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?

Ans. We say a variable as categorical if it has limited unique values, and in the dataset provides, there were 7 such columns. Those are 'season','year','month','holiday','weekday','workingday','weathersit'.

These columns had the following effect on our target column 'cnt':

- Surge in demand can be observed in summer, fall season
- Drastic increase in booking count from 2018 to 2019
- booking peaked in the month of September
- As expected, bookings are high when the weather is clear
- Wednesday, Thursday, Friday, Saturday have more bookings than Sunday, Monday, Tuesday
- more bookings are observed on non holidays
- more bookings are observed on working day

2 Why is it important to use drop_first=True during dummy variable creation?

Ans. We create dummy variable to convert categorical values into numerical(Boolean).
For example, if a variable has 3 unique data values, the get_dummies() command creates 3 new columns as 0 0 0, 0 0 1, 0 1 0. But here if observe carefully, 2 new columns are enough to represent 3 unique values,
So in order reduce the excess column we drop the first column created using get_dummies() command.

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

Ans. temp and atemp variables has the highest correlation with our target cnt variable and also temp and atemp are also highly correlated

4 How did you validate the assumptions of Linear Regression after building the model on the training set?

Ans. We can validate our assumptions of linear regression using following techniques:

- Normality of error terms: Error terms should be normally distributed
- Multi-collinearity Check: There shouldn't high correlation between variables in the trained model
- Linear Relationship Validation: Linearity should be visible among variable
- Homoscedasticity: No visible pattern in residual variables must be visible
- Independence of residual: no auto correlation be visible

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

Ans. Based on coefficients, Top 3 features are:

- Temp
- Winter(season)
- Sat(weekday)

General Subjective Questions:

1 Explain the linear regression algorithm in detail.

Ans. Linear regression may be defined as a statistical algorithm used to model the relationship between the independent variable/s (also known as the predictor variable) and the dependent variable (also known as the target variable). The goal of linear regression is to find the line of best-fit.

The linear regression algorithm works by minimizing the sum of the squared distances between the predicted values and the actual values of the dependent variable. This is achieved by calculating the slope and intercept of the line that best fits the data.

The equation of the line is typically represented as:

$y = mx + c$

Where y is the dependent variable, x is the independent variable, m is the slope, and c is the intercept.
If x=0, then y would be equal to 'c'

The algorithm uses a training set of data to estimate the values of the slope and intercept, and then applies these values to a test set of data to make predictions. The accuracy of the model is typically measured using mean squared error or R-squared.

Linear regression can be used for both simple regression (with only one independent variable) and multiple regression (with multiple independent variables).

Further linear relationship can be positive or negative as explained below:

- Positive Linear Relationship: When dependant variable increase with increase in independent variable then the relationship is said to be positive
- Negative Linear Relationship: When dependant variable decrease with increase in independent variable then the relationship is said to be negative

The following assumptions are made during building the model:

- Multi-Collinearity
- Auto-Corelation
- Linear Relationship between variables
- Normality of error terms
- Homoscedasticity

2 Explain the Anscombe's quartet in detail.

Ans. Anscombe's quartet is a set of four datasets that were created by the statistician Francis Anscombe in 1973 to illustrate the importance of visualizing data before drawing conclusions. Despite having identical summary statistics, each of the datasets has a different pattern when plotted, highlighting the limitations of relying on summary statistics alone.

The four datasets each have 11 observations of two variables, x and y. The first dataset has a linear relationship between x and y, the second dataset has a non-linear relationship, the third dataset has a clear outlier, and the fourth dataset has a relationship that is heavily influenced by a single observation.

