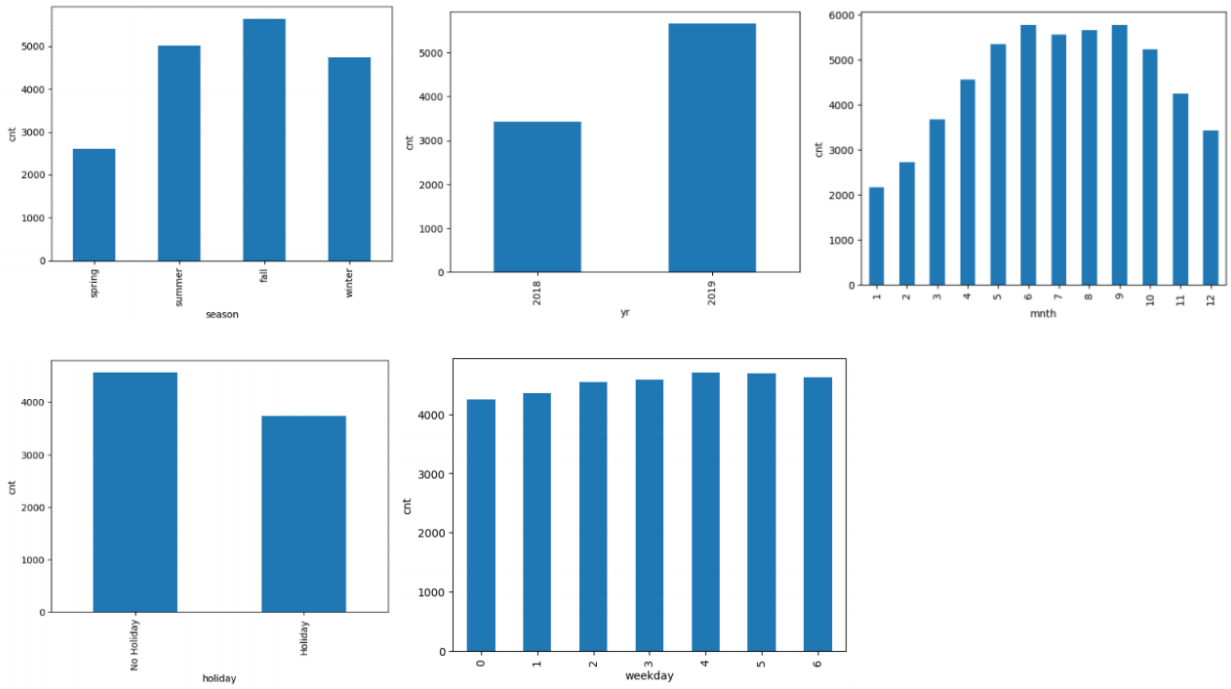
Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

We infer that some of the categorical variables such as season, year, month, holiday and weathersit
have significant impact on the output variable ‘cnt’ but certain other categorical variables such as
weekday, workingday do not have much impact. Same can be seen from the below plots.



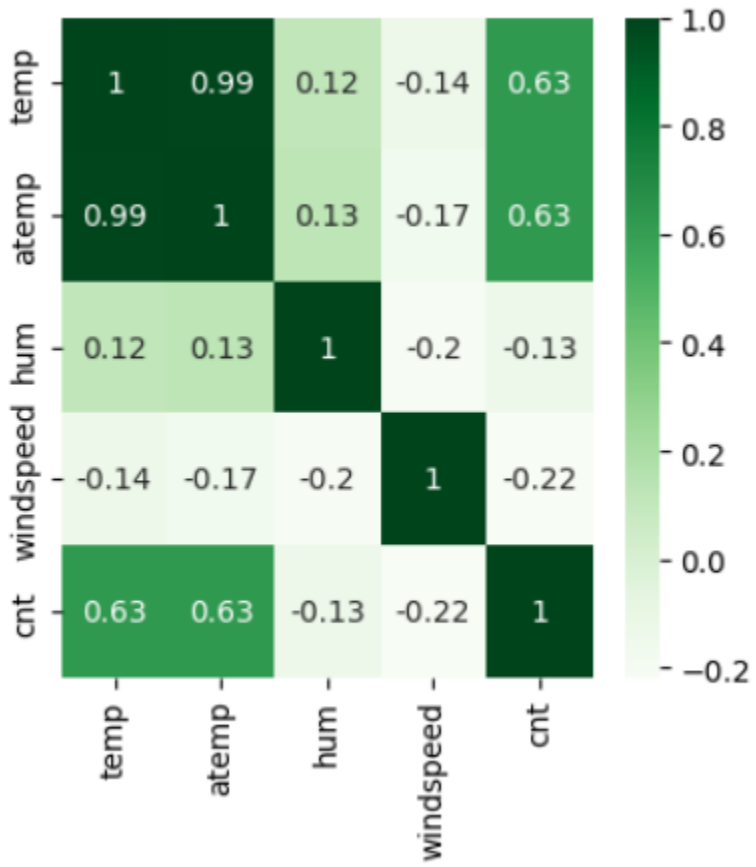
2. Why is it important to use **drop_first=True** during dummy variable creation? (2 mark)

