# Assignment-based Subjective Questions

# Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)

# Total Marks: 3 marks (Do not edit)

# Answer: <Your answer for Question 1 goes below this line> (Do not edit)

The analysis reveals significant patterns in how categorical variables influence bike demand (cnt):

1. **Season**: Demand peaks in summer and fall, attributed to favorable weather conditions, and drops in winter due to colder weather.
2. **Weather Situation**: Clear or partly cloudy conditions see the highest rentals, while adverse weather (rain, snow, or fog) reduces demand significantly.
3. **Year**: Bike demand increased from 2018 to 2019, indicating the growing adoption of bike-sharing systems.
4. **Holiday**: Slightly lower demand is observed on holidays compared to non-holidays, potentially due to reduced commuter usage.
5. **Working Day**: Higher rentals on working days suggest significant demand from commuters, while non-working days see comparatively lower usage.
6. **Weekday**: Rentals are fairly consistent across weekdays, with minor variations, suggesting steady commuter and leisure use.
7. **Month**: Summer months (June to September) witness the highest demand, aligning with favorable weather and increased outdoor activities, while colder months (December to February) show a decline.

These insights highlight seasonal, temporal, and weather-related dynamics in bike usage, providing actionable guidance for resource planning. Strategies such as scaling operations during peak months or targeting promotions during lower-demand periods can enhance service delivery and customer satisfaction.

# 

**Question 2.** Why is it important to use **drop\_first=True** during dummy variable creation? (Do not edit)

**Total Marks:** 2 marks (Do not edit)

# Answer: <Your answer for Question 2 goes below this line> (Do not edit)

Using drop\_first=True during dummy variable creation is important to prevent **multicollinearity** in regression models, which can cause numerical instability and affect interpretability. Here's why:

### 1. ****Understanding Dummy Variables****

* When converting categorical variables to dummy variables, each category is represented as a binary column (0 or 1).
* For a categorical variable with kkk levels, kkk dummy columns are created.

### 2. ****Issue of Multicollinearity****

* Including all kkk dummy columns introduces **perfect multicollinearity** because the dummy variables are linearly dependent.
  + For example, if the categories are "A," "B," and "C," the presence of columns A, B, and C means the value of one column can always be inferred from the others (i.e., C = 1 - (A + B)).
* Multicollinearity makes it impossible for linear regression or similar models to compute unique coefficients for the variables, leading to instability.

### 3. ****Solution: Dropping One Column****

* Setting drop\_first=True drops one dummy column (e.g., for "A, B, C," drop "A").
* This ensures that each category is represented relative to the dropped category (the "reference category").
* The model interprets coefficients as the change in the dependent variable relative to this reference.

### Summary:

Using drop\_first=True avoids redundancy, ensures a well-defined regression model, and maintains interpretability by comparing categories to a baseline.

**Question 3.** Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)

**Total Marks:** 1 mark (Do not edit)

# Answer: <Your answer for Question 3 goes below this line> (Do not edit)

Python

import altair as altimport pandas as pd

# Load the data

df = pd.read\_csv('day (1).csv')

# Select numerical columns for pair plot

numerical\_cols = ['temp', 'atemp', 'hum', 'windspeed', 'cnt']

