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)

Insights on Categorical Variables:

1. **The categorical variables available in the assignment include:**
   - "season", "workingday", "weathersit", "weekday", "yr", "holiday", and "mnth".

2. **Season**:
   - Based on the available data, the most favorable seasons for biking are **summer** and **fall**.
     - Higher rental targets can be planned for summer and fall, along with strategic advertising.
   - **Spring** has a significantly lower consumption ratio compared to the other seasons.

3. **Working Day**:
   - The **workingday** variable represents whether the day is a weekday or a weekend/holiday.
   - **Registered users** tend to rent bikes on working days, while **casual users** prefer renting bikes on non-working days. This effect is neutralised when considering the total count, due to the opposing behaviours of registered and casual users.
   - Understanding the identity and behaviour of **registered** and **casual users** in relation to working and non-working days will help create targeted strategies to increase bike rentals.

4. **Weather Situation (weathersit)**:
   - The most favourable weather condition for bike rentals is **clear weather or days with few clouds**.
   - The count of **registered users** remains relatively high even on **light rainy days**, suggesting that bikes are being used for **daily commuting** to workplaces.
   - There is no data available for **heavy rain or snow days**, so the impact of extreme weather on bike rentals is unclear.

5. **Weekday**:
   - When analysing the "cnt" column, there is no significant pattern observed with respect to the weekday.

- However, when examining bike usage by **registered users**, we observe higher usage on **working days**. Conversely, **casual users** tend to rent bikes more on **non-working days**.

6. **Year (yr)**:
   - Data for **two years** is available, and there is a noticeable increase in bike rentals from **2018** to **2019**, indicating growth in the business.
7. **Holiday**:
   - Comparing bike consumption on **holidays** between registered and casual users reveals that **casual users** rent bikes more frequently on holidays than registered users.
8. **Month (mnth)**:
   - The bike rental ratio is higher during the months of **June, July, August, September**, and **October**.
   - The **75th percentile** of rentals grows significantly during these months.

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)

One-hot encoding is used to create dummy variables that represent the different categories of a categorical variable. Each dummy variable takes values of 1 or 0, where 1 indicates the presence of the respective category and 0 indicates its absence. For example, if a categorical variable has three categories, three dummy variables will be created.

By setting drop_first=True when creating the dummy variables, the base or reference category is dropped. This is done to avoid **multicollinearity**, which could arise if all the dummy variables were included in the model. The reference category can easily be identified by the rows where the value is 0 across all the dummy variables for that particular category.

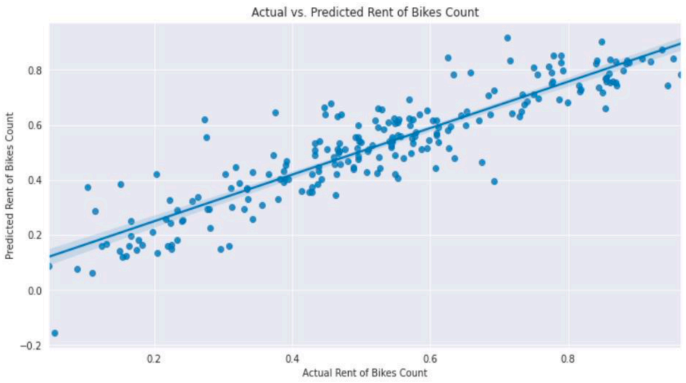
Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)
Total Marks: 1 mark (Do not edit)
Answer: <Your answer for Question 3 goes below this line> (Do not edit)

1. **"temp"** is the variable with the highest correlation with the target variable, at **0.63**.
2. The **"casual"** and **"registered"** variables are components of the target variable, as their values sum to form the target variable. Therefore, the correlation for these two variables is not considered.

