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)

There are various categorical variables such as mnth, weekday, season, weathersit, yr, workingday, holiday, etc. The scatter plot between these variables and dependent variable 'cnt' indicates that there is a linear relation between them. The heatmap depicting the co-relations also give some in insights -

a. There is a high correlation between cnt and months January and February as compared to other months
b. There is no correlation between weekdays and the independent variable cnt.
c. Season such as spring has a significant negative impact on the independent variable cnt.
d. Weather situations such light snow and mist have negative correlation with cnt.
e. Variable yr has a significant impact on cnt with high correlation.
f. Variables such as holiday and working day have relatively lesser impact on the cnt variable.

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

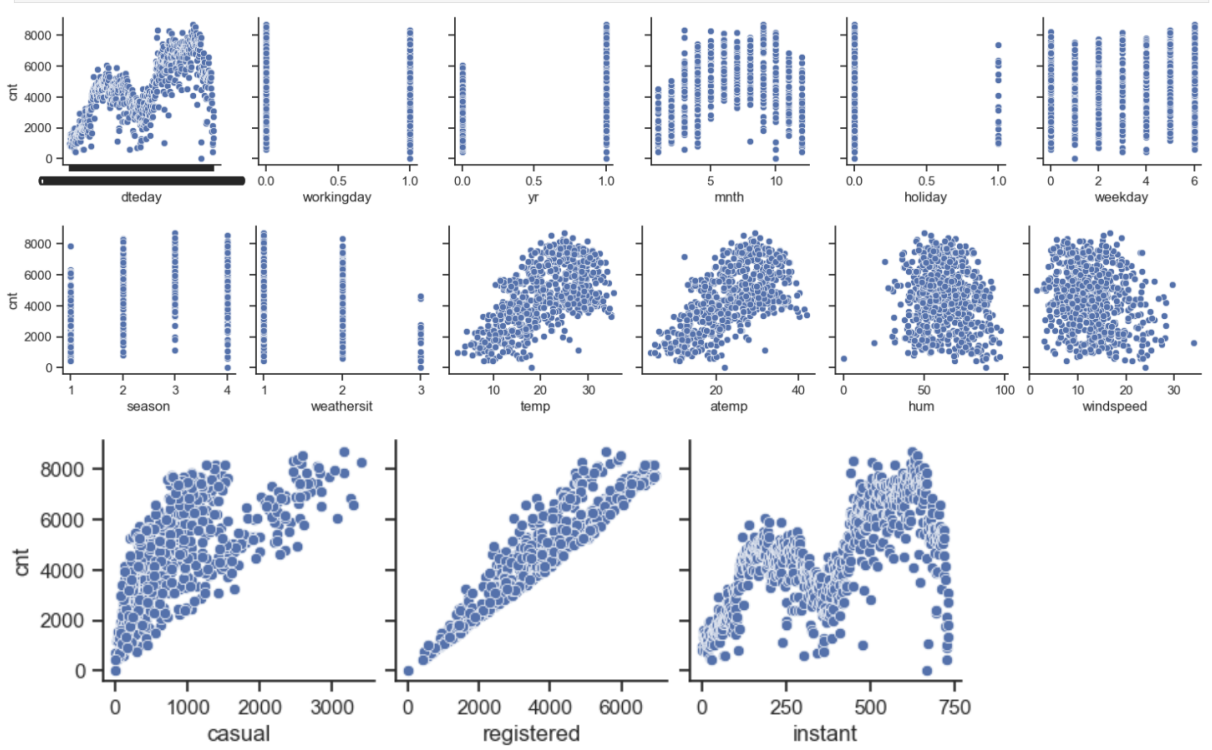
When creating dummy variables for categorical features with k levels, the conventional approach is to create $k-1$ dummy variables. This is because if all the levels are included it can lead to -

a. Multicollinearity issue - This is because the information in the k-th dummy variable can be perfectly predicted from the first $k-1$ dummy variables.
b. Model complexity - Including unnecessary dummy variables in a model increases its complexity without adding valuable information.

The method get_dummies in panda creates k dummy variables, hence there is a need to use drop_first=True to reduce to k-1 dummy variables.

