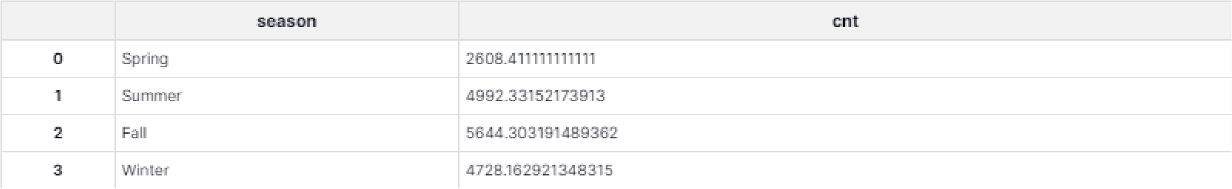
Amit Ranjan ML c59-UPGRADE IITB, EPGP in ML & AI

Assignment –based Subjective Questions

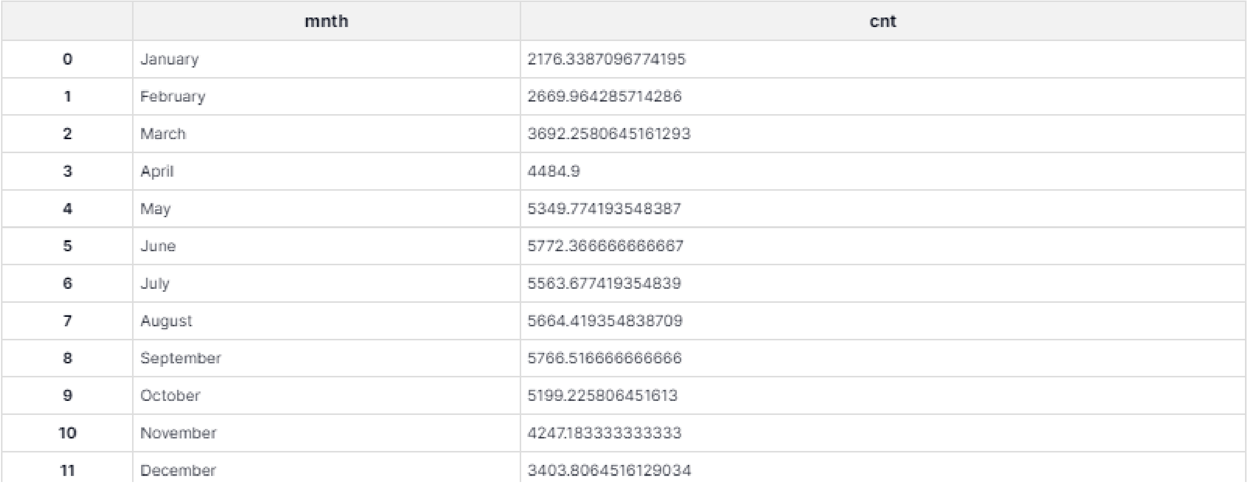
Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?



	season	cnt
0	Spring	2608.411111111111
1	Summer	4992.33152173913
2	Fall	5644.303191489362
3	Winter	4728.162921348315

The summarized bike demand across different seasons shows that spring has the lowest average demand for bikes, with an average count of 2608.41. In contrast, fall has the highest average demand (5644.30), followed by summer (4992.33) and winter (4728.16). This analysis confirms the observation that bike demand is indeed lower in spring compared to other seasons.

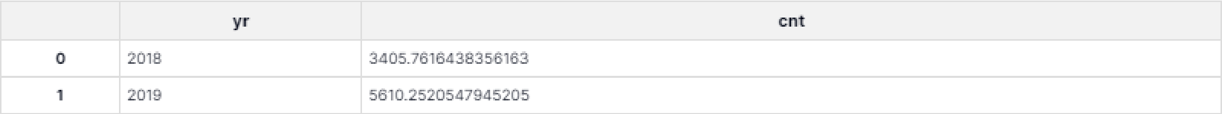
After examine the influence of other categorical variables (mnth, yr, weekday, workingday, and weathersit) on bike demand. We'll start analyzing the monthly demand to see how it varies throughout the year.



	mnth	cnt
0	January	2176.3387096774195
1	February	2669.964285714286
2	March	3692.2580645161293
3	April	4484.9
4	May	5349.774193548387
5	June	5772.366666666667
6	July	5563.677419354839
7	August	5664.419354838709
8	September	5766.516666666666
9	October	5199.225806451613
10	November	4247.183333333333
11	December	3403.8064516129034

The summarized bike demand across different months reveals a clear seasonal pattern in bike usage. Demand starts lower in January, gradually increases through the spring and peaks in the summer months (June, July, August, September), before declining again towards the end of the year. This pattern aligns with the earlier observation that demand is lower in spring (March shows an increase but is still relatively low compared to peak summer months) and confirms that weather and seasonality significantly influence bike demand.