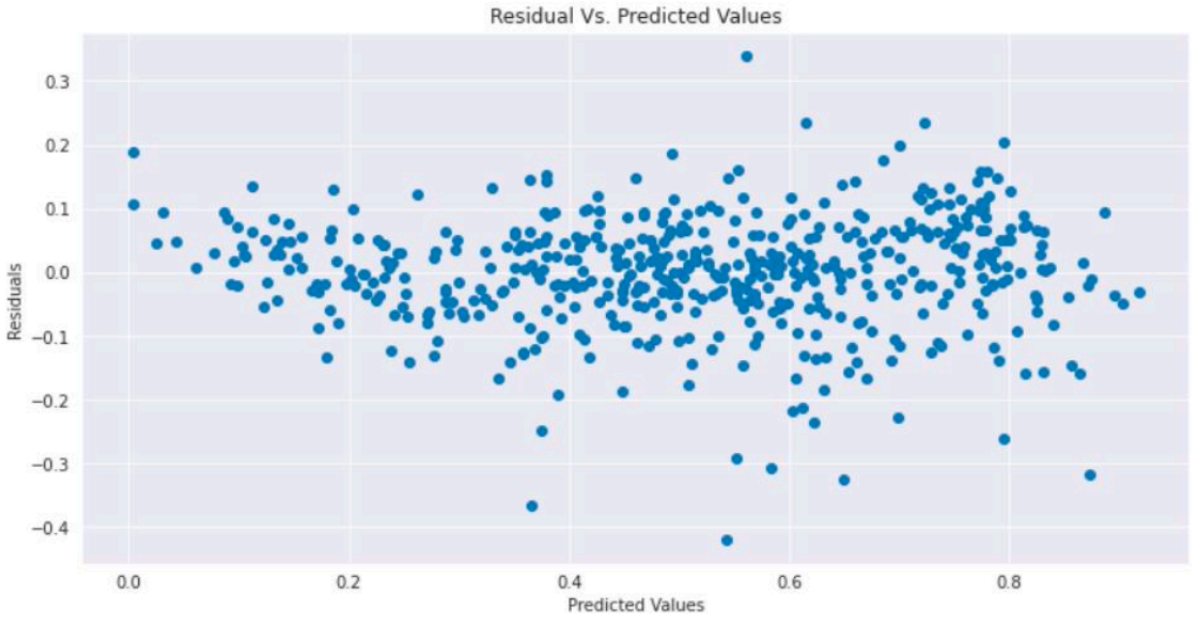
3. **"atemp"** is a derived parameter based on **temp**, **humidity**, and **windspeed**, and is therefore excluded from the model preparation process.

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)

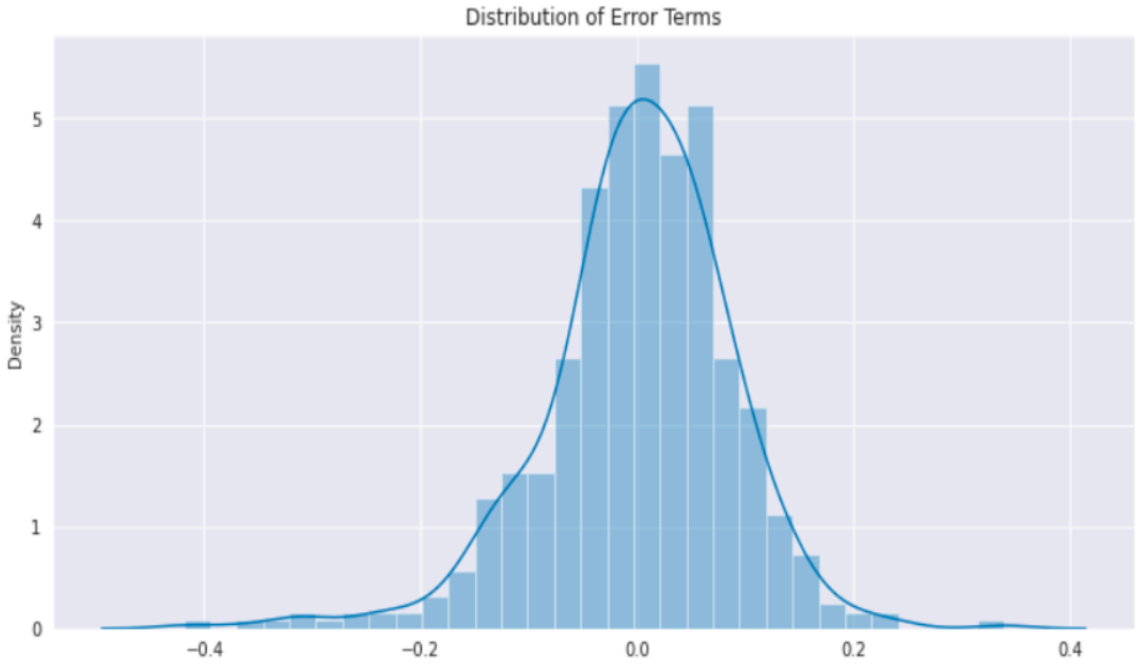
1. **Linear relationship between independent and dependent variables**: The linearity is validated by observing the points distributed symmetrically around the diagonal line in the actual vs. predicted plot, as shown in the figure below.



2. **Error terms are independent of each other**: There is no specific pattern observed in the error terms with respect to the predictions, which suggests that the error terms are independent of each other.