For creating dummy variables using pandas
pd.get_dummies()
We get the same number of new columns as the number of categories in the feature. For Ex- if we have ‘4 ‘
categories in season, we will get’ 4’ new columns. But we don’t need 4 new columns to express all ‘4 ‘
categories of season and we need only ‘3’ (n-1). This option of **drop_first=True** removes one redundant
column thus reducing the complexity and time needed to model the data and make predictions.

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

with the target variable?
(1 mark)

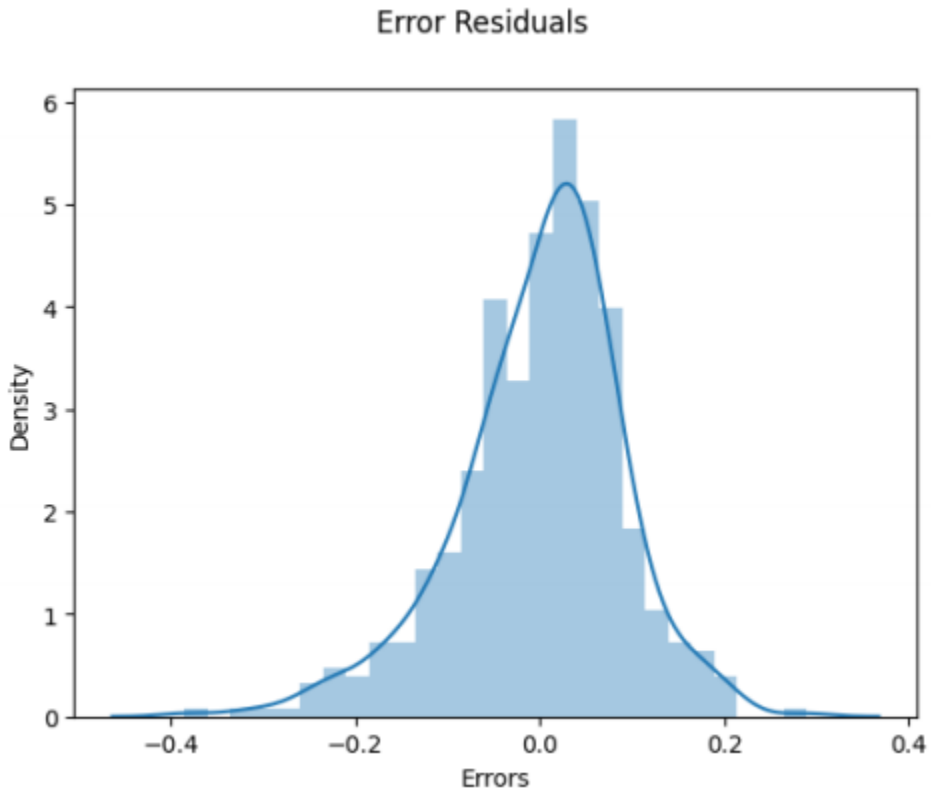
The variable *atemp* has the highest correlation but *temp* is very close second as shown below



