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? (3 marks)

   - **Seasonal Demand**: Among the seasons, Spring is observed to have the lowest demand for bike sharing compared to other seasons.
   - **Weather Conditions**: The presence of light snow and rain leads to a noticeable decrease in the demand for bike sharing. Adverse weather conditions deter users from opting for bike sharing.
   - **Annual Trends**: The year 2019 witnessed a significant increase in the demand for bike sharing, indicating a growing popularity or expanded availability during this period.
   - **Impact of Holidays**: There is an observed increase in bike sharing demand during holidays. This could be attributed to more recreational and leisure activities during these times.
   - **Monthly Trends**: There is an increasing trend in the number of bikes rented from January to September. However, there is a noticeable dip in demand during the months of November and December, possibly due to colder weather conditions affecting user preference.
   - **Daily Variations**: The variation in the number of bikes rented across different days of the week is relatively minor, suggesting a consistent daily demand.
   - **Working vs. Non-Working Days**: On non-working days, the demand for bike sharing is higher, possibly due to more people engaging in recreational activities or using bike sharing for leisure trips rather than commuting.

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

   It helps in maintaining the model's robustness, interpretability, and computational efficiency by eliminating unnecessary redundancy in the feature set.

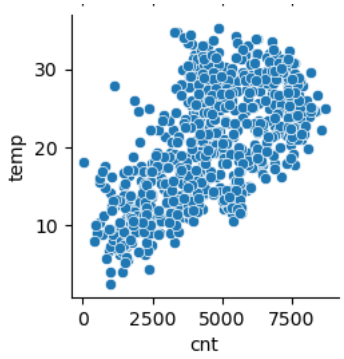
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

   Temperature has the highest correlation.

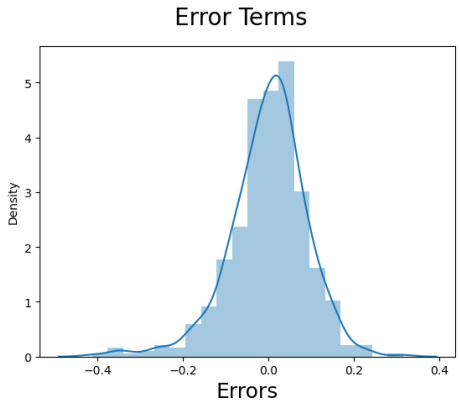
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

   The validation has been done by plotting various graphs for the assumptions:

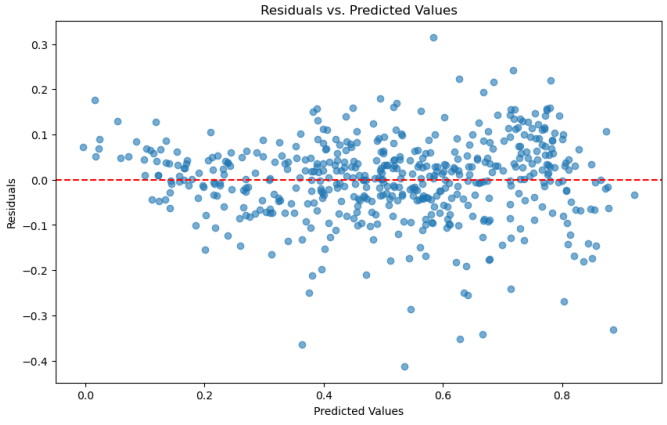
- Linear relationship between X and Y : Plotted scatter plot and the relationship appears to be linear.



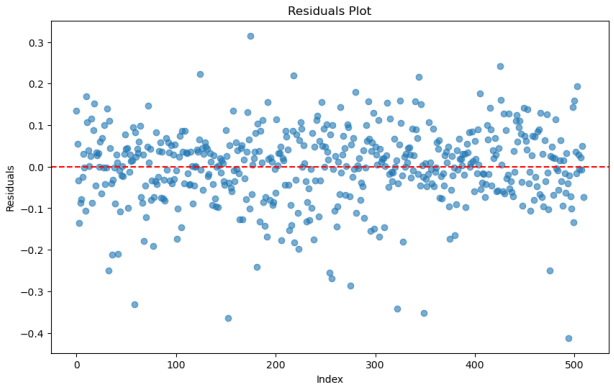
- Error terms are normally distributed: **Histogram** of residuals. Residuals form a bell-shaped curve, this suggests normality.



- Error terms have constant variance (homoscedasticity) : Scatter plot of the residuals against the predicted value. The points are evenly distributed around the horizontal line at zero, it indicates that the variance of the error terms is constant (homoscedasticity)



- Error terms are independent of each other: Scatter plot of the residuals from a regression model. Residuals appear randomly scattered around zero without forming any discernible pattern, this suggests that the error terms are independent.



5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

The demand of bikes were highly explained by features below :

- **temp (0.549936):** This coefficient is the largest positive value, indicating that temperature has a significant positive impact on bike demand. As the temperature increases, more people are likely to rent bikes.
- **Light Snow & Rain (-0.288021):** This has a significant negative impact on bike demand. The negative coefficient indicates that light snow and rain reduce the demand for bikes, which makes sense as less favorable weather conditions would discourage outdoor activities like biking.
- **windspeed (-0.1552389):** This variable also has a considerable negative impact on bike demand. Higher wind speeds likely make cycling less comfortable or safe, thereby reducing the demand.