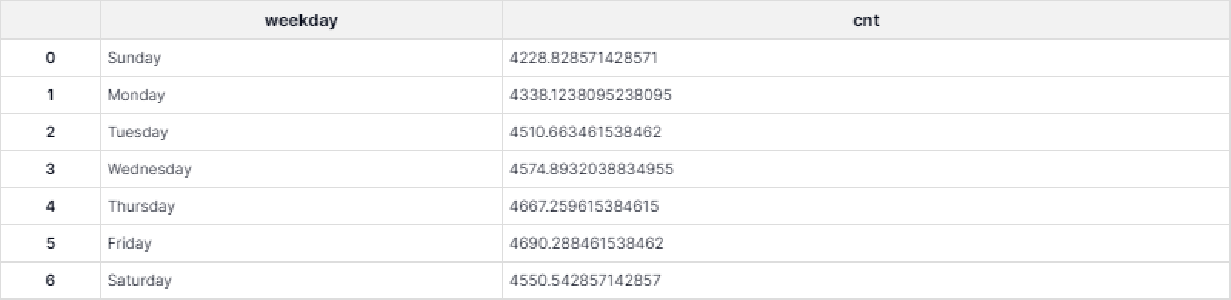
Next, we explore the impact of other categorical variables such as yr (year), weekday, workingday, and weathersit on bike demand. This help us to understand if there are specific days of the week, types of weather, or differences between years that significantly affect bike usage.



	yr	cnt
0	2018	3405.7616438356163
1	2019	5610.2520547945205

The summarized bike demand across the years 2018 and 2019 shows a significant increase in bike usage from 2018 to 2019. The average daily count of bike rentals in 2018 was 3405.76, which increased to 5610.25 in 2019. This indicates a growing trend in bike demand over the years covered in the dataset.

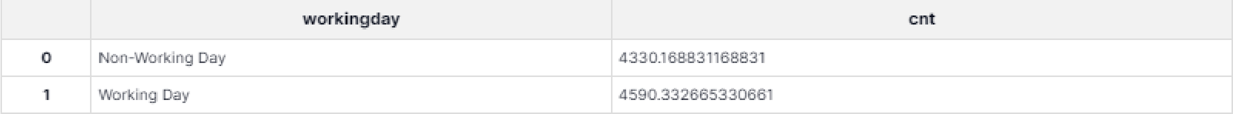
Next, analyze the impact of weekday, workingday, and weathersit on bike demand to complete our exploration of the categorical variables' influence. We'll start by examining the demand variation across different weekdays.



	weekday	cnt
0	Sunday	4228.828571428571
1	Monday	4338.1238095238095
2	Tuesday	4510.663461538462
3	Wednesday	4574.8932038834955
4	Thursday	4667.259615384615
5	Friday	4690.288461538462
6	Saturday	4550.542857142857

