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)

- There are more users using bike sharing when the weather is clear, or partly cloudy.
- Bike sharing usage increases when the wind speed is between 8 and 15.
- Bike sharing is more popular on working days (not holidays).
- Bike sharing usage is higher when the temperature is between 10°C and 30°C.
- There are more bike-sharing users when the humidity is between 50% and 80%.
- Users tend to use bike sharing more during the fall season, while the spring season sees the least usage.
- Most bookings occur between May and September.
- The year 2019 saw a higher number of bike usage compared to the previous year.
- The 'temp' and 'atemp' variables show the highest correlation with the target variable (cnt).

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)

- Multicollinearity:
    - Including all the dummy variables leads to multicollinearity, which can affect the performance of linear models (e.g., linear regression) by inflating standard errors and making it harder to interpret the coefficients.
- Efficiency:
    - Dropping one category reduces the number of variables and ensures the model is more efficient and interpretable.
- Model Interpretation:
    - By dropping the first category, the model interprets the remaining dummy variables in relation to the dropped category, which serves as the baseline or reference group.
- Reduces redundancy
- In Python, when using libraries like pandas, we can set `drop_first=True` while creating dummy variables to automatically drop one of the dummy variables to adhere to this $n - 1$ rule.

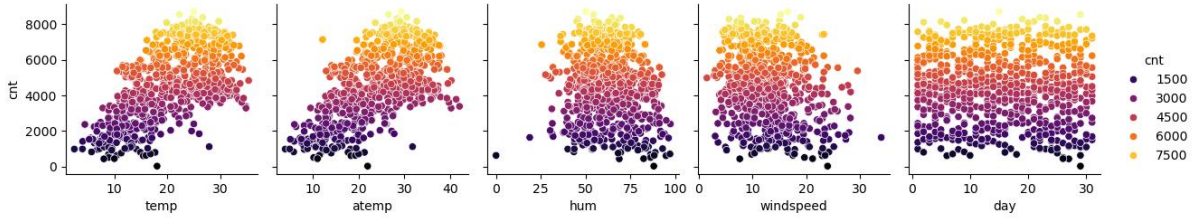
Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)

Total Marks: 1 mark (Do not edit)

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

- The variable 'temp' and atemp having the strongest correlation with the target variable, as depicted in the graph below. ['atemp' and 'temp' are redundant variables, removed one

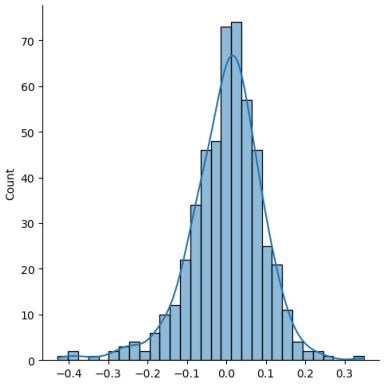
during modelling]

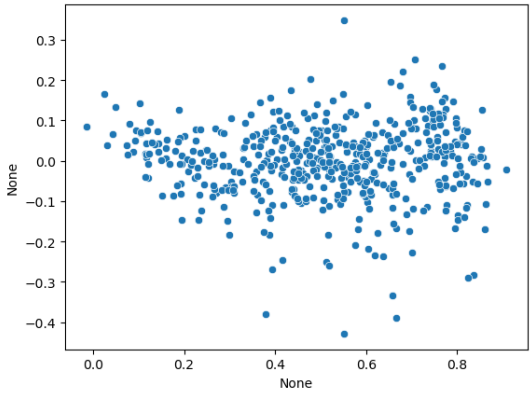


Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)

Total Marks: 3 marks (Do not edit)

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

- Assumption 1 – There should be a linear relation between independent variable and dependent variable.
    - Here dependent variable is cnt and independent variables are temp, atemp, hum, windspeed, etc. And from the graph above – there is a linear relationship.
- Assumption 2 – The residuals/error terms (Y – y predicted) should be normally distributed (mean at 0). It is clearly visible from the graph below.
    - 

- Assumption 3 - Homoscedasticity check (There should be no visible pattern in residual values.)
    - ```
      sns.scatterplot(x=y_train_predicted,y=(y_train - y_train_predicted))
      ```
    - ```
      plt.show()
      ```
    - 

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)

- Temperature (temp)
- Year
- Month (September)

```
1  lr_model_final.params
```
✓ 0.0s

```
const        0.130983
yr           0.233089
temp         0.512421
windspeed   -0.153827
2-summer     0.103252
4-winter     0.125201
2-misty     -0.081842
3-snow      -0.284333
10-oct       0.035694
8-aug        0.058868
9-sep        0.118525
dtype: float64
```

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)

