MLR questions

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?

Categorical Variables:
The variables provided in the assignment include "season," "workingday," "weathersit," "weekday," "yr," "holiday," and "mnth."

Season:

- **Summer and Fall:** These seasons are the most favorable for biking, suggesting potential for higher marketing targets and strategic advertising during these times.
- **Spring:** Exhibits a significantly lower usage rate.

Working Day:

- The term "workingday" denotes whether the day is a weekday or falls on a weekend/holiday.
- **Registered Users:** Predominantly rent bikes on working days.
- **Casual Users:** Prefer non-working days, though the total bike rental count balances out despite differing patterns between user types.

Weather Situation (Weathersit):

- **Favorable Conditions:** Clear or few clouds days are optimal for biking.
- **Light Rain:** Registered users frequently rent bikes, indicating usage for commuting.
- **Heavy Rain/Snow:** No data available.

Weekday:

- No notable pattern in overall bike rentals (cnt) across weekdays.
- **Registered Users:** More likely to rent bikes on weekdays.
- **Casual Users:** Tend to rent less on weekdays.

Year (Yr):

- Data covers two years, showing an increase in bike rentals from 2018 to 2019.

Holiday:

- **Casual Users:** More likely to rent bikes on holidays compared to registered users.

Month (Mnth):

- **High Rental Period:** Bike rentals are higher from June to October.
- **Quantile Data:** The 75th quantile of rentals increases during these months.

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

One-hot encoding is a method used to convert categorical variables into dummy variables (binary values 0 and 1). Each dummy variable represents one category of the original variable, where '1' indicates the presence and '0' indicates the absence of that category. For example, if a categorical variable has three categories, it will be transformed into three dummy variables.

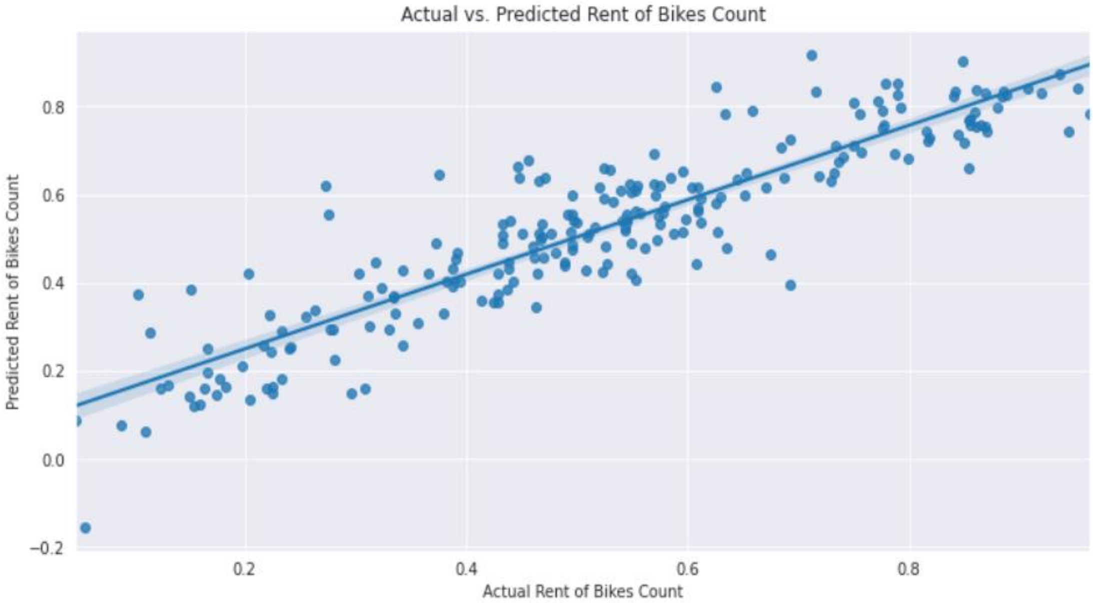
When creating dummy variables, the option **drop_first=True** is often used. This approach drops the first category, which serves as the base or reference category. This is done to prevent multicollinearity, which can complicate the model if all categories are included as dummies. The dropped category is inferred by the absence (all zeros) in the remaining categories for any given observation.

