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 data set were: seasons, weather, working day, month, year, weekday. After the model building and analyzing the predictors below are the findings.

Season: Summer and Winter were favorable for more bike demand compared to other months like spring and Fall.

Weather: Clear sky shows high demand of bike, misty cloudy weather shows a negative correlation.

Month: Preferred months are Aug, Sep and Oct. Peak summer and winter months like Jan, July are not preferred.

Holiday: Holidays with and without shows a similar peak, but the mean is more when there is no holiday.

Weekday: The mean is almost similar over the week expect Sunday, the demand goes little higher on the Sundays compared to other days.

Year: 2018 shows higher demand compared to previous month.

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)

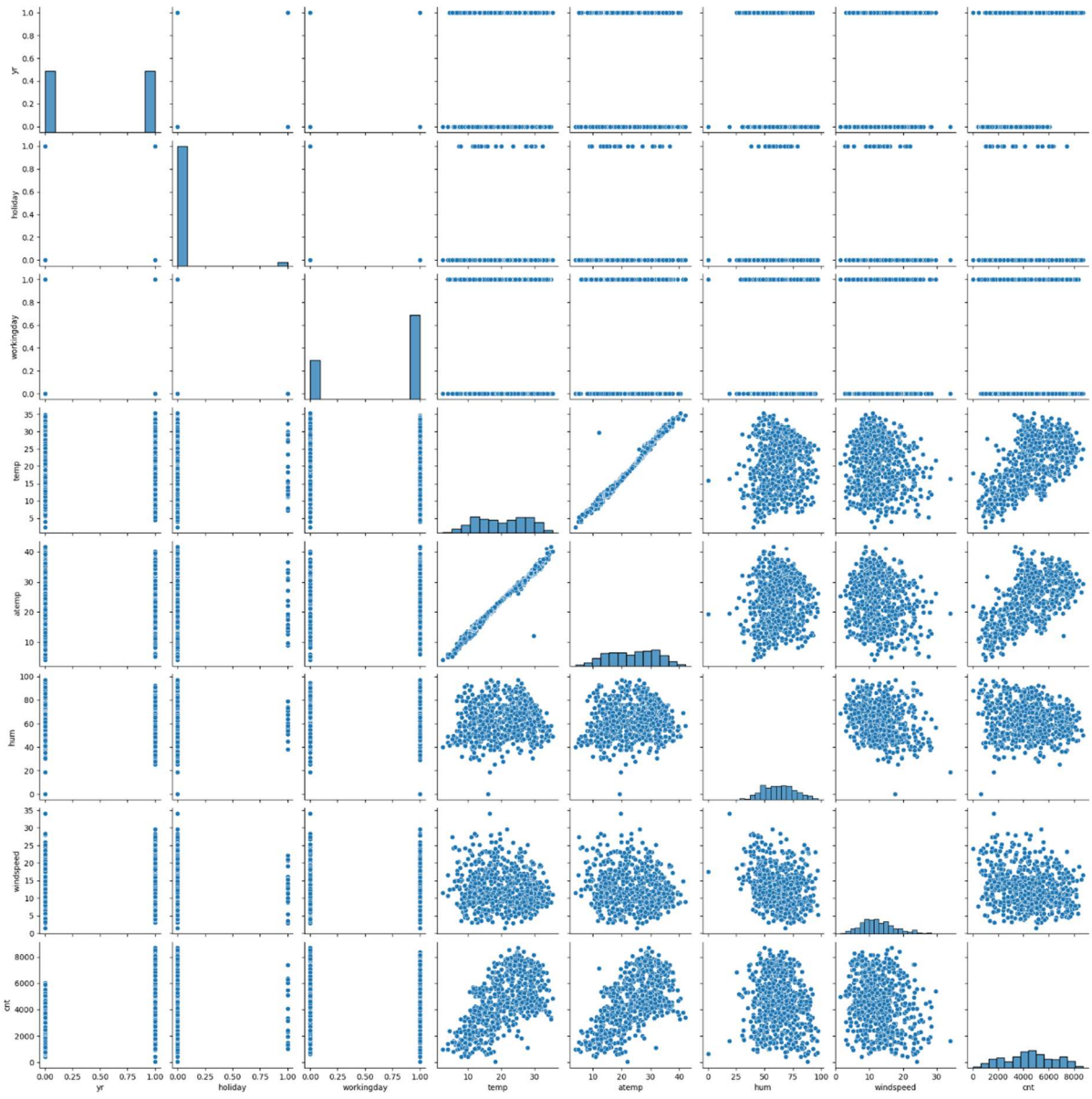
drop_first=True during dummy variable creation because the same information can be retained in one less column. Say we have three values for a categorical value, Mon Tue Wed. When both Tue and Wed is 0, 0 it means it is Monday. This is used to reduce the multicollinearity as the same information is already available using other predictors.

