Assignment-based Subjective Questions

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

Based on the regression analysis, we can infer the following about the effects of the categorical variables (like 2019_yr, season, mist, and light_snow) on the dependent variable (cnt, which represents the bike demand):

Year (2019_yr)

- **Positive Effect:** The coefficient for 2019_yr is positive (0.2473), indicating that bike demand increased significantly in 2019 compared to the baseline year (2018). This suggests a growing trend in bike usage, potentially due to increased popularity, improvements in biking infrastructure, or other socio-economic factors encouraging biking in 2019.

Seasons (summer, fall, winter)

- **Positive Effect for Summer, Fall, and Winter:**
  - **Summer (`0.2573`)**: Biking demand is higher in summer, likely due to favorable weather conditions and longer daylight hours.
  - **Fall (`0.3162`)**: Fall has the highest positive impact on bike demand, possibly due to mild temperatures and fewer weather-related disruptions.
  - **Winter (`0.2268`)**: Although winter generally has colder temperatures, bike demand still shows a positive effect compared to the omitted season (likely spring), suggesting that biking remains popular even in colder months, perhaps due to the availability of proper gear or cultural habits.

Weather Conditions (mist, light_snow)

- **Negative Effect:**
  - **Mist (`-0.0857`)**: Misty weather negatively affects bike demand, as reduced visibility and damp conditions make biking less appealing.
  - **Light Snow (`-0.2877`)**: Light snow has the most substantial negative impact on bike demand among the weather-related variables, as snow makes biking more challenging and unsafe, leading to a significant reduction in usage.

And the other categorical variables have no impact or very little impact on the target variable.

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

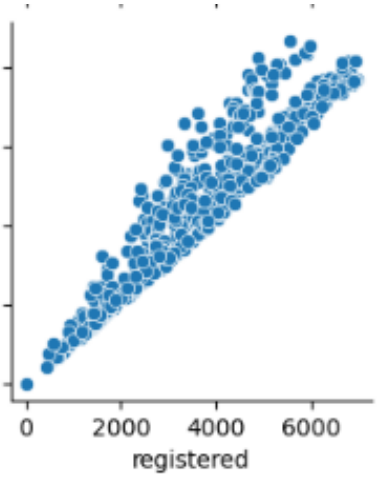
Using drop_first=True during the creation of dummy variables is important for the reasons below ensuring that the regression model remains statistically valid.

- Avoiding Multicollinearity
- Reducing Redundancy.

It also helps in interpretation of Coefficients and is important for avoiding the dummy variable trap.

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

Find the below scatter plot from the pair-plots for cnt and registered variables



And from the corr() method, we have found that registered has 0.945 correlation with the target variable 'cnt'.

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

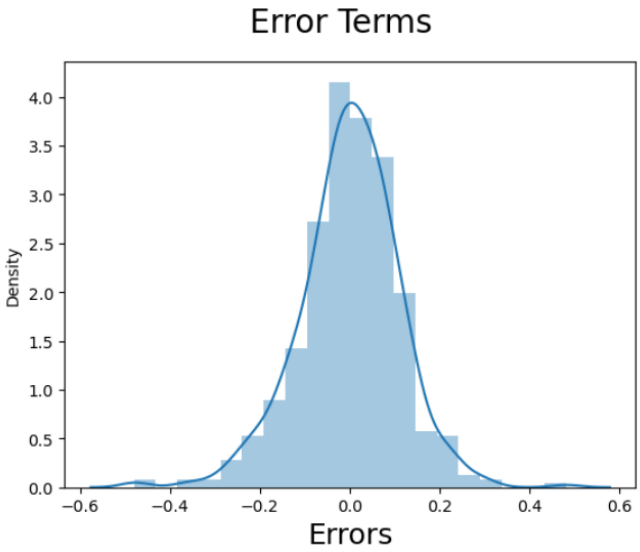
The assumptions of Linear Regression are,

- There is a linear relationship between X and y.
- Error terms are normally distributed with mean zero