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

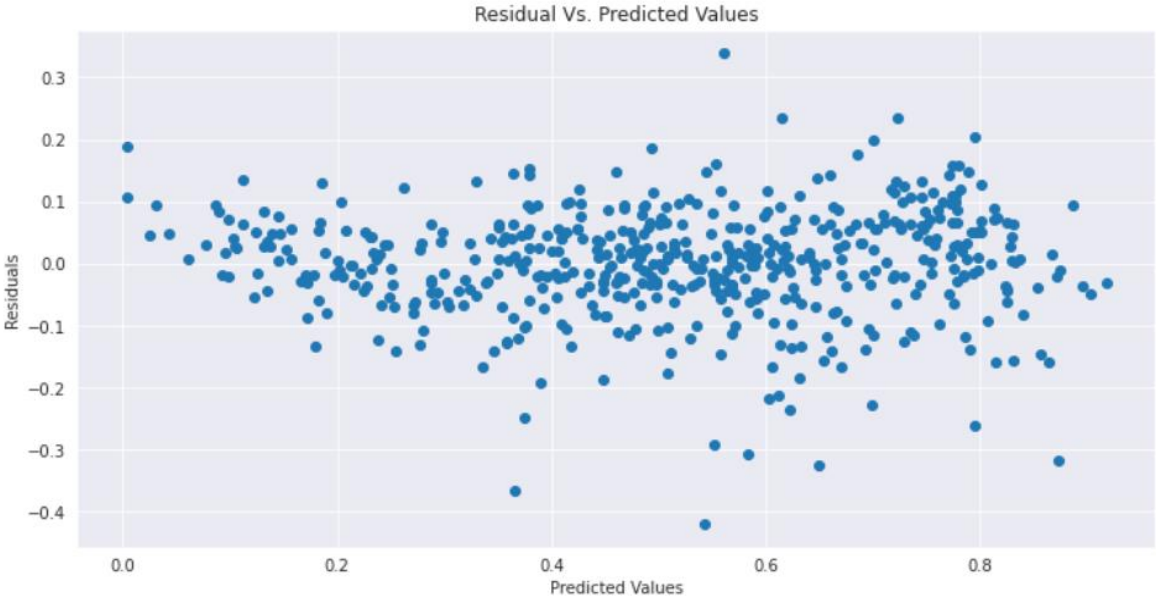
- **Temperature (temp):** This variable has the strongest correlation with the target variable, with a correlation coefficient of 0.63.
- **Casual and Registered:** These two variables are components of the target variable since their values sum up to form the target variable. Therefore, their correlations are not considered separately.
- **Apparent Temperature (atemp):** This is a derived variable calculated from temperature, humidity, and windspeed. It is excluded from the model to avoid redundancy and simplify the model preparation.

