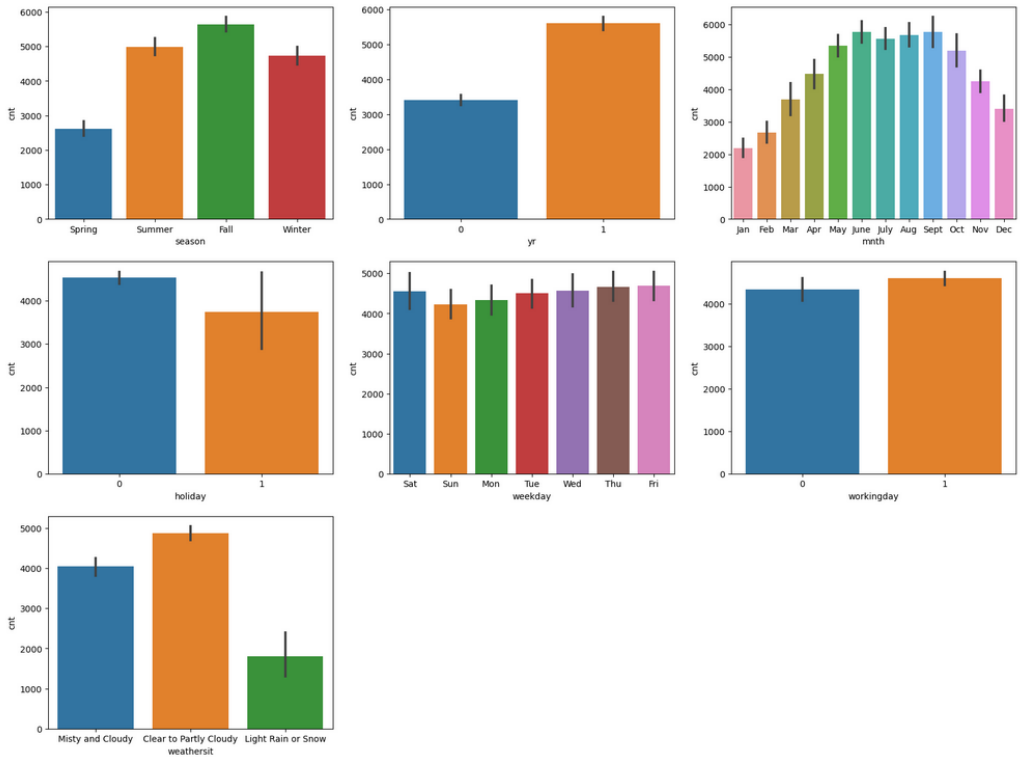
Assignment-based Subjective Questions

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

Ans - Below are the finding :

1)Bike demand more in Fall season followed by summer season
2)In 2019 there are more rental bike users than in 2018
3)There are more booking in the months May, June, July, Aug, Sept and Oct.
4)There are more rental bike users when wheather is clear to partly cloudy
5)There is not even single day on which heavy rain/snow has occured
6)Usage of bike on weekdays is slighlty highrer than holidays
7)Usage of bike is similiar irrespective of being working day or not



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

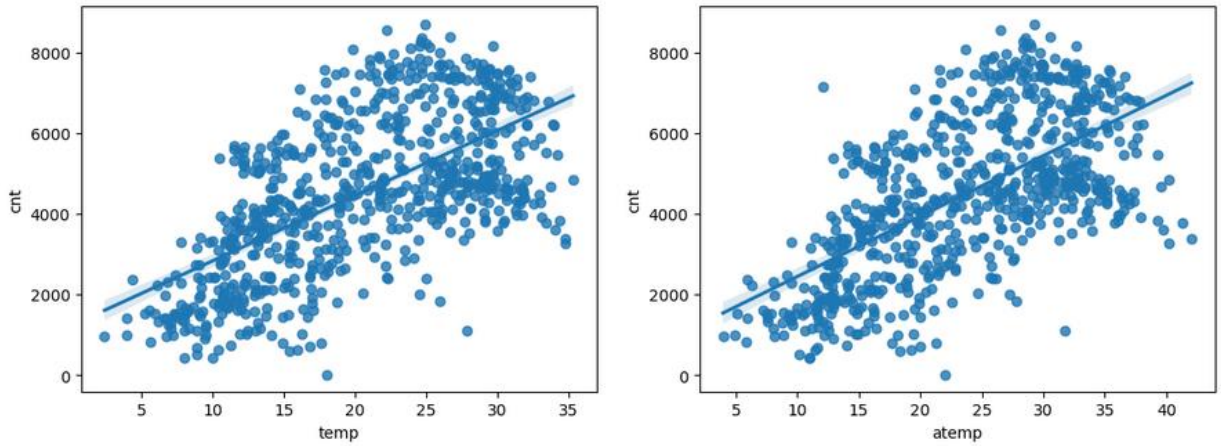
Ans - It is important in order to achieve n-1 dummy variables as it can be used to delete extra column while creating dummy variables.

For Example: We have three variables: Furnished,Semi-furnished and un-furnished. The urgument will drop Furnished variable and will keep only semi-furnished and un-furnished and assign values 0 and 1 as shown below. So in below example first row shows 0-0 for Semi-furnished and un-furnished so it means it is Furnished.

