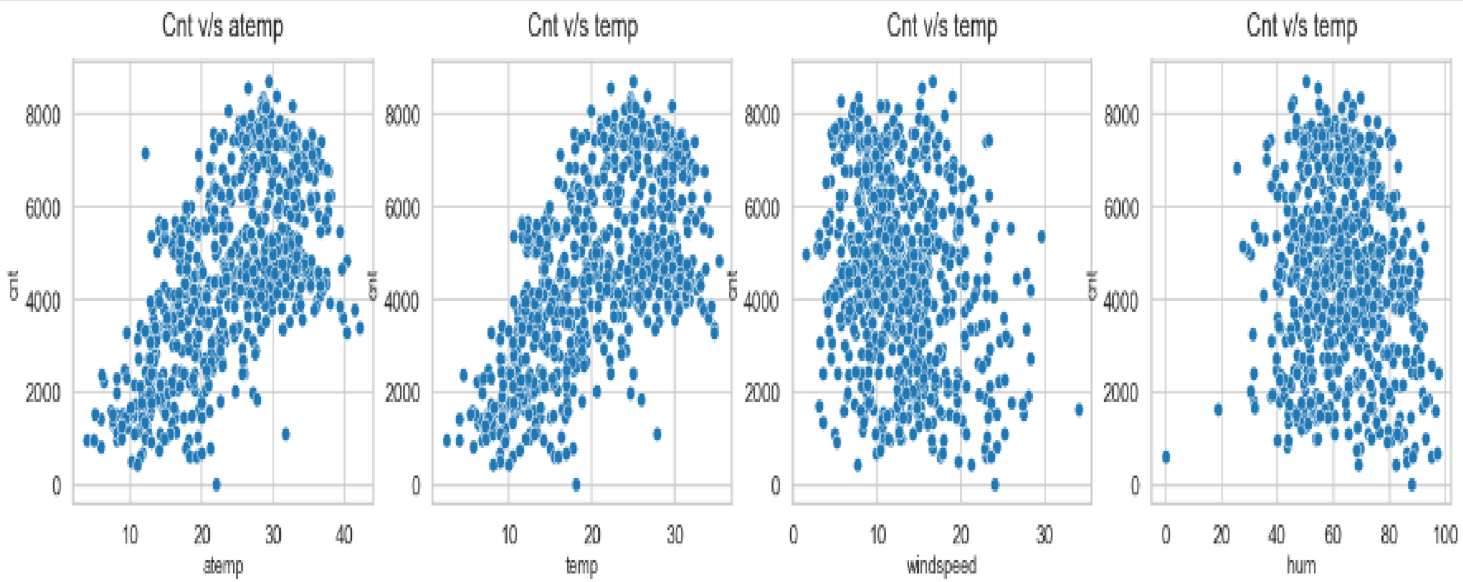
- In pandas **.get_dummies** parameter i.e. drop_first allows you Whether to get **k-1 dummies** out of **k categorical levels** by **removing the first level**.

- When you set drop_first = True ,then it will drop the first level

- Since one of the columns can be generated completely from the others, and hence retaining this extra level does not add any new information for the modelling process,

