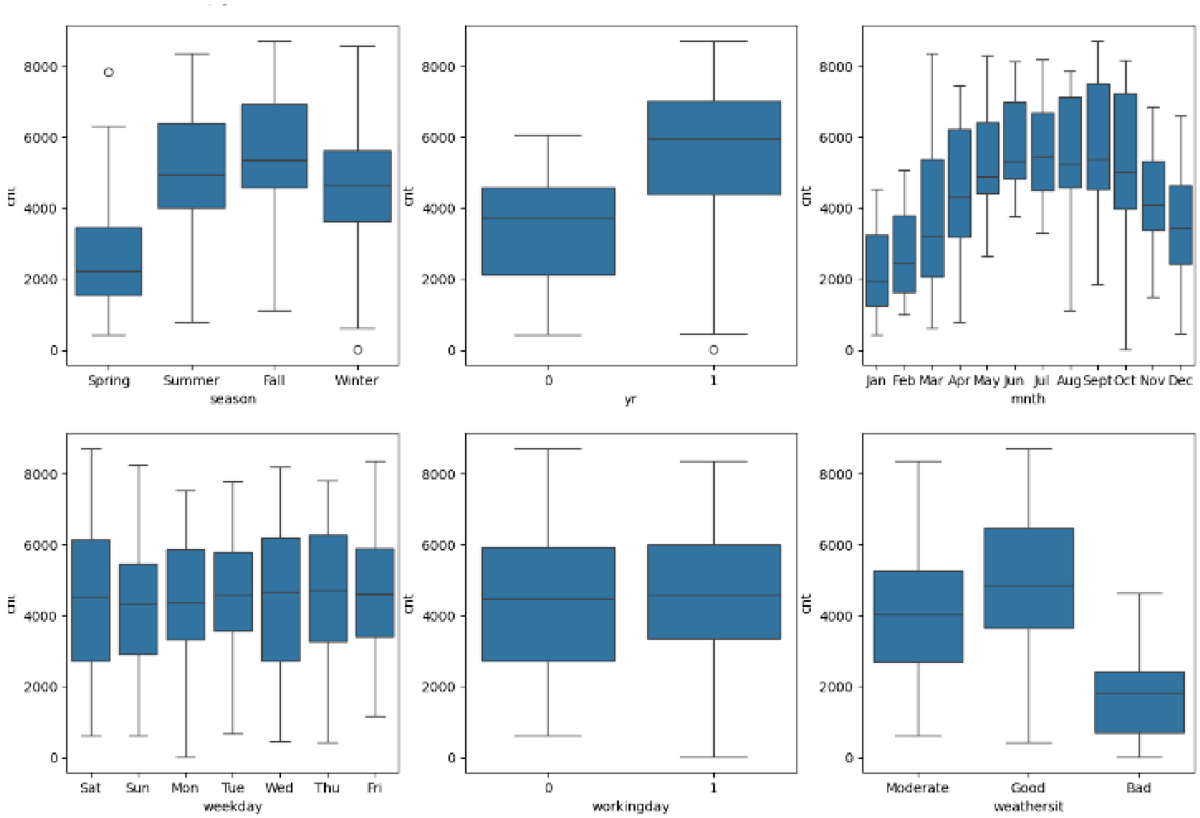
Assignment-based Subjective Questions

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)

Season, month, year, week, working day, and weather are some of the categorical factors. The dependent variable “cnt” is significantly impacted by these category variables. The association between the same is displayed in the figure below.