<Your answer for Question 6 goes here>

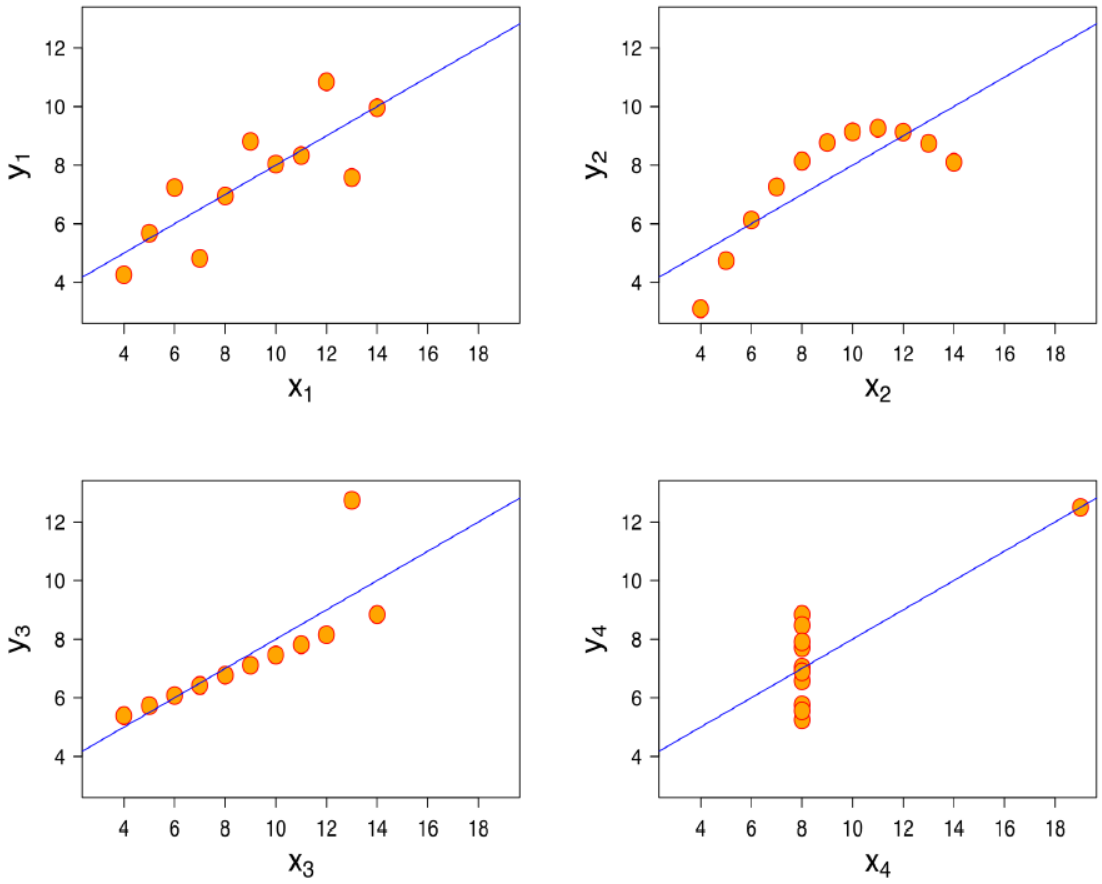
- Linear regression is one of the most fundamental and widely used algorithms in statistics and machine learning. It is used to model the relationship between a dependent variable (target) and one or more independent variables (predictors or features). In linear regression, the relationship between the variables is modeled using a straight line.
- Linear Regression categories
    - Simple Linear Regression
        - Simple linear regression involves a single predictor variable and the target variable. The goal is to find the line that best fits the data points.
        - The equation for simple linear regression is:
            - $y=\beta 0+\beta 1x+\epsilon$
    - Multi Linear Regression
        - In multiple linear regression, the target variable y is modelled as a linear combination of multiple predictors:
        - Equation
            - $y=\beta 0+\beta 1x1+\beta 2x2+\cdots+\beta pxp+\epsilon$

        - Simple linear regression involves a single predictor variable and the target variable. The goal is to find the line that best fits the data points.
        - The equation for simple linear regression is:
            - $y=\beta 0+\beta 1x+\epsilon$
    - Linear regression steps
        - Preparing data for modeling
            - Encoding
                - Convert binary variable to 1/0
                - Convert the categorical to dummy variables.
            - Rescaling of variables
            - Split into train-test
        - Train the model
            - Bottom-up approach
            - Top-down approach - Add all at a time, remove one by one.
        - Remove based on p value and VIF
            - Find out p-value and VIF for all variables
        - There are multiple scenarios
            - High p-value (statistically insignificant), High VIF (independent variable correlation) - Remove
            - High-Low
                - High p-value - low VIF
                    - Remove high p-value first, check it again
                - Low p-value - High VIF
                    - Remove high VIF first, check it again
            - Low-low - Keep
        - Repeat the steps - check VIF, p-value - remove - calculate mode summary, VIF- repeat.
        - Evaluate the model (Using Train data and test data)- Residual analysis

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)

<Your answer for Question 7 goes here>

- Anscombe's Quartet is a set of four datasets that have nearly identical simple descriptive statistics, such as mean, variance, correlation, and regression line, yet are visually very different from one another.
- It was created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data before drawing conclusions from statistical analysis.
- The key lesson of Anscombe's Quartet is that statistical measures like mean, variance, correlation, and regression line alone may not be sufficient to understand the underlying patterns in data. Visual inspection of data often reveals important details that summary statistics can miss, such as outliers, non-linearity, and heteroscedasticity.
- Graphical representa8on of Anscombe's quartet
    - All four sets are identical when examined using simple summary statistics but vary considerably when graphed.