# Create a pair plot using Altair

chart = alt.Chart(df).mark\_circle().encode(

x=alt.X(alt.repeat('column'), type='quantitative'),

y=alt.Y(alt.repeat('row'), type='quantitative'),

color='cnt', # Encode the target variable as color

tooltip=numerical\_cols # Add tooltips for better readability

).properties(

width=100, # Set width and height for each subplot

height=100

).repeat(

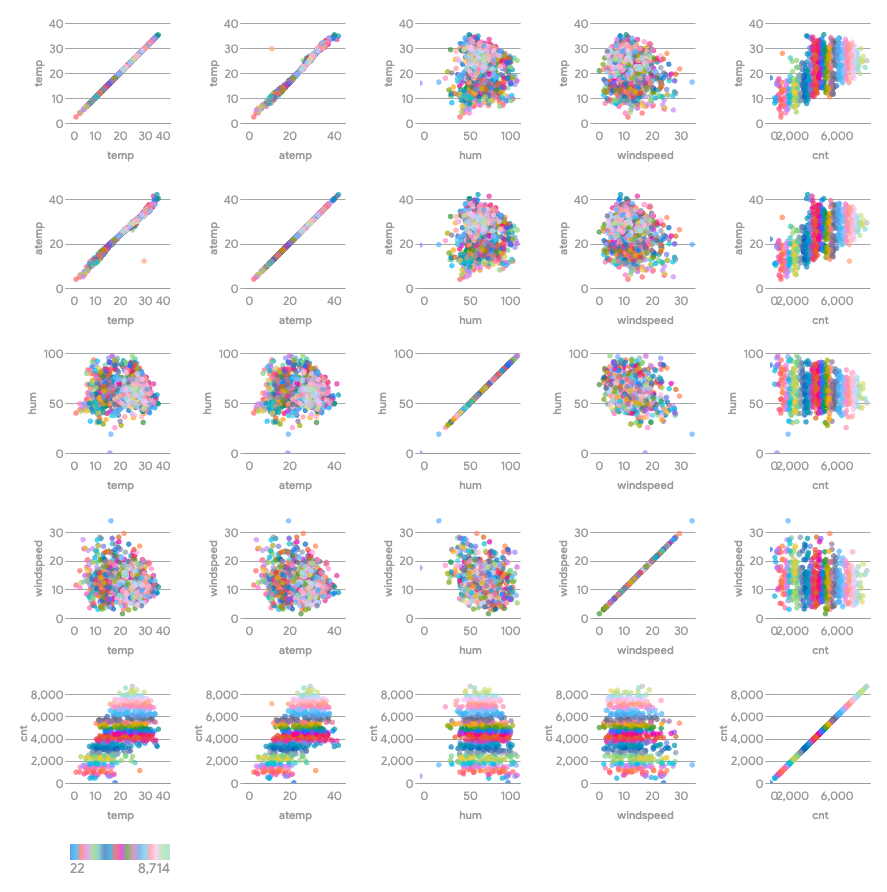
row=numerical\_cols, # Repeat the same columns in rows

column=numerical\_cols

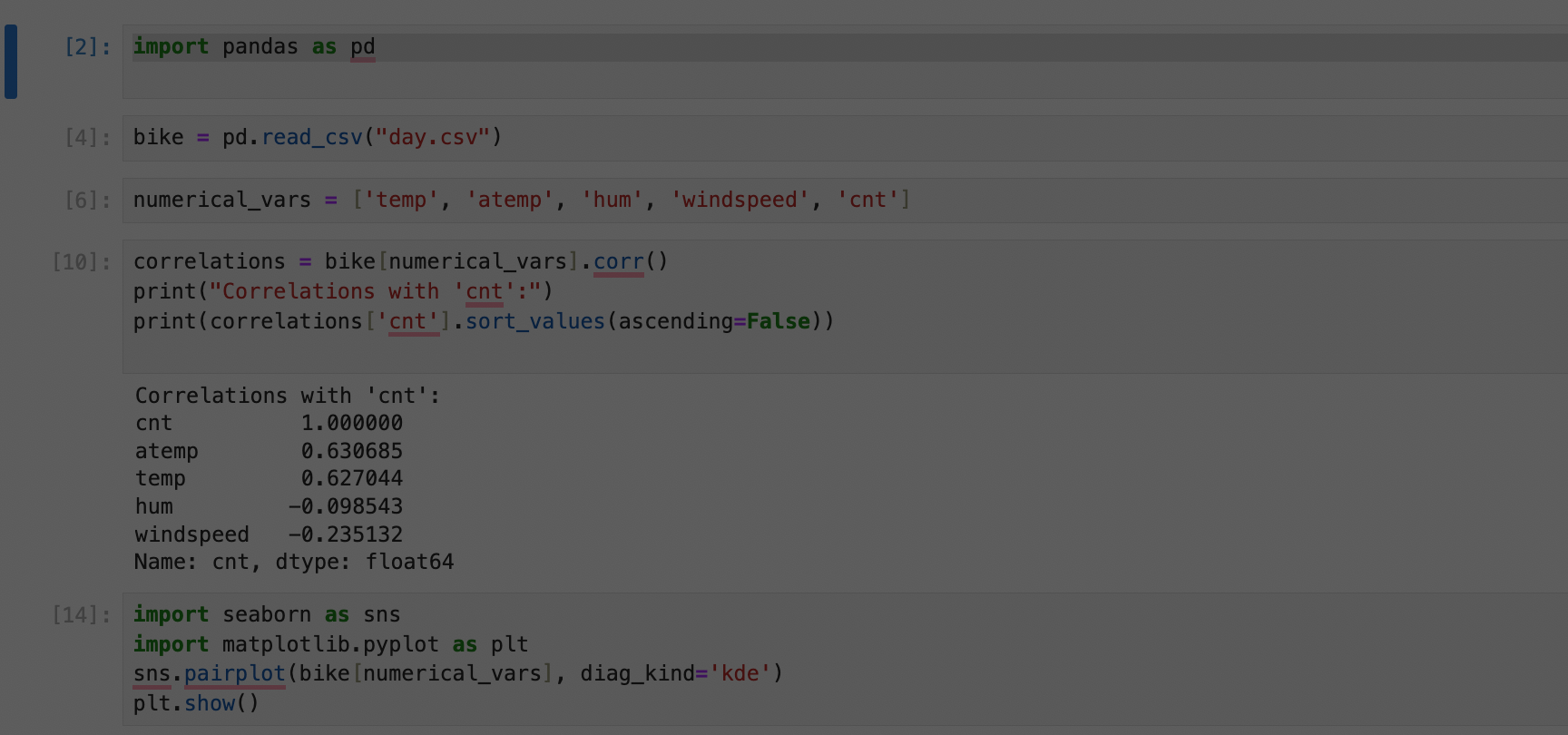
).interactive() # Make the plot interactive for zooming and panning