  would it be good practice to always drop the first level

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

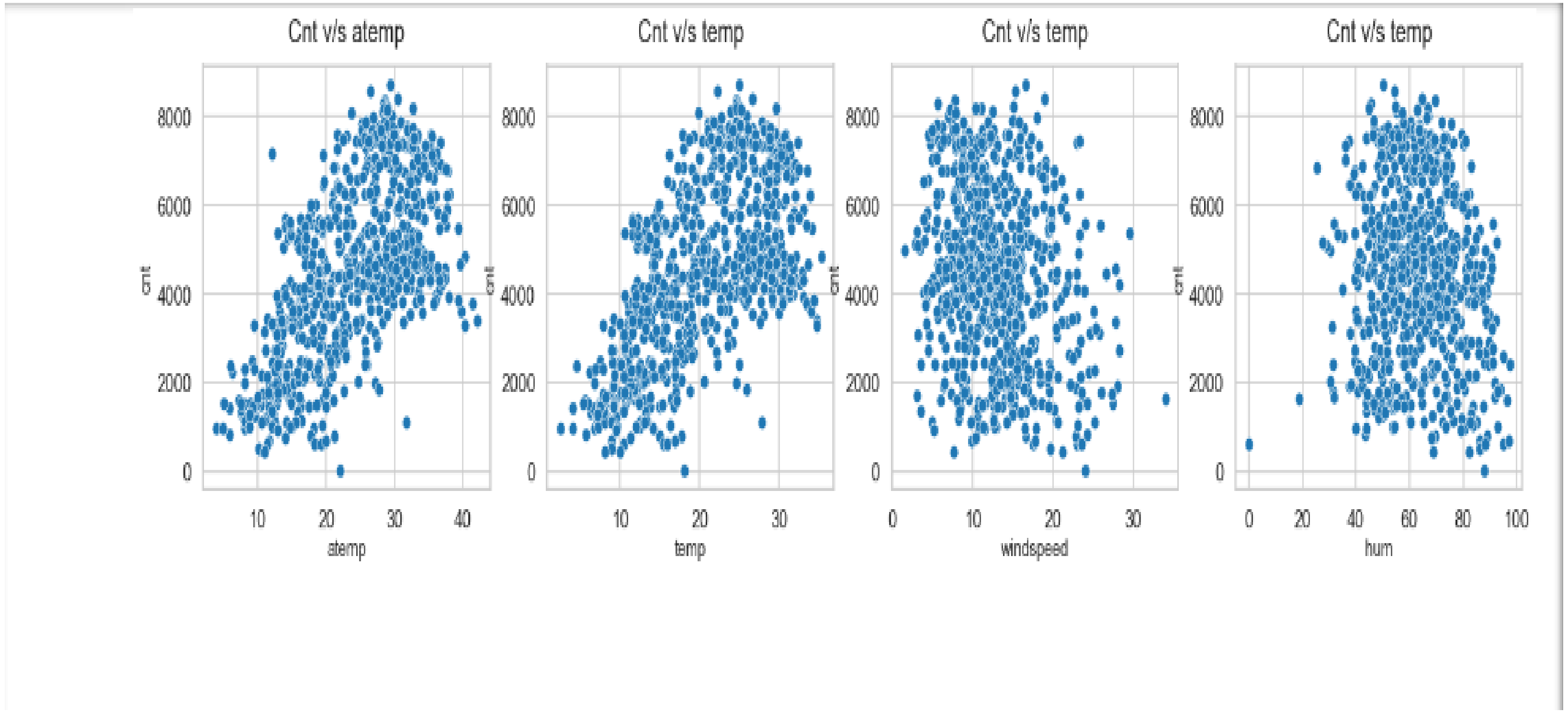


Inference

- atemp veriable has highest correlation with the target variable cnt

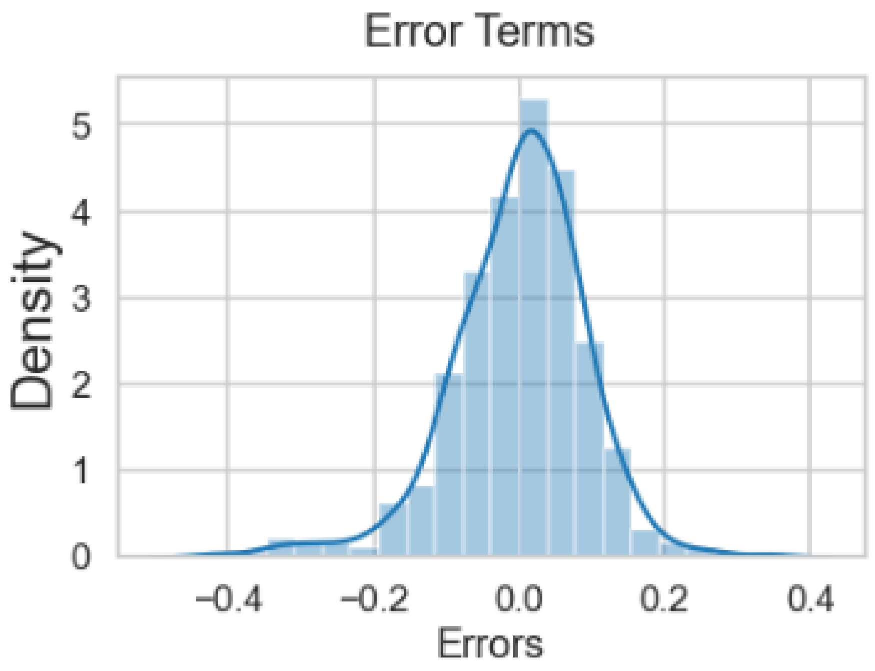
❖ Q4.How did you validate the assumptions of Linear Regression after building the model on the training set?

- **Linear Relationship :** Linear regression assumes that there exist a linear relationship between the dependent variables and predictor
  - **To validate : Pair-wise scatterplots** are used in validating the **linearity assumption** as it is easy to visualize a linear relationship on a plot.