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

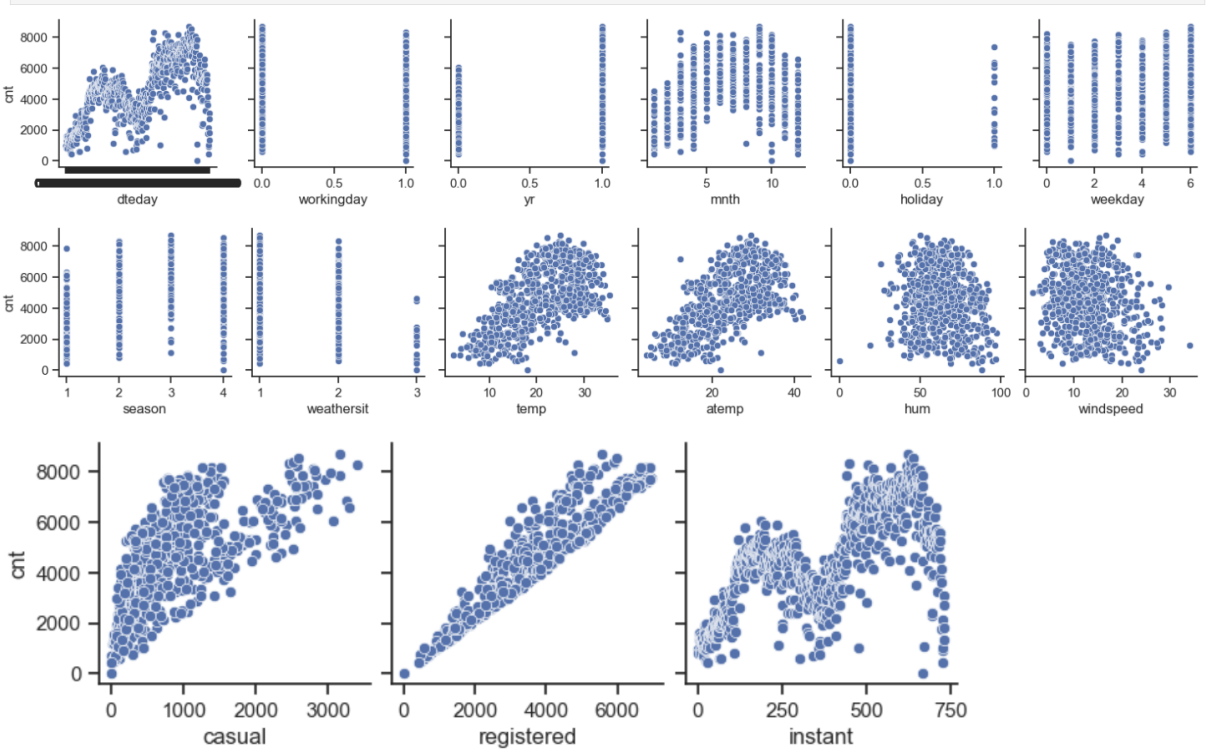
The variables temp and atemp have the hight correlation with the target variable apart from casual and registered which can be ignored as cnt is a direct sum of casual and registered.


cnt
dteday
workingday
yr
mnth
holiday
weekday
season
weathersit
temp
atemp
hum
windspeed
casual
registered
instant