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 pairplot in my analysis was below. In this case two variables temp and atemp shows high correlation also the heatmap shows a value for 0.63. After RFE the atemp was dropped and further prediction was done with only temp.



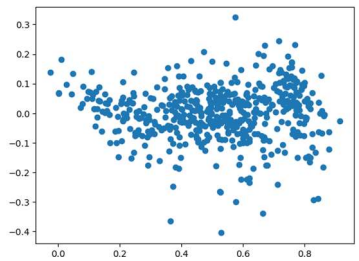
Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: <Your answer for Question 4 goes below this line> (Do not edit)

Below checks were performed.

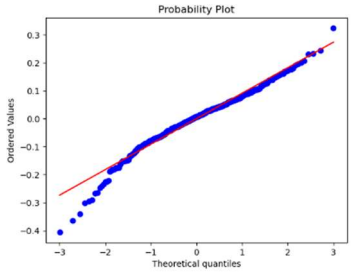
1. Linearity:
    - Plot Residuals vs. predicted Values: Check for any patterns. A random scatter suggests linearity.



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 2.065

3. Homoscedasticity:
    - Plot Residuals vs. Fitted Values: Look for constant variance. If the spread of residuals 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)

Rainy cloudy weather reduces the bike demand.
Increase in demand in 2019 compared to 2018
Temperature is important factor, low temp reduces the rental, high temp increases it.

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)

Definition

Linear regression is a supervised learning algorithm used for predicting a continuous target variable based on one or more predictor variables. It assumes a linear relationship between the input variables (features) and the single output variable (target).

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:

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

Evaluating the Model

R-squared (R^2): 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 (R^2) 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.

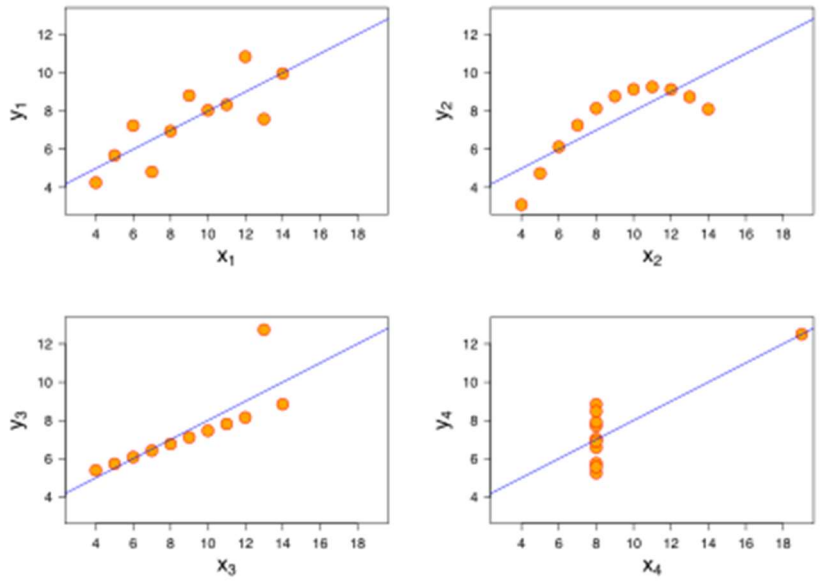
Question 7. Explain the Anscombe's quartet in detail. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Anscombe's quartet is a set of four datasets, which has same statistical properties but when visualized it shows different visualization.