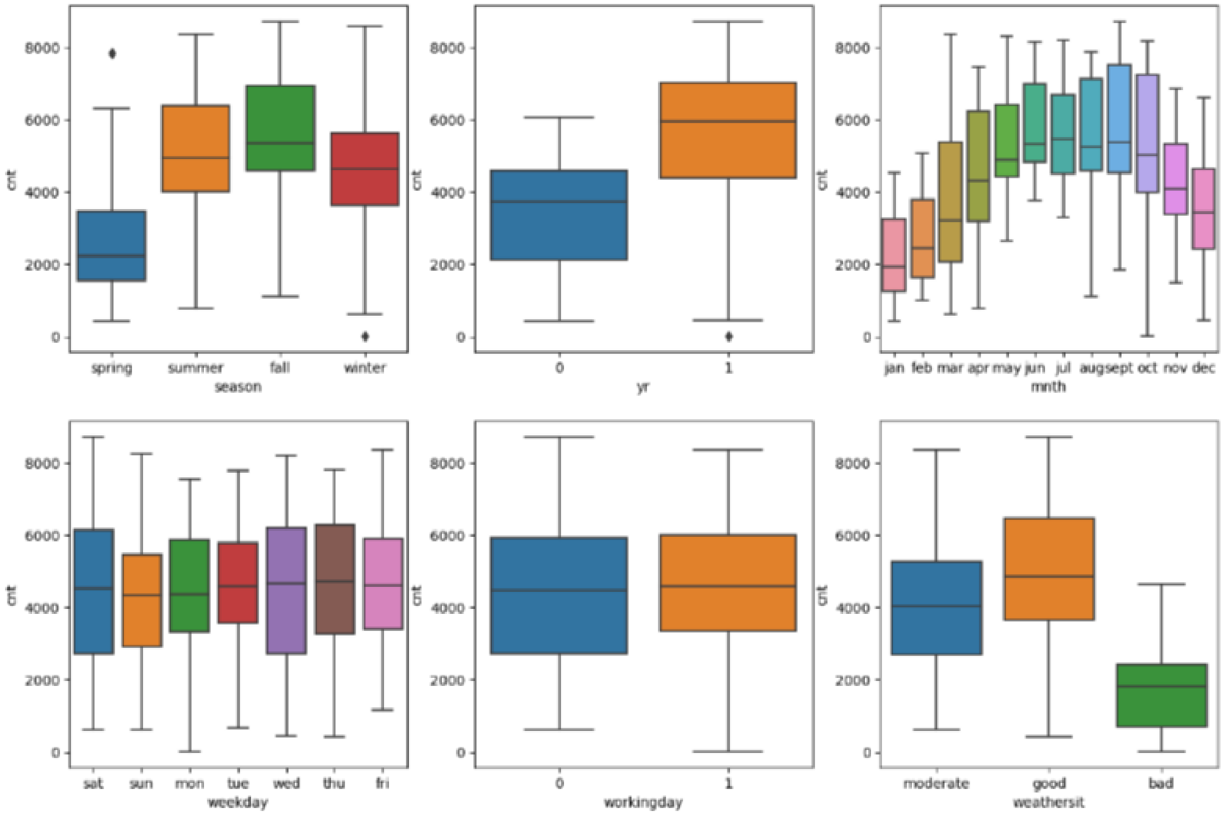
The summarized bike demand across different weekdays shows a relatively consistent demand throughout the week, with slight variations. The demand is somewhat higher on weekdays compared to weekends, with Thursday and Friday showing the highest average demand. This suggests that bike usage is slightly more prevalent during the workweek, possibly for commuting purposes.

We, examine the impact of working day and weathersit on bike demand to complete our exploration. We started by analyzing the demand on working days versus non-working days.



	workingday	cnt
0	Non-Working Day	4330.168831168831
1	Working Day	4590.332665330661

Weather conditions play a crucial role in bike demand, with clear weather conditions being the most conducive to higher bike usage.
These variables are visualized using bar plot and box plot below.



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

Using `drop_first=True` during dummy variable creation is important for several reasons, primarily to avoid the issue of multicollinearity in the dataset when applying linear models. Multicollinearity occurs when one predictor variable in the model can be linearly predicted from the others with a substantial degree of accuracy. This can lead to unreliable and unstable estimates of regression coefficients, which can affect the interpretability and predictive power of your model.

Here's why `drop_first=True` is crucial:

Reduces Multicollinearity: By dropping one of the dummy variables, you eliminate the perfect multicollinearity that arises because the dropped category can be inferred from the remaining categories. For instance, if you have a categorical variable for seasons with four categories (Spring, Summer, Fall, Winter) and you create dummy variables for all four, knowing the values of three of the dummy variables (0 or 1) will always let you determine the value of the fourth. This perfect predictability is what we aim to avoid.

