1. Assignment-based Subjective Questions

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

Ans :

- As per the season is concerned, the descending order of the booking is as follows:
  Fall > summer > winter > spring
- As per the season is concerned, maximum booking is done for the month between May and October and minimum booking is observed for the rest months.
- Clear weather gives maximum booking and as the weather deteriorates, the booking becomes lesser.
- As per the weekday is concerned, last four days, i.e., Thu, Fri, Sat and Sun book a greater number of bikes as compared to the first three days of the week.
- Holiday shows less booking.
- Working day has no effect on bike booking as working day and non-working day give similar result.
- 2019 shows more booking compared to 2018.

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

Ans:

- It is used to get k-1 dummies out of k categorical levels by removing the first level.
- It is important to reduce the columns in the dataset in order to reduce correlations created among dummy variables.

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

Ans:

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

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

Ans:

- By using 'sns.distplot' over residuals we can assume the residuals are normally distributed with mean = 0.
- By using 'LinearRegression.fit' and 'RFE.fit' we can assume the relation between Feature variables and dependent variable is linear in nature.
- By dropping few variables (temp,hum, etc) multicollinearity can be eliminated and assume the rest variables are in insignificant multicollinearity state.
- By using 'plt.scatter' over residuals and y_train value, we can assume there is no visible pattern in residuals.

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

Ans:

- June Month
- August Month
- September Month

2. General Subjective Questions

2.1 Explain the linear regression algorithm in detail.

Ans :

linear regression is a statistical technique to understand the relationship between one dependent variable and one or several independent variables (Feature variables).
The objective of Linear regression is to find a linear equation that can best determine the value of dependent variable Y for different values independent variables in X.
Mathematically the relationship can be represented with the help of following equation:

$$Y = mx + C$$

Here, 'Y' is the dependent variable that we are trying to predict.
'x' is the independent variable that we are using to make predictions.
'm' is the slope of the regression line which represents the effect 'x' has on 'Y'
'c' is the Y interception which gives the result of 'Y' when all 'x' = 0.
Therefore, in order to find the best fitted line, the following process is followed step by step.
Step 1: Check the correlations between all feature variables (x) with the dependent variable (Y) using sns.heatmap.
Step 2: Pick the most correlated variable only and build the model with one variable only using statsmodels.api.OLS over the train dataset.
Step 3: Observe R-squared value of the model, p-value of all feature variables using statsmodels.api.OLS().summery() and VIF using variance_inflation_factor
Step 4: Add another variable which gives a better correlation results than all other variables present in step 1.
Step 5: Repeat step 2,3,4 until all feature variables are considered by either dropping the variable or accepting it.
Step 6: Once final model has been established, conclude the model by a linear equation as follows:

$$Y = m_1x_1 + m_2x_2 + \cdots \ldots \ldots \ldots + m_nx_n \ + C$$

Where m_i = Coefficient of ith variable, x_i = ith variable

2.2 Explain the Anscombe's quartet in detail.

Ans:

Anscombe's quartet is a collection of four datasets that, when plotted as a scatter plot on a graph, have various interpretations while having similar descriptive statistical qualities in terms of means, variance, R-Squared, correlations, and linear regression lines. The datasets were developed by statistician Francis Anscombe in 1973 to highlight the value of data visualisation and to illustrate how summary statistics by themselves may be deceptive.

Anscombe's quartet consists of four datasets, each of which has 11 x-y pairings of data. Each dataset seems to have a different relationship between x and y when it is plotted, with various variability patterns and correlation strengths. In spite of these differences, each dataset has the identical summary statistics, including the same x and y mean and variance, x and y correlation coefficient, and linear regression line.

The significance of exploratory data analysis and the pitfalls of relying just on summary statistics are demonstrated using Anscombe's quartet. It also highlights how vital data visualisation is for identifying patterns, outliers, and other critical features that may not be readily apparent from summary statistics alone.

	x1	x2	x3	x4	y1	y2	y3	y4
0	10	10	10	8	8.04	9.14	7.46	6.58
1	8	8	8	8	6.95	8.14	6.77	5.76
2	13	13	13	8	7.58	8.74	12.74	7.71
3	9	9	9	8	8.81	8.77	7.11	8.84
4	11	11	11	8	8.33	9.26	7.81	8.47
5	14	14	14	8	9.96	8.10	8.84	7.04
6	6	6	6	8	7.24	6.13	6.08	5.25
7	4	4	4	19	4.26	3.10	5.39	12.50
8	12	12	12	8	10.84	9.13	8.15	5.56
9	7	7	7	8	4.82	7.26	6.42	7.91
10	5	5	5	8	5.68	4.74	5.73	6.89

Fig 1: Four Datasets

	I	II	III	IV
Mean_x	9.000000	9.000000	9.000000	9.000000
Variance_x	11.000000	11.000000	11.000000	11.000000
Mean_y	7.500909	7.500909	7.500000	7.500909
Variance_y	4.127269	4.127629	4.122620	4.123249
Correlation	0.816421	0.816237	0.816287	0.816521
Linear Regression slope	0.500091	0.500000	0.499727	0.499909
Linear Regression intercept	3.000091	3.000909	3.002455	3.001727

Fig 2: Statistical Results of Four Datasets

From the Fig 1, we can see there are four different datasets but they give us the similar statistical results shown in Fig 2. Therefore we can assume the datasets are identical in behavior. But data visualization using scatter plot tell us the different story as shown in Fig 3.