The datasets were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and to show that summary statistics alone can be misleading.

The example was a 4 dataset with each 11 point (x,y). When plotted, each dataset seems to have a unique connection between x and y, with unique variability patterns and distinctive correlation strengths. Despite these variations, each dataset has the same summary statistics, such as the same x and y mean and variance, x and y correlation coefficient, and linear regression line.

Anscombe's quartet is used to illustrate the importance of exploratory data analysis and the drawbacks of depending only on summary statistics. It also emphasizes the importance of using data visualization to spot trends, outliers, and other crucial details that might not be obvious from summary statistics alone. An example data set.



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)

The Pearson correlation coefficient (*r*) is the most common way of measuring a linear correlation. It is a number between –1 and 1 that measures the strength and direction of the relationship between two variables.

When to use the Pearson correlation coefficient

The Pearson correlation coefficient (r) is one of several correlation coefficients that you need to choose between when you want to measure a correlation. The Pearson correlation coefficient is a good choice when all of the following are true:

Both variables are quantitative: You will need to use a different method if either of the variables is qualitative.

The variables are normally distributed: You can create a histogram of each variable to verify whether the distributions are approximately normal. It's not a problem if the variables are a little non-normal.

The data have no outliers: Outliers are observations that don't follow the same patterns as the rest of the data. A scatterplot is one way to check for outliers—look for points that are far away from the others.

The relationship is linear: "Linear" means that the relationship between the two variables can be described reasonably well by a straight line. You can use a scatterplot to check whether the relationship between two variables is linear

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)

Scaling is the process of transforming the features of your data so that they are on a similar scale. This is particularly important in machine learning algorithms that are sensitive to the scale of the data, such as those that rely on distance calculations.

Reasons for performing scaling:

1. **Improves Model Performance**: Many machine learning algorithms perform better when the features are on a similar scale. This is because features with larger ranges can dominate the learning process, leading to biased models.
2. **Speeds Up Convergence**: In optimization algorithms like gradient descent, scaling can help speed up

convergence by ensuring that the steps taken towards the minimum are more uniform.

3. **Prevents Numerical Instability**: Scaling can help prevent numerical instability in algorithms that involve matrix operations, where large differences in feature scales can lead to large condition numbers and unstable solutions.

Types of Scaling

Normalized Scaling

Normalization (also known as min-max scaling) rescales the features to a fixed range, usually [0, 1] or [-1, 1]. 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.

When to Use: 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.

Standardized Scaling

Standardization (also known as z-score normalization) rescales the features so that they have the properties of a standard normal distribution with a mean of 0 and a standard deviation of 1. 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.

Normalization	Standardization
This method scales the model using minimum and maximum values.	This method scales the model using the mean and standard deviation.
When features are on various scales, it is functional.	When a variable's mean and standard deviation are both set to 0, it is beneficial.
Values on the scale fall between [0, 1] and [-1, 1].	Values on a scale are not constrained to a particular range.
Additionally known as scaling normalization.	This process is called Z-score normalization.
When the feature distribution is unclear, it is helpful.	When the feature distribution is consistent, it is helpful.

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)

An infinite VIF value indicates perfect multicollinearity, meaning that one predictor variable is a perfect linear combination of one or more other predictor variables. This situation arises when there is exact redundancy among the predictors.
Causes of Infinite VIF

1. Perfect Collinearity: If one predictor can be perfectly predicted from the others, the denominator in the VIF calculation (which involves the R-squared value of the regression of one predictor on the others) becomes zero, leading to an infinite VIF.
2. Dummy Variable Trap: In categorical variables, if dummy variables are not correctly coded (e.g., including all categories without dropping one), it can lead to perfect multicollinearity.
3. Duplicate Variables: Including the same variable more than once in the model, either directly or through transformations that do not add new information, can cause infinite VIF.

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)

A Q-Q plot, or quantile-quantile plot, is a graphical tool used to assess whether a dataset follows a particular theoretical distribution, most commonly the normal distribution. It compares the quantiles of the dataset with the quantiles of the theoretical distribution.
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

Interpreting 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.