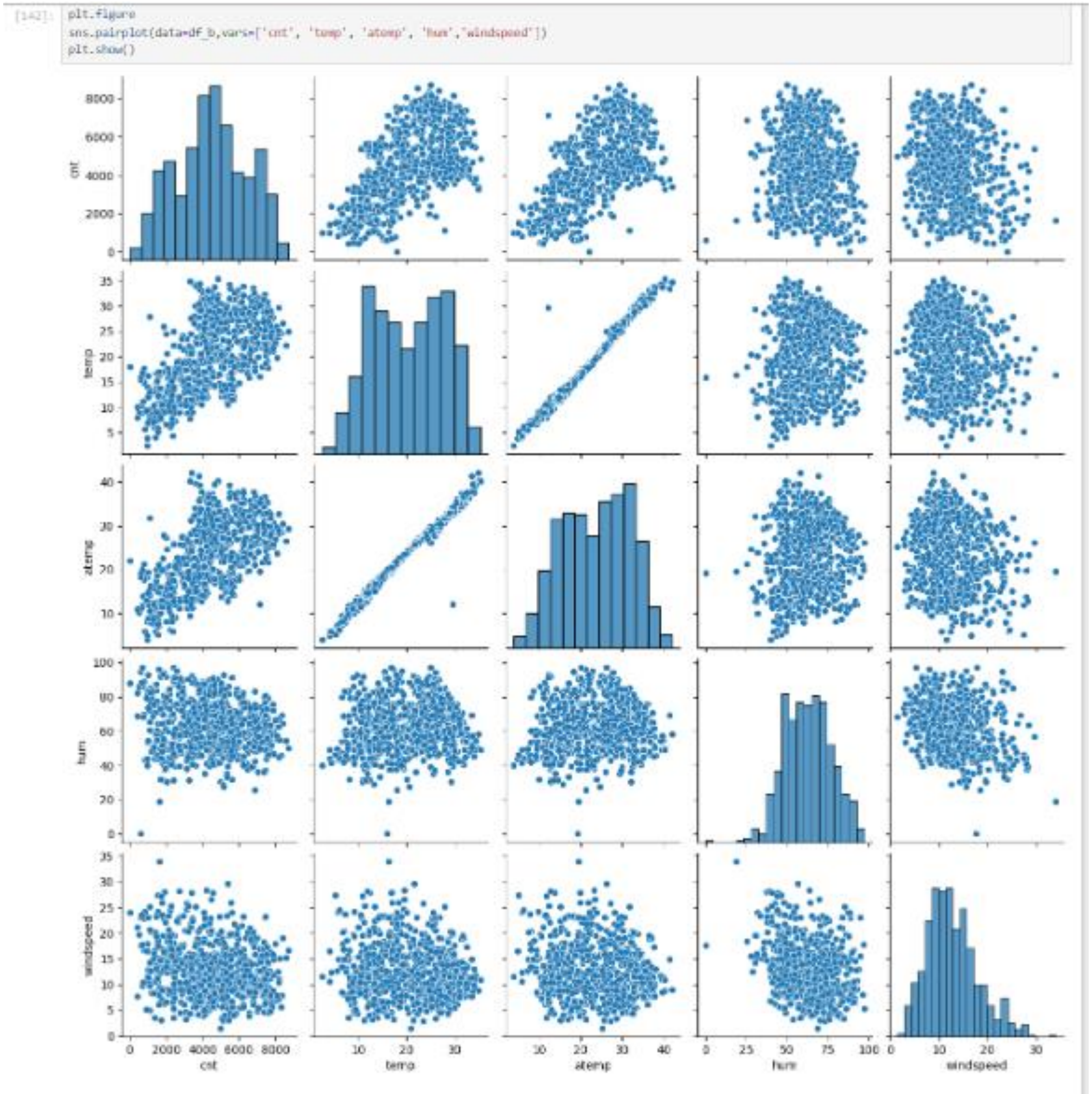
Question 2. Why is it important to use **drop_first=True** during dummy variable creation? (Do not edit)

Total Marks: 2 marks (Do not edit)

Answer: <Your answer for Question 2 goes below this line> (Do not edit)

The purpose of the dummy variable is to allow you to generate "n-1" new columns for a category variable with "n" levels, each of which uses a zero or one to indicate whether or not that level exists. In order for the outcome to match up n-1 levels, drop_first=True is needed. As a result, the correlation between the dummy variables is decreased. For instance, drop_first will drop the first column if there are three levels.

