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?

Season

- Based on the available data, summer and fall are the most favorable seasons for biking. This suggests that higher targets can be set and strategic advertising can be planned during these periods.
- Spring, however, shows a significantly low bike consumption ratio.

Working Day

- "Workingday" indicates whether a day is a weekday or a weekend/holiday.
- Registered users tend to rent bikes on working days, while casual users prefer non-working days. When considering the total count, this opposing behavior nullifies the individual patterns.
- Understanding the distinct behaviors of registered and casual users, and developing relevant strategies for working and non-working days, could help increase overall bike rentals.

Weather Situation (weathersit)

- The most favorable weather condition is clear skies with few clouds.
- Registered user counts remain relatively high even on lightly rainy days, suggesting bikes are used for daily commutes.
- No data is available for heavy rain or snow days.

Weekday

- When examining the "cnt" (total count) column, no significant pattern is observed with the weekday.
- However, plotting the relationship with registered users reveals higher bike usage on working days, while the opposite is true for casual users.

Year (yr)

- Two years of data are available, showing an increase in bike rentals from 2018 to 2019.

Holiday

- When comparing bike consumption between registered and casual users on holidays, casual users are observed to use bikes more frequently on holidays.

Month (mnth)

- Bike rental ratios are higher for June, July, August, September, and October.
- The 75th percentile of bike rentals also increases during these months.

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

When using one-hot encoding, **dummy variables** are created to represent the range of values within a categorical variable. Each dummy variable is binary, taking on a value of **1 to indicate the presence of a specific category and 0 to indicate its absence**. For instance, if a categorical variable has three distinct categories, three corresponding dummy variables will be generated.

The parameter drop_first = True is employed during the creation of these dummy variables to **exclude the base or reference category**. This practice is crucial to **prevent multicollinearity** from being introduced into the model if all dummy variables were included. The reference category can be easily inferred: it is the category for which all other dummy variables in a given row are simultaneously 0.

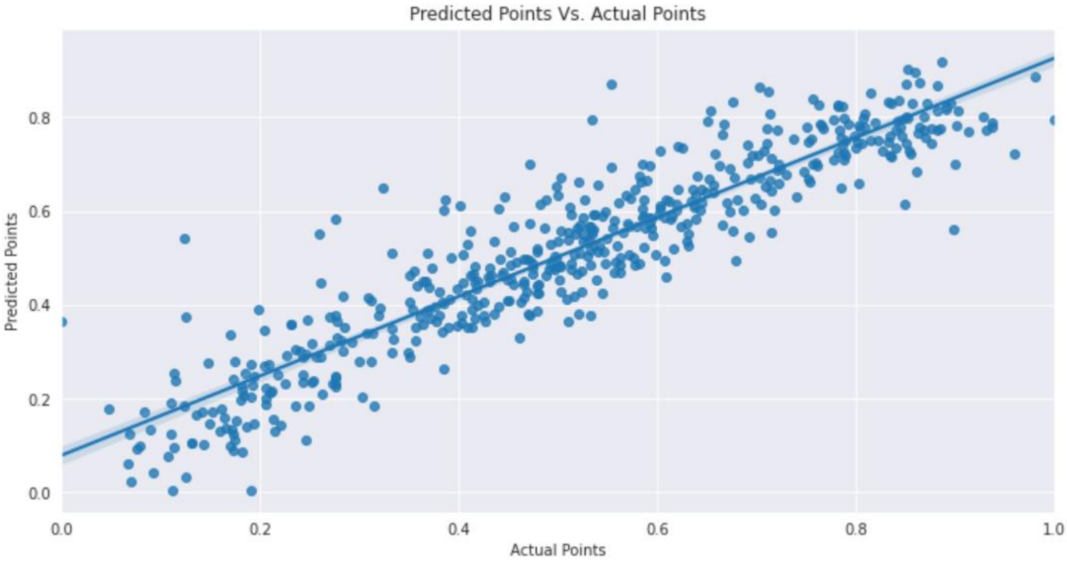
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable?

The variable "temp" shows the highest correlation with the target variable, standing at 0.63. We are not considering the correlation of "casual" and "registered" variables as they are direct components that sum up to the target variable. Furthermore, "atemp" is a derived parameter from "temp," "humidity," and "windspeed," and thus it is being excluded from consideration as it will be eliminated during model preparation.