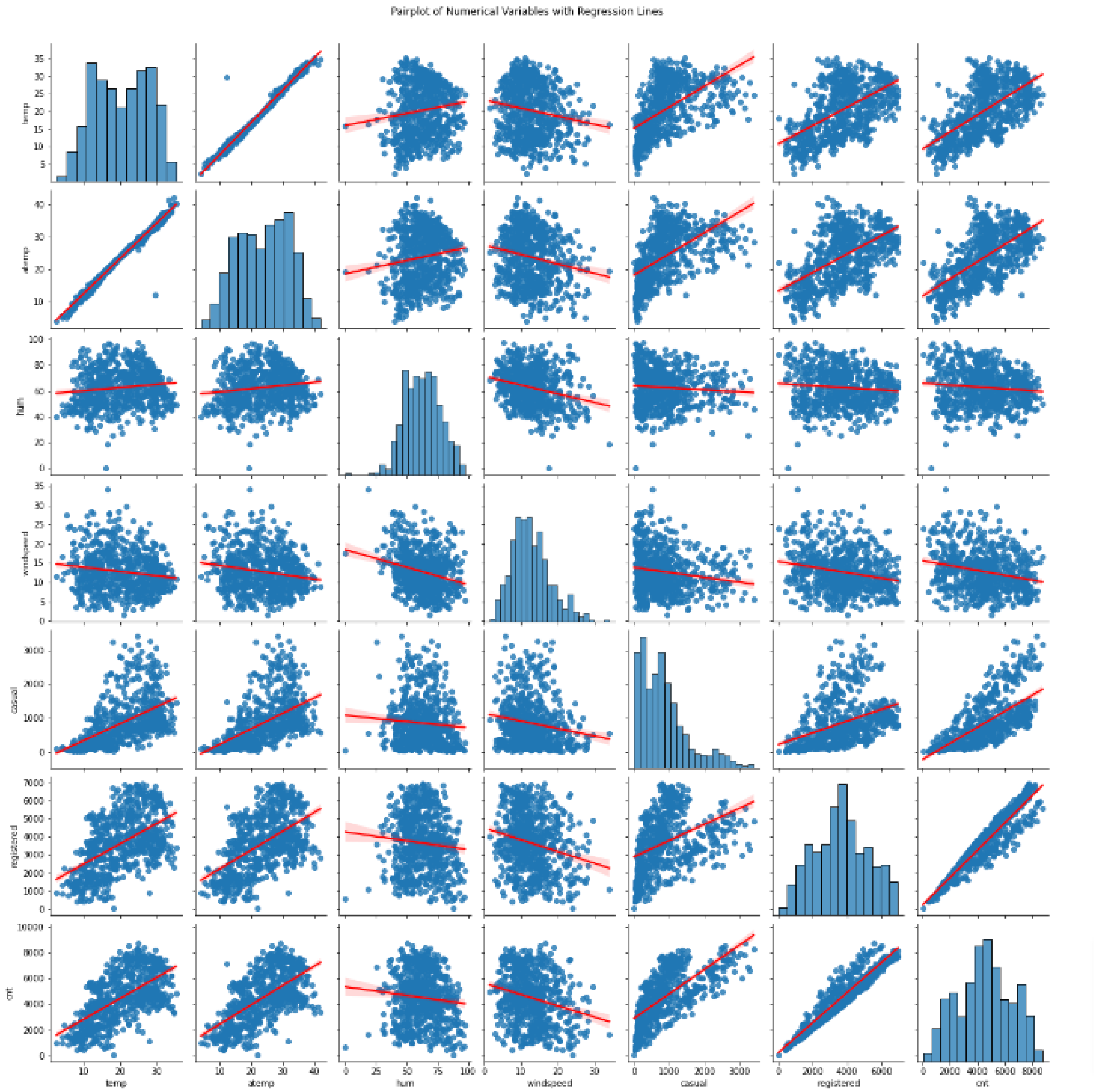
Simplifies the Model: Dropping one dummy variable reduces the number of predictor variables in the model, which can help in simplifying the model. This does not reduce the amount of information the model captures about the original variable, because the dropped category's effect is captured in the intercept term.

Reference Category: The category that is dropped serves as a reference or baseline against which the effects of the other categories are measured. This can be useful for interpretation purposes, as the coefficients of the remaining dummy variables represent the change from this baseline category.

In summary, using `drop_first=True` is a best practice when creating dummy variables for linear regression models and other statistical models where multicollinearity can be an issue. It helps in avoiding multicollinearity, simplifies the model, and aids in interpretation by establishing a reference category.

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

The ‘temp’ and ‘atemp’ variables have highest correlation when compared to the rest with target variable as ‘cnt’.


Pairplot of Numerical Variables with Regression Lines

The pairplot above visualizes the relationships among the numerical variables in the dataset, including the target variable cnt (count of total bike rentals). Each plot shows the relationship between cnt and another numerical variable, with a regression line indicating the trend.

Based on the regression lines and the scatter plot patterns, the variable registered appears to have the highest correlation with the target variable cnt. This is evident from the tight clustering of points along the regression line, indicating a strong linear relationship between the number of registered users and the total count of bike rentals.

This analysis suggests that the number of registered users is a significant predictor of bike rental demand, potentially more so than other numerical variables like temperature (temp and atemp), humidity (hum), or windspeed.

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

Validating the assumptions of linear regression after building the model involves several key steps to ensure the model's reliability and accuracy. Here's how you can validate these assumptions:

Linearity: The relationship between the independent variables and the dependent variable should be linear. This can be checked using scatter plots of observed vs. predicted values or residuals vs. predicted values.

Normality of Residuals: The residuals (differences between observed and predicted values) should be normally distributed. This assumption can be validated using a histogram or a Q-Q (quantile-quantile) plot of the residuals.

Homoscedasticity: The residuals should have constant variance across all levels of the independent variables. This means the spread of residuals should be similar across all values of the independent variables. Scatter plots of residuals vs. predicted values are used to check for homoscedasticity.

