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)

The categorical variables in the dataset included seasons, weather, working day, month, year, and weekday.

After building the model and analyzing the predictors, the following insights were observed:

Season: Summer and winter seasons were associated with higher bike demand compared to spring and fall months.

Weather: Clear skies correlated with increased bike demand, while misty or cloudy conditions showed a negative relationship.

Month: The months of August, September, and October were preferred, whereas peak summer and winter months like January and July were less favored.

Holiday: Demand patterns during holidays and non-holidays were similar, but the average demand was higher when there were no holidays.

Weekday: Demand remained relatively consistent throughout the week, except for Sundays, which experienced a slight increase in demand.

Year: The year 2018 exhibited higher demand compared to the previous year.

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)

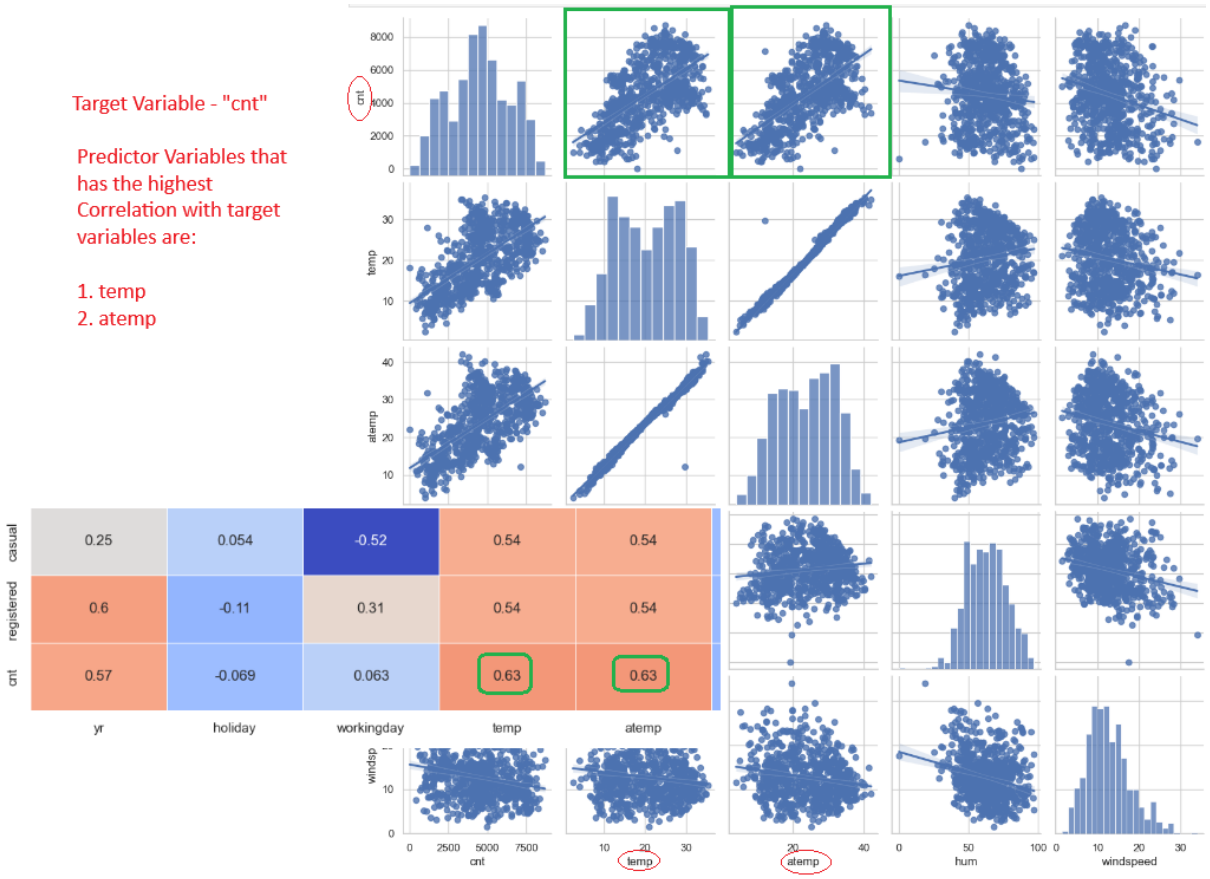
- Prevents multicollinearity by eliminating perfect correlation among dummy variables.
- Reduces the number of dummy variables from n to n−1, maintaining linear independence.
- Simplifies the regression model by reducing the complexity associated with redundant variables.
- Improves interpretability of the model by focusing on the impact of the remaining categories.
- Keeps all necessary information about the categorical variable while eliminating redundancy.