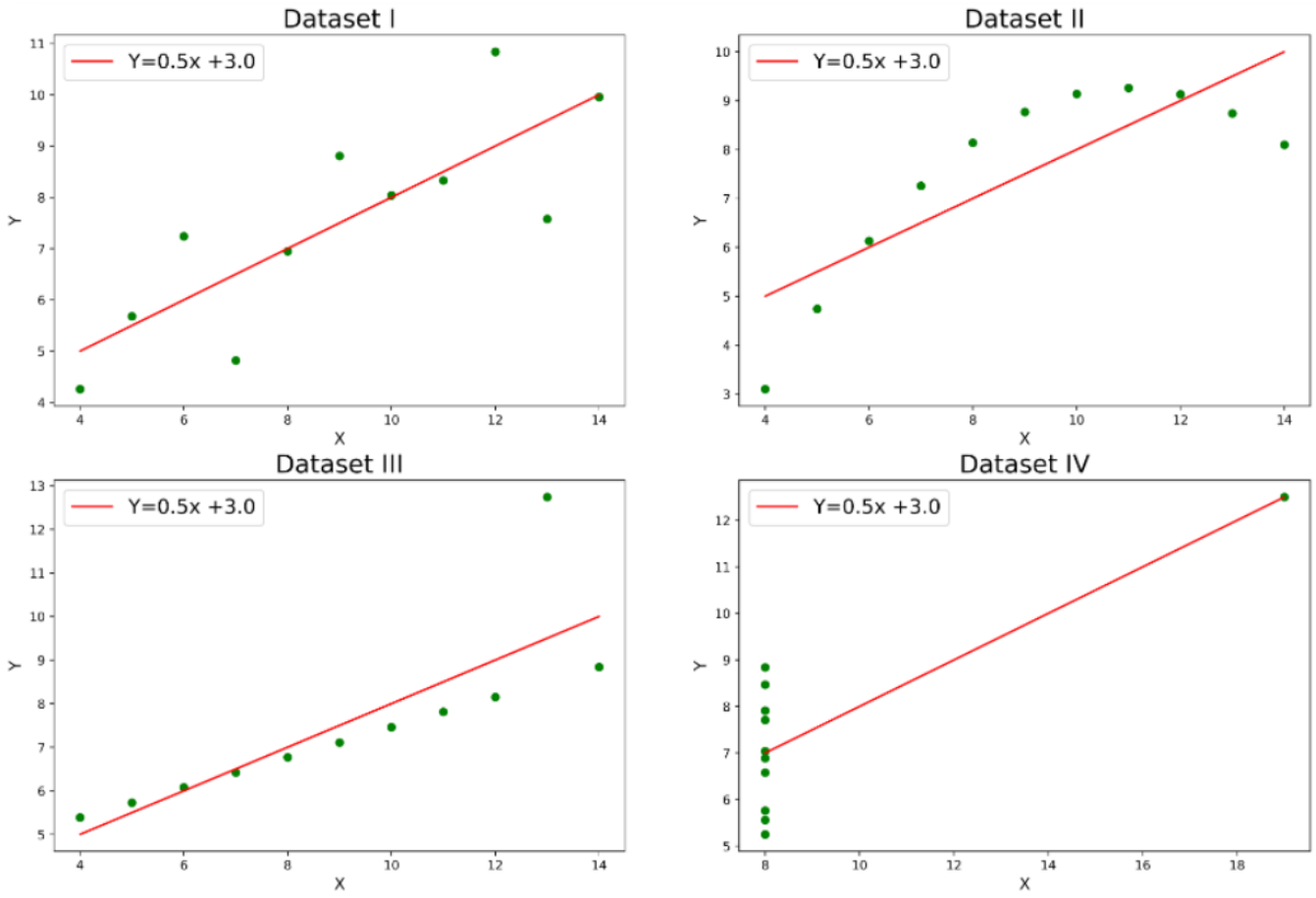

Fig 3: Visualisation of Four Datasets

In the first one(top left of Fig 3) if you look at the scatter plot you will see that there seems to be a linear relationship between x and y.
In the second one(top right of Fig 3) if you look at this figure you can conclude that there is a non-linear relationship between x and y.
In the third one(bottom left of Fig 3) you can say when there is a perfect linear relationship for all the data points except one which seems to be an outlier which is indicated be far away from that line.
Finally, the fourth one(bottom right of Fig 3) shows an example when one high-leverage point is enough to produce a high correlation coefficient.

