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?**

The assignment includes the following categorical variables for analysis: "season," "workingday," "weathersit," "weekday," "yr," "holiday," and "mnth."

i. season:

- Summer and fall are identified as the most favourable seasons for biking based on the available data.
- Strategic advertising and targeting during these seasons could lead to higher rental targets.
- Spring shows significantly lower bike rental ratios.

ii. workingday:

- This variable distinguishes between weekdays and weekends/holidays.
- Registered users predominantly rent bikes on weekdays, whereas casual users prefer weekends and holidays.
- Addressing the differing preferences of registered and casual users could optimize rental strategies.

iii. weathersit:

- Clear or partly cloudy days are favoured for bike rentals.
- Even on lightly rainy days, registered user counts remain relatively high, suggesting bike usage for daily commuting.
- Data on heavy rain or snow days are unavailable.

iv. weekday:

- Analysis of the "cnt" column does not reveal significant patterns with weekdays.
- However, when examining "registered" users, bike usage is higher on weekdays, whereas "casual" users show the opposite trend.

v. yr:

- Data spans two years, with bike rentals increasing notably from 2018 to 2019.

vi. holiday:

- Comparing bike usage between "registered" and "casual" users on holidays reveals that "casual" users utilize bikes more frequently on holidays.

vii. mnth:

- Months with higher bike rental ratios include June, July, August, September, and October.
- The 75th percentile of bike rentals increases notably during these months.
- This refined analysis provides actionable insights into seasonal trends, user behaviour based on weekdays and weather conditions, and the impact of holidays on bike rentals. These insights can guide targeted marketing strategies and operational planning for maximizing bike rental revenues.

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

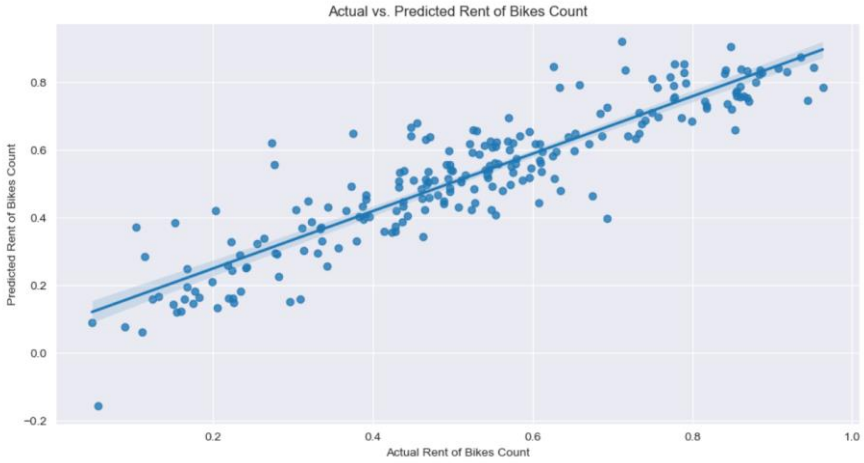
One-hot encoding creates dummy variables for each category of a categorical variable. Each dummy variable indicates whether a specific category is present (1) or absent (0) in a given observation. By setting drop_first=True, we exclude one dummy variable (typically the first one) to prevent multicollinearity. This approach allows us to infer the presence of the omitted category when all remaining dummy variables for that category are 0 in a particular observation.

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

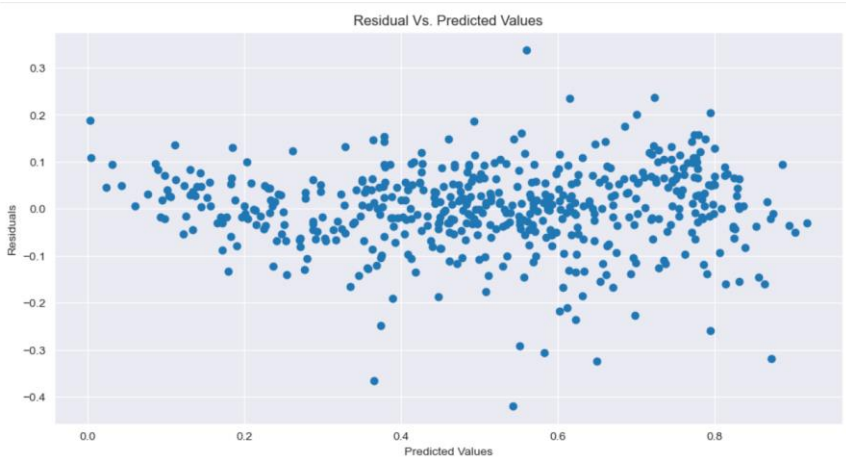
- **"temp"** exhibits the highest correlation with the target variable at 0.63.
- **"casual" and "registered"** variables are components of the target variable, as their values sum up to form the total count. Therefore, their correlations are not considered separately.
- **"atemp"** is derived from "temp," humidity, and windspeed, and has been excluded from the model preparation process. Thus, its correlation with the target variable is not taken into account.