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)

- The pair plot from my analysis is displayed below.
- In this instance, the variables **temp** and **atemp** exhibit a strong correlation, with the heatmap indicating a correlation coefficient of **0.63**.
- Following the Recursive Feature Elimination (RFE) process, **atemp** was removed, and subsequent predictions were carried out using only temp.

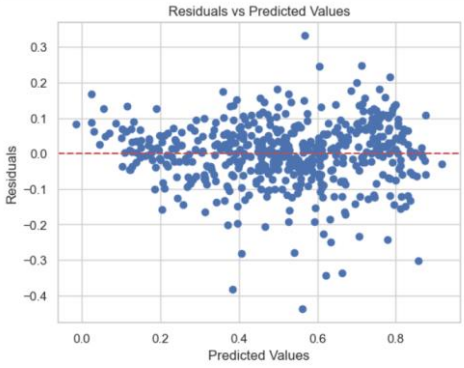


Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: <Your answer for Question 4 goes below this line> (Do not edit)

After performing below tests, I have confirmed on the assumptions of Linear Regression.

1. **<u>Linearity</u>**:

The scatter plot below, which depicts the relationship between the residuals and the predicted values (y_train_pred), indicates linearity, as the points appear to be randomly dispersed.



2. **Independence**:

Durbin-Watson Test: This test checks for autocorrelation in the residuals. A value close to 2 indicates no autocorrelation.

Durbin-Watson of the model is 2.0317

Durbin-Watson Test

```
import statsmodels.api as sm
import statsmodels.stats.stattools as stattools

# Calculate the Durbin-Watson statistic
dw_stat = stattools.durbin_watson(lm1.resid)

print("Durbin-Watson statistic:", dw_stat)
```

```
Durbin-Watson statistic: 2.031745578116921
```

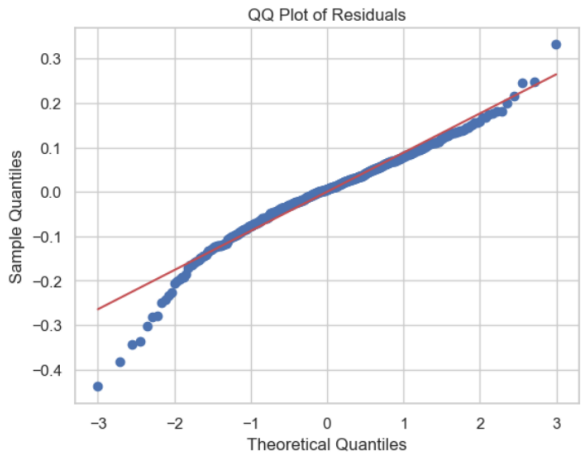
3. **Homoscedasticity**:

Residuals vs. Fitted Values:

Look for constant variance. If the spread of residual is consistent across all levels of the fitted values, homoscedasticity is satisfied.

4. **Normality of Residuals**:

Q-Q Plot: Plot the quantiles of the residuals against the quantiles of a normal distribution. If the points lie on a straight line, the residuals are normally distributed.



5. **No Multicollinearity:**

Variance Inflation Factor (VIF): Calculate VIF for each predictor. A VIF value greater than 10 indicates high multicollinearity.

6. Check correlation matrix, there should not be any highly correlated variables.

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)

- There is an increase in demand for bike rentals in 2019 compared to 2018.
- Rainy and cloudy weather leads to a decrease in bike demand.
- Temperature plays a crucial role, where low temperatures tend to reduce rentals, while higher temperatures lead to increased demand.

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 is a statistical method used to understand the relationship between two or more variables.
- It helps us predict the value of one variable based on the value of another.
- Independent Variable (X): This is the variable we use to make predictions.
- Dependent Variable (Y): This is the variable we want to predict.
- The goal of linear regression is to find the best-fitting straight line (called a regression line) that represents the relationship between the independent and dependent variables.

Types of Linear Regression:
Simple Linear Regression: Involves a single predictor variable. The model is:

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

Multiple Linear Regression: Involves multiple predictor variables. The model is as shown below.

$$y=\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_n x_n+\epsilon$$

- Where the β are coefficient and x are the independent variable.
- Y is the dependent variable.
- ϵ is the residual term.