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

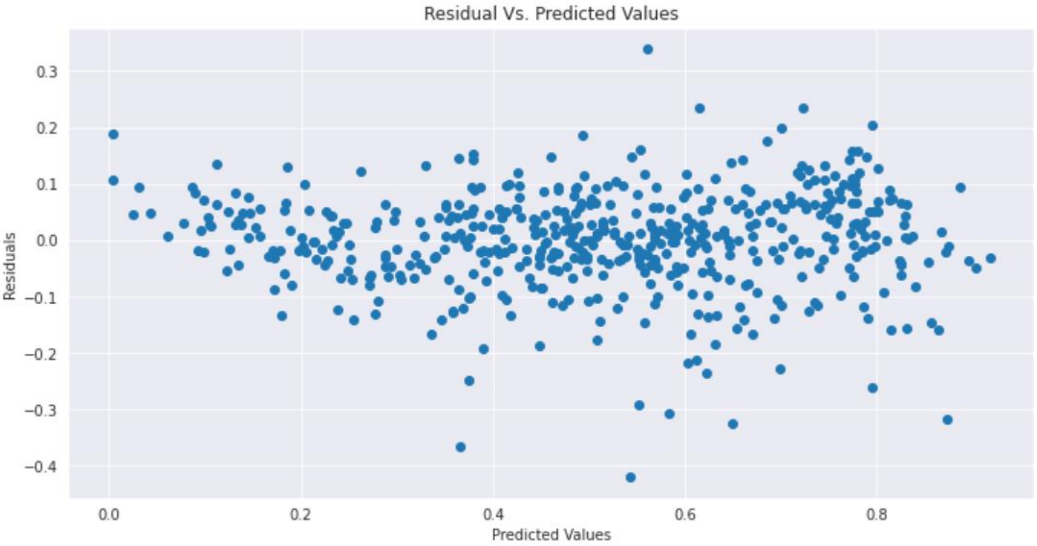
Linear Relationship Between Independent and Dependent Variables

Linearity is confirmed by observing the symmetrical distribution of points around the diagonal line in the actual vs. predicted plot .



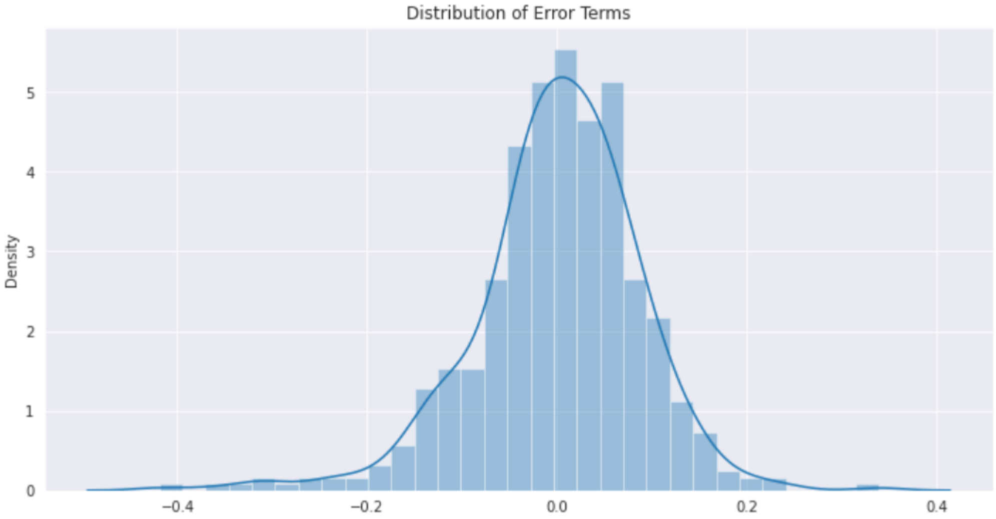
Independence of Error Terms

The absence of any discernible pattern in the error terms with respect to predictions indicates that the error terms are independent of each other.