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

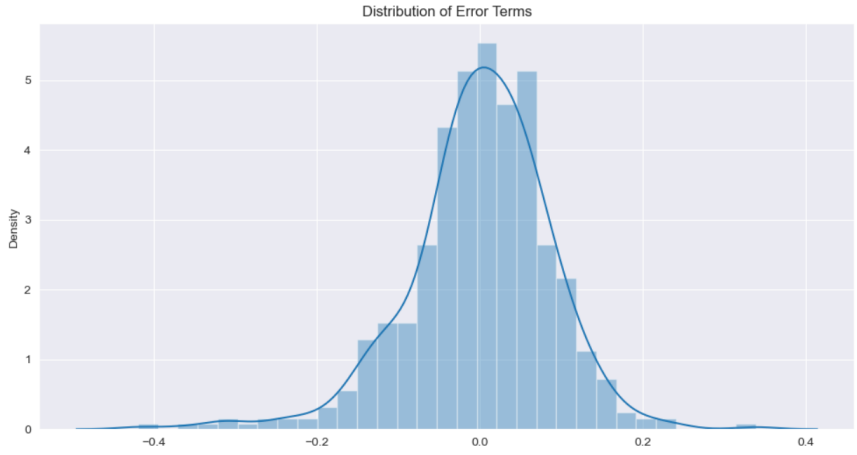
1. Linear relationship between independent and dependent variables – The linearity is validated by looking at the points distributed symmetrically around the diagonal line of the actual vs predicted plot as shown in the below figure.



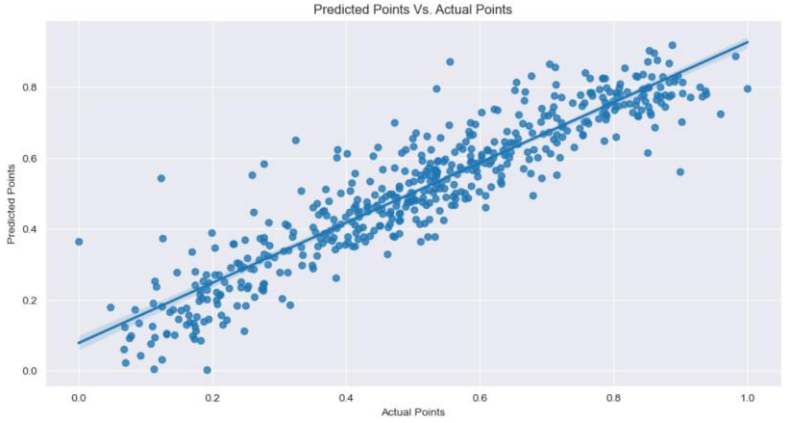
2. Independence of Error Terms: The absence of discernible patterns in the error terms relative to predictions suggests that the error terms are independent of each other.



3. Normal Distribution of Error Terms: The histogram and distribution plot illustrate the normal distribution of error terms, centred around a mean of 0. The figure below clearly demonstrates this characteristic.



4. Constant Variance of Error Terms (Homoscedasticity): The error terms exhibit approximately constant variance, indicating adherence to the assumption of homoscedasticity.



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

The Top 3 Variables:

- **weathersit:** Temperature emerges as the most significant feature positively impacting business performance. Conversely, adverse environmental conditions such as rain, humidity, windspeed, and cloudy skies have a negative effect.
- **'yr' (Year):** The year-over-year growth appears to be organic, likely influenced by geographical factors.
- **'season' (Season):** The winter season plays a crucial role in driving demand for shared bikes.

These insights highlight the influential factors affecting bike-sharing demand based on the dataset provided.