How Linear Regression Works:

Assumptions: Linear regression makes several key assumptions:
Linearity: The relationship between the predictors and the target is linear.
Independence: Observations are independent of each other.
Homoscedasticity: Constant variance of the errors.
Normality: The errors are normally distributed.

Fitting the Model: The goal is to find the best-fitting line through the data points. This is done by minimizing the sum of the squared differences between the observed values and the values predicted by the model (least squares method).

Coefficient Estimation: The coefficients β are estimated using the least squares method, which

- minimizes the sum of the squared residuals:

$$\text{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Where:
yi = observed value (actual value)
yi^ = predicted value (from the model)
n = number of observations

Evaluating the Model

R-squared (R2):
Measures the proportion of the variance in the dependent variable that is predictable from the independent variables. It ranges from 0 to 1.

Adjusted R-squared:
Adjusts the (R2) value based on the number of predictors in the model, providing a more accurate measure when multiple predictors are used.

Mean Squared Error (MSE):
The average of the squared differences between the observed and predicted values.

Limitations of Linear Regression

Only applicable to linear model, which cannot be true sometimes in reality.
Overfitting, multicollinearity can affect the model.
Outliers can after the model.

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)

<Your answer for Question 7 goes here>

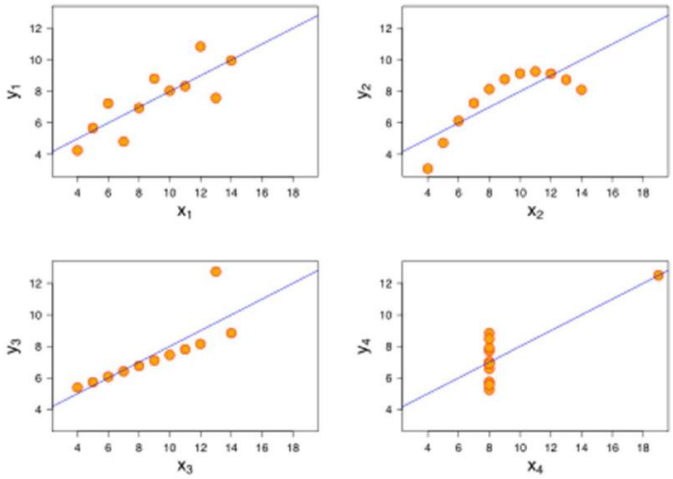
- **Anscombe's Quartet** is a collection of four sets of data that look very different when you plot them, but all have the same basic statistics (like average and correlation).
- It teaches us that looking at the actual data through graphs is really important. Just because numbers look the same doesn't mean the data behaves the same way.
- If we only look at numbers (like averages and correlations), we might think the data is telling us one thing, but the graphs can show a completely different story.
- So, basically reminds us to always visualize our data because it can reveal important details that numbers alone might hide!

Here are the key points: Each dataset consists of 11 pairs of (x,y) values.
The four datasets are:

Dataset I: A linear relationship with a positive correlation.
Dataset II: A nonlinear relationship with a parabolic shape.
Dataset III: A linear relationship with an outlier that heavily influences the slope.
Dataset IV: A vertical line indicating a constant xxx value with varying y values, showing no correlation.

Anscombe's quartet

Dataset I		Dataset II		Dataset III		Dataset IV	
x	*y*	*x*	*y*	*x*	*y*	*x*	*y*
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89



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)

<Your answer for Question 8 goes here>

- Pearson's R is also known as Pearson correlation coefficient.
- It is a statistical measure that evaluates the strength and direction of a linear relationship between two continuous variables.

Pearson's R ranges from -1 to 1.

- -1 indicates a perfect negative linear relationship.
- 0 indicates no linear relationship.
- 1 indicates a perfect positive linear relationship.

Values close to 1 or -1 signify a strong relationship, while values close to 0 suggest a weak relationship.

Pearson correlation coefficient (*r*) value	Strength	Direction
Greater than .5	Strong	Positive
Between .3 and .5	Moderate	Positive
Between 0 and .3	Weak	Positive
0	None	None
Between 0 and –.3	Weak	Negative
Between –.3 and –.5	Moderate	Negative
Less than –.5	Strong	Negative

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)

<Your answer for Question 9 goes here>

- If there is one feature that has huge values (like income in thousands) compared to another feature that has smaller values (like age), it can dominate the model's predictions.
- Scaling is a way of adjusting data values so they're all on a similar level.
- It's like making everything fit within a certain range, so that no single feature overpowers others just because of its size.