	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	8	5,56
	7	4,82	7	7,26	7	6,42	8	7,91
	5	5,68	5	4,74	5	5,73	8	6,89
SUM	99,00	82,51	99,00	82,51	99,00	82,50	99,00	82,51
AVG	9,00	7,50	9,00	7,50	9,00	7,50	9,00	7,50
STDEV	3,32	2,03	3,32	2,03	3,32	2,03	3,32	2,03

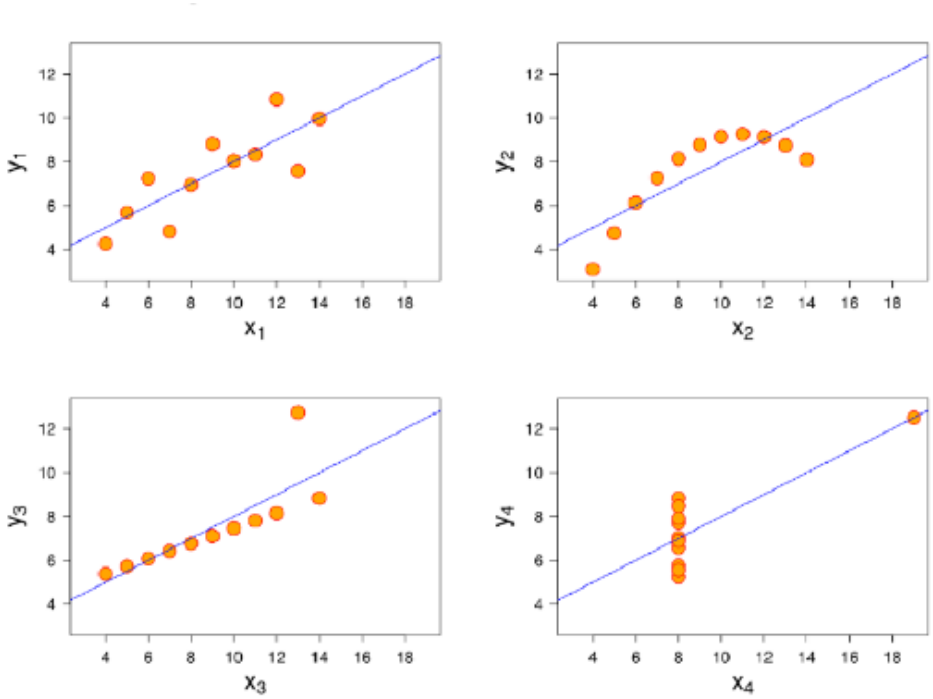
When plotted, the datasets appear as follows:

Dataset 1(plot y1): A linear relationship with a slight positive slope and little variability around the line.

Dataset 2(plot y2): A non-linear relationship with a clear quadratic pattern.

Dataset 3(plot y3): A linear relationship with a clear outlier that is far from the rest of the data.

Dataset 4(plot y4): A relationship that appears to be mostly linear, except for a single outlier that has a large influence on the line of best fit.



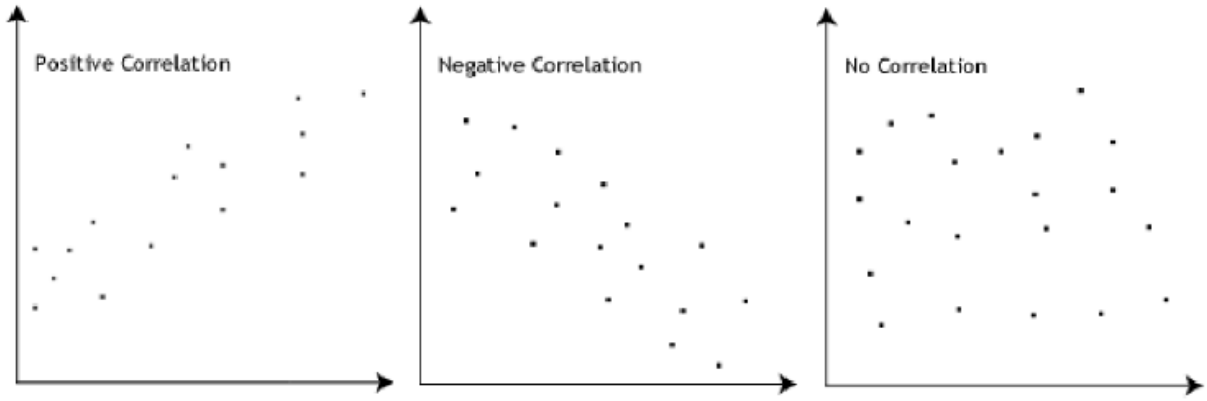
Despite having identical means, variances, and correlation coefficients, the datasets demonstrate the importance of visualizing data to understand the underlying patterns and relationships. It also highlights the limitations of relying solely on summary statistics, which can obscure important details about the data.