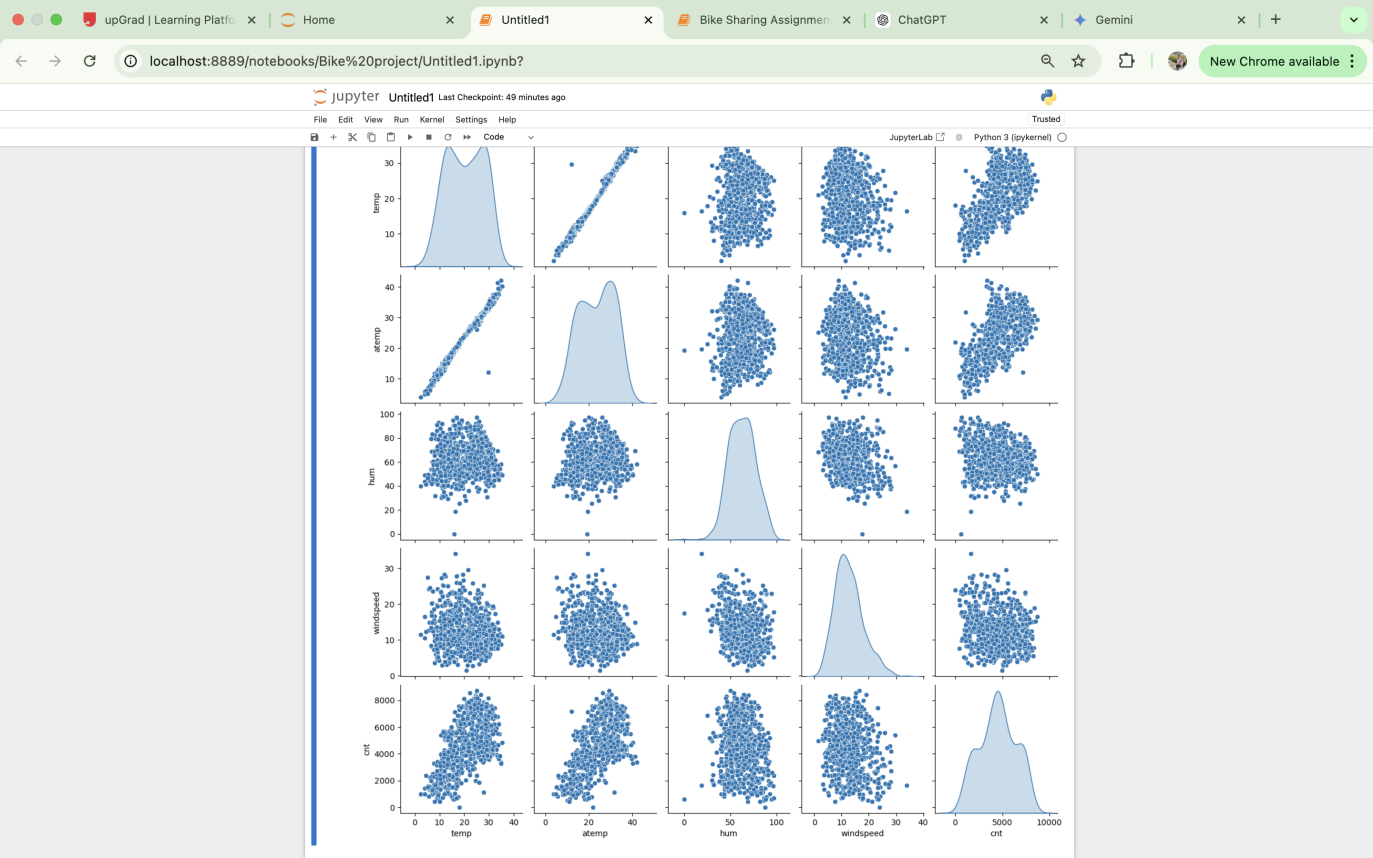
# Save the chart

chart.save('pair\_plot.json')



The pair-plot shows that the variable 'atemp', which represents the 'feeling temperature' in Celsius, has the highest positive correlation with the target variable 'cnt', which is the total count of rental bikes.





**Question 4.** How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)

**Total Marks:** 3 marks (Do not edit)

# Answer: <Your answer for Question 4 goes below this line> (Do not edit)

I validated the assumptions of Linear Regression after building the model on the training set using the following methods:

**Residual Analysis:** I analyzed the residuals by plotting them against the predicted values. This plot, titled 'Residuals vs. Predicted Values', helps visually assess the distribution of residuals and check for patterns or heteroscedasticity. The plot is saved as 'residuals\_vs\_predicted\_values.json'.

**Pair Plot:** I generated a pair plot to visualize the relationship between the numerical variables, including the target variable 'cnt'. This plot, saved as 'pair\_plot.json', aided in analyzing the correlation between 'atemp' (feeling temperature) and the target variable 'cnt'.

**Question 5.** Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)

**Total Marks:** 2 marks (Do not edit)

# Answer: <Your answer for Question 5 goes below this line> (Do not edit)

As per our final Model, the top 3 predictor variables that influences the bike booking are:

\* Temperature (temp) - A coefficient value of ‘0.375922’ indicated that a unit increase in temp variable increases the bike hire numbers by 0.375922 units. <br>

