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?

Based on the analysis of the categorical variables here are some findings

- Holidays seem to have lesser demand.
- Spring season has the least demand - it means the dependent variable is -ve affected by this. Fall season has the most demand followed by summer - dependent variable is +ve affected.
- Days of the week do not have too much correlation. The effect of these on dependent variable is very less.

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

When creating dummy variables from categorical features, drop_first=True is an important parameter in order to avoid multicollinearity in regression models and to maintain model interpretability.

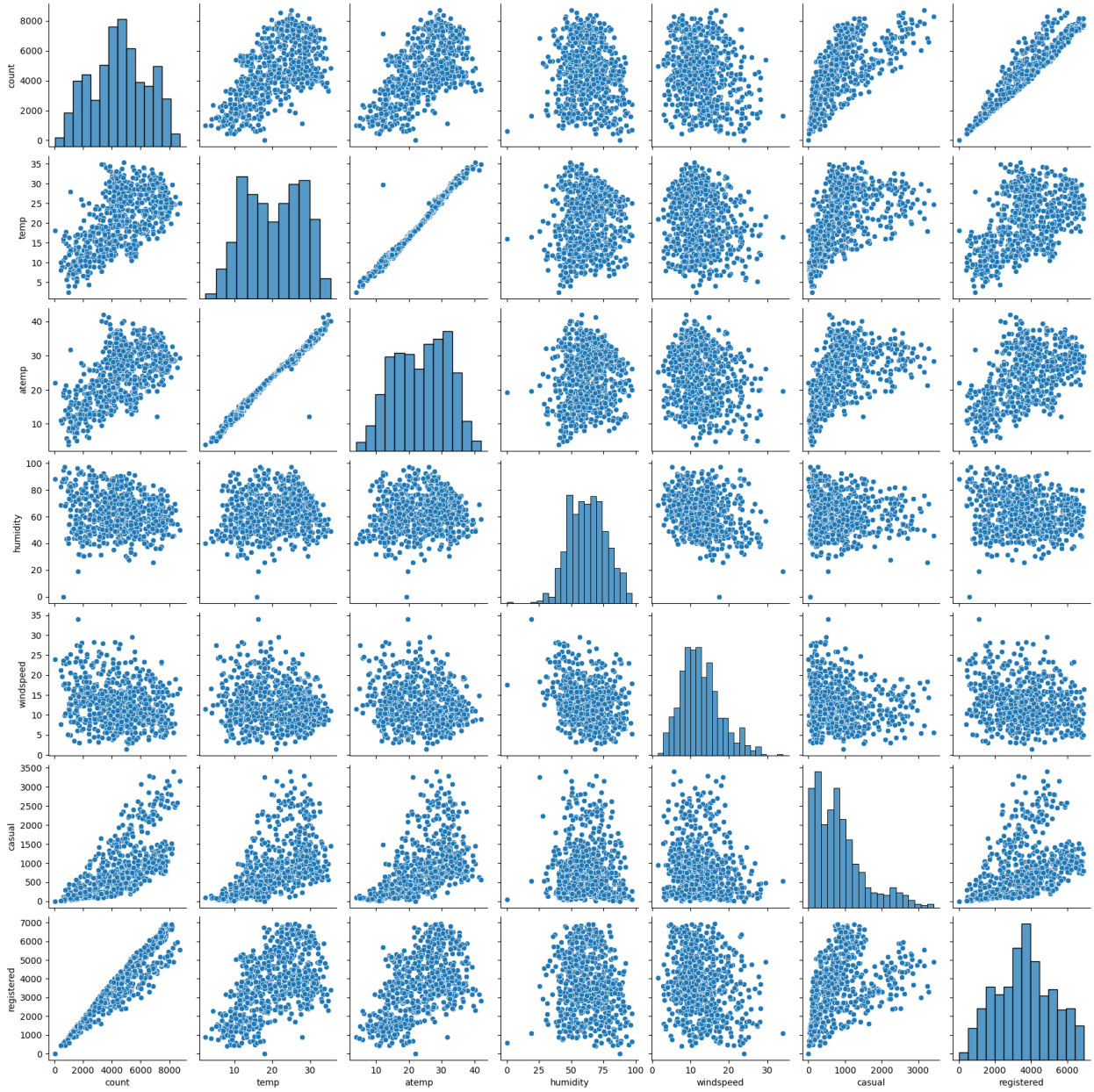
1. Multicollinearity:
    1.1. Multicollinearity occurs when predictor variables are highly correlated with each other.
    1.2. In dummy variable encoding, if you create n dummies for a categorical variable with n categories, you only need n−1 dummies to represent all the information about that categorical variable.
    1.3. By setting drop_first=True, one dummy variable is dropped, preventing perfect multicollinearity among the dummy variables.
    1.4. Perfect multicollinearity can adversely affect regression models as it hampers the matrix inversion process necessary for calculating coefficients.
2. Interpretability:
    2.1. Dropping one dummy variable retains interpretability. The dropped category becomes the reference category.