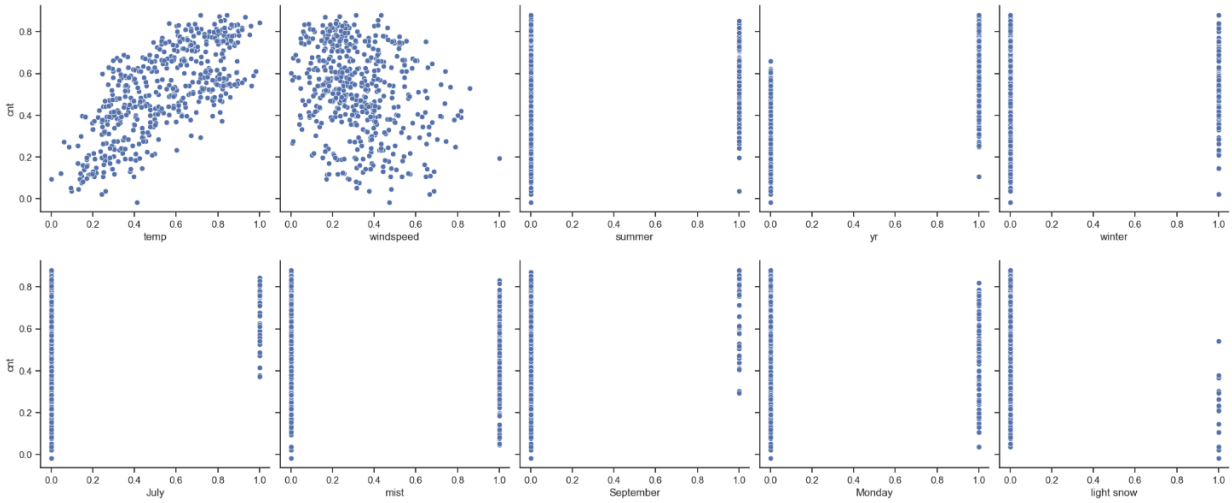
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

The different assumptions of Linear regression are validated as follows -

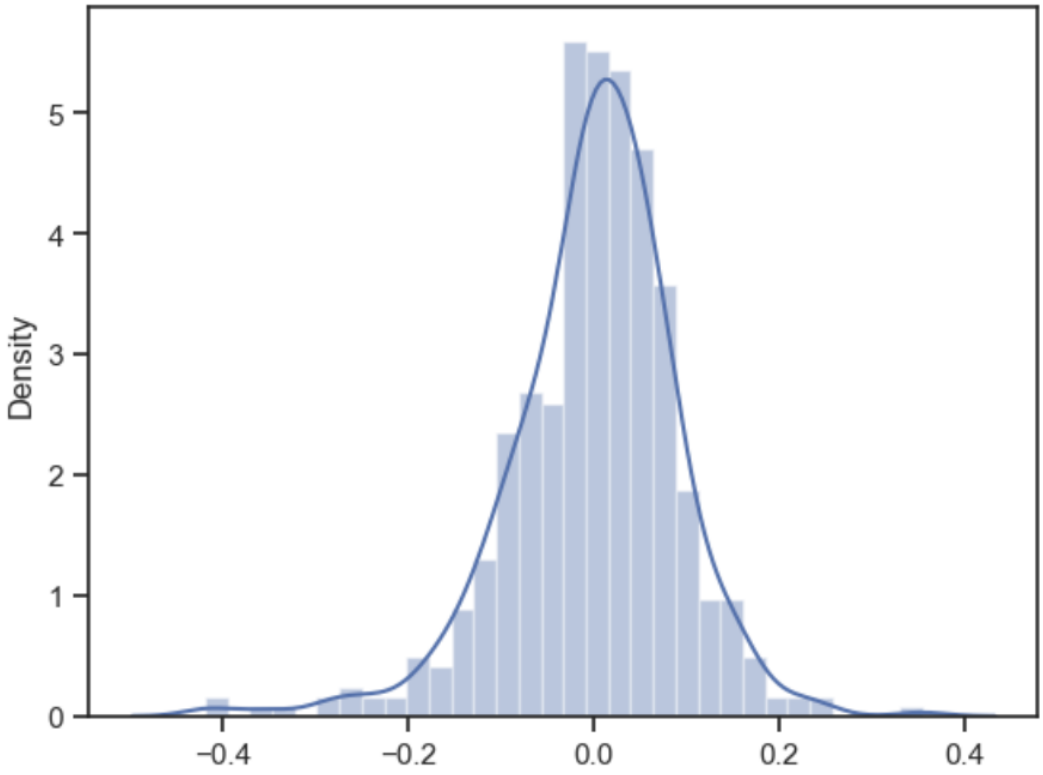
1. **Linear relationship between features and independent variable** - The various pair plots between cnt and different features from the datasets indicates that the relationship is linear.