Independence of Residuals: Residuals should be independent of each other, which means the residuals should not show any patterns when plotted in time order (for time series data) or against the independent variables.

No Multicollinearity: Independent variables should not be too highly correlated with each other. This can be checked using Variance Inflation Factor (VIF) scores, where a VIF score of 10 or above indicates high multicollinearity.

To validate these assumptions in practice, we would typically:
Split data into training and testing sets.
Build linear regression model using the training set.
Generate predictions on the training set and calculate residuals.
Perform the above checks (1-5) using the residuals and the independent variables.

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

Below are the top three features contributing significantly towards explain the demand of the shared bikes.

- Temperature: It's a critical factor as more people are likely to rent bikes in comfortable weather conditions. Both too hot and too cold temperatures can deter people from outdoor activities like biking.

- Season: This ties closely with temperature but also includes the impact of seasonal changes on people's outdoor activities. For instance, spring and summer might see higher bike rentals due to more favorable weather conditions compared to fall and winter.

- Year: This could reflect a trend of increasing or decreasing bike usage over time, possibly due to changes in public awareness, infrastructure improvements, or societal shifts towards more sustainable modes of transportation.

These factors are essential for understanding and forecasting bike rental demand, allowing for better resource allocation and marketing strategies to meet user needs effectively.

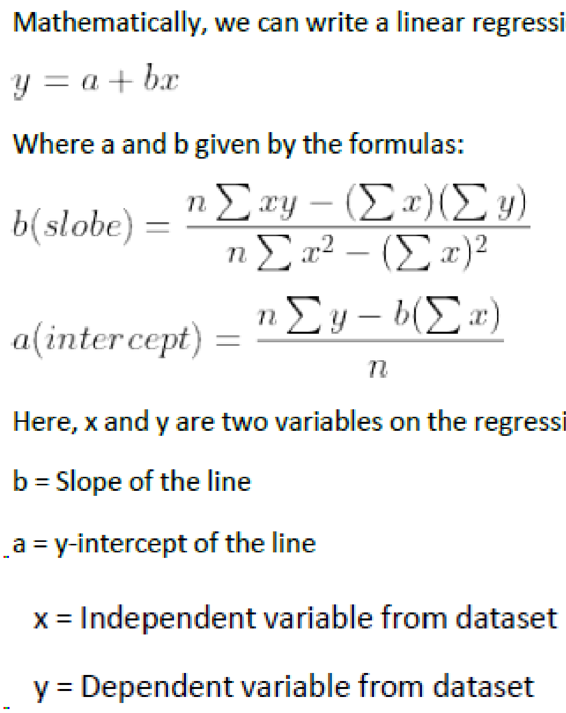
General Subjective Questions

Question 1. Explain the linear regression algorithm in detail.

Linear Regression is a fundamental algorithm in machine learning and statistics used for predicting a continuous outcome variable (dependent variable) based on one or more predictor variables (independent variables). The goal is to find a linear relationship between the predictors and the outcome. Here's a detailed explanation of the Linear Regression algorithm:

1. The Model

Linear regression is one of the very basic form of machine learning where we train a model to predict the behavior of your data based on some variables. In the case of linear regression as you can see the name suggests linear that means the two variables which are on the x-axis and y-axis should be linearly correlated.



Mathematically, we can write a linear regression equation as:

$$y = a + bx$$

Where a and b given by the formulas:

$$b(slobe) = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

$$a(intercept) = \frac{n\sum y - b(\sum x)}{n}$$

Here, x and y are two variables on the regression line.

b = Slope of the line

a = y-intercept of the line

x = Independent variable from dataset

y = Dependent variable from dataset

2. Assumptions

For linear regression to provide reliable predictions, certain assumptions must be met:

Linearity: The relationship between the independent and dependent variables is linear.
Independence: Observations are independent of each other.
Homoscedasticity: Constant variance of error terms.
Normality: The error terms are normally distributed.

3. Fitting the Model

Fitting a linear regression model involves finding the values of the coefficients (β) that minimize the difference between the predicted and actual values. This is typically done using the Least Squares method, which minimizes the sum of the squared differences between the observed and predicted values.

4. Evaluation

After fitting the model, it's evaluated to determine its effectiveness. Common metrics include:

R-squared: Represents the proportion of the variance for the dependent variable that's explained by the independent variables.

Adjusted R-squared: Adjusts the R-squared value based on the number of predictors in the model.
Mean Squared Error (MSE): The average of the squares of the errors between the actual and predicted values.