The significance of visualisation in data analysis is emphasised by this quartet. A thorough understanding of the dataset's structure may be obtained by looking at the data.

2.3 What is Pearson's R?

Ans:

A numerical evaluation of the strength of the linear connection between the variables is provided by Pearson's r. The correlation coefficient will be positive if the variables tend to rise and fall together. The correlation coefficient will be negative if the variables have a tendency to rise and fall in opposition, with low values of one variable correlated with high values of the other.

Between +1 and -1 are the possible values for the Pearson correlation coefficient, or r. There is no link between the two variables, as shown by a value of 0. A number larger than 0 denotes a positive connection, meaning that when one variable's value rises, the value of the other variable also rises. A result that is less than 0 denotes a negative connection, meaning that when one variable's value rises, the value of the other variable falls. As may be seen in the diagram below:

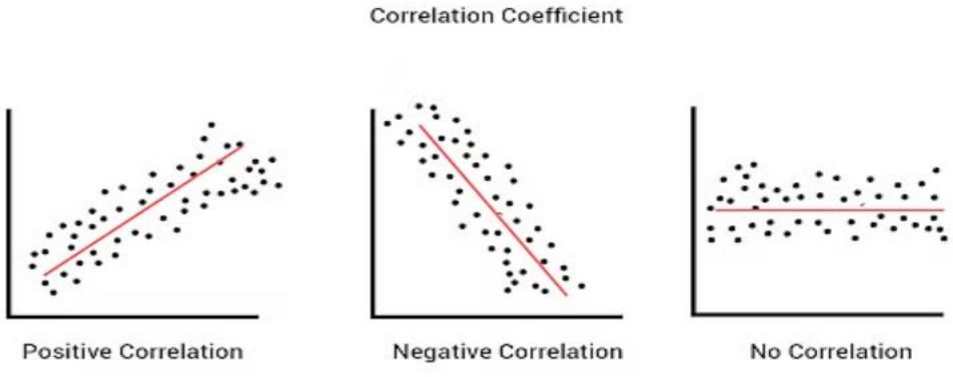

Fig 4: Scatter Plots for different Correlations

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

Ans:

- By The technique of normalising the variety of characteristics in a dataset is known as feature scaling. Real-world datasets frequently include characteristics that vary in size, scope, and units. We must thus do feature scaling in order for machine learning models to comprehend these characteristics on the same scale.
- The reason behind the scaling in dataset is as follows:
    1. Ease of interpretation
    2. Faster convergence for gradient descent methods

- Difference between Normalized and Standardized Scaling:

Normalization	Standardization
This method scales the model using minimum an d maximum values.	This method scales the model using the mean and standar d deviation.
When features are on various scales, it is functio nal.	When a variable's mean and standard deviation are both s et to 0, it is beneficial.
Values on the scale fall between [0, 1] and [-1, 1].	Values on a scale are not constrained to a particular range.
Additionally known as scaling normalization.	This process is called Z-score normalization.
When the feature distribution is unclear, it is helpful.	When the feature distribution is consistent, it is helpful.

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

Ans:

A high VIF score denotes a strong correlation between the variables. Therefore, VIF = ∞ means there is a highest possible relation between the feature variable with the dependent variable while building the model.

When R-squared (R2) is 1 , then $VIF = \frac{1}{1-R2} = \frac{1}{0} = \infty$, which represents a very strong relationship between feature variables and dependent variable.

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

Ans:

- A graphical approach for detecting whether two samples of data come from the same population or not is the quantile-quantile plot. The quantiles of the first data set are shown against the quantiles of the second data set in a q-q graphic. A quantile is the percentage of points that fall below the specified number.
- The quantiles of the first data set are shown against the quantiles of the second dataset in a q-q figure. A quantile is the percentage of points that fall below the specified number. In other words, the 0.3 (or 30%) quantile is the value at which 30% of the data are below it and 70% are above it. Additionally, a 45-degree reference line is drawn. The points should roughly lie along this reference line if the two sets are drawn from a population with the same distribution. The further the two data sets deviate from this reference line, the more evidence there is that they came from populations with distinct distributions.
- It is frequently desirable to determine if the assumption of a common distribution is supported when there are two data samples. If so, location and scale estimators can combine the two sets of data to derive estimates for the shared position and scale. If two samples do differ, it is also helpful to comprehend the variations. More information about the nature of the difference may be gleaned from the q-q plot than from analytical techniques like the chi-square and Kolmogorov-Smirnov 2-sample tests. the two data sets have originated from populations with various distributions, leading to the conclusion that.