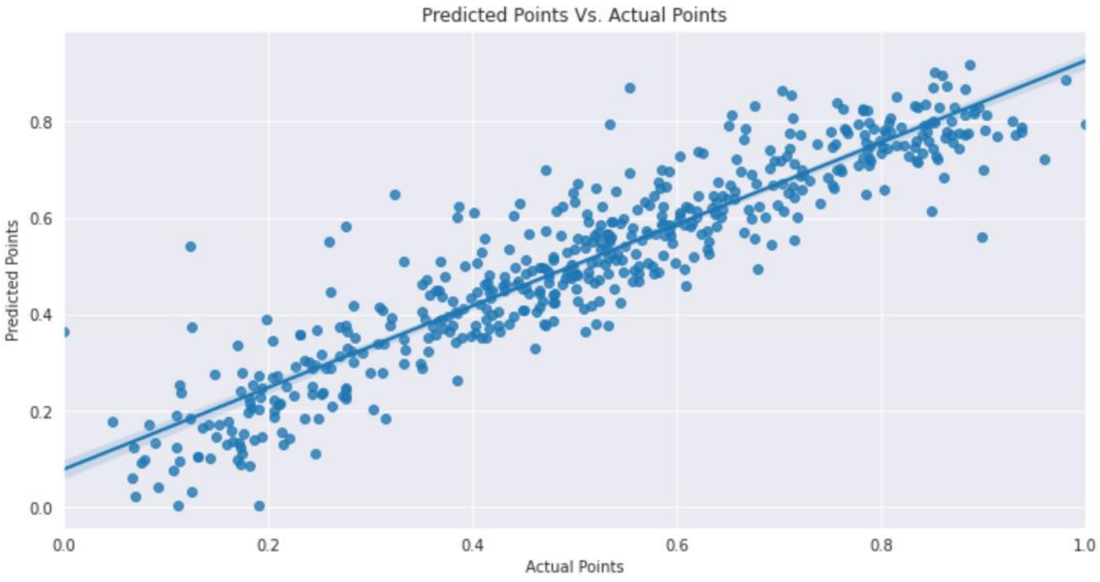
Normal Distribution of Error Terms

Histograms and distribution plots facilitate the assessment of the normal distribution of error terms, typically with a mean close to zero. This is clearly illustrated in the figure below.



Constant Variance of Error Terms (Homoscedasticity)

The error terms exhibit approximately constant variance, thereby satisfying 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 three influential variables are:

- **Weathersit:** Temperature significantly and positively impacts the business. Conversely, other environmental conditions such as rain, humidity, wind speed, and cloud cover negatively affect it.
- **Year ('yr'):** The year-on-year growth appears to be organic, aligning with the geographical attributes.
- **Season:** The winter season plays a crucial role in the demand for shared bikes.

General Subjective Questions

1. Explain the linear regression algorithm in detail.

Linear regression is a statistical method used to model the best linear relationship between independent and dependent variables. The algorithm identifies a "best-fitting" line to represent this association.

There are two primary types of linear regression algorithms:

- **Simple Linear Regression (SLR):** This method involves a single independent variable. The line equation for SLR is Y=β0+β1X.
- **Multiple Linear Regression (MLR):** This method utilizes multiple independent variables. The line equation for MLR is Y=β0+β1X1+⋯+βpXp+ϵ.
- In both equations, β0 represents the Y-intercept (the value of Y when X equals 0), and β1,β2 ,…,βp represent the slopes or gradients.

Cost functions are instrumental in determining the optimal values for β0,β1,β2,…,βp, which in turn enable accurate predictions of the target variable. The objective is to minimize the cost function to achieve the best-fitting line for predicting the dependent variable. There are two categories of cost function minimization approaches: unconstrained and constrained.

- The **Sum of Squared Errors (SSE)** function is commonly employed as a cost function to identify the best-fit line.
    - Given a straight-line equation $Y=\beta0+\beta1X$, the predicted Y value for a given xi is Ypred $=\beta0+\beta1xi$, and the actual Y value is Yi.
    - The cost function is typically represented as $J(\beta1,\beta0)=\sum(yi-\beta1xi-\beta0)2$.
- Unconstrained minimization problems can be solved using two main methods:
    - Closed-form solutions
    - Gradient descent

During the process of finding the best-fit line, discrepancies between the actual and predicted values, known as **residuals**, inevitably arise. To minimize the sum of squared errors, the **Ordinary Least Squares (OLS)** method is employed.

- The error for each data point is calculated as $ei=yi-Ypred$.
- OLS aims to minimize the total squared error, which is referred to as the **Residual Sum of Squares (RSS)**.
- $RSS=\sum i=1n(yi-Ypred)2$.

In essence, the **Ordinary Least Squares method is utilized to minimize the Residual Sum of Squares and estimate the beta coefficients.**

2.Explain the Anscombe's quartet in detail.

Anscombe's Quartet is a famous illustration in statistics, created in 1973 by statistician Francis Anscombe. It consists of four distinct datasets that, despite being visually very different when plotted, share nearly identical basic descriptive statistics. This quartet serves as a powerful demonstration of the crucial importance of data visualization before drawing conclusions or building models, and it highlights how summary statistics alone can be profoundly misleading.

The Four Datasets: Each of Anscombe's four datasets has nearly identical summary statistics (mean, variance for X & Y, correlation ≈ 0.816, and the same linear regression line $y=0.5x+3.0$). However, their scatter plots reveal starkly different patterns:

1. **Dataset I:** Classic linear relationship, ideal for linear regression.
2. **Dataset II:** Clearly non-linear (parabolic), making linear regression unsuitable.
3. **Dataset III:** Strong linear trend with one significant outlier skewing the regression line.
4. **Dataset IV:** Data mostly clustered vertically, with a single influential high-leverage point driving the entire correlation.