5. Prediction
With the model fitted and evaluated, it can be used to make predictions on new data. The new data is fed into the model, and it outputs the predicted values based on the learned coefficients.

6. Validation
It's crucial to validate the model using unseen data to ensure it generalizes well. This is often done by splitting the data into training and testing sets, where the model is trained on the training set and validated on the testing set.

Linear regression is widely used due to its simplicity and interpretability, making it a good starting point for regression tasks. However, it's important to check the assumptions and evaluate the model to ensure it's suitable for your specific dataset and problem.

Question 2. Explain the Anscombe’s quartet in detail.

Anscombe's quartet comprises four datasets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (x, y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties. Here's a detailed explanation:

Key Points of Anscombe’s Quartet:
Statistical Similarities: The four datasets have nearly identical mean, variance, correlation, and linear regression lines (y = 3.00 + 0.500x) when rounded to two decimal places. This suggests that they share the same statistical properties at a glance.

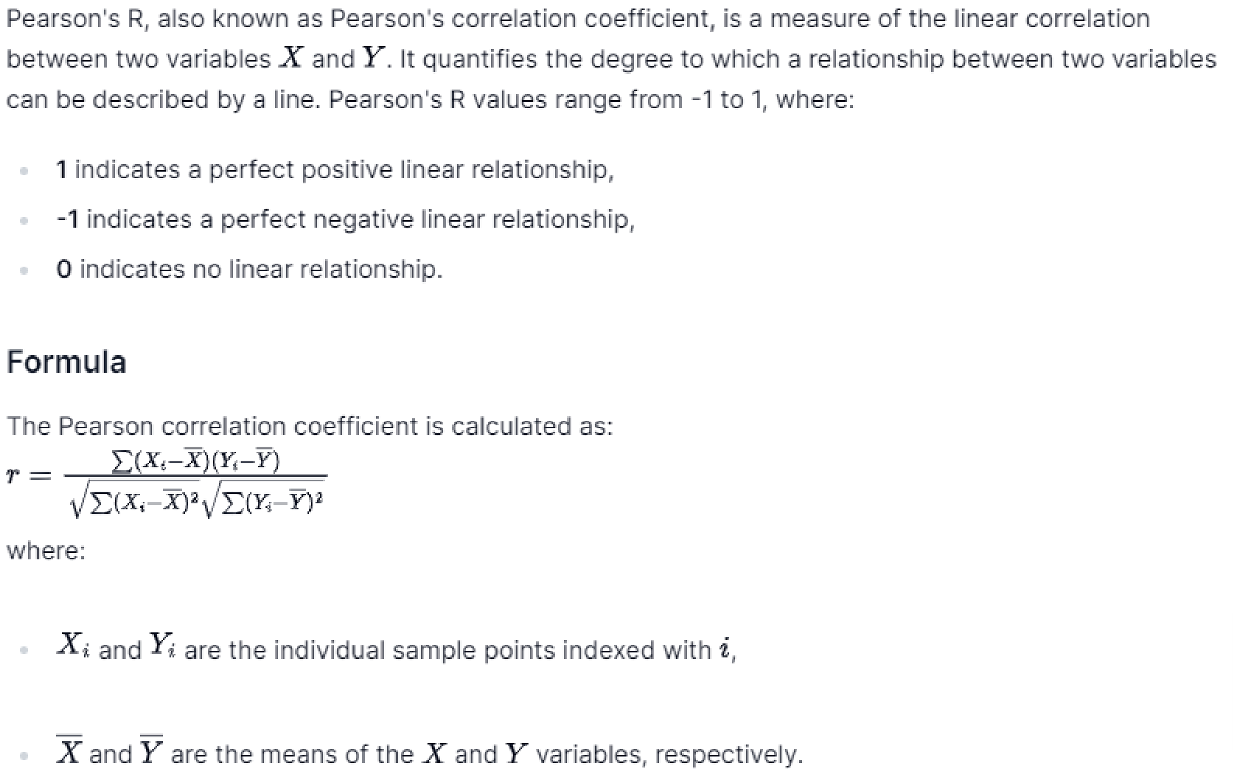
Visual Differences: When plotted, each dataset looks very different. One is linear, another is nonlinear, the third appears to be a tight linear relationship except for an outlier, and the fourth has a low variance in x except for one outlier.