3.**Error terms are normally distributed**: The histogram and distribution plot help to visualise the normal distribution of the error terms, with a mean of 0. The figure below



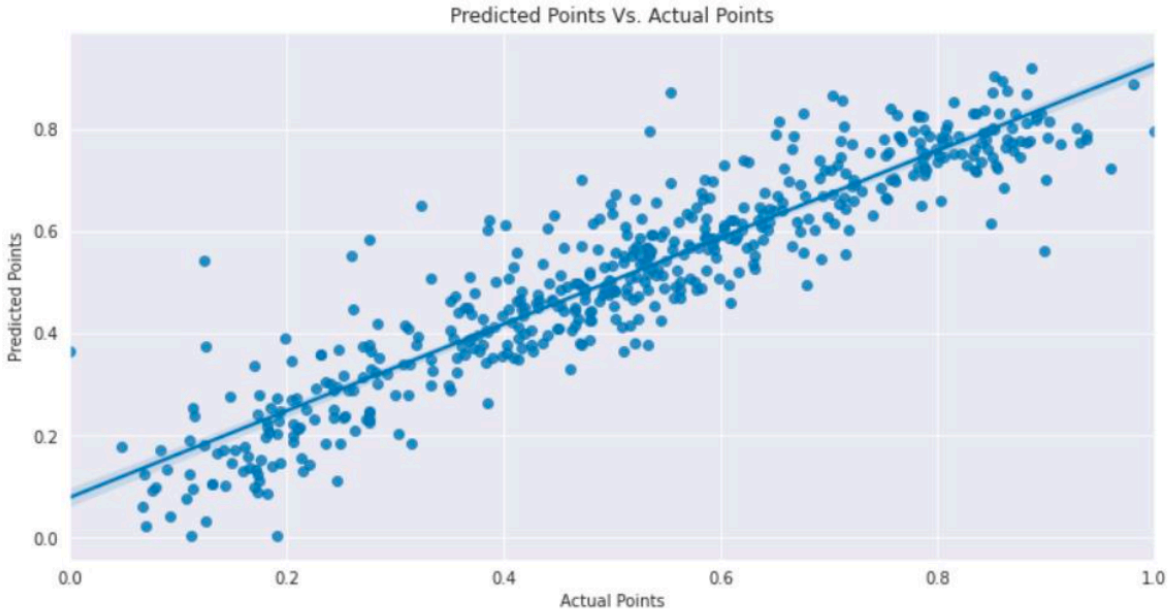
clearly illustrates this.

4. **Error terms have constant variance (homoscedasticity)**:

The error terms exhibit approximately constant variance, which satisfies the assumption of homoscedasticity.

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)

1. **Weathersit**:
   Temperature is the most significant feature that positively affects the business, whereas other environmental conditions, such as rain, humidity, wind speed, and cloudiness, negatively impact the business.



2. **Yr**:
   The year-on-year growth appears to be organic, driven by geographical factors.
3. **Season**:
   The winter season plays a crucial role in the demand for shared bikes.

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:

Linear regression is a method for finding the best linear relationship between independent variables and a dependent variable.

- The algorithm uses the best-fitting line to map the relationship between independent variables and the dependent variable.

There are two types of linear regression algorithms:

Simple Linear Regression (SLR):

- A single independent variable is used.
- The equation for the line is : $Y = \beta_0 + \beta_1 X$

Multiple Linear Regression (MLR):

- Multiple independent variables are used.
- The equation for the line is: $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \in$

where $\beta_0 = value\ of\ the\ Y\ when\ X = 0\ (Y\ intercept)$ and $\beta_1, \beta_2, \ldots, \beta_p = Slope\ or\ the\ gradients\ of\ the\ line\ with\ respect\ to\ each\ independent\ variable\ X_1, X_2, X_3 \ldots X_p$.

Cost Function:

The cost function is used to identify the best possible values for $\beta_1, \beta_2, \ldots, \beta_p$ that minimise the error in predicting the target variable. The goal is to minimise the cost function to get the best-fitting line.

- The most commonly used cost function is the **Sum of Squared Errors (SSE)**, which is minimised to find the optimal parameters for the model.
- The cost function can be represented as:

$$J(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

Where:
- $Ypred = \beta 0 + \beta 1 xi$ is the predicted value.
- yi is the actual value.
- x_i is the independent variable.

Cost Function Minimisation Approaches:

There are two main types of cost function minimisation approaches:

1. **Unconstrained Minimisation**:
   - Solved using methods like **Closed Form** or **Gradient Descent**.
2. **Constrained Minimisation**:
   - Typically involves adding constraints to the minimisation problem, although this is less common in basic linear regression.

Ordinary Least Squares (OLS):

When fitting the regression line, we encounter errors while mapping actual values to predicted values. These errors are called **residuals**. To minimize the sum of squared errors (or residuals), **Ordinary Least Squares (OLS)** is used.

- The residual for each data point is calculated as: $ei = yi - ypred$

  Where yi is the actual value and $ypred$ is the predicted value.
- The **Residual Sum of Squares (RSS)** is the sum of the squared residuals:

  $RSS = = \sum (y_i - y_{pred})^2$

The goal of OLS is to minimise this residual sum of squares (RSS) to find the optimal values of the parameters $\beta_1, \beta_2, \ldots, \beta_p$.