Anscombe's quartet is often used in statistics courses to illustrate the importance of data visualization and the limitations of summary statistics.

3 What is Pearson's R?

Ans. Pearson's correlation coefficient, often denoted by "r", is a measure of the strength and direction of the linear relationship between two variables.

The Pearson correlation coefficient ranges from -1 to +1, where a value of -1 indicates a perfect negative linear relationship, a value of +1 indicates a perfect positive linear relationship, and a value of 0 indicates no linear relationship between the two variables.



The formula for calculating Pearson's correlation coefficient is:

r = (nΣxy - ΣxΣy) / √[(nΣx^2 - (Σx)^2)(nΣy^2 - (Σy)^2)]

where:
- n is the number of observations
- Σxy is the sum of the products of the corresponding values of the two variables
- Σx and Σy are the sums of the values of the two variables, respectively
- Σx^2 and Σy^2 are the sums of the squared values of the two variables, respectively

The Pearson correlation coefficient is sensitive to outliers and assumes that the relationship between the two variables is linear.

4 What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Ans. Scaling is a data preprocessing technique that involves transforming the values of numerical variables to a specific range or distribution. The objective of scaling is to make the data more suitable for analysis or modeling by reducing the impact of differences in the scale and magnitude of the variables.

Scaling is performed for several reasons:
- Scaling can help to avoid numerical overflow or underflow issues that can arise when dealing with variables with different magnitudes.
- Scaling can make the interpretation of coefficients in linear models more meaningful by ensuring that the variables have a similar scale.

There are two common types of scaling: normalized scaling and standardized scaling.

Normalized scaling involves transforming the data so that it has a range between 0 and 1. The formula for normalized scaling is:

x_norm = (x - min(x)) / (max(x) - min(x))

where x is the original value of the variable, and x_norm is the normalized value of the variable.

Standardized scaling involves transforming the data so that it has a mean of 0 and a standard deviation of 1. The formula for standardized scaling is:

x_std = (x - mean(x)) / std(x)

where x is the original value of the variable, and x_std is the standardized value of the variable.

The main difference between normalized scaling and standardized scaling is that normalized scaling preserves the original distribution of the data, while standardized scaling transforms the data to have a standard normal distribution. Standardized scaling is often preferred in statistical analysis and modeling because it ensures that the variables are on the same scale, and makes it easier to compare the relative importance of the variables. However, normalized scaling may be preferred in some cases, such as when the original distribution of the data is important for analysis or visualization purposes.

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

Ans.

- VIF: variance inflation factor is sometimes inf(infinite) when there is a perfect correlation.
- In case of perfect correlation, R2 becomes 1, which leads to 1/(1-R2) becoming infinite.
- This can be solved by dropping those column/s

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

Ans. A Q-Q (quantile-quantile) plot is a graphical technique for comparing the distribution of a sample of data to a theoretical distribution. It is often used to assess whether a set of data follows a particular distribution, such as the normal distribution, and to identify departures from that distribution.

In linear regression, Q-Q plots are used to assess the normality of the residuals, which are the differences between the observed values of the dependent variable and the predicted values from the regression model. If the residuals are normally distributed, the points on the Q-Q plot will lie on a straight line. Departures from the straight line can indicate that the residuals are not normally distributed, which can violate the assumptions of the linear regression model.