General Subjective Questions

1. **Explain the linear regression algorithm in detail.**

- Linear regression is the method of finding the best linear relationship within the independent variables and dependent variables.
- The algorithm uses the best fitting line to map the association between independent variables with dependent variable.
- There are 2 types of linear regression algorithms
    - o Simple Linear Regression – Single independent variable is used.
        - $Y = \beta_0 + \beta_1 X$ is the line equation used for SLR. o Multiple Linear Regression – Multiple independent variables are used.
        - $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \in$ is the line equation for MLR.
    - o $\beta_0 = value\ of\ the\ Y\ when\ X = 0\ (Y\ intercept)$
    - o $\beta_1, \beta_2, \ldots, \beta_p = Slope\ or\ the\ gradient$.
- Cost functions – The cost functions help to identify the best possible values for the $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ which helps to predict the probability of the target variable. The

minimization approach is used to reduce the cost functions to get the best fitting line to predict the dependent variable. There are 2 types of cost function minimization approaches – **Unconstrained and constrained**.

- o Sum of squared function is used as a cost function to identify the best fit line. The cost functions are usually represented as
  - The straight-line equation is $Y = \beta_0 + \beta_1 X$
  - The prediction line equation would be $Y_{pred} = \beta_0 + \beta_1 x_i$ and the actual Y is as Y_i. $Nowth e\ cost\ function\ will\ be\ J(\beta_1, \beta_0) = \sum (y_i - \beta_1 x_i - \beta_0)^2$
- o The unconstrained minimization is solved using 2 methods
  - Closed form
  - Gradient descent

- While finding the best fit line we encounter that there are errors while mapping the actual values to the line. These errors are nothing but the residuals. To minimize the error squares OLS (Ordinary least square) is used. o $e_i = y_i - y_{pred}$ is provides the error for each of the data point. o OLS is used to minimize the total e^2 which is called as Residual sum of squares.
  - o $\text{RSS} = = \sum_{i=1}^{n} (y_i - y_{pred})^2$
- Ordinary Lease Squares method is used to minimize Residual Sum of Squares and estimate beta coefficients.

Linear regression relies on several assumptions:

1. **Linearity:** The relationship between x and y is linear.
2. **Independence:** Observations are independent of each other.
3. **Homoscedasticity:** The variance of the error terms ϵ\epsilon is constant across all levels of x
4. **Normality:** The error terms ϵ\epsilonϵ are normally distributed.
5. **No Multicollinearity:** The independent variables x is not highly correlated with each other.

Advantages of Linear Regression:

- **Interpretability:** Coefficients indicate the strength and direction of relationships between variables.
- **Simplicity:** Easy to implement and understand.
- **Efficiency:** Computationally efficient for large datasets.
- **Versatility:** Can be extended to handle non-linear relationships through transformations.

Limitations:

- **Assumption Sensitivity:** Performance degrades if assumptions (linearity, normality, etc.) are violated.
- **Overfitting:** Prone to overfitting with many predictors and insufficient data.

- **Underperformance:** May not capture complex relationships well compared to non-linear models.

Linear regression serves as a foundational tool in data analysis and predictive modelling, providing valuable insights into relationships between variables when assumptions are met.

2. **Explain the Anscombe's quartet in detail.**

Anscombe's quartet is a famous dataset in statistics that consists of four sets of x-y data pairs. Each dataset has the same statistical properties when analysed individually, yet they exhibit vastly different characteristics when graphed. This illustrates the importance of visualizing data and not relying solely on summary statistics.

Francis Anscombe, an English statistician, created this dataset in 1973 to emphasize the significance of graphing data before making conclusions based on statistical measures alone. Despite their identical statistical properties, the datasets reveal strikingly different relationships when plotted.

Characteristics of Anscombe's Quartet:

1. Dataset 1:

- x: [10, 8, 13, 9, 11, 14, 6, 4, 12, 7]
- y: [8.04, 6.95, 7.58, 8.81, 8.33, 9.96, 7.24, 4.26, 10.84, 4.82]
- Summary Statistics
- Mean of x: 9.0, Mean of y: 7.50
- Variance of x: 11.0, Variance of y: 4.12
- Correlation coefficient: 0.816
- Linear regression: \(y = 3.00 + 0.50x \)
- Relationship Roughly linear with a slight upward trend.

2. Dataset 2

- x: [10, 8, 13, 9, 11, 14, 6, 4, 12, 7]
- y: [9.14, 8.14, 8.74, 8.77, 9.26, 8.10, 6.13, 3.10, 9.13, 7.26]
- Summary Statistics
- Mean of x: 9.0, Mean of y: 7.50
- Variance of x: 11.0, Variance of y: 4.12
- Correlation coefficient: 0.816
- Linear regression: \(y = 3.00 + 0.50x \)
- Relationship Roughly linear, but with a noticeable outlier (x=8, y=8.14).

