Assignment-based Subjective Questions:

Ques 1: From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?
Answer: Have used pair bar plot for doing my analysis on categorical variables/columns like season, months, weekday etc. Below are some points from the visualizations:

- Wed, Thu, Fri and Sat have a greater number of bookings as compared to other week days.
- Most of the bookings have been done during the month of Jun, Jul, Aug and Sep. Also seems the trend increasing starting the year and decreasing coming end of the year.
- Clear weather also attracts more bookings as per the graphs.
- Year 2019 have a greater number of bookings from previous year which also a good progress as per business.
- Booking also seems equals on working and no-working days.
- Fall season also have a greater number of bookings then summer comes second.

Ques 2: Why is it important to use drop_first=True during dummy variable creation?
Answer: drop_first=True is useful because it reduces the extra number of columns, created during dummy variable creation. Which is useful in sense of reducing correlations created among dummy variables.

Hence if we have categorical variable with n-levels, then we need to use n-1 columns to represent the dummy variables.

In case we have 3 types of Categorical column values and we want to create dummy variable. If one variable is not X and Y, then It is obvious Z. So, we do not need 3rd variable to identify the Z.

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

Ques 4: How did you validate the assumptions of Linear Regression after building the model on the training set?
Answer: I have validated the assumptions of Linear Regression after building the model on the training set as below:

- **Multicollinearity check:** It is essential to detect and deal with the multicollinearity present in the model. So always should have an insignificant multicollinearity between variables.
- **Normality of Error terms:** Error terms should be normally distributed with mean zero.
- **Homoscedasticity:** There should be no visible pattern in residual values.
- **Independence of residuals:** Should have no auto-correlation
- **Linear relationship validation:** Linearity should be visible among variables.

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

i) temp
ii) sep
iii) Fall

General Subjective Questions:

Ques 1: Explain the linear regression algorithm in detail.