The Datasets:

Dataset I shows a simple linear relationship, corresponding closely to the assumptions of linear regression.
Dataset II shows a curve; linear regression is not appropriate for this dataset.
Dataset III shows a linear relationship similar to Dataset I, but with one outlier that significantly affects the regression line.
Dataset IV shows a relationship where x values are almost constant except for one outlier, which again significantly affects the regression line.
Lessons from Anscombe’s Quartet:
Visual Analysis is Crucial: Before using statistical models, it's essential to visually inspect the data. Graphs can reveal anomalies, patterns, or characteristics that numbers alone may not show.
Impact of Outliers: Outliers can have a significant impact on statistical summaries and models. It's crucial to identify and understand them before drawing conclusions.
Limitations of Summary Statistics: Summary statistics can be misleading. Datasets with the same statistical properties can have very different distributions and relationships.

Anscombe's quartet is a powerful demonstration of why data scientists and statisticians need to look beyond numerical summaries and perform exploratory data analysis (EDA) to understand their data fully.

Question 3. What is Pearson’s R?



Pearson's R, also known as Pearson's correlation coefficient, is a measure of the linear correlation between two variables X and Y. It quantifies the degree to which a relationship between two variables can be described by a line. Pearson's R values range from -1 to 1, where:

- 1 indicates a perfect positive linear relationship,
- -1 indicates a perfect negative linear relationship,
- 0 indicates no linear relationship.

Formula

The Pearson correlation coefficient is calculated as:

$$r = \frac{\sum(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum(X_i - \overline{X})^2}\sqrt{\sum(Y_i - \overline{Y})^2}}$$

where:

- X_i and Y_i are the individual sample points indexed with i,
- $\overline{X}$ and $\overline{Y}$ are the means of the X and Y variables, respectively.

Interpretation

Values closer to 1 or -1 indicate a stronger linear relationship between the two variables.

A positive value suggests a positive association; as one variable increases, the other tends to increase.

A negative value indicates a negative association; as one variable increases, the other tends to decrease.

Values near 0 suggest a weak linear relationship.

Use Cases

Pearson's R is widely used in statistics to measure the strength and direction of a linear relationship between two continuous variables. It's important in fields such as finance, medicine, and social sciences for tasks like feature selection, exploratory data analysis, and validating hypotheses about relationships between variables.

Limitations

It only measures linear relationships; nonlinear relationships are not well captured.

It is sensitive to outliers, which can significantly affect the correlation coefficient.

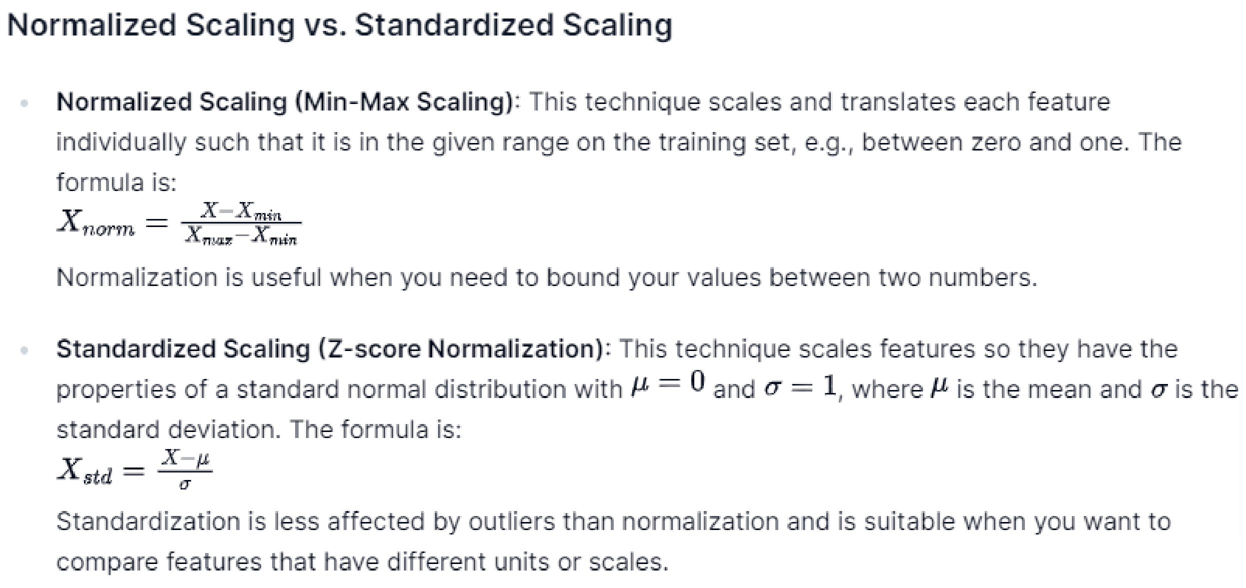
It assumes that both variables are normally distributed.

Understanding Pearson's R is crucial for analyzing and interpreting relationships between variables in data science and statistics.