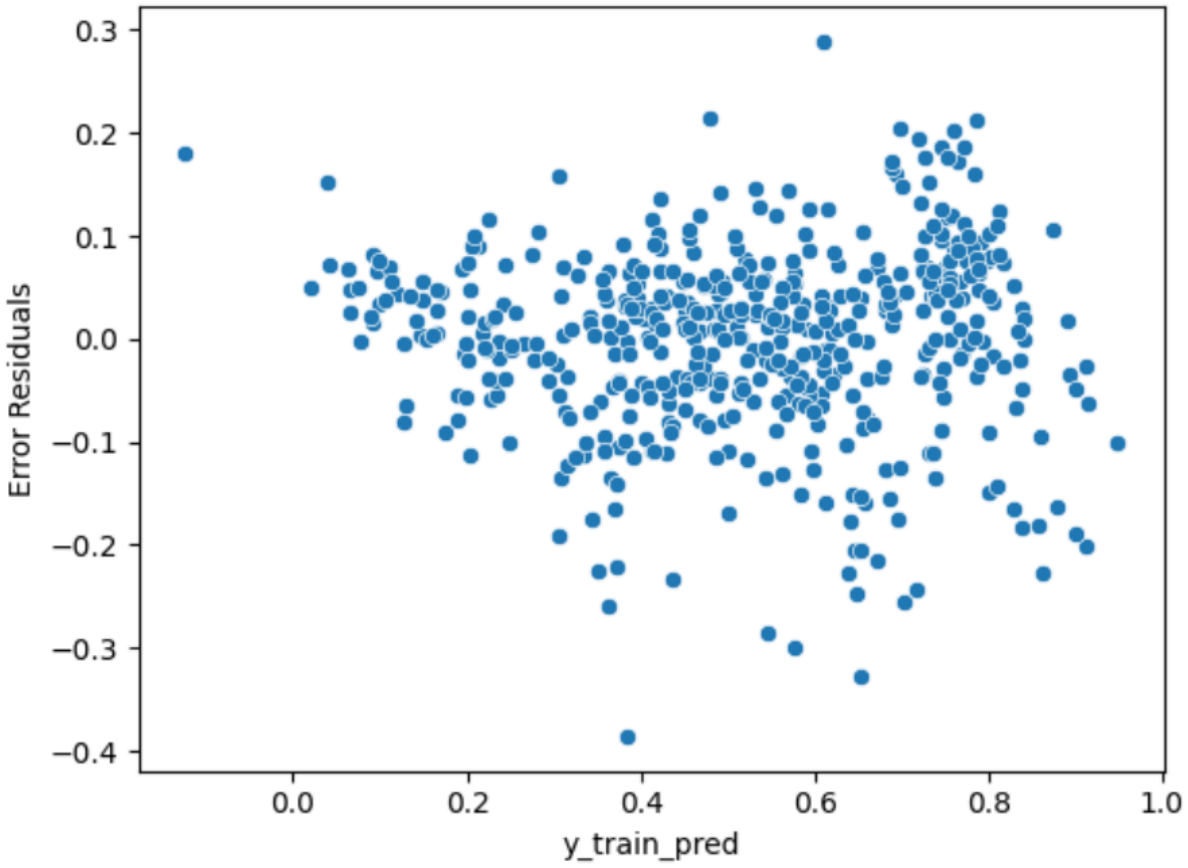
4. How did you validate the assumptions of Linear Regression after building the model on the training set?
(3 marks)

There are 4 assumptions of linear regression model
1. There is a linear relationship between X and Y
2. Error terms are normally distributed with mean zero(not X, Y)
3. Error terms are independent of each other
4. Error terms have constant variance (homoscedasticity)
The first assumption is something that has to be looked at before building the model by looking at the
scatter and bar plots to check if the linear model can be used. But the question talks about AFTER
building the model where the remaining 3 assumptions need to be validated.
To validate the 2nd assumption, we calculate the residual and plot the histogram of residuals as shown

below. This looks very close to a normal distribution with mean 0.



The second assumption is validated by plotting the residual against the y value to see if there is any
trend, and we don’t see any major trend.



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 top 3 features contributing significantly towards the demand are:

S.No	Variable name	Variable Explanation	Coefficient value
1.	temp	Temperature	0.601
2.	yr	Year	0.224
3.	hum	Humidity	-0.202

General Subjective Questions

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

Linear regression is a supervised machine learning algorithm used for modeling the relationship between a dependent variable (also known as the target) and one or more independent variables (also known as predictors or features). It assumes a linear relationship between the variables. The algorithm aims to find the best-fitting line (or hyperplane in multiple dimensions) that minimizes the sum of squared differences between the observed and predicted values.