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)



```
plt.figure
sns.pairplot(data=df_b,vars=['cnt', 'temp', 'atemp', 'hum','windspeed'])
plt.show()
```

By looking at Pairplot temp and atemp highest correlation compare to other variables when target variable is ‘cnt’.

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)

The following criteria are used to validate linear regression models:

1. Multicollinearity should be little or none.
2. No-Autocorrelation or little.
3. Relationship between variables should be linear in nature
4. Homoscedasticity : No visible pattern in residual values
5. Error terms to be normally distributed.

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)

Top 3 features contributing significantly towards explaining the demand of the shared bikes –

1. Year
2. Season
3. Sept

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)

Linear regression is statistical model which defines the linear replationship between dependent variable with set of of independent variables. The linear relationship between variables means when the value of one or more independent variables will change (increase or decrease) then dependent variables will also change (increase or decrease) accordingly.

Mathematically the relationship can be represented with the help of following equation
$Y = mX + c$

Here, Y is the dependent variable we are trying to predict. X is the independent variable we are using to make predictions. m is the slope of the regression line which represents the effect X has on Y c is a constant, known as the Y-intercept. If X = 0, Y would be equal to c. Furthermore, the linear relationship can be positive or negative in nature as explained below−
Positive Linear Relationship: A linear relationship will be called positive if both independent and dependent variable increases.

Negative Linear relationship: A linear relationship will be called negative if independent increases and dependent variable decreases.

Linear regression are two type:

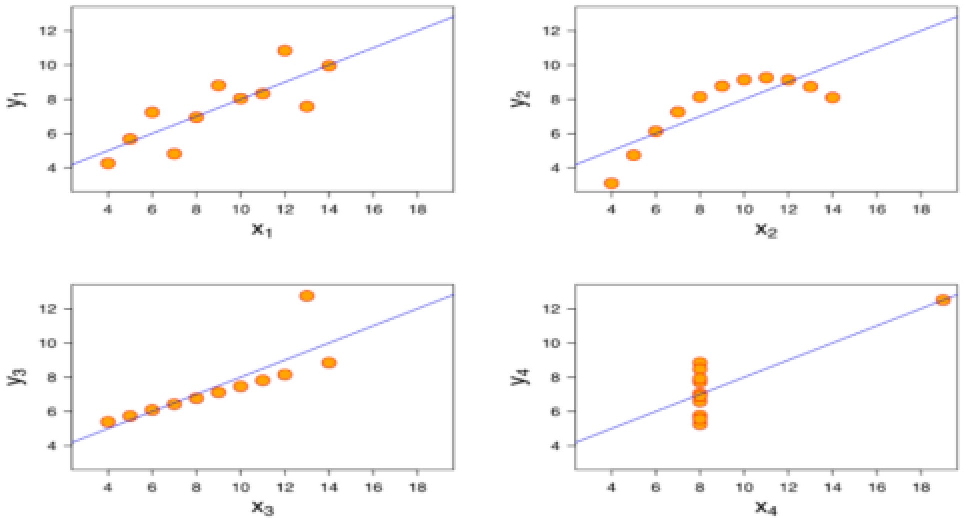
1. Simple LR
2. Multiple LR

Assumptions on LR model:

6. Multicollinearity should be little or none.
7. No-Autocorrelation or little.
8. Relationship between variables should be linear in nature
9. Homoscedasticity : No visible pattern in residual values
10. Error terms to be normally distributed.

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 collection of four data sets that, when analyzed using basic descriptive statistics, are essentially the same; nonetheless, they contain some oddities that could trick a regression model if it were constructed. They show up differently on scatter plots and have wildly disparate distributions. It was made to demonstrate the significance of plotting the graphs prior to analysis and model construction, as well as how other observations impact statistical features. Plots of these four data sets share essentially identical statistical observations, offering identical statistical information about variance and the mean of all x and y points across the four datasets.



A) The first data set matches the linear regression model, there appears to be a linear relationship between X and Y,

B) The second data set does not match the linear regression model since it does not demonstrate a linear connection between X and Y.