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

Scaling is a preprocessing step in data processing and machine learning that involves adjusting the scale of the data so that different features contribute equally to the analysis or model performance. It's crucial because many machine learning algorithms, especially those that compute distances or assume normality, perform better or converge faster when features are on a similar scale.

Why is Scaling Performed?
Equal Contribution: Ensures all features contribute equally to the result, preventing features with larger scales from dominating the model's behavior.
Improved Algorithm Performance: Many algorithms, like gradient descent, converge faster when features are scaled.
Distance-Based Algorithms: Algorithms like K-Nearest Neighbors (KNN) and K-Means clustering that calculate distances between data points are sensitive to the scale of data.
Regularization: Scaling is essential when regularization is used, as it assumes all features are centered around zero and on the same scale.



Normalized Scaling vs. Standardized Scaling

- **Normalized Scaling (Min-Max Scaling):** This technique scales and translates each feature individually such that it is in the given range on the training set, e.g., between zero and one. The formula is:
  $X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$
  Normalization is useful when you need to bound your values between two numbers.

- **Standardized Scaling (Z-score Normalization):** This technique scales features so they have the properties of a standard normal distribution with $\mu = 0$ and $\sigma = 1$, where μ is the mean and σ is the standard deviation. The formula is:
  $X_{std} = \frac{X - \mu}{\sigma}$
  Standardization is less affected by outliers than normalization and is suitable when you want to compare features that have different units or scales.

In summary, both scaling techniques are essential for preprocessing data in machine learning, but the choice between normalization and standardization depends on the algorithm you're using and the specific requirements of the dataset.

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

The Variance Inflation Factor (VIF) measures how much the variance of an estimated regression coefficient increases if your predictors are correlated. If no factors are correlated, the VIFs will all be equal to 1.

Why VIF Might Be Infinite:
An infinite VIF value typically occurs when there is perfect multicollinearity in the data, meaning at least one predictor variable can be perfectly predicted from the others. This situation arises when:

Exact Linear Relationship:
There's an exact linear relationship between some of the predictor variables. For example, if one variable is the sum of two others, those variables are perfectly multicollinear.
Duplicate Columns:
If the dataset accidentally includes duplicate columns or variables with identical values, it will lead to an infinite VIF.
Highly Correlated Variables: While not always resulting in an infinite VIF, variables that are very highly correlated can lead to extremely high VIF values, approaching infinity as the correlation approaches 1.
Implications:
Model Interpretation:
Infinite or very high VIF values indicate that the model coefficients are not reliable due to multicollinearity. The standard errors of the coefficients become large, making it difficult to determine the individual effect of predictor variables on the response variable.
Model Prediction:
While multicollinearity affects the interpretation of coefficients, it does not necessarily affect the model's ability to predict the response variable. However, it can make the model sensitive to changes in the model's makeup or the data, potentially affecting its robustness.
Solutions:
To address infinite VIF values, you might consider:

Removing Variables:
Eliminate variables that are causing multicollinearity.
Combining Variables:
Combine highly correlated variables into a single predictor.
Principal Component Analysis (PCA):
Use PCA or another dimensionality reduction technique to reduce the feature space to a set of uncorrelated components.

In summary, an infinite VIF is a clear indication of perfect multicollinearity that needs to be addressed for reliable model interpretation and potentially more robust predictions.

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

A Q-Q (quantile-quantile) plot is a graphical tool to assess if a set of data plausibly came from some theoretical distribution such as a Normal, Exponential, or Uniform distribution. It compares the quantiles of the data to the quantiles of the theoretical distribution, providing a visual means to evaluate how well the data matches the specified distribution.

Use and Importance in Linear Regression:
Assumption Checking: Linear regression assumes that the residuals (errors) are normally distributed. A Q-Q plot of the residuals can be used to visually check this assumption. If the residuals are normally distributed, the points on the Q-Q plot will approximately lie on a straight line.

Identifying Skewness and Outliers:
The Q-Q plot can help identify skewness in the data. If the plot deviates from a straight line in a systematic way, it suggests that the data may be skewed. Additionally, outliers may appear as points that are far away from the line.

Improving Model Accuracy:
By identifying deviations from normality, a Q-Q plot can guide data transformation or the use of alternative models that do not assume normality, potentially improving model accuracy.

Comparative Analysis:
It can be used to compare the distribution of residuals across different models or subgroups within the data, providing insights into model performance and the appropriateness of model assumptions.

In summary, the Q-Q plot is a simple yet powerful tool for assessing the distribution of data relative to a theoretical distribution. In the context of linear regression, it is particularly useful for checking the normality of residuals, which is a key assumption for the validity of regression analysis.