- In python
  - anscombe_data = sns.load_dataset('data') # Create a pairplot to visualize the quartet sns.pairplot(anscombe, hue='dataset')
  - plt.show()

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)

<Your answer for Question 8 goes here>

- Pearson's r, also known as Pearson's correlation coefficient, is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables.
- It is one of the most widely used methods for assessing the degree of correlation between two variables in statistics.

The formula for Pearson's correlation coefficient is:

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

Where:

- x_i and y_i are individual data points in the two variables x and y,
- $\bar{x}$ and $\bar{y}$ are the means (averages) of the x and y variables, respectively,
- The summation $\sum$ runs over all the data points in the dataset.

- The value of Pearson's r lies between -1 and 1, and its interpretation is as follows:
    - r=1: Perfect positive correlation. As one variable increases, the other variable increases in a perfectly linear manner.
    - r=−1: Perfect negative correlation. As one variable increases, the other decreases in a perfectly linear manner.
    - r=0: No linear correlation. The two variables do not have any linear relationship. However, this does not mean there is no relationship at all—it just means that there is no linear relationship.
    - 0<r<1: Positive correlation. As one variable increases, the other tends to increase, but not necessarily in a perfectly linear way. The closer r is to 1, the stronger the positive linear relationship.
    - −1<r<0: Negative correlation. As one variable increases, the other tends to decrease, but not necessarily in a perfectly linear way. The closer r is to -1, the stronger the negative linear relationship.

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>

- Scaling refers to the process of transforming the features or variables in a dataset to a specific range or distribution.
- This is typically done to bring different variables onto a comparable scale, which is particularly important when using machine learning algorithms that are sensitive to the magnitude of input features.
- Types of Scaling
    - There are two common types of scaling: **Normalization** and **Standardization**.
- Normalization (Min-Max Scaling)
    - Normalization, also known as **Min-Max Scaling**, involves transforming the features so that they lie within a specified range, typically between **0 and 1**, but can also be scaled to other ranges, such as [-1, 1].

The formula for Min-Max scaling is:

$$X_{\text{norm}} = \frac{X - \min(X)}{\max(X) - \min(X)}$$

Where:

- X is the original value of the feature.
- $\min(X)$ is the minimum value of the feature.
- $\max(X)$ is the maximum value of the feature.

- Standardization (Z-Score Scaling)
    - Standardization, also known as Z-score scaling, involves transforming the data so that it has a mean of 0 and a standard deviation of 1. It centres the data around 0 and scales it based on how much each feature deviates from the mean.

The formula for standardization is:

$$X_{\text{std}} = \frac{X - \mu}{\sigma}$$

Where:

- X is the original value of the feature.
- μ is the mean of the feature.
- σ is the standard deviation of the feature.

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)

<Your answer for Question 10 goes here>

- A VIF value becomes infinite when the R2 value of the regression of the given variable on all other predictors equals 1. This means the variable in question is perfectly linearly correlated with one or more of the other variables. Here's why this happens:
- Perfect Multicollinearity:
    - When two or more variables are perfectly collinear (i.e., one variable is an exact linear combination of others), the R2 value of the regression becomes 1. This means that the predictor variable can be perfectly predicted using other predictors, leaving no independent variance to explain.
    - If R2=1, then the denominator of the VIF formula becomes 1−1=01 - 1 = 01−1=0, leading to a division by zero, which results in an infinite VIF.
- Mathematical Explanation:
    - VIF is calculated using the formula VIF=1/1−1= 1/0 = infinity

- Implication of Infinite VIF:
    - An infinite VIF indicates that there is perfect multicollinearity. In this case, it is impossible to determine the unique contribution of the variables involved because they provide redundant information. The model cannot reliably estimate the coefficients for those variables because they are perfectly correlated.

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)

<Your answer for Question 11 goes here>

- A Q-Q plot (Quantile-Quantile plot) is a graphical tool used to assess whether a dataset follows a particular theoretical distribution, such as the normal distribution.
- It compares the quantiles of the observed data against the quantiles of a specified reference distribution (often the normal distribution). The points in the plot represent the relationship between the observed data and the theoretical distribution.
- In a Q-Q plot:
    - X-axis: Represents the quantiles from the theoretical distribution (e.g., the normal distribution).
    - Y-axis: Represents the quantiles from the observed data.
- If the data points lie on or near a straight line (usually a 45-degree line), it suggests that the data follows the distribution being compared to (e.g., normal distribution). Deviations from the line suggest discrepancies between the observed and expected distributions.
- Use of Q-Q plot in regression
    - Checking Normality of Residuals
    - Identifying Outliers
    - Diagnosing Skewness or Kurtosis in Residuals
    - Validating Homoscedasticity (Constant Variance)