2.2. The coefficients of the remaining dummy variables represent the change in the dependent variable compared to the reference category.

2.3. This simplifies interpretation as it provides a clear comparison between each category and the reference category.

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

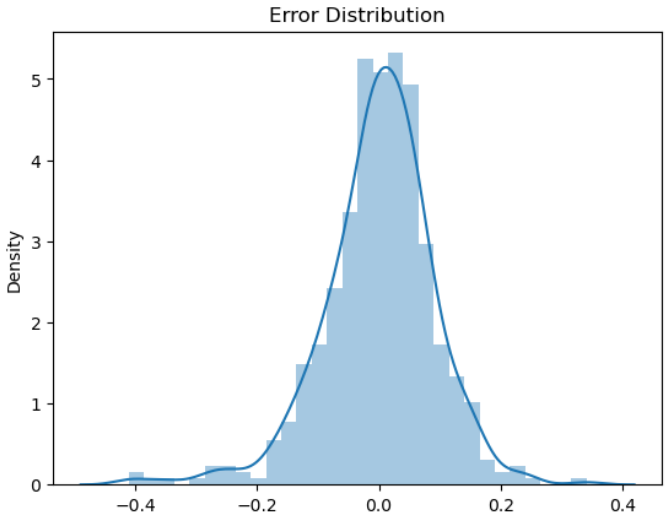
Here is the pair plot of the numerical variables.

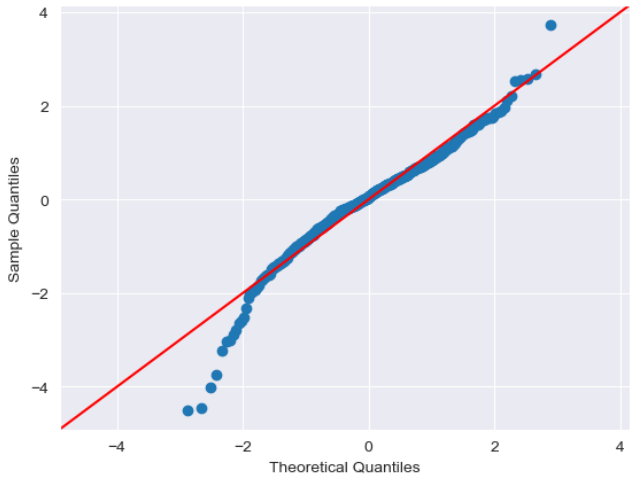


Looking at the plots, temperature looks to be correlated well with the count.