Here's how linear regression works:

- Simple Linear Regression: In the case of a single independent variable, the algorithm calculates the slope and intercept of the line that best fits the data points.
- Multiple Linear Regression: When there are multiple independent variables, the algorithm estimates the coefficients for each variable, representing the impact of that variable while holding others constant.

The equation for a simple linear regression line can be represented as:

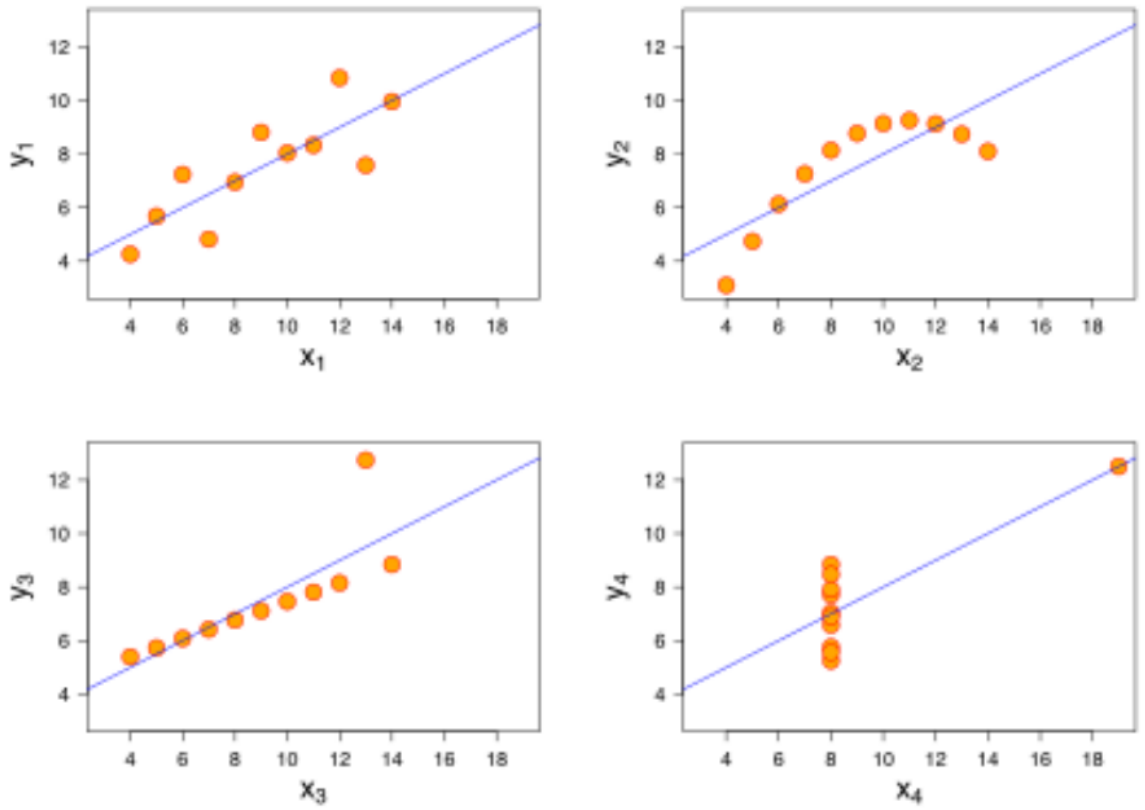
$y = mx + c$

where:

- y is the dependent variable.
- x is the independent variable.
- m is the slope (coefficient).
- b is the intercept.

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

Anscombe's quartet is a set of four datasets that have nearly identical statistical properties (such as means, variances, correlations, and linear regression lines), yet they have very different distributions and visual representations. This quartet emphasizes the importance of visualizing data and not relying solely on summary statistics. It illustrates how data that appears similar in a statistical sense can lead to different conclusions when graphed.