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

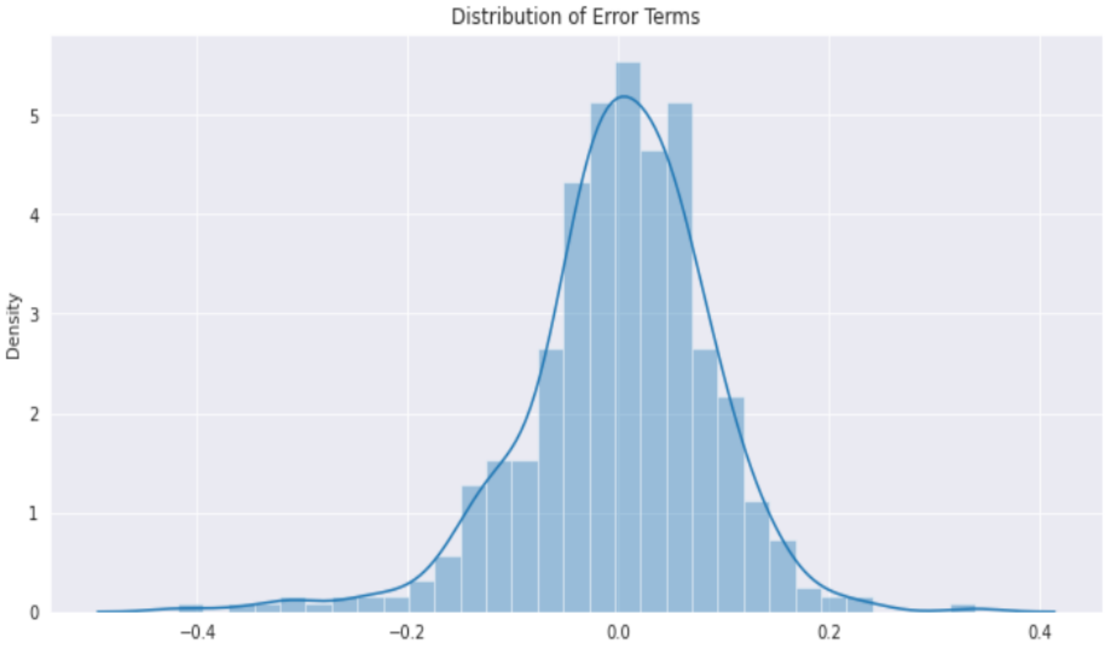
- **Linear Relationship Validation:** The linear relationship between independent and dependent variables is confirmed by examining how the points are symmetrically distributed around the diagonal line in the actual versus predicted plot, as illustrated in the figure below.



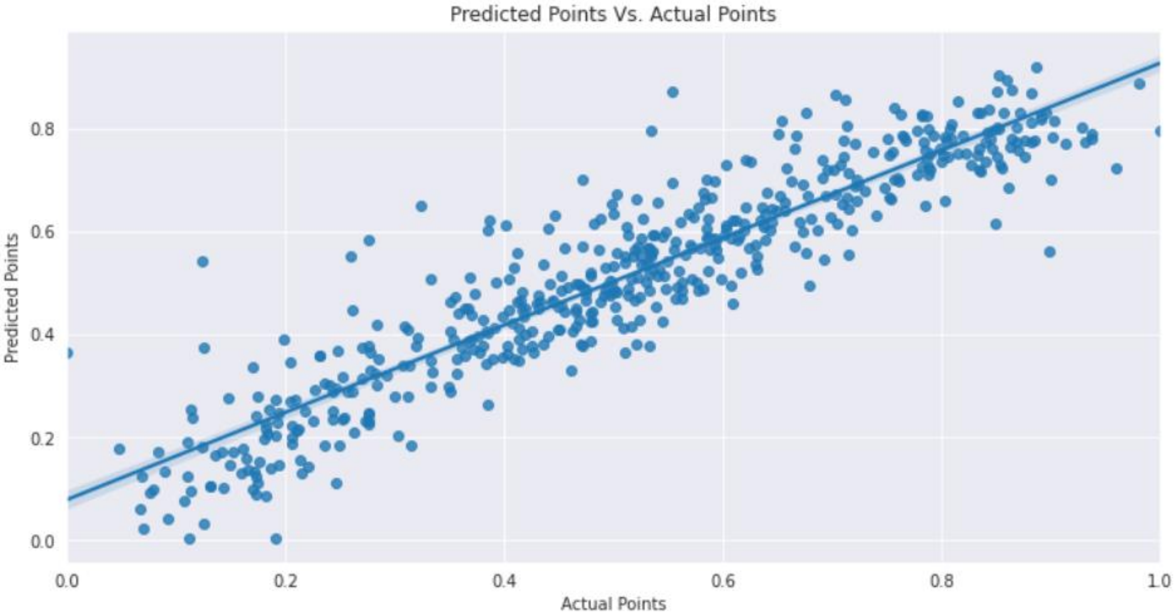
- **Independence of Error Terms:** There is no discernible pattern in the error terms relative to the predictions, indicating that the error terms are independent of each other.