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

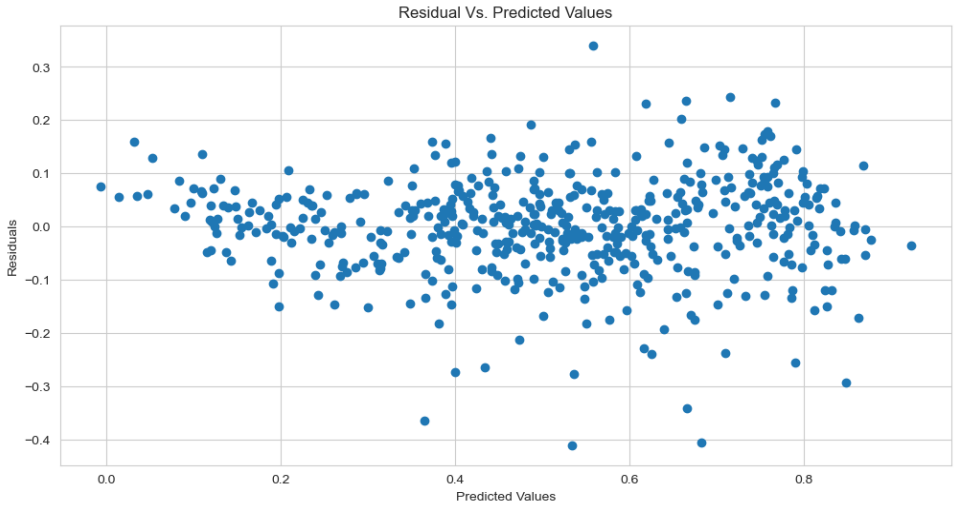
Residual analysis was done to validate the assumptions of linear regression after the model was built.





We checked for normality of distribution of residuals using the distribution plot and Q-Q plot. See above charts.

We also checked that there are no patterns in the residuals using a scatter plot. See below chart.



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

Based on the final model that we built, the top 3 features contributing significantly are:

Sl. No	Parameter	Correlation Direction
1.	temp	Positive correlation
2.	Light Snow	Negative correlation
3.	Year	Positive correlation

General Subjective Questions

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

Linear regression is a fundamental supervised learning algorithm used for modeling the relationship between a dependent variable (target) and one or more independent variables (features). It assumes a linear relationship between the predictors and the target variable. Here's a detailed breakdown:

Assumptions of Linear Regression:

1. Linearity: The relationship between the independent and dependent variables is linear.
2. Independence: Observations are independent of each other.
3. Homoscedasticity: Residuals (the differences between predicted and actual values) have constant variance across all levels of predictors.
4. Normality: Residuals are normally distributed.
5. No or Little Multicollinearity: The predictors are not highly correlated.

Components of Linear Regression

1. Model Representation:
   The model equation for simple linear regression with one independent variable is:

   $$y=\beta_0+\beta_1 x$$

   - (y) is the dependent variable.
   - (x) is the independent variable.
   - β_0 is the intercept (where the line crosses the y-axis).
   - β_1 is the slope (the change in (y) for a unit change in (x)).
2. Objective:
   Minimize the sum of squared differences between the observed and predicted values. This is known as Ordinary Least Squares (OLS) estimation.
3. Training:
   Using the given dataset, the algorithm estimates the coefficients (β_0) and (β_1) that best fit the data by minimizing the sum of squared errors.
4. Prediction:
   Once the model is trained, it can be used to predict the dependent variable ((y)) for new or unseen data by applying the learned coefficients.

Steps Involved:

1. Data Preprocessing:
   - Handling missing values.
   - Feature scaling (if necessary).
   - Splitting data into training and testing sets.
2. Model Fitting:
   - Computing the coefficients using the training data.
   - The algorithm finds the line that best fits the data by adjusting the intercept and slope.
3. Model Evaluation:
   - Assessing the model's performance using metrics like Mean Squared Error (MSE), R-squared, etc., on the test set.
   - Checking for assumptions' validity (linearity, homoscedasticity, etc.).
4. Prediction and Interpretation:
   - Making predictions on new data.

    - Interpreting the coefficients to understand the impact of predictors on the target variable.

Linear regression is versatile and serves as a basis for more complex models. Extensions like multiple linear regression handle multiple predictors.