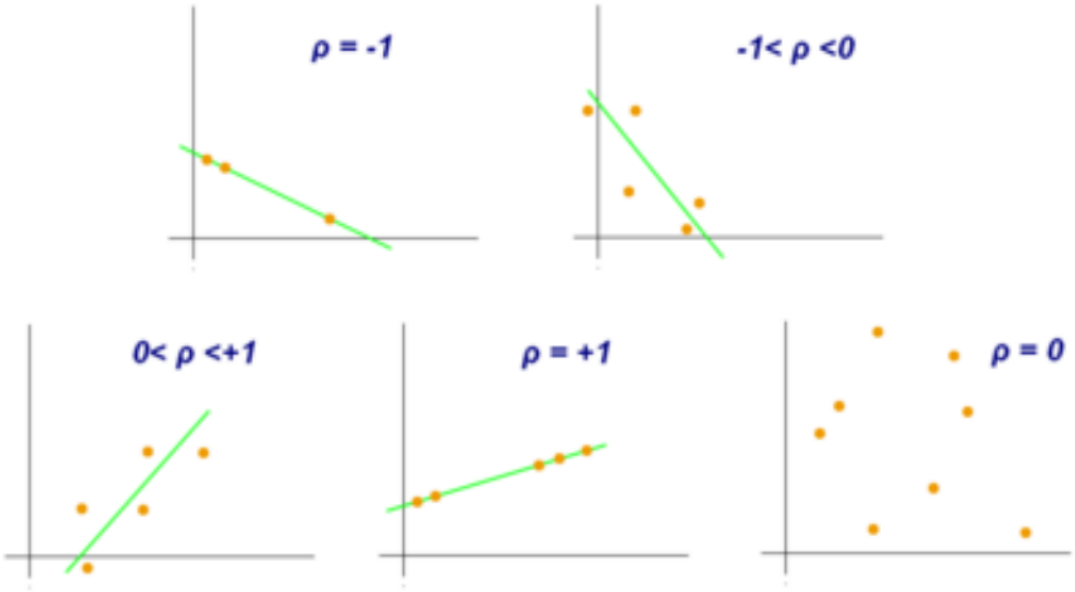
This quartet demonstrates the importance of visualizing the data when analyzing it and shows the effect
of outliers on statistical properties.

For all 4 datasets
1. The first dataset has linear relationship between input and output.
2. The second dataset is non-linear in nature and a linear line cannot be fitted even though it has same correlation coefficient as the first data
3. The third data is linear but the entire relationship gets affected by just one outlier point and we can see that the fitted line is not representing the actual dataset
4. The fourth dataset shows that there is no relationship between input and output, but a highly skewed data points results in a high correlation and simply fitting a linear model is meaningless.

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

Pearson's correlation coefficient (often denoted as Pearson's R) is a measure of the linear relationship between two continuous variables. It quantifies the strength and direction of the linear association between the variables. The value of Pearson's R ranges between -1 and 1:

- A positive value (close to 1) indicates a strong positive linear correlation.
- A negative value (close to -1) indicates a strong negative linear correlation.
- A value close to 0 indicates a weak or no linear correlation.



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 variables to have a specific scale or range. It is performed to ensure that all variables have a similar scale, which can help improve the performance of certain algorithms, such as gradient descent-based optimization in linear regression.

- Normalized Scaling: This scales the variables to a range of [0, 1]. It is useful when features have different ranges and you want to bring them to a common scale.
- Standardized Scaling: This scales the variables to have a mean of 0 and a standard deviation of 1. It preserves the shape of the distribution and is useful when variables have different units.

The advantage of standardized scaling over normalized scaling is that it doesn't compress the data between
a particular range and this is useful when extreme data point outliers. In case of extreme data points,
MinMax scaling can result in most of the data points getting compressed very close together.

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

VIF measures the multicollinearity between predictor variables in a regression model. A high VIF indicates that a predictor is highly correlated with other predictors, which can affect the model's interpretability and stability.

- Infinite VIF: This occurs when one or more predictor variables are perfectly correlated, leading to a division by zero in the VIF formula. It indicates severe multicollinearity and suggests that the predictor can be expressed as a linear combination of other predictors.

(3 marks)
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 is a graphical tool used to assess if a dataset follows a specific theoretical distribution, such as the normal distribution. It compares the quantiles of the dataset against the quantiles of the chosen theoretical distribution. If the points on the plot roughly form a straight line, it suggests that the dataset follows the chosen distribution. Q-Q plots help detect deviations from the expected distribution and guide decisions about data transformations.