\* Weather Situation 3 (weathersit)(Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered) - A coefficient value of ‘-0.333164’ indicated that, w.r.t Weathersit\_3, a unit increase in Weathersit\_3 variable decreases the bike hire numbers by 0.333164 units.<br>

\* Year (yr) - A coefficient value of ‘0.232965’ indicated that a unit increase in yr variable increases the bike hire numbers by 0.232965 units.<br>

# General Subjective Questions

**Question 6.** Explain the linear regression algorithm in detail. (Do not edit)

**Total Marks:** 4 marks (Do not edit)

**Answer:** Please write your answer below this line. (Do not edit)

Linear Regression is a statistical and machine learning algorithm used to model the relationship between one dependent variable (target) and one or more independent variables (features). Here's a detailed breakdown of how the algorithm works:

### 1. ****Understanding the Linear Regression Model****

Linear Regression assumes a linear relationship between the independent variables (XXX) and the dependent variable (YYY). The model can be expressed as:

Y=β0+β1X1+β2X2+⋯+βpXp+ϵY = \beta\_0 + \beta\_1 X\_1 + \beta\_2 X\_2 + \dots + \beta\_p X\_p + \epsilonY=β0​+β1​X1​+β2​X2​+⋯+βp​Xp​+ϵ

* YYY: Dependent variable (target).
* X1,X2,...,XpX\_1, X\_2, ..., X\_pX1​,X2​,...,Xp​: Independent variables (features).
* β0\beta\_0β0​: Intercept, the value of YYY when all XXX values are 0.
* β1,β2,...,βp\beta\_1, \beta\_2, ..., \beta\_pβ1​,β2​,...,βp​: Coefficients of the independent variables, representing the change in YYY for a one-unit change in XXX.
* ϵ\epsilonϵ: Error term, representing the variation in YYY not explained by the model.