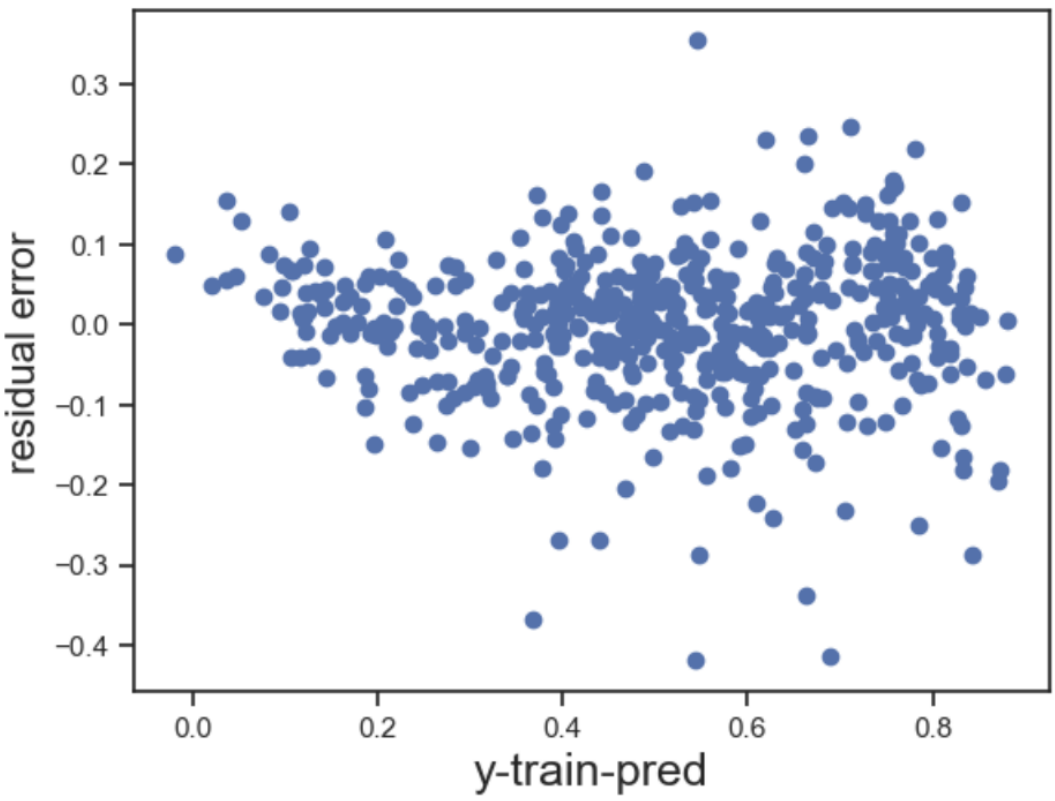
Also, the pairplot between y-train-pred and different other features is still linear indicating that the linear model is a true fit.