Semi-furnished	un-furnished
0	0
1	0
0	1

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Ans-temp and atemp varibales are having highest linear correlation with target variable cnt.



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

Ans - From below results I can validate the assumptions of Linear Regression :

1)Linear Model - There is linear relationship between dependent and independent variable
2)P-value - All values should be less than 0.05
3)VIF- All values should be less than 5
4)Number of observations Greater than the number of predictors - test_train_split is 70-30
5)Error is normally distributed
6)R^2 value - High value of R^2 (Same for both test and train data within 5% of difference)
7)No Multicollinearity and Homoscedasticity

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Ans - 1)temp
2)yr
3)Season

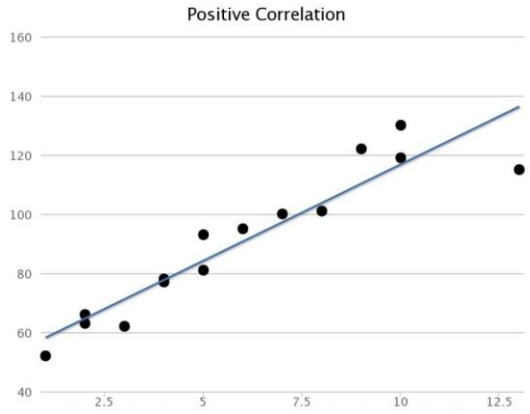
General Subjective Questions

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

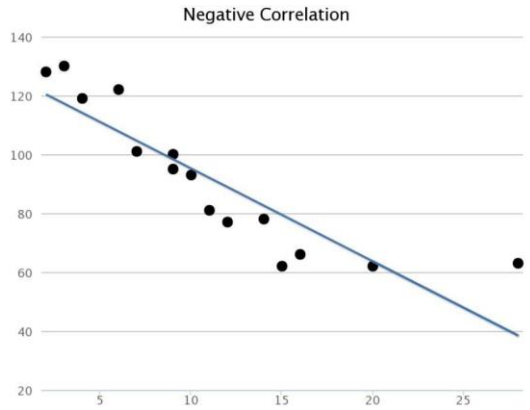
Ans-Linear regression is a type of machine-learning algorithm more specifically a supervised machine-learning algorithm that learns from the labelled datasets and maps the data points to the most optimized linear functions. which can be used for prediction on new datasets.

First we should know what supervised machine learning algorithms is. It is a type of machine learning where the algorithm learns from labelled data. Labeled data means the dataset whose respective target value is already known. Supervised learning has two types:
Classification: It predicts the class of the dataset based on the independent input variable. Class is the categorical or discrete values. like the image of an animal is a cat or dog?
Regression: It predicts the continuous output variables based on the independent input variable. like the prediction of house prices based on different parameters like house age, distance from the main road, location, area, etc.
Mathematically the relationship can be represented with the help of following equation:
$Y = mX + c$

Positive Linear Relationship:



Negative Linear Relationship:



Mainly there are 7 assumptions taken while using Linear Regression:

- Linear Model
- No Multicolinearlity in the data
- Homoscedasticity of Residuals or Equal Variances
- No Autocorrelation in residuals
- Number of observations Greater than the number of predictors
- Each observation is unique
- Predictors are distributed Normally

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

Ans - Anscombe's quartet comprises a set of four datasets, having identical descriptive statistical properties in terms of means, variance, R-squared, correlations, and linear regression lines but having different representations when we scatter plots on a graph.

The datasets were created to demonstrate the importance of visualizing data and to show that summary statistics alone can be misleading.

The four datasets that make up Anscombe's quartet each include 11 x-y pairs of data. 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

Purpose of Anscombe's Quartet
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.

Anscombe's Quartet Dataset

The four datasets of **Anscombe's quartet.**

I		II		III		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

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

Ans - The Pearson correlation coefficient (r) is the most widely used correlation coefficient and is known by many names:
Pearson's r
Bivariate correlation
Pearson product-moment correlation coefficient (PPMCC)
The correlation coefficient

The Pearson correlation coefficient is a descriptive statistic, meaning that it summarizes the characteristics of a dataset. Specifically, it describes the strength and direction of the linear relationship between two quantitative variables. Although interpretations of the relationship strength (also known as effect size) vary between disciplines, the table below gives general rules of thumb:

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

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Ans - Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units.
If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.
In normalized scaling values are scaled between 0 and 1 whereas in standardized scaling values are normally distributed by calculating mean and standard deviation

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Ans - This shows a perfect correlation between two variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity

6. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.(3 marks)

Ans - The quantile-quantile(q-q plot) plot is a graphical method for determining if a dataset follows a certain probability distribution or whether two samples of data came from the same population or not. Q-Q plots are particularly useful for assessing whether a dataset is normally distributed or if it follows some other known distribution. They are commonly used in statistics, data analysis, and quality control to check assumptions and identify departures from expected distributions.

Quantile-Quantile (Q-Q) plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a Normal, exponential or Uniform distribution. Also, it helps to determine if two data sets come from populations with a common distribution.
This helps in a scenario of linear regression when we have training and test data set received separately and then we can confirm using Q-Q plot that both the data sets are from populations with same distributions.