3. Dataset 3

- x: [10, 8, 13, 9, 11, 14, 6, 4, 12, 7]
- y: [7.46, 6.77, 12.74, 7.11, 7.81, 8.84, 6.08, 5.39, 8.15, 6.42]
- Summary Statistics
- Mean of x: 9.0, Mean of y: 7.50
- Variance of x: 11.0, Variance of y: 4.12
- Correlation coefficient: 0.816
- Linear regression: \(y = 3.00 + 0.50x \)
- Relationship Non-linear, driven by an outlier (x=13, y=12.74).

4. Dataset 4

- x: [8, 8, 8, 8, 8, 8, 8, 8, 8, 8]
- y: [6.58, 5.76, 7.71, 8.84, 8.47, 7.04, 5.25, 5.56, 7.91, 6.89]
- Summary Statistics
- Mean of x: 9.0, Mean of y: 7.50
- Variance of x: 11.0, Variance of y: 4.12
- Correlation coefficient: 0.816
- Linear regression: \(y = 3.00 + 0.50x \)
- Relationship Perfectly linear, except for an extreme outlier (x=8, y=[5.76 to 8.84]).

Key Observations:

- Statistical Similarity All datasets have identical mean, variance, correlation coefficient, and regression line parameters.
- Visual Differences Each dataset presents a different visual pattern (linear, non-linear, outlier influence) when graphed.
- Implications Anscombe's quartet underscores the importance of data visualization in understanding relationships and identifying outliers or patterns that statistical measures alone might overlook.

Anscombe's quartet serves as a compelling demonstration of why visual exploration of data is essential in statistical analysis. Despite their identical summary statistics, the datasets exhibit distinct characteristics when plotted, highlighting the limitations of relying solely on numerical summaries in data analysis.

3. **What is Pearson's R?**

The Pearson's R (also known as Pearson's correlation coefficients) measures the strength between the different variables and the relation with each other. The Pearson's R returns values between -1 and 1. The interpretation of the coefficients are:

- *-1 coefficient indicates strong inversely proportional relationship.*
- *0 coefficient indicates no relationship.*
- *1 coefficient indicates strong proportional relationship.*

$$r = \frac{n(\Sigma x * y) - (\Sigma x) * (\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2] * [n\Sigma y^2 - (\Sigma y)^2]}}$$

Where:

N = the number of pairs of scores

Σxy = the sum of the products of paired scores

Σx = the sum of x scores

Σy = the sum of y scores

Σx^2 = the sum of squared x scores

Σy^2 = the sum of squared y scores

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

- What - The scaling is the data preparation step for regression model. The scaling normalizes these varied datatypes to a particular data range.
- Why – In many cases, feature data is sourced from public domains where definitions of variables and their units can vary widely. This diversity leads to significant differences in data ranges and units. Without scaling these datasets, processing them without proper unit alignment becomes likely. Moreover, wider ranges increase the difficulty of comparing coefficients against the variance of the dependent variable.
- Scaling primarily influences coefficients, not the accuracy or predictive precision of models.
- Normalization/Min-Max scaling – The Min max scaling normalizes the data within the range of 0 and 1. The Min max scaling helps to normalize the outliers as well.

$$MinMaxScaling: x = \frac{x - \min(x)}{\max(x) - \min(x)}$$

- Standardization converges all the data points into a standard normal distribution where mean is 0 and standard deviation is 1.

$$Standardization: x = \frac{x - mean(x)}{sd(x)}$$

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

$$VIF = \frac{1}{1 - R^2}$$

The VIF formula clearly signifies when the VIF will be infinite. If the R^2 is 1 then the VIF is infinite. The reason for R^2 to be 1 is that there is a perfect correlation between 2 independent variables.

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

- Q-Q plots, short for quantile-quantile plots, serve as graphical tools to assess whether two datasets originate from a common distribution. The theoretical distributions can include normal, exponential, or uniform types. In linear regression, Q-Q plots are valuable for determining if the training dataset and test dataset are drawn from populations with similar distributions. This method visually examines if the data points align along a straight line, indicating conformity to a normal distribution.
- Interpretations o Similar distribution: If all the data points of quantile are lying around the straight line at an angle of 45 degree from x-axis.
    - Y values < X values: If y-values quantiles are lower than x-values quantiles. o X values < Y values: If x-values quantiles are lower than y-values quantiles.
    - Different distributions – If all the data points are lying away from the straight line.
- Advantages
    - Distribution aspects like loc, scale shifts, symmetry changes and the outliers all can be daintified from the single plot.
    - The plot has a provision to mention the sample size as well.