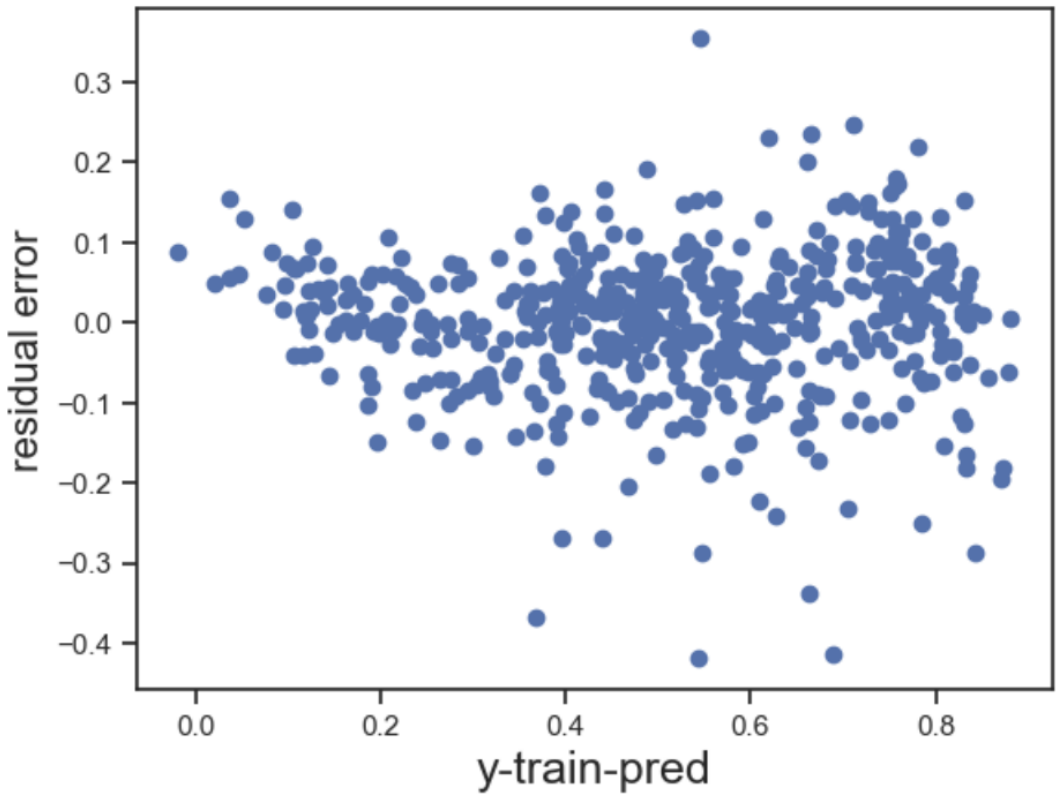
2. Error terms are normally distributed with mean at 0 - the PDF plotted for residual error indicates that error terms are normally distributed with mean =0



3. Error terms are independent of each other - The scatter plot between error terms and y_train_pred clearly indicates that the error terms are independent of each other with no visible pattern.



4. Error terms have constant variance - The scatter plot between error terms and y_train_pred also indicates that the spread of error terms is same and consistent across all predicted values indicating homoscedasticity.



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

Year (yr):
Coefficient: 0.2338
p-value: 0.000 (highly significant)
Since the coefficient is positive, it means from 2018 to 2019 the demand for bikes has increased.

Light Snow (light snow):
Coefficient: -0.2856
p-value: 0.000 (highly significant)
Weather situation - "Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds" is also a significant contributor.
The negative coefficient indicates that during light snow conditions, the demand for shared bikes decreases. This is understood, as snowy weather can lead to less number of bikes used.

Temperature (temp):
Coefficient: 0.4923
p-value: 0.000 (highly significant)
The positive coefficient implies that as the temperature increases, the demand for shared bikes also increases. Warm days can see increased demand for bikes. This is in sync with light snow weather situations, as temp generally decreases in such situations leading to drop in bikes demand.

General Subjective Questions

1. **Explain the linear regression algorithm in detail. (4 marks)**
Linear regression is a supervised learning algorithm used for predicting a continuous outcome variable (also called the dependent variable) based on one or more predictor variables (independent variables). The algorithm assumes a linear relationship between the predictor variables and the target variable.

The goal is to find the best-fit line that minimizes the sum of squared differences between the predicted and actual values. Here's a detailed explanation of the linear regression algorithm:

Model Representation:

The linear regression model is represented by the equation:

$Y=\beta0 + \beta1X1 + \beta2X2 +\ldots+ \beta nXn + \varepsilon$

- Y is the dependent variable (target/outcome).
- $X1$ $X2,\ldots,Xn$ are the independent variables (features/predictors).
- $\beta0$ is the intercept (the value of Y when all X's are zero).
- $\beta1, \beta2,\ldots,\beta n$ are the coefficients (slope) representing the change in Y for a one-unit change in X.
- ε is the error term, representing the unobserved factors affecting Y.

Cost Function:

The goal is to minimize the sum of squared differences between the predicted ($\hat{Y}$) and actual(Y) values:

$\text{Minimize}\sum i=1,m(Yi - \hat{Y}i)^2$

Finding the Coefficients:

The coefficients $\beta0,\beta1,\ldots,\beta n$ are estimated using the method of least squares.