2. Explain the Anscombe's quartet in detail

Anscombe's quartet is a famous example in statistics that consists of four datasets that have nearly identical simple descriptive statistics (e.g., mean, variance, correlation) but have vastly different distributions when graphed. This quartet was created by the statistician Francis Anscombe in 1973 to emphasize the importance of graphically visualizing data before analyzing it and to showcase the limitations of relying solely on summary statistics.

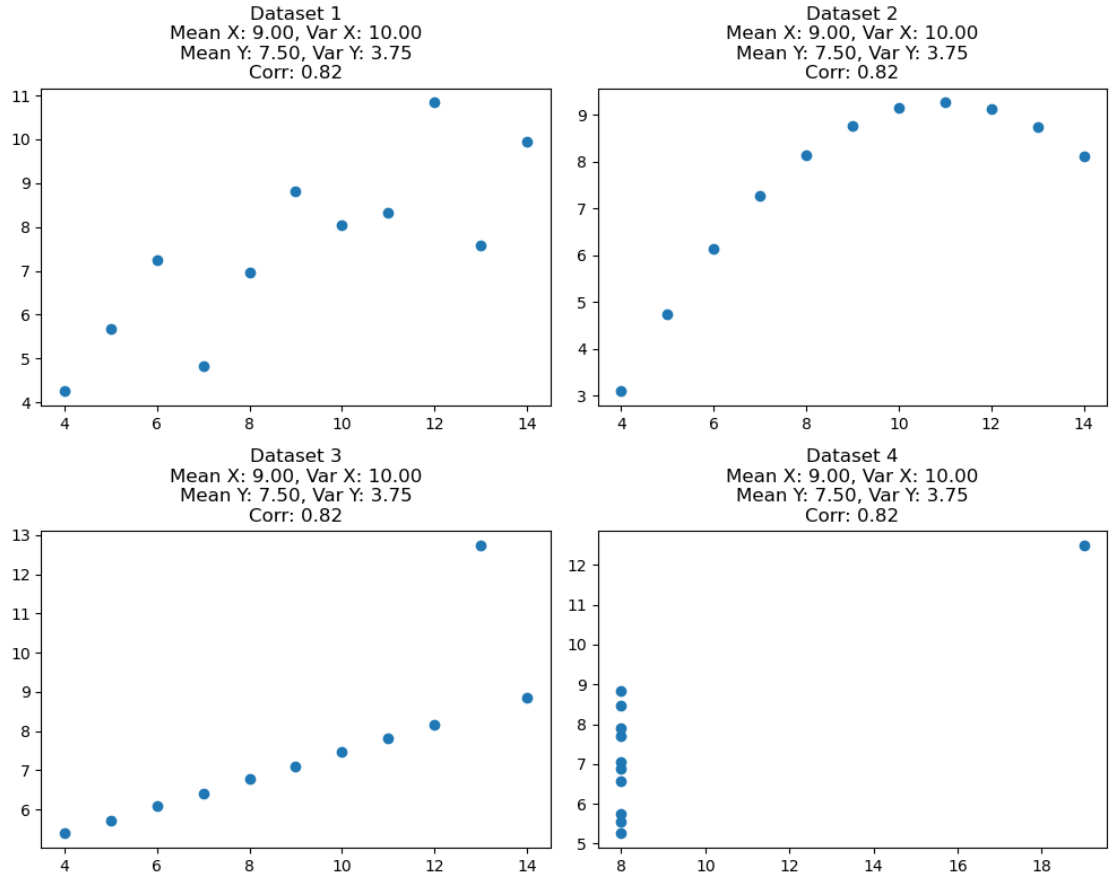
Characteristics of Anscombe's Quartet:

- Four Datasets
- Different Relationships:
    - Dataset I: Linear relationship.
    - Dataset II: Non-linear relationship.
    - Dataset III: Perfectly linear but with an outlier.
    - Dataset IV: No obvious relationship but heavily influenced by an outlier.
- Similar Summary Statistics:
    - Mean of x and y variables, variance, correlation coefficient, and regression line parameters (slope and intercept) are almost the same across all four datasets.