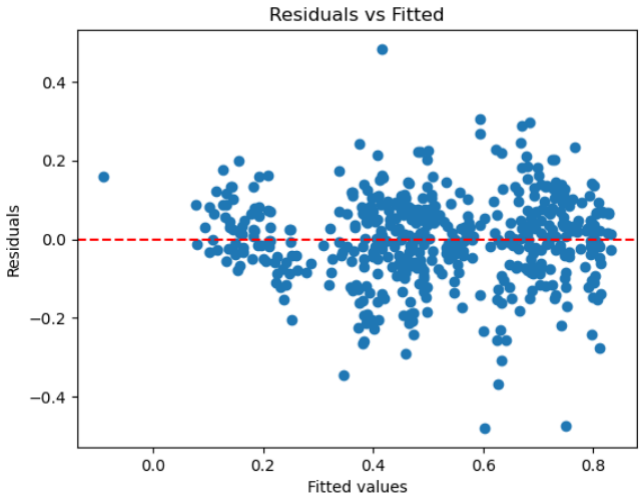
- Error terms are independent of each other
- Error terms have constant variance (homoscedasticity)

As part of our residual analysis, we have validated the below assumptions

- Error terms are normally distributed with mean zero



- Error terms are independent of each other. No visible patterns observed.



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

Based on the final model's statistics, the top 3 features that contribute significantly towards explaining the demand for shared bikes are:

- **Year (2019)**:
    - Coefficient: 0.2473

    - This feature shows the year 2019, and it has a positive and significant impact on bike demand, indicating that the demand for shared bikes increased in 2019 compared to the baseline year (2018).
- **Season (Fall)**:
    - Coefficient: 0.3162
    - The fall season has the highest positive impact on bike demand. This suggests that the demand for shared bikes was significantly higher during the fall compared to other seasons.
- **Season (Summer)**:
    - Coefficient: 0.2573
    - Summer also has a significant positive impact on bike demand, following fall. This indicates that the warmer months contribute to an increase in bike usage.

Season (winter) is the next feature variable that significantly impacts the demand for shared bikes.

General Subjective Questions

Explain the linear regression algorithm in detail?

Linear regression is a statistical technique used to model and analyze the relationship between a dependent variable (often denoted as y) and one or more independent variables (denoted as X1,X2,…,Xn). The primary goal is to find the best-fit line (also called a regression line) that minimizes the difference between the observed data points and the predictions made by the model. The best-fit line is described by the equation:

$$y = \beta 0 + \beta 1X1 + \beta 2X2 + ... \beta nXn + \epsilon$$

- y is the dependent variable.
- X1 ,X2 ,…,Xn are the independent variables.
- β0 is the intercept.
- β1 ,β2 ,…,βn are the coefficients (slopes) that represent the change in y for a one-unit change in each Xi.
- ϵ is the error term, which accounts for the variability in y that cannot be explained by the independent variables.

Objective:

The objective of linear regression is to find the values of the coefficients $\beta 0, \beta 1, \ldots, \beta n$ that minimize the error term. This is done by minimizing the sum of squared residuals (errors). The residuals are the differences between the observed values and the predicted values.

Assumptions:

- **Linearity**: The relationship between the dependent and independent variables is linear.
- **Independence**: The residuals (errors) are independent of each other.
- **Homoscedasticity**: The residuals have constant variance at all levels of the independent variables.
- **Normality**: The residuals are normally distributed.

Steps involved in building the Linear Regression Model:

Building a linear regression model involves several key steps:

- **Data Collection and Preprocessing**
  - **Data Collection**: Gather data that includes both the dependent and independent variables. The data should be relevant, accurate, and sufficient in size.
  - **Data Cleaning**: Handle missing values, outliers, and any inconsistencies in the data.
  - **Exploratory Data Analysis (EDA)**: Understand the data's distribution, relationships, and patterns through visualizations and summary statistics.

- **Feature Selection by Visualizing data**
  - This helps in finding the variables with multicollinearity.
  - It also identifies if some predictors directly correlate strongly with the outcome variable.
  - Select the predictors (independent variables) that you believe have a relationship with the dependent variable with respective to business.

- **Dummy Variable Creation**
  - Convert categorical variables into numerical form using dummy variables. For example, if you have a categorical variable Season with values Winter, Spring, Summer, and Fall, create dummy variables for each.

- Use drop_first=True to avoid the dummy variable trap, which can cause multicollinearity.

- **Splitting the Data**
  - Divide the dataset into a training set (used to build the model) and a testing set (used to evaluate the model's performance). Typically, an 80-20 or 70-30 split is used.

