Assignment-based Subjective Questions

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

Here are a few takeaways from the visualisation:

- The fall season appears to have drawn more bookings. And, in each season, the number of bookings increased dramatically between 2018 and 2019.
- The majority of reservations were made in the months of May, June, July, August, September, and October.
- The trend increased at the beginning of the year until the middle of the year, when it began to decline as we approached the conclusion of the year.
- Clear weather attracted more bookings, which seems clear.
- There are higher bookings on Thursday, Friday, Saturday, and Sunday than at the beginning of the week.
- When it is not a holiday, the count of bookings appears to be lower, which seems sensible given that on holidays, people may wish to spend time at home and enjoy time with family.
- Booking seems to be nearly equal on working and non-working days.
- 2019 saw an increase in bookings over the previous year, indicating a good rise in the demand

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

By dropping one of the one-hot encoded columns from each categorical feature, we ensure there are no reference columns—the remaining columns become linearly independent. Moreover, if we have three variable two can serve the purpose of prediction in order to make our model look clean and clutter free, we remove the first dummy variable.

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

'temp' variable has the highest correlation with the target variable

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

I have validated the assumption of Linear Regression Model based on below 5 assumptions –

- Error terms should be normally distributed
- There should be insignificant multicollinearity among variables.
- Linearity should be visible among variables
- Homoscedasticity :There is no visible pattern in residual values.
- Independence of residuals: No auto-correlation

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

Top three features are as follows:

- temp
- light-snow
- Windspeed

General Subjective Questions

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

Regression is a parametric technique used to predict continuous (dependent) variable given a set of independent variables. It is parametric in nature because it makes certain assumptions based on the data set. If the data set follows those assumptions, regression gives incredible results. Otherwise, it struggles to provide convincing accuracy. Mathematically, regression uses a linear function to approximate (predict) the dependent variable given as:

$$Y = \beta o + \beta 1X + \in$$

where, Y - Dependent variable
X - Independent variable
βo - Intercept
β1 - Slope
∈ - Error
the equation above is nothing but a line equation (y = mx + c) we studied in schools.

βo and β1 are known as coefficients. This is the equation of simple linear regression. It's called 'linear' because there is
just one independent variable (X) involved. In multiple regression, we have many independent variables (Xs).

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 positive if independent increases and dependent variable decreases.

Assumptions

- Error terms should be normally distributed.
- There should be insignificant multicollinearity among variables.
- Linearity should be visible among variables .
- Homoscedasticity: There is no visible pattern in residual values.
- Independence of residuals: No auto-correlation.

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

Anscombe's Quartet was developed by statistician Francis Anscombe. It comprises four datasets, each containing eleven (x, y) pairs. The essential thing to note about these datasets is that they share the same descriptive statistics. But things change completely, and I must emphasize COMPLETELY, when they are graphed. Each graph tells a different story irrespective of their similar summary statistics.

	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

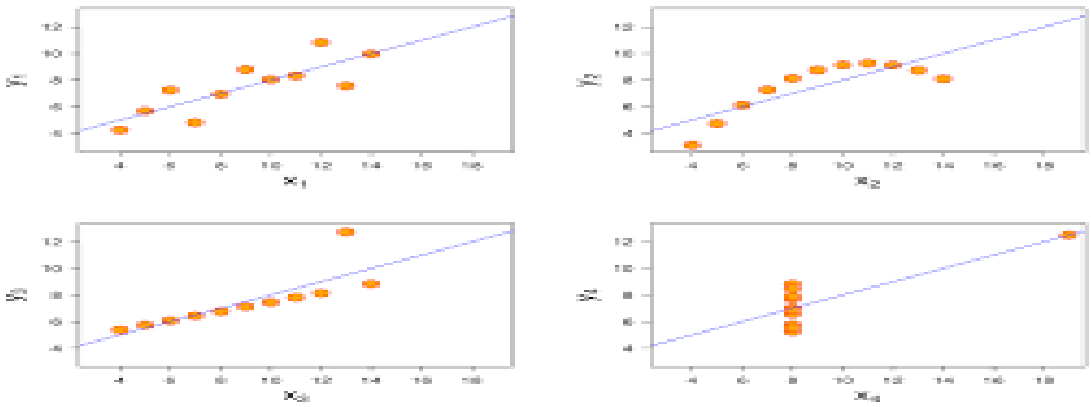
The summary statistics show that the means and the variances were identical for x and y across the groups:

• Mean of x is 9 and mean of y is 7.50 for each dataset.

• Similarly, the variance of x is 11 and variance of y is 4.13 for each dataset

• The correlation coefficient (how strong a relationship is between two variables) between x and y is 0.816 for each dataset

When we plot these four datasets on an x/y coordinate plane, we can observe that they show the same regression lines as well but each dataset is telling a different story:



- Dataset I appears to have clean and well-fitting linear models.
- Dataset II is not distributed normally.
- In Dataset III the distribution is linear, but the calculated regression is thrown off by an outlier.
- Dataset IV shows that one outlier is enough to produce a high correlation coefficient.

This quartet emphasizes the importance of visualization in Data Analysis. Looking at the data reveals a lot of the structure and a clear picture of the dataset.

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

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.

Pearson correlation coefficient (*r*)	Correlation type	Interpretation	Example
Between 0 and 1	Positive correlation	When one variable changes, the other variable changes in the **same direction**.	Baby length & weight: The longer the baby, the heavier their weight.
0	No correlation	There is **no relationship** between the variables.	Car price & width of windshield wipers: The price of a car is not related to the width of its windshield wipers.
Between 0 and –1	Negative correlation	When one variable changes, the other variable changes in the **opposite direction**.	Elevation & air pressure: The higher the elevation, the lower the air pressure.

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

Feature scaling is a technique for standardizing independent features present in the data to a fixed range. This is done during data pre-processing.

Techniques to perform Feature Scaling
Consider the two most important ones:

- **Min-Max Normalization:** This technique re-scales a feature or observation value with distribution value between 0 and 1.

- **Standardization:** It is a very effective technique which re-scales a feature value so that it has distribution with 0 mean value and variance equals to 1.

S.NO.	Normalization	Standardization
1.	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2.	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3.	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
4.	It is really affected by outliers.	It is much less affected by outliers.
5.	Scikit-Learn provides a transformer called `MinMaxScaler` for Normalization.	Scikit-Learn provides a transformer called `StandardScaler` for standardization.
6.	This transformation squishes the n-dimensional data into an n-dimensional unit hypercube.	It translates the data to the mean vector of original data to the origin and squishes or expands.
7.	It is useful when we don't know about the distribution	It is useful when the feature distribution is Normal or Gaussian.
8.	It is a often called as Scaling Normalization	It is a often called as Z-Score Normalization.

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

If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables.

$$\mathrm{VIF} = \frac{1}{1 - R_i^2}$$

A **rule of thumb** for interpreting the variance inflation factor:

- 1 = not correlated.
- Between 1 and 5 = moderately correlated.
- Greater than 5 = highly correlated.

To solve this, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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

As the name suggests, this plot is used to determine the normal distribution of errors. It uses standardized values of residuals. Ideally, this plot should show a straight line. If you find a curved, distorted line, then your residuals have a non-normal distribution (problematic situation).

Importance of Q-Q plot:

When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified. If so, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale. If two samples do differ, it is also useful to gain some understanding of the differences. The q-q plot can provide more insight into the nature of the difference than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample tests.