Making Predictions:

Once the coefficients are determined, predictions for new data points can be made using the linear equation:

$Y\hat{} = \beta 0 + \beta 1X1 + \beta 2X2 + \ldots + \beta nXn$

Assumptions of Linear Regression:

- **Linearity**: The relationship between the variables is assumed to be linear.
- **Independence**: Observations are assumed to be independent.
- **Homoscedasticity**: Residuals should have constant variance.
- **Normality of Residuals**: Residuals are assumed to be normally distributed.
- **No Multicollinearity**: Predictor variables should not be highly correlated.

Evaluation:

Common metrics for evaluating the performance of a linear regression model include mean squared error (MSE), R-squared, and adjusted R-squared.

Implementation:

Linear regression can be implemented using various programming languages (Python, R, etc.) and libraries (Scikit-Learn, StatsModels, etc.). The process involves data reading, data preparation, model fitting, and evaluation.

Linear regression is a foundational algorithm and forms the basis for more complex regression techniques.

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

Anscombe's quartet is a set of four datasets that have nearly identical simple descriptive statistics (mean, variance, correlation, and regression coefficients), yet they have very different distributions and appear very different when graphed.

This set of datasets was created by the statistician Francis Anscombe in 1973 to emphasize the importance of visualizing data and not relying solely on summary statistics. The quartet highlights the limitations of only relying on summary statistics without inspecting the actual data.

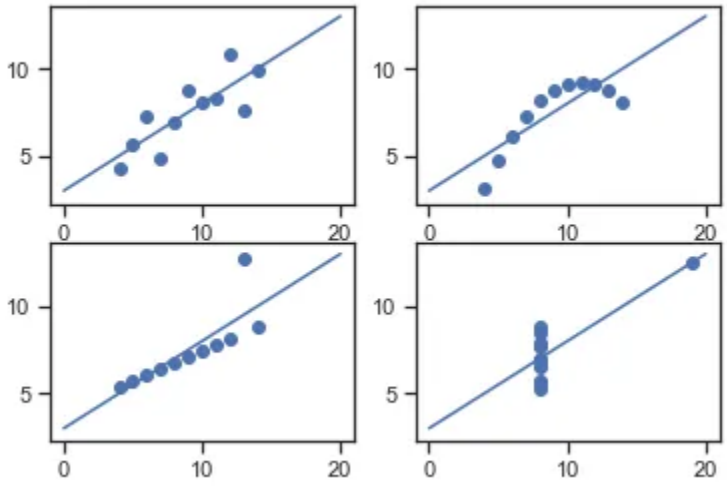
The four datasets in Anscombe's quartet are labeled I, II, III, and IV. Each dataset consists of 11 (x, y) pairs. Let's discuss each dataset in detail:

x1	y1	x2	y2	x3	y3	x4	y4
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

Apply the statistical formula on the above data-set:

- Average Value of x = 9,
- Average Value of y = 7.50
- Variance of x = 11
- Variance of y =4.12
- Correlation Coefficient = 0.816
- Linear Regression Equation : y = 0.5 x + 3

However, the statistical analysis of these four data-sets are pretty much similar. But when we plot these four data-sets across the x & y coordinate plane, we get the following results & each pictorial view represents a different behavior.



- Data-set I — consists of a set of (x,y) points that represent a linear relationship with some variance.
- Data-set II — shows a curve shape but doesn't show a linear relationship.
- Data-set III — looks like a tight linear relationship between *x* and *y*, except for one large outlier.
- Data-set IV — looks like the value of *x* remains constant, except for one outlier as well.
- In short Anscombe's quartet emphasizes the importance of visualizing data to gain insights and detect patterns that may not be apparent in summary statistics alone.
- It serves as a cautionary example against overreliance on summary statistics without exploring the actual data.

3. What is Pearson's R? (3 marks)

Pearson's correlation coefficient, often denoted as *r*, is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables. It is named after Karl Pearson, who introduced the concept.

The formula for Pearson's correlation coefficient between variables *X* and *Y* with *n* data points is given by:

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

- r = correlation coefficient
- Xi, yi are individual data points
- X-bar, y-bar are means of x, y respectively.

The value of *r* ranges between -1 and 1:

- $r=1$: Perfect positive linear correlation.
- $r=-1$: Perfect negative linear correlation.
- $r=0$: No linear correlation.