- **Rescaling the Data**
  - Rescale the data to ensure all features contribute equally to the model.
  - **Normalization**: Rescales data to a range of [0, 1].
  - **Standardization**: Rescales data to have a mean of 0 and a standard deviation of 1.

- **Dividing into X and y sets**
  - Get the train data X without the target variable and y with only the target variable.

- **Model Building**
  - Use the Ordinary Least Squares (OLS) method to fit the linear regression model to the training data. This method minimizes the sum of squared residuals (the differences between observed and predicted values).

- **Feature Selection using Recursive Feature Elimination (RFE)**
  - Systematically remove features with the least importance, retrain the model, and rank the features based on their significance.

- **Building model using statsmodel**
  - Build the model using statsmodel for the detailed statistics.
  - As statsmodel, does not support the intercept, we add it using add_constant method.
  - Find the features with p value greater than 0.05 and eliminate them one by one and rebuild the model.
  - Once all the p values are in acceptable range, find the Variance Inflation Factor and remove the features one by one which is greater than 10 and rebuild the model.

- Check for p values and then VIF values, eliminate if not in acceptable range and rebuild as mentioned in the above steps.
- Get the statistics of the final model built.

- **Model Evaluation**
  - Analyze residuals (the differences between observed and predicted values) to ensure the assumptions of linear regression are met.
  - **Normality**: Residuals should be normally distributed.
  - **Homoscedasticity**: Residuals should have constant variance.
  - **Homoscedasticity**: Residuals should have constant variance.

- **Model Testing and Validation**
  - **Prediction**: Use the testing set to make predictions and compare them against actual values.
  - Calculate metrics like R-squared to evaluate the model's performance.

- **Final Model and Equation**
  - **Coefficients Interpretation**: The model's coefficients indicate the relationship between each predictor and the dependent variable.
  - **Best-Fit Line Equation**: Use the final coefficients to construct the equation of the best-fit line. Like below

$$y=\beta 0 +\beta 1\ X1 +\beta 2\ X2 +\cdots+\beta n\ Xn$$

Linear regression is a powerful tool for understanding the relationship between variables. By following the steps outlined above ranging from data preprocessing to model validation, you can build a reliable linear regression model that accurately predicts the dependent variable while adhering to the assumptions of the linear regression framework.

Explain the Anscombe's quartet in detail.

Anscombe's Quartet is a group of four datasets designed to have nearly identical statistical properties but appear different when graphed. It was created by the statistician Francis Anscombe in 1973 to demonstrate the importance of data visualization and the potential pitfalls of relying solely on summary statistics when analyzing data.

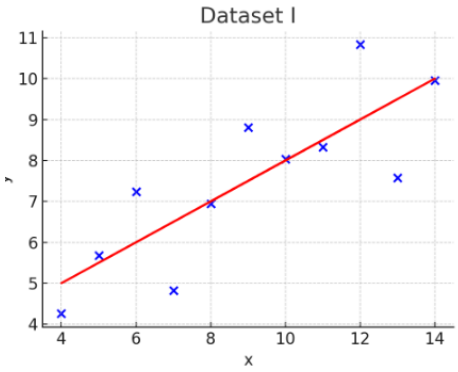
Overview of the Quartet

Each dataset in Anscombe's Quartet consists of 11 (x, y) pairs. Despite the differences in their graphical appearance, all four datasets have the following identical statistical properties:

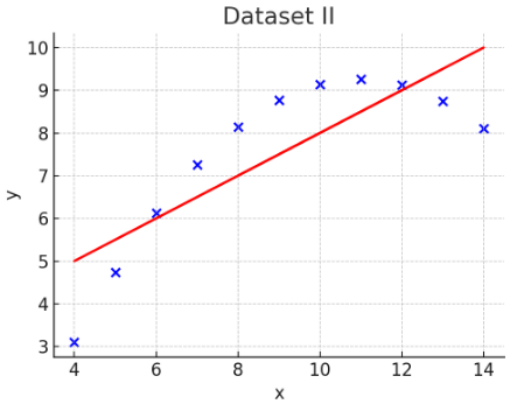
- **Mean of x**: 9
- **Variance of x**: 11
- **Mean of y**: 7.5
- **Variance of y**: 4.12 (approximately)
- **Correlation between x and y**: 0.816 (approximately)
- **Linear regression line**: $y=3+0.5xy = 3 + 0.5x$y=3+0.5x
- **Coefficient of determination ($R2R^2$R2)**: 0.67