### 2. ****Assumptions of Linear Regression****

To ensure the reliability of the model, the following assumptions must hold:

1. **Linearity**: The relationship between XXX and YYY is linear.
2. **Independence**: Observations are independent of each other.
3. **Homoscedasticity**: The variance of residuals (errors) is constant across all levels of XXX.
4. **Normality of Residuals**: The residuals are normally distributed.
5. **No Multicollinearity**: Independent variables should not be highly correlated with each other.

### 3. ****Goal of Linear Regression****

The goal is to find the best-fitting line that minimizes the difference between the observed and predicted values of YYY. This is achieved by minimizing the **Sum of Squared Errors (SSE)**:

SSE=∑i=1n(Yi−Y^i)2SSE = \sum\_{i=1}^{n} (Y\_i - \hat{Y}\_i)^2SSE=i=1∑n​(Yi​−Y^i​)2

Where:

* YiY\_iYi​: Actual value.
* Y^i\hat{Y}\_iY^i​: Predicted value from the model.

### 4. ****Steps in Linear Regression****

#### a. ****Formulating the Hypothesis****

The hypothesis is the equation of the line, Y=β0+β1XY = \beta\_0 + \beta\_1 XY=β0​+β1​X.

#### b. ****Optimization (Least Squares Method)****

* The coefficients β0,β1,…,βp\beta\_0, \beta\_1, \dots, \beta\_pβ0​,β1​,…,βp​ are estimated by minimizing the SSE.
* The normal equation is used to calculate the coefficients: β^=(XTX)−1XTY\hat{\beta} = (X^T X)^{-1} X^T Yβ^​=(XTX)−1XTY Where:
  + XXX: Matrix of features.
  + YYY: Vector of target values.
  + XTX^TXT: Transpose of XXX.
  + XTXX^T XXTX: A square matrix, invertible if XXX has full rank.

#### c. ****Model Training****

The model learns the coefficients β0,β1,…,βp\beta\_0, \beta\_1, \dots, \beta\_pβ0​,β1​,…,βp​ from the training data.

#### d. ****Prediction****

Using the learned coefficients, predictions are made using:

Y^=β0+β1X1+β2X2+⋯+βpXp\hat{Y} = \beta\_0 + \beta\_1 X\_1 + \beta\_2 X\_2 + \dots + \beta\_p X\_pY^=β0​+β1​X1​+β2​X2​+⋯+βp​Xp​

### 5. ****Model Evaluation****

#### a. ****R-squared (****R2R^2R2****)****

Measures the proportion of variance in the dependent variable explained by the independent variables:

R2=1−SSETSSR^2 = 1 - \frac{\text{SSE}}{\text{TSS}}R2=1−TSSSSE​

Where:

* TSSTSSTSS: Total Sum of Squares (variation in YYY around the mean).

#### b. ****Adjusted**** R2R^2R2

Accounts for the number of predictors in the model, penalizing overfitting.

#### c. ****Residual Analysis****

Residuals (Yi−Y^iY\_i - \hat{Y}\_iYi​−Y^i​) should be randomly distributed with no patterns.

### 6. ****Advantages of Linear Regression****

* Simple to implement and interpret.
* Fast to train and predict.
* Works well for linearly separable data.

### 7. ****Limitations****

* Assumes a linear relationship between variables.
* Sensitive to outliers, which can significantly affect predictions.
* Assumes independence of predictors and no multicollinearity.
* May underperform when relationships are nonlinear or complex.

**Question 7.** Explain the Anscombe’s quartet in detail. (Do not edit)

**Total Marks:** 3 marks (Do not edit)

**Answer:** Please write your answer below this line. (Do not edit)

### ****Anscombe’s Quartet****

Anscombe’s Quartet is a set of four datasets that are intentionally designed to have nearly identical summary statistics but display vastly different data distributions when visualized. These datasets were created by the statistician Francis Anscombe in 1973 to illustrate the importance of visualizing data before analyzing it.