In practice, Anscombe's quartet serves as a cautionary tale against blindly relying on summary statistics and highlights the importance of data visualization to gain a deeper understanding of data distributions, relationships, and potential outliers. It emphasizes the need for exploratory data analysis and graphical tools to complement statistical analysis for more accurate insights. See table below for a sample Anscombe's quartet.

Q1		Q2		Q3		Q4	
X	Y	X	Y	X	Y	X	Y
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

As you can see from the graph below, the four data sets look very different in shape



3. What is Pearson's R?

Pearson's correlation coefficient, often denoted as *r*, is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. It assesses the degree to which two variables are linearly related to each other.

Range:

- Pearson's r ranges from -1 to +1.
- $r = 1$ indicates a perfect positive linear relationship (as one variable increases, the other variable also increases proportionally).
- $r = -1$ indicates a perfect negative linear relationship (as one variable increases, the other variable decreases proportionally).
- $r = 0$ suggests no linear relationship between the variables.

Strength of Relationship:

- The closer r is to +1 or -1, the stronger the linear relationship.
- Values near 0 suggest a weak or no linear relationship between the variables.

Direction of Relationship:

- If $r > 0$, it indicates a positive linear relationship (as one variable increases, the other tends to increase).
- If $r < 0$, it indicates a negative linear relationship (as one variable increases, the other tends to decrease).

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

Scaling is a data preprocessing technique used to transform the values of features or variables in a dataset to a similar scale. It is performed to ensure that the variables contribute equally and no one variable dominates analyses, especially in gradient descent-based algorithms (e.g. linear regression). It also helps in the interpretation of the beta-coefficients since they will be comparable.

There are two types of scaling

Normalization (Min-Max Scaling):

- Range: Scales the values to a range between 0 and 1.
- All values are linearly transformed to fit within the defined range.

Standardization (Z-score Scaling):

- Mean & Standard Deviation: Scales the values to have a mean of 0 and standard deviation of 1.
- Preserves the shape of the original distribution but centers the data around 0 with a standard deviation of 1.

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

A variance inflation factor (VIF) is a measure of the amount of multicollinearity in regression analysis. Multicollinearity exists when there is a correlation between multiple independent variables in a multiple regression model.

VIF (Variance Inflation Factor) value being infinite is often observed when there's perfect multicollinearity in the dataset. When one predictor variable in a regression model can be exactly predicted by a linear combination of other predictor variables, it results in perfect multicollinearity. This situation causes the determinant of the matrix in the VIF calculation to be zero, resulting in an infinite VIF value.

$$VIF_i = \frac{1}{1 - R_i^2}$$

If the r-squared in the above equation is 1, then the value of VIF for that variable will be infinity. The interpretation of this is that the ith variable is perfectly correlated to another variable and hence can be predicted by that variable.

Handling multicollinearity is crucial in regression analysis to ensure the stability and reliability of the model. Identifying and resolving perfect multicollinearity issues are essential steps to avoid infinite VIF values and obtain accurate model estimates

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 used to assess whether a given dataset follows a particular probability distribution, such as the normal distribution. It compares the quantiles of the dataset against the quantiles of a specified theoretical distribution, allowing us to visually inspect if the data aligns with the assumed distribution.

Use and Importance in Linear Regression:

Normality Assumption Check: Q-Q plots help assess the assumption in linear regression (that the residuals (the differences between observed and predicted values) follow a normal distribution) by plotting the quantiles of the residuals against the quantiles of a normal distribution. Deviations from a straight line indicate departures from normality.

How to Interpret a Q-Q Plot:

If the plotted points fall close to a straight line, it indicates that the residuals follow a normal distribution, supporting the assumption of normality.

Additionally, outliers might appear as points significantly far away from the line.