However, when you plot these datasets, they look drastically different:

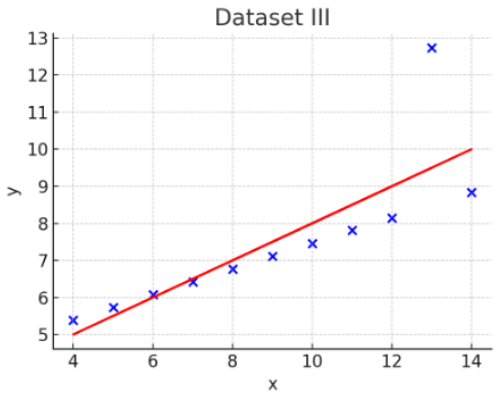
Dataset I: This dataset represents a typical linear relationship with a good fit to the regression



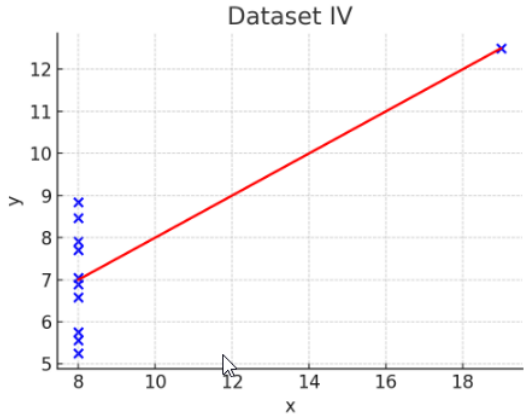
Dataset II: In this dataset, the data points are arranged in a nearly perfect vertical line, but with one outlier that drags the regression line up to fit the overall trend.



Dataset III: This dataset has a strong non-linear relationship, but the regression line doesn't fit well. Despite this, the summary statistics are identical to the other datasets.



Dataset IV: This dataset contains a single vertical cluster of points except for one outlier, which significantly influences the regression line.



Anscombe's Quartet highlights several important lessons in data analysis:

- **The Importance of Visualization**: While summary statistics like mean, variance, and correlation are useful, they can be misleading without visualizing the data. Visual representations can reveal patterns, relationships, or outliers that are not evident from the statistics alone.
- **Outliers and Leverage Points**: Outliers or unusual points can have a significant impact on the statistical analysis, such as distorting the regression line or correlation coefficient.
- **Non-Linearity**: Even when data have the same linear regression line, the underlying relationship between the variables can be non-linear, as shown in Dataset III.
- **Statistical Measures Can Be Deceptive**: Identical statistical summaries can arise from very different data distributions. This underscores the need to look beyond summary statistics when analyzing data.

What is Pearson's R?

Pearson's R, also known as the Pearson correlation coefficient, is a statistical measure that quantifies the linear relationship between two variables. It is denoted by r and ranges from -1 to 1.

Value Range:

- r=1: Perfect positive linear relationship. As one variable increases, the other also increases proportionally.
- r=−1: Perfect negative linear relationship. As one variable increases, the other decreases proportionally.
- r=0: No linear relationship between the variables.

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

r = correlation coefficient
x_i = values of the x-variable in a sample
$\bar{x}$ = mean of the values of the x-variable
y_i = values of the y-variable in a sample
$\bar{y}$ = mean of the values of the y-variable

Example:
If you're analyzing the relationship between the number of study hours and exam scores, Pearson's r would tell you whether an increase in study hours is associated with an increase in exam scores, and how strong that relationship is.

Limitations:
Pearson's R only measures linear relationships. It does not capture non-linear relationships. It can be influenced by outliers, which may distort the perceived relationship.

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

Scaling is a process of adjusting the range or distribution of data values in a dataset so that different features or variables are brought to a similar scale. This is particularly important in

machine learning algorithms where the distance between data points is important, such as in regression, clustering, and neural networks.

Why is Scaling Performed?

- **Algorithm Efficiency**: Many algorithms work better when features are on a similar scale. For example, in gradient descent optimization, scaling helps achieve faster convergence.
- **Fairness**: Features with larger ranges can dominate the distance calculations or model coefficients, leading to biased results. Scaling ensures that each feature contributes equally.