### ****Key Purpose****

Anscombe’s Quartet highlights how relying solely on statistical summaries like mean, variance, correlation, or regression coefficients can be misleading. Visualizing data provides critical context that helps in understanding underlying patterns, relationships, and anomalies.

### ****The Datasets****

The four datasets in Anscombe’s Quartet share the following statistical properties:

1. **Mean of** XXX: Identical across datasets.
2. **Mean of** YYY: Identical across datasets.
3. **Variance of** XXX and YYY: Identical.
4. **Correlation between** XXX **and** YYY: Identical.
5. **Linear regression line**: Similar slope and intercept for all datasets.

However, the datasets have very different distributions and relationships when visualized.

### ****Data Summary****

Here’s a table summarizing the common statistics:

| **Statistic** | **Value (same for all datasets)** |
| --- | --- |
| Mean of XXX | 9.0 |
| Mean of YYY | 7.5 |
| Variance of XXX | 11.0 |
| Variance of YYY | 4.12 |
| Correlation between XXX and YYY | 0.816 |

**Lessons from Anscombe’s Quartet**

**Importance of Visualization**:

* 1. Numerical summaries alone are insufficient; visualizations reveal the true nature of data.
  2. Plots like scatterplots, histograms, and boxplots are critical.

**Context Matters**:

* 1. Similar summary statistics can describe vastly different datasets.
  2. Always interpret numbers in the context of visual data.

**Model Appropriateness**:

* 1. Blindly applying linear regression or other techniques can lead to incorrect conclusions.
  2. Understand data distribution and relationships before modeling.

**Question 8.** What is Pearson’s R? (Do not edit)

**Total Marks:** 3 marks (Do not edit)

**Answer:** Please write your answer below this line. (Do not edit)

Pearson's R, also known as the Pearson correlation coefficient, is a statistical measure that quantifies the linear relationship between two continuous variables. It provides information about both the strength and direction of the relationship.

**Key characteristics of Pearson's R:**

* **Range:** It ranges from -1 to +1.
* **Direction:**
  + A positive value indicates a positive linear association (as one variable increases, the other tends to increase).
  + A negative value indicates a negative linear association (as one variable increases, the other tends to decrease).
* **Strength:**
  + Values closer to +1 or -1 indicate a stronger linear relationship.
  + A value of 0 indicates no linear relationship.

**Interpretation:**

* r = +1: Perfect positive linear correlation.
* r = -1: Perfect negative linear correlation.
* r = 0: No linear correlation.
* **0 < r < 0.3:** Weak positive correlation.
* **-0.3 < r < 0:** Weak negative correlation.
* **0.3 < r < 0.7:** Moderate positive correlation.
* **-0.7 < r < -0.3:** Moderate negative correlation.
* **0.7 < r < 1:** Strong positive correlation.
* **-1 < r < -0.7:** Strong negative correlation.

**Formula:**

The formula for calculating Pearson's R is:

r = Σ[(xi - x̄)(yi - ȳ)] / √[Σ(xi - x̄)² \* Σ(yi - ȳ)²]

where:

* xi and yi are individual data points
* x̄ and ȳ are the means of the respective variables
* Σ denotes the sum

**Use Cases:**

Pearson's R is widely used in various fields to analyze relationships between variables, such as:

* Finance: Examining the correlation between stock prices.
* **Healthcare:** Studying the relationship between lifestyle factors and disease risk.
* **Social Sciences:** Investigating the correlation between education level and income.

**Limitations:**

* Pearson's R only measures linear relationships. It may not accurately capture non-linear associations.
* It is sensitive to outliers, which can significantly influence the correlation coefficient.
* Correlation does not imply causation. Even if two variables are highly correlated, it doesn't necessarily mean that one causes the other.

Overall, Pearson's R is a valuable tool for understanding the linear relationship between two variables, but it's essential to be aware of its limitations and interpret the results in conjunction with other analyses and contextual information.

**Question 9.** What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)

**Total Marks:** 3 marks (Do not edit)

**Answer:** Please write your answer below this line. (Do not edit)

# <Your answer for Question 9 goes here>

**Question 10.** You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)

**Total Marks:** 3 marks (Do not edit)

**Answer:** Please write your answer below this line. (Do not edit)

The value of VIF (Variance Inflation Factor) can be infinite when there is perfect multicollinearity between one or more predictor variables in a regression model. Here's why:

**Understanding** Multicollinearity: Multicollinearity occurs when two or more independent variables in a regression model are highly correlated with each other. This means that one variable can be linearly predicted from the others with a high degree of accuracy.