- **Normal Distribution of Error Terms:** The histogram and distribution plot demonstrate that the error terms are normally distributed with a mean of zero, as shown in the figure below.



- **Constant Variance of Error Terms (Homoscedasticity):** The error terms appear to have a roughly constant variance, indicating that they meet 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?

Top 3 Variables:

- **Weathersit:** Temperature significantly boosts business, while other environmental conditions like rain, humidity, windspeed, and cloudiness have a negative impact.
- **Year (Yr):** Year-over-year growth appears organic, influenced by geographic attributes.
- **Season:** The winter season crucially influences the demand for shared bikes.

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
  - Simple Linear Regression – Single independent variable is used.
    - $Y = \beta_0 + \beta_1 X$ is the line equation used for SLR.
  - 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.
  - $\beta_0 = value\ of\ the\ Y\ when\ X = 0\ (Y\ intercept)$
  - $\beta_1, \beta_2, \ldots, \beta_p = Slope\ or\ the\ gradient.$
- Cost functions – The cost functions helps 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**.
  - 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.
    - $Now\ the\ cost\ function\ will\ be\ J(\beta_1, \beta_0) = \sum (y_i - \beta_1 x_i - \beta_0)^2$
  - The unconstrained minimization are 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.
  - $e_i = y_i - y_{pred}$ is provides the error for each of the data point.
  - OLS is used to minimize the total e^2 which is called as Residual sum of squares.
  - $\mathrm{RSS} = \ = \sum_{i=1}^{n} (y_i - y_{pred})^2$

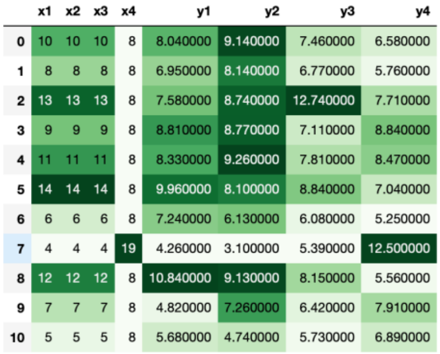
- Ordinary Lease Squares method is used to minimize Residual Sum of Squares and estimate beta coefficients.

2. Explain the Anscombe’s quartet in detail.

Statistics such as variance and standard deviation are commonly used to understand the variation in data without examining each data point individually. These statistics are effective for describing general trends and characteristics of the data.

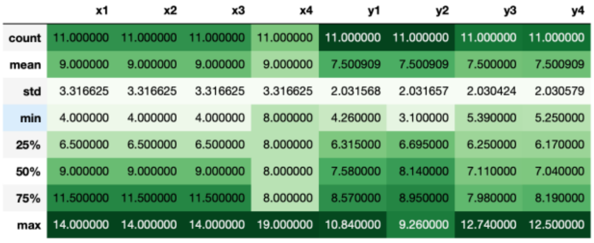
In 1973, Francis Anscombe demonstrated that statistical measures alone are insufficient to fully represent data sets. He created several datasets that, while sharing identical statistical properties, illustrated the limitations of relying solely on these statistics.

- Illustrations
    - One of the data set is as follows:



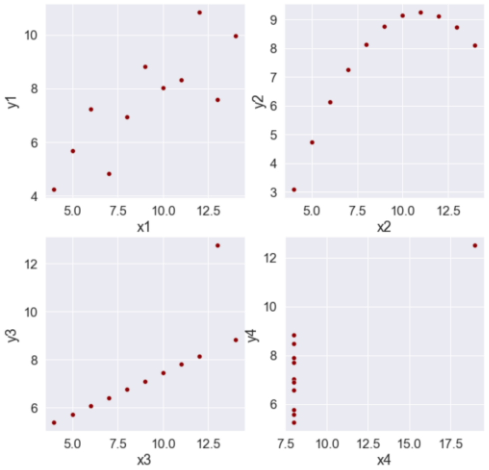
	x1	x2	x3	x4	y1	y2	y3	y4
0	10	10	10	8	8.040000	9.140000	7.460000	6.580000
1	8	8	8	8	6.950000	8.140000	6.770000	5.760000
2	13	13	13	8	7.580000	8.740000	12.740000	7.710000
3	9	9	9	8	8.810000	8.770000	7.110000	8.840000
4	11	11	11	8	8.330000	9.260000	7.810000	8.470000
5	14	14	14	8	9.960000	8.100000	8.840000	7.040000
6	6	6	6	8	7.240000	6.130000	6.080000	5.250000
7	4	4	4	19	4.260000	3.100000	5.390000	12.500000
8	12	12	12	8	10.840000	9.130000	8.150000	5.560000
9	7	7	7	8	4.820000	7.260000	6.420000	7.910000
10	5	5	5	8	5.680000	4.740000	5.730000	6.890000

    - If the descriptive statistics are checked for above data set then all looks same:



	x1	x2	x3	x4	y1	y2	y3	y4
count	11.000000	11.000000	11.000000	11.000000	11.000000	11.000000	11.000000	11.000000
mean	9.000000	9.000000	9.000000	9.000000	7.500909	7.500909	7.500000	7.500909
std	3.316625	3.316625	3.316625	3.316625	2.031568	2.031657	2.030424	2.030579
min	4.000000	4.000000	4.000000	8.000000	4.260000	3.100000	5.390000	5.250000
25%	6.500000	6.500000	6.500000	8.000000	6.315000	6.695000	6.250000	6.170000
50%	9.000000	9.000000	9.000000	8.000000	7.580000	8.140000	7.110000	7.040000
75%	11.500000	11.500000	11.500000	8.000000	8.570000	8.950000	7.980000	8.190000
max	14.000000	14.000000	14.000000	19.000000	10.840000	9.260000	12.740000	12.500000

    - However, when plotted these points, the relation looks completely different as depicted below.



- Anscombe's Quartet signifies that multiple data sets with many similar statistical properties could still be different from one another when plotted.
- The dangers of outliers in data sets are warned by the quartet. Check the bottom 2 graphs. If those outliers would have not been there the descriptive stats would have been completely different in that case.
- Important points
  - Plotting the data is very important and a good practice before analysing the data.
  - Outliers should be removed while analysing the data.
  - Descriptive statistics do not fully depict the data set in its entirety.

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 Pearon'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.*

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

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

- **What:** Scaling is a crucial data preparation step in regression modeling. It involves normalizing various data types to a specific range, ensuring consistency across the dataset.
- **Why:** Often, feature data is gathered from public domains where variable interpretations and units may vary widely. This can lead to significant differences in the units and ranges of data. Without scaling, processing this data might proceed without proper unit conversions, increasing the likelihood of discrepancies. Furthermore, larger data ranges can disproportionately influence the regression coefficients, making it difficult to compare variations in the dependent variable.
- 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 coon 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 (quantile-quantile) plots are graphical tools used to evaluate whether two datasets originate from the same distribution. These plots are particularly useful in linear regression for verifying that training and test datasets come from populations with similar distributions. They can also be employed to assess the normality of datasets by checking if the data points form a straight line with specific patterns:

Interpretations:

- **Similar Distribution:** Data points align closely with a straight line at a 45-degree angle from the x-axis, indicating that the distributions are similar.
- **Y Values < X Values:** If the quantiles of y-values are consistently lower than those of the x-values.
- **X Values < Y Values:** If the quantiles of x-values are consistently lower than those of the y-values.
- **Different Distributions:** Data points deviate significantly from the straight line, suggesting differing distributions.

Advantages:

- **Comprehensive Insights:** A Q-Q plot can reveal shifts in location or scale, changes in symmetry, and the presence of outliers, all within a single visual representation.
- **Sample Size Display:** The plot can include information about the sample size, enhancing its analytical usefulness.