Types of Scaling:

- **Normalized Scaling (Min-Max Scaling)**:
  It rescales the data so that all feature values fall within a specified range, usually [0, 1].

$$X_{\text{scaled}} = \frac{X_i - X_{\text{mean}}}{X_{\text{max}} - X_{\text{min}}}$$

- **Standardized Scaling**
  It rescales the data to have a mean of 0 and a standard deviation of 1

$$X_{\text{scaled}} = \frac{X_i - X_{\text{mean}}}{\sigma}$$

Key Differences:

- **Range**:
  - **Normalized Scaling**: Typically transforms data to a fixed range [0, 1].
  - **Standardized Scaling**: Transforms data to have a mean of 0 and a standard deviation of 1, but the range is not fixed.
- **Impact of Outliers**:
  - **Normalized Scaling**: More sensitive to outliers since it depends on the minimum and maximum values in the data.
  - **Standardized Scaling**: Less sensitive to outliers, especially when the data has a normal distribution.

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

A Variance Inflation Factor (VIF) can become infinite when there is perfect multicollinearity among the predictors in a regression model. Let's break down why this occurs:

VIF is a measure that quantifies how much the variance of a regression coefficient is inflated due to multicollinearity among the predictors.

$$\mathrm{VIF}(X_i) = \frac{1}{1 - R_i^2}$$

Where Ri2 is the coefficient of determination of a regression model that predicts the ith predictor using all the other predictors.

Why VIF Becomes Infinite?

Perfect Multicollinearity: This occurs when one predictor variable is a perfect linear combination of one or more other predictor variables. In this case, the Ri2 for that predictor will be 1.0.

- When Ri2 =1: The denominator in the VIF formula becomes zero, then the VIF becomes infinite.
- Interpretation: An infinite VIF indicates that the predictor is perfectly collinear with other predictors, making it impossible to estimate its coefficient uniquely. This situation is problematic because it means that the regression model cannot differentiate between the effects of the collinear predictors.

To address infinite VIF, we need to remove redundant variables, which are perfectly collinear.

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

A **Q-Q plot** (Quantile-Quantile plot) is a graphical tool used to compare the distribution of a dataset with a theoretical distribution, typically the normal distribution. It helps to assess whether the data follows a particular distribution.

What is Q-Q Plot

- **Axes**: A Q-Q plot plots the quantiles of the sample data on the y-axis against the corresponding quantiles of the theoretical distribution on the x-axis.
- **Quantiles**: Quantiles are points in your data below which a certain percentage falls. For instance, the 25th percentile is the value below which 25% of the data lies.
- **Interpretation**:
    - **Straight Line**: If the data follows the theoretical distribution (e.g., normal distribution), the points on the Q-Q plot will lie roughly along a straight line.
    - **Deviations from Line**: Deviations from this line indicate departures from the expected distribution.

Use of Q-Q Plot

In the context of linear regression, a Q-Q plot is primarily used to check the assumption of **normality** of residuals (errors). The assumptions of linear regression include:

1. **Linearity**: The relationship between the independent and dependent variables should be linear.
2. **Independence**: The residuals (errors) should be independent.
3. **Homoscedasticity**: The residuals should have constant variance.
4. **Normality**: The residuals should be normally distributed.

Importance of Q-Q Plot:

- **Assessing Normality of Residuals**:
    - One of the key assumptions in linear regression is that the residuals (the differences between observed and predicted values) should be normally distributed.
    - A Q-Q plot helps to visually assess this assumption. If the residuals are normally distributed, the points on the Q-Q plot will lie along a straight line.
- **Identifying Skewness or Kurtosis**:
    - If the points on the Q-Q plot deviate from the line at the tails, it indicates that the residuals have skewness (asymmetry) or kurtosis (heaviness of tails).
    - For example, if the points form an S-shape, it suggests that the residuals have more extreme values than would be expected under a normal distribution (leptokurtic).
- **Diagnosing Model Fit**:
    - The normality of residuals is essential for making valid statistical inferences, such as confidence intervals and hypothesis tests on the regression coefficients.

- If the residuals are not normally distributed, the validity of these inferences is compromised, and the model might need to be revised (e.g., by transforming the data).

A Q-Q plot is an essential diagnostic tool in linear regression, used to check whether the residuals are normally distributed. Normality of residuals is a key assumption in linear regression, and the Q-Q plot provides a simple visual way to verify this assumption. Proper assessment and interpretation of a Q-Q plot help ensure the validity of the regression model and its predictions.