- **OLS** helps estimate the beta coefficients that minimise the RSS, thereby providing the best-fitting line for the data.

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)

Statistics such as variance and standard deviation are often considered sufficient for understanding the variation in data without examining every individual data point. These statistics are useful for describing general trends and key characteristics of the data.

However, in 1973, **Francis Anscombe** realised that relying solely on statistical measures is not always enough to fully capture the nature of a data set. To illustrate this point, he created several data sets, each with identical statistical properties, demonstrating how the same statistics can describe very different data patterns.

Illustrations

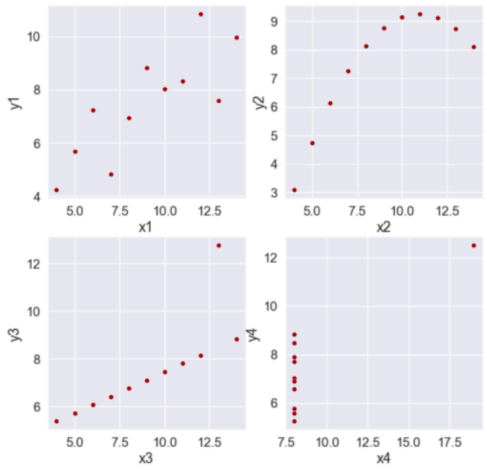
- One of the data set is as follows:

	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 demonstrates that multiple data sets with identical statistical properties can still differ significantly when plotted visually.

- The quartet also highlights the dangers of **outliers** in data sets. For example, if the outliers in the bottom two graphs were removed, the descriptive statistics would be completely different.

Important Points:

- **Plotting the data** is an essential and good practice before analyzing it.
- **Outliers** should be addressed or removed during data analysis.
- **Descriptive statistics** alone do not fully represent the data set, as they may overlook underlying patterns or anomalies.

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)

Pearson's R (also known as Pearson's correlation coefficient) measures the strength and direction of the linear relationship between two variables. The value of Pearson's R ranges from -1 to 1, and the interpretation of the coefficients is as follows:

- A coefficient of **-1** indicates a **strong inverse relationship** (perfect negative correlation).
- A coefficient of **0** indicates **no linear relationship** between the variables.
- A coefficient of **1** indicates a **strong direct relationship** (perfect positive correlation).

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

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

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)

What: Scaling is a crucial data preparation step for regression models. It normalises variables with different units and ranges into a consistent scale, making the data suitable for analysis.

Why: In many cases, feature data is collected from public domains where the interpretation of variables and their units may not be standardised. This often leads to high variance in the units and ranges of the data. Without proper scaling, there is a high risk of processing the data without appropriate unit conversions, which can lead to inaccurate model results. Furthermore, when the data has a wide range, it increases the possibility that the regression coefficients may be skewed, impairing the model's ability to compare the variance of the dependent variable.

Scaling primarily affects the coefficients of the model. However, it does not impact the overall prediction or the precision of predictions.

Normalisation/Min-Max Scaling: Min-Max scaling normalises the data to a fixed range, typically between 0 and 1. This technique is also effective in reducing the impact of outliers, ensuring that they do not distort the analysis.

$MinMaxScaling$: $x = x - \min(x)/\max(x) - \min(x)$

Standardisation transforms the data so that it follows a standard normal distribution, with a mean of 0 and a standard deviation of 1.

$Standardization$: $x = x - mean(x)/sd(x)$

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)

The VIF (Variance Inflation Factor) formula clearly shows that the VIF will be infinite when the R-squared is equal to 1. This occurs because an R-squared 1 indicates perfect correlation between two independent variables.

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

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)

Q-Q plots (Quantile-Quantile plots) are graphical tools used to assess whether two data sets come from the same distribution. The theoretical distributions could be normal, exponential, uniform, or others. Q-Q plots are particularly useful in linear regression to verify that the training and test data sets come from populations with the same distribution. They are also an effective method for checking if a data set follows a normal distribution, where the points should align along a straight line, as explained below:

Interpretations:

- **Similar distribution**: If all the quantile points lie along a straight line at a 45-degree angle from the x-axis, the distributions are similar.
- **Y-values < X-values**: If the y-values' quantiles are lower than the x-values' quantiles, it indicates a deviation where the second data set is "smaller" or has lower values.
- **X-values < Y-values**: If the x-values' quantiles are lower than the y-values' quantiles, the opposite pattern is observed.
- **Different distributions**: If the points deviate significantly from the straight line, it suggests that the distributions are different.

Advantages:

- A Q-Q plot can reveal various aspects of the data distribution, including location shifts, scale changes, symmetry issues, and the presence of outliers.
- The plot also allows for the inclusion of sample size information, providing more context for interpretation.