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Linear regression is a statistical technique used to model the relationship between a dependent variable (target) and one or more independent variables (features). The algorithm fits a linear equation to the data, with the goal of predicting the target variable based on the input features.

- **Equation**: The linear regression equation is $y = b0 + b1x1 + b2x2 + \cdots + bnxn + e$, where:
  - y is the dependent variable (target),
  - x1,x2,… are independent variables (features),
  - b0 is the intercept,
  - b1,…,bn are the coefficients (slopes), and
  - e is the error term (difference between actual and predicted values).

- **Steps in Linear Regression**:

  1. Fitting the Model: The algorithm estimates the coefficients b0,b1,…,bn by

minimizing the sum of squared errors (SSE), which is the sum of squared differences between actual and predicted values.

2. Prediction: Once the coefficients are determined, predictions are made by plugging in the values of the independent variables into the equation.

3. Evaluation: Model performance is typically evaluated using metrics like R-squared, mean squared error (MSE), or root mean squared error (RMSE).

Linear regression assumes a linear relationship between the variables, normally distributed error terms, and independence of errors.

- **Assumptions of simple linear regression are** :

1. Linear relationship between X and Y
2. Error terms are normally distributed (not X, Y)
3. Error terms are independent of each other
4. Error terms have constant variance (homoscedasticity)

Also, there is no assumption on the distribution of Xand Y, just that the error terms have to have a normal distribution

2. Explain the Anscombe's quartet in detail. (3 marks)

Anscombe's Quartet is a collection of four different datasets that have nearly identical summary statistics (mean, variance, correlation, and linear regression line) but appear very different when visualized. It was created by statistician Francis Anscombe in 1973 to demonstrate the importance of data visualization in data analysis. All four datasets have the same mean, variance, and regression line, yet they exhibit different patterns when plotted.

1. One dataset follows a typical linear pattern.
2. Another dataset has a non-linear relationship.
3. A third dataset has an outlier that influences the regression line.
4. The fourth dataset has identical xxx-values and a vertical line.

The lesson is that relying solely on summary statistics can be misleading. Visualizing data helps identify outliers, non-linearity, or other important trends that summary statistics might overlook

3. What is Pearson's R? (3 marks)

Pearson's R, also known as the **Pearson correlation coefficient**, is a measure of the strength and direction of the linear relationship between two variables

- **Pearson Correlation Coefficient (r) Between 0 and 1 (**Positive Correlation**):**
  - As one variable increases, the other variable also increases, indicating a direct relationship.
  - **Example:** Baby Length & Weight: Longer babies tend to weigh more.

- **Pearson Correlation Coefficient (r) = 0 (**No Correlation**):**
  - No relationship exists between the variables; changes in one do not affect the

other.
    - **Example:** Car Price & Windshield Wiper Width: Car price is unrelated to wiper width.

- **Pearson Correlation Coefficient (r) Between 0 and -1 (**Negative Correlation)**:**
    - As one variable increases, the other variable decreases, indicating an inverse relationship.
    - **Example:** Elevation & Air Pressure: Higher elevation corresponds to lower air pressure.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Scaling is the process of transforming features to a specific range or distribution.

Scaling is Performed so as to:

- To ensure that all features contribute equally to the model.
- To avoid dominance of features with larger magnitudes.
- To improve the convergence of optimization algorithms.

Types of Scaling:

1. **Normalized Scaling**:
    - In normalized scaling, features are scaled to a range, typically [0, 1] or [-1, 1].
      Formula: X'= x -xmin/ xmax – xmin
2. **Standardized Scaling**:
    - Features are scaled to have a mean of 0 and a standard deviation of 1.
    - **Formula**: X'= x - xmean/ xstd

It is important to note that scaling just affects the coefficients and none of the other parameters like t-statistic, F statistic, p-values, R-square, etc.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

An infinite **Variance Inflation Factor (VIF)** occurs when there is **perfect multicollinearity** among the independent variables in the model. This means one or more variables are exact linear combinations of others.

Reasons for Infinite VIF:

- **Duplicate Variables**: Including the same variable multiple times.
- **Derived Variables**: When a variable is a mathematical combination of others (e.g., a sum or ratio of existing variables).
- **Perfect Correlation**: When two variables are perfectly correlated (correlation coefficient = 1 or -1).

An infinite VIF signals that the model cannot differentiate the impact of perfectly correlated variables, leading to computational issues.

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

(3 marks)

A **Q-Q Plot** (Quantile-Quantile plot) is a graphical tool to assess whether a set of data follows a particular distribution, typically the normal distribution. It plots the quantiles of the residuals against the quantiles of a normal distribution.

Use in Linear Regression:

In linear regression, one key assumption is that the residuals (errors) are normally distributed. A Q-Q plot helps to visually check this assumption.

How to Interpret:

- If the residuals follow a normal distribution, the points will approximately lie on the 45-degree line.
- If the points deviate significantly from the line (e.g., curve or form an S-shape), it indicates deviations from normality, such as skewness or heavy tails.

Importance:

- A normally distributed error term is crucial for making valid inferences from the model, such as hypothesis testing and confidence interval calculations.
- If the residuals are not normally distributed, it suggests potential model issues or the need for transformation.