C) The third dataset has certain outliers that a linear regression model is unable to manage.

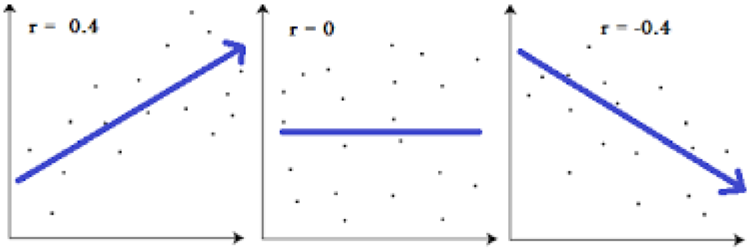
D) The fourth data set generates a high correlation coefficient because it has a high leverage point.
It concludes that while regression algorithms can be fooled, data visualization is crucial when

developing a machine learning model.

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)

Pearson correlation coefficient (r) is a number that ranges from -1 to 1 and indicates how strongly and in which direction two variables are related.

A positive relationship is shown by a value greater than 0; that is, as one variable's value rises, the other variable's value rises as well. A negative relationship is shown by a value less than 0; that is, when one variable's value rises, the other variable's value falls.



Positive Correlation No correlation 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 converting your data to fall into a predetermined range. This kind of data pre-processing stage involves fitting data to a predetermined scale and accelerating algorithmic computations. Features in the collected data vary in range, magnitude, and unit. Without scaling, the algorithm has a tendency to overlook other factors and weigh large values, which leads to inaccurate modeling.
Example: Without employing the feature scaling method, an algorithm may interpret 2500 meters as being more than 6 kilometers, which is untrue and will result in inaccurate predictions. In order to address this problem, we employ feature scaling to bring all values to the same magnitudes.

The difference between standardizing scaling and normalizing scaling
1. Standardized scaling uses the mean and standard deviation for scaling, while normalized scaling uses the minimum and maximum value of features.
2. Standardized scaling is used to guarantee zero mean and unit standard deviation, while normalized scaling is utilized when features are of different scales.
3. While standardized scaling lacks or is not limited in a certain range, normalized scaling scales values between (0,1) or (-1,1).
4. Outliers have an impact on normalized scaling, but they have no influence on standardized scaling.
5. Standardized scaling is used when the distribution is normal, while normalized scaling is used when the distribution is unknown.
6. Standardized scaling is known as Z Score Normalization, while normalized scaling is known as

scaling normalization.

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)

VIF (VarianceInflationFactor) = infinity if the correlation is perfect. There is a correlation between the variables when the VIF score is high. If the VIF is 4, it indicates that multicollinearity has caused the variance of the model coefficient to be exaggerated by a factor of 4.
Two independent variables exhibit complete correlation when the VIF value is infinite. R-squared (R2) = 1 in the event of perfect correlation results in 1/ (1-R2) infinity. We must remove one of the factors creating this perfect multicollinearity from the dataset in order to resolve this.

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 graphical method for figuring out whether two data sets are from populations with a similar distribution is the quantile-quantile (q-q) plot.

Use of Q-Q Plot:-

Plotting the quantiles of the first and second data sets against one another is known as a q-q plot. The fraction (or percentage) of points below the specified value is referred to as a quantile.
In other words, 25% of the data falls below the 0.25 (or 25%) quantile, while 75% of the data falls above it. Additionally, a 45-degree reference line is plotted. The points should roughly lie along this reference line if the two sets originate from a population with the same distribution. The evidence supporting the conclusion that the two data sets originate from populations with distinct distributions increases with the degree of divergence from this reference line.

Importance of Q-Q Plot:-
It is frequently desirable to determine if the assumption of a common distribution is justified when there are two data samples. If so, estimates of the shared position and scale can be obtained by pooling the two data sets using location and scale estimators. If two samples do differ, it is also useful to gain some understanding of the differences. More information about the nature of the difference can be obtained from the q-q plot than from analytical techniques like the chi-square and Kolmogorov-Smirnov 2-sample tests.