3. What is Pearson's R?

Pearson's R, also known as the **Pearson Product-Moment Correlation Coefficient (PPMCC)**, is a widely used statistical measure that quantifies the **strength and direction of a linear relationship between two continuous variables.**

Here's a breakdown of its key aspects:

1. **Range and Interpretation:**
   - **Value:** Pearson's R (denoted by 'r') ranges from **-1 to +1**.
   - **Direction:**
     - **+1:** Indicates a **perfect positive linear correlation**. As one variable increases, the other increases proportionally.
     - **-1:** Indicates a **perfect negative linear correlation**. As one variable increases, the other decreases proportionally.
     - **0:** Indicates **no linear relationship** between the two variables.
   - **Strength (Magnitude):** The absolute value of 'r' (how close it is to -1 or +1) indicates the strength of the linear relationship:
     - Values closer to ±1 indicate a **stronger** linear relationship.
     - Values closer to 0 indicate a **weaker** linear relationship.
     - Common guidelines for interpretation (though these can vary by discipline):
       - ±0.1 to ±0.3: Weak correlation
       - ±0.3 to ±0.5: Moderate correlation
       - ±0.5 to ±1.0: Strong correlation
2. **What it Measures:** Pearson's R specifically measures the **linear association**. It tells you how well a straight line can describe the relationship between two variables. It **does not imply causation**, meaning a high correlation doesn't mean one variable causes the other. It only suggests that they tend to move together in a predictable linear fashion.
3. **Assumptions:** For Pearson's R to be an appropriate and reliable measure, certain assumptions about the data should ideally be met:
   - **Linearity:** The relationship between the two variables should be linear.
   - **Continuous Variables:** Both variables should be continuous (interval or ratio scale).
   - **Normality:** The variables should ideally be approximately normally distributed (though Pearson's R can be somewhat robust to minor deviations).
   - **No Significant Outliers:** Outliers can heavily influence the coefficient

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

Scaling in machine learning is a crucial data preprocessing technique that transforms numerical features to a common range or scale. This is done to ensure that all features contribute fairly to a model, preventing those with inherently larger magnitudes (like a house's price) from disproportionately influencing the learning process compared to features with smaller magnitudes (like the number of rooms). Scaling is vital for several reasons: it significantly improves the performance and convergence speed of many algorithms, especially distance-based ones such as K-Nearest Neighbors and Support Vector Machines, and those that rely on gradient descent like neural networks and linear regression. By preventing features from dominating purely due to their size, scaling ensures more stable and robust model training.

While both **normalized scaling (Min-Max scaling)** and **standardized scaling (Z-score normalization)** aim to bring features to a comparable range, they achieve this differently. **Normalized**

scaling rescales features to a fixed, predefined range, typically between 0 and 1, using the formula Xnorm=Xmax−XminX−Xmin. This method is particularly useful when you need values within a strict boundary, but it can be quite sensitive to outliers, as extreme values will compress the rest of the data. In contrast, **standardized scaling** transforms features to have a mean of 0 and a standard deviation of 1, using the formula Xstd=σX−μ. This approach does not bound the values to a specific range, but it is less sensitive to outliers and is often preferred when the data is approximately normally distributed or when working with algorithms that assume such a distribution, like Principal Component Analysis or linear regression. The choice between normalization and standardization ultimately depends on the specific machine learning algorithm being used, the nature of your data, and the problem's unique requirements.

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

The Variance Inflation Factor (VIF) quantifies the severity of multicollinearity among independent variables in a regression model. The formula for VIF for a given independent variable Xj is:

VIFj=1−Rj21

where Rj2 is the R2 value obtained from regressing Xj on all the other independent variables in the model.

The value of VIF becomes infinite when Rj2 is exactly 1. This happens in scenarios of **perfect multicollinearity**, meaning that the independent variable Xj can be perfectly predicted by a linear combination of one or more of the other independent variables in the model. In such a situation, the denominator (1−Rj2) becomes 1−1=0, leading to division by zero and an infinite VIF. This indicates that the independent variable is completely redundant, as its information is entirely captured by other predictors, making it impossible for the model to uniquely estimate the regression coefficient for that variable.

6.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 that compares the quantiles of a dataset, typically residuals from a statistical model, against the quantiles of a theoretical distribution, most commonly the normal distribution. In linear regression, its primary use is to **visually assess the crucial assumption that the error terms (residuals) are normally distributed**. If the residuals follow a normal distribution, the points on the Q-Q plot will align closely with a straight diagonal line; significant deviations, such as S-shapes or points veering off at the ends, indicate departures from normality.

This check is of paramount importance because the **validity of statistical inferences** (like p-values and confidence intervals for regression coefficients) relies heavily on this normality assumption. If the assumption is violated, these inferences can be inaccurate, leading to potentially incorrect conclusions about the predictors' significance. Therefore, the Q-Q plot acts as a vital diagnostic tool, guiding analysts to consider data transformations, robust regression methods, or alternative modeling approaches if non-normality is detected, ensuring the reliability of the linear regression model's outputs.