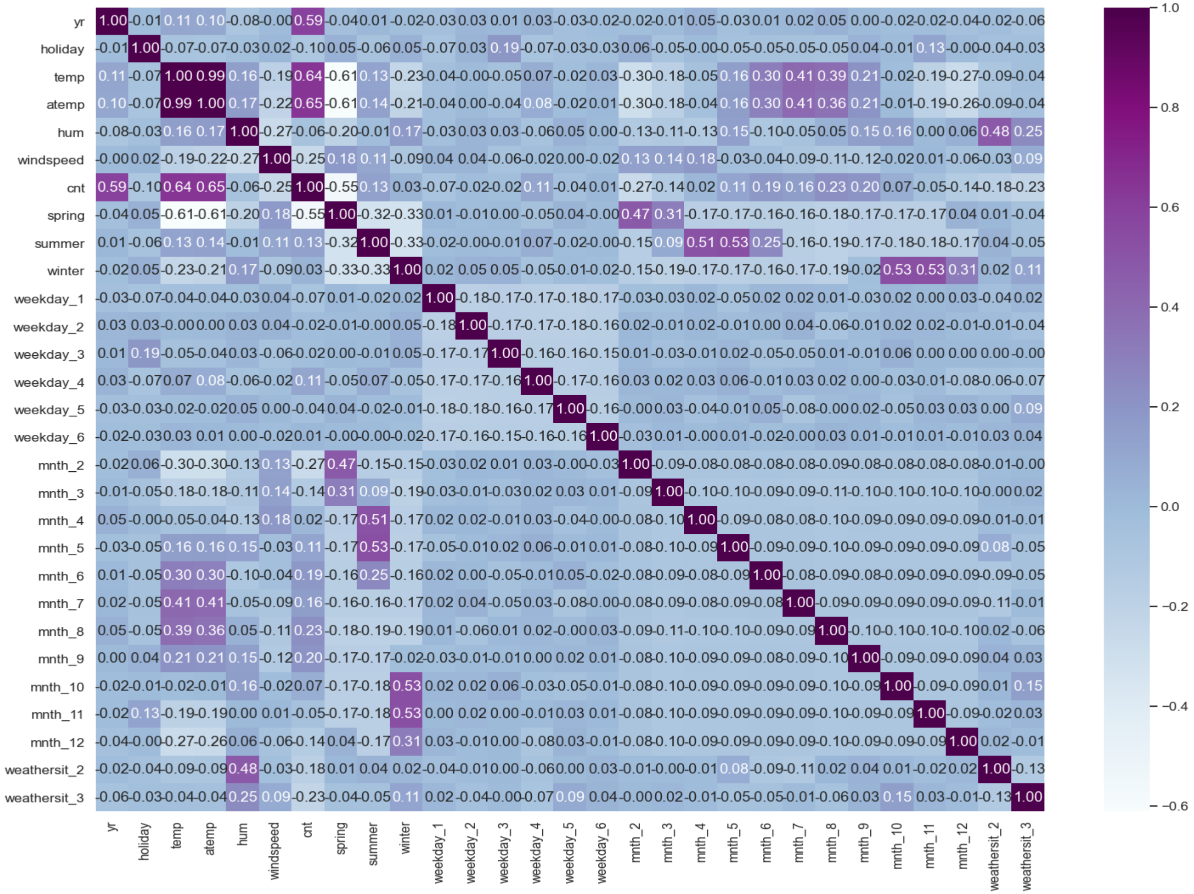
❖ Q4.continued..

- **Homoscedasticity :** Homoscedasticity means that the residuals have constant variance no matter the level of the dependent variable.
  - **To Validate** :**residual plot (displot)** is used to verify that the variance of the error terms is constant across the values of the dependent variable.



❖ Q4.continued..

- **Absence of Multicollinearity : Multicollinearity** refers – two or more independent variables are highly correlated
  - **To Validate : correlation heatmap** and **VIF** calculation (from statsmodels.stats.outliers_influence import variance_inflation_factor) is used



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

- **The equation of best fitted surface based on final model :**

cnt = 0.2649+yr(0.2649)-holiday(0.0896)+atemp(0.4253)-windspeed(0.1277)-spring(0.1274)+winter(0.0438)+mnth_3(0.0380)+mnth_9(0.0731)-weathersit_2(0.0792)-weathersit_3(0.2867)

- **Top 3 features** contributing towards explaining demand of the shared bikes are as below :

1. **atem** with coefficient value 0.4253

i.e a unit increase in atemp variable, increases the bike hire numbers by 0.4253 units.

2. weathersit_3 with coefficient value -0.2867

i.e bike rental is decreased when weather conditions are Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds

3. Yr with coefficient value 0.236

i.e a unit increase in yr variable, increases the bike hire numbers by 0.2649 units

General Subjective Questions

❖ Q1. Explain the linear regression algorithm in detail

- Linear Regression is a machine learning algorithm which is based on **supervised learning** category. It finds a best linear-fit relationship on any given data, between independent (Target) and dependent (Predictor) variables. In other words, it creates the best straight-line fitting to the provided data to find the best linear relationship between the independent and dependent variables. Mostly it uses **Sum of Squared Residuals** Method.

- Linear regression is of the 2 types:
  - **Simple Linear Regression** - It explains the relationship between a dependent variable and only one independent variable using a straight line.
    Formula : $Y=\beta0+\beta1X1 +\epsilon$
  - **Multiple Linear Regression-**shows the relationship between one dependent variable and several independent variables.
    Formula : $Y=\beta0+\beta1X1+\beta2X2+\ldots+\beta pXp+\epsilon$
    Two methods :
    - ✓ Differentiation
    - ✓ Gradient descent

...Continued

We can use statsmodels or SKLearn libraries in python for the linear regression.

Below pointers describes the algorithm of linear regression :

1. Retrieve independent variable (x) and dependent variable y
2. **Split the Data into Training and Testing Sets**
3. **RESCALING THE FEATURES -** If we don't have comparable scales, then some of the coefficients as obtained by fitting the regression model might be very large or very small as compared to the other coefficients. This might become very annoying at the time of model evaluation. So it is advised to use standardization or normalization so that the units of the coefficients obtained are all on the same scale. As we know, there are two common ways of rescaling:
    - Min-Max scaling
    - Standardisation (mean-0, sigma-1)

4. we fit the model on the training data and predict the values for the testing data. The method of least squares is used to minimize the residual.

5. However over-fitting occurs when the model is not efficient anymore. To identify whether the model is fitted efficiently a corrected R^2 is calculated .We use R2 score to measure the accuracy of our model.

6. The last step for the linear regression analysis is the test of significance. Linear regression uses two tests to test whether the found model and the estimated coefficients can be found in the general population the sample was drawn from. Firstly, the F-test tests the overall model. The null hypothesis is that the independent variables have no influence on the dependent variable. In other words the F-tests of the linear regression tests whether the $R^2=0$. Secondly, multiple t-tests analyze the significance of each coefficient and the intercept. The t-test has the null hypothesis that the coefficient/intercept is zero.

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

- **Anscombe’s quartet** comprises four datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven (x,y) points. which provides same statistical information that involves **variance**, and **mean** of all x,y points in all four datasets

- **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.

- This tells us about the importance of visualizing the data before applying various algorithms out there to build models out of them which suggests that the data features must be plotted in order to see the distribution of the samples that can help you identify the various anomalies present in the data like outliers, diversity of the data, linear separability of the data, etc. 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?

- Correlation between sets of data is a measure of how well they are related. The most common measure of correlation in stats is the Pearson Correlation.
- It shows the linear relationship between two sets of data.
- **Pearson's correlation coefficient (r)** is a **measure of the strength of the association** between the two variables.
- Two letters are used to represent the Pearson correlation: Greek letter rho (ρ) for a population and the letter "r" for a sample.
- The first step in studying the relationship between two continuous variables is to draw a scatter plot of the variables to check for linearity. The correlation coefficient should not be calculated if the relationship is not linear.
- The nearer the scatter of points is to a straight line, the higher the strength of association between the variables. Also, it does not matter what measurement units are used.
- Pearson's correlation coefficient (r) for continuous (interval level) data ranges from -1 to +1:
  - ✓ r = -1 : data lie on a perfect straight line with a negative slope
  - ✓ r = 0 : no linear relationship between the variables
  - ✓ r = 1 : data lie on a perfect straight line with a positive slope

- Positive correlation indicates that both variables increase or decrease together, whereas negative correlation indicates that as one variable increases, so the other decreases, and vice versa.

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

What is scaling : Feature scaling is a method used to normalize the range of independent variables or features of data.
It is performed during the data pre-processing to handle highly varying magnitudes or values or units

Why is scaling performed : Most of the times, 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 hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

Difference between normalized scaling and standardized scaling

normalized scaling	standardized scaling
It brings all of the data in the range of 0 and 1	Standardization replaces the values by their Z scores. It brings all of the data into a standard normal distribution which has mean (μ) zero and standard deviation one (σ).
sklearn.preprocessing.MinMaxScaler helps to implement normalization in python	sklearn.preprocessing.scale helps to implement standardization in python
	One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers
MinMaxScaling: x=X-min(x)/max(x)-min(x)	Standardisation: x=x-mean(x)/sd(x)

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

- VIF is an index that provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity.

VIF	Conclusion
1	No multicollinearity
4 - 5	Moderate
10 or greater	Severe

- If there is perfect correlation, then VIF = infinity

- In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity.

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

Q-Q plot : A Q-Q plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight

Use and importance of a Q-Q plot in linear regression :

- Q-Q(quantile-quantile) plots play a very vital role to graphically analyze and compare two probability distributions by plotting their quantiles against each other. If the two distributions which we are comparing are exactly equal then the points on the Q-Q plot will perfectly lie on a straight line y = x.
- Q-Q plots are used to find the type of distribution for a random variable whether it be a Gaussian Distribution, Uniform Distribution, Exponential Distribution or even Pareto Distribution, etc.