Interpretation of the magnitude of *r*:

- Close to 1: Strong positive correlation.
- Close to -1: Strong negative correlation.
- Around 0: Weak or no linear correlation.

It's important to note that Pearson's correlation coefficient measures only linear relationships. If the relationship between variables is nonlinear, Pearson's *r* may not accurately reflect the strength and direction of the association.

In Python, you can compute Pearson's correlation coefficient using libraries such as NumPy or with functions provided in libraries like SciPy or pandas. For example, in pandas:

correlation = df['X'].corr(df['Y'])

print(f"Pearson's correlation coefficient: {correlation}")

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

Scaling is the process of transforming the numerical features of a dataset to a specific range or distribution. The goal is to bring all features to a similar scale, preventing some features from dominating others in situations where the numerical ranges differ widely. Scaling is a crucial preprocessing step in many machine learning algorithms.

Why Scaling is Performed:

Equal Weight: Scaling ensures that all features contribute equally to the model training process. Features with larger numerical ranges might otherwise dominate the learning algorithm.
Convergence: Some machine learning algorithms, particularly those based on distances (e.g., k-nearest neighbors, k-means clustering), are sensitive to the scale of features. Scaling helps these algorithms converge faster.
Regularization: Regularized linear models (e.g., Lasso, Ridge regression) are sensitive to the scale of features. Scaling prevents certain features from having disproportionately large effects on the regularization term.
Interpretability: Scaling aids in the interpretability of coefficients in linear models. Without scaling, coefficients might not be directly comparable.

Difference between Normalized Scaling and Standardized Scaling:

Normalized Scaling (Min-Max Scaling):

- Formula: Xnormalized = (X − Xmin) / (Xmax - Xmin).
- Range: Scales data to a specific range, usually [0, 1].
- Advantages: Simple and intuitive. Preserves the shape of the original distribution.

Standardized Scaling (Z-score Scaling):

- Formula: Xstandardized = $X-\mu$ /σ
- Range: Centers the data around zero with a standard deviation of 1.
- Advantages: Works well when the data follows a Gaussian distribution.
- Preserves information about outliers.

In summary, both normalized and standardized scaling aim to bring features to a comparable scale, but they do so in different ways and are suitable for different scenarios. Normalized scaling is often preferred when the distribution of the data is not Gaussian, while standardized scaling is more suitable when Gaussian assumptions hold and when preserving information about outliers is important.

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

The Variance Inflation Factor (VIF) is calculated to understand the correlation between the independent variables and is denoted by

VIF = 1/(1 - R^2) , where R^2 represents the square of correlation.

An infinite VIF indicates that the denominator is zero, implying 1 - R^2 = 0. Consequently, R^2 equals 1, signifying a perfect correlation with a pearson correlation coefficient of 1. A correlation coefficient of 1 implies a high level of correlation between independent variables, signifying **perfect multicollinearity.**

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

Q-Q plots:

Q-Q plots are graphical tools that help to assess the validity of some assumptions in regression models, such as normality, linearity, and homoscedasticity. It is a short for quantile-quantile plot, a scatterplot that compares the quantiles of two distributions. One distribution is usually the observed data, and the other is a theoretical or reference distribution, such as the normal distribution. The idea is to see how well the data fit the expected distribution by checking if the points lie on or near a straight line.

Create Q-Q plot:

To create a Q-Q plot, we need to sort the data from smallest to largest and assign them ranks. Then calculate the expected quantiles of the reference distribution for each rank. For example, if we use the normal distribution, we can use the inverse cumulative distribution function (CDF) to find the expected quantiles. Finally, we need to plot the observed data on the y-axis and the expected quantiles on the x-axis.

Interpret Q-Q plot:

To interpret a Q-Q plot, we need to look at the shape and pattern of the points. If the points lie on or close to a 45-degree line, it means that the data follow the reference distribution closely. If the points deviate from the line, it means that there are some differences between the data and the reference distribution. For example, if the points are curved, it means that the data are skewed or have heavy tails. If the points are scattered or have gaps, it means that the data have outliers or are multimodal.