In Linear Regression, we have 2 types of Scaling:

1. Standardized Scaling
    - Centers values around 0, with most values between -1 and 1.
    - Best when data roughly follows a bell-shaped, normal distribution, which helps many models make accurate predictions.
    - The formula for standardization is: **$x' = x - \mu / \sigma$**

    Where:
    - (x) is the original value.
    - (μ) is the mean of the feature.
    - (σ) is the standard deviation of the feature.
    - (x') is the standardized value.

When to Use: Standardization is useful when the data follows a Gaussian distribution and you want to compare features that have different units or scales.

2. Normalized Scaling / MinMax Scaler
    - Adjusts values to be between 0 and 1.
    - Useful when we don't assume the data follows a particular pattern, and just need everything in a small, consistent range.
    - The formula for normalization is: **$x' = (x - \min(x)) / (\max(x) - \min(x))$**

    Where:
    - (x) is the original value.
    - (min(x)) is the minimum value of the feature.
    - (max(x)) is the maximum value of the feature.
    - (x') is the normalized value.

Normalization is useful when you know that the distribution of your data does not follow a Gaussian distribution and you want to bound the values within a specific range.

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)

<Your answer for Question 10 goes here>

From my observation, I can say that there are 3 reasons why we see VIF is Infinite sometimes. They are:

1. Perfect Multicollinearity
    - When one independent variable is an exact linear combination of others, it will lead to perfect multicollinearity Issue.
    - Example: Age, Age (In Months). Both are same. Just 2 different variables exist.
2. Impact of absence of Constant (Intercept) term
    - While building a model and fitting it, if we are using statsmodel package, we need to separately add constant to it, as it won't come by default.
    - If we miss that step, some variables may become linearly dependent, causing high or infinite VIF values.
    - Regression through the origin can distort relationships between variables causing very high VIF values.
3. Mistake during Dummy Variable Creation
    - While creating dummies for categorical columns, dropping one column along with removing original column is mandatory.
    - No. of dummies should always be 1 less than the Unique values of that categorical variable.
    - Example: Encoding a "Region" variable into dummy variables for "North," "South," "East," and "West" without dropping one category.

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)

<Your answer for Question 11 goes here>

- In the term Q-Q Plot, Q-Q Stands for Quantile-Quantile.
- It is a graph that helps us see if data follows a specific distribution, usually a normal distribution.
- In Linear Regression, Q-Q Plot is used in Residual Analysis. It helps to check if the Residuals are Normally distributed.
- We will plot the quantiles of the residuals against the quantiles of a normal distribution.
- If the points mostly line up in a straight line, it indicates the data is normally distributed.
- Deviations from a straight line suggest the data might have issues like skewness, which

could affect the accuracy of the regression model.

How a Q-Q Plot Works:

1. **Quantiles**: Quantiles are points taken at regular intervals from the cumulative distribution function (CDF) of a random variable. For example, the median is the 0.5 quantile.
2. **Plotting**: In a Q-Q plot, the quantiles of the sample data are plotted against the quantiles of the theoretical distribution. If the data follows the theoretical distribution, the points should lie approximately on a straight line

Intercepting a Q-Q Plot:

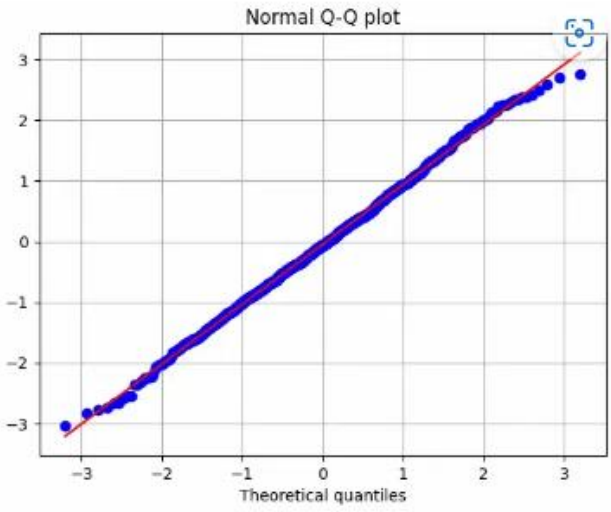
Straight Line: If the points lie on or near the straight line, the residuals are approximately normally distributed.

S-shaped Curve: Indicates heavy tails

Inverted S-shaped Curve: Indicates light tails

Deviations at Ends: Suggest skewness in the data.

Sample Q-Q Plot:



Q-Q plot