**How Multicollinearity Affects VIF:** VIF measures how much the variance of an estimated regression coefficient is inflated due to multicollinearity in the model. It is calculated as:

VIF = 1 / (1 - R^2)

where R^2 is the coefficient of determination of the regression of that predictor variable on all the other predictor variables.

**Perfect Multicollinearity:** In the case of perfect multicollinearity, one predictor variable can be perfectly predicted from the others. This means R^2 will be equal to 1. When R^2 is 1, the denominator in the VIF formula becomes 0, resulting in an infinite VIF value.

**In practical terms, an infinite VIF indicates that the independent variable is redundant and can be calculated directly from one or more other variables in the model, providing no new or unique information to the analysis.**

**Common Causes of Infinite VIF:**

* **Redundant Variables:** Including variables that are mathematical transformations of other variables (e.g., including both 'temperature' and 'temperature squared').
* **Categorical Variables:** Creating dummy variables for all categories of a categorical variable without dropping one as a reference category.
* **Linear Combinations:** Including variables that are linear combinations of other variables (e.g., including 'total sales' and also including 'sales from channel A' and 'sales from channel B' where 'total sales' is the sum of the other two).

**Addressing Infinite VIF:**

* **Identify the Cause:** Carefully examine the variables with infinite VIF and their relationships with other variables in the model.
* **Remove Redundant Variables:** The most common solution is to remove one of the variables causing the multicollinearity.
* **Combine Variables:** If appropriate, consider combining highly correlated variables into a single composite variable.
* **Regularization Techniques:** In some cases, regularization techniques like Ridge Regression or Lasso Regression can help mitigate the effects of multicollinearity.

By addressing infinite VIF, you can improve the stability and interpretability of your regression model.

**Question 11.** What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

(Do not edit)

**Total Marks:** 3 marks (Do not edit)

**Answer:** Please write your answer below this line. (Do not edit)

A Q-Q plot (Quantile-Quantile plot) is a graphical tool used to assess the distribution of data. It helps determine if a dataset follows a particular theoretical distribution (such as the normal distribution) by plotting the quantiles of the data against the quantiles of the theoretical distribution.

**How it works:**

1. The data is sorted and ranked.
2. Quantiles (percentiles) are calculated for the data points.
3. These quantiles are plotted against the corresponding quantiles of the theoretical distribution.

**Interpretation:**

* If the data follows the theoretical distribution, the points on the Q-Q plot will fall approximately along a straight line.
* Deviations from the straight line indicate departures from the theoretical distribution.

**Use and Importance in Linear Regression:**

In linear regression, a Q-Q plot is primarily used to examine the normality assumption of the residuals. The residuals should be normally distributed for the model to be valid and for accurate inferences.

**Here's how it's used:**

1. **Model Building:** A linear regression model is built using the training data.
2. **Residual Calculation:** The residuals (differences between actual and predicted values) are calculated.
3. **Q-Q Plot of Residuals:** A Q-Q plot of the residuals is generated, comparing their distribution to a normal distribution.

**Importance:**

* **Assessing Normality:** The Q-Q plot provides a visual check for the normality of residuals. If the points deviate substantially from a straight line, it suggests that the residuals are not normally distributed.
* **Identifying Outliers:** Q-Q plots can help identify outliers in the residuals. Outliers often appear as points that are far away from the straight line.
* **Model Validation:** Checking the normality assumption is an essential part of model validation. If the normality assumption is violated, it might be necessary to transform the data or consider a different type of regression model.
* **Accurate Inferences:** Valid inferences about the model parameters (coefficients, p-values, confidence intervals) depend on the normality of residuals.

**Example:**

If the Q-Q plot of residuals shows an upward curve, it indicates that the tails of the residual distribution are heavier than those of a normal distribution. This information can guide further analysis and model adjustments.

In summary, a Q-Q plot is a valuable tool in linear regression for assessing the normality of residuals, which is crucial for model validation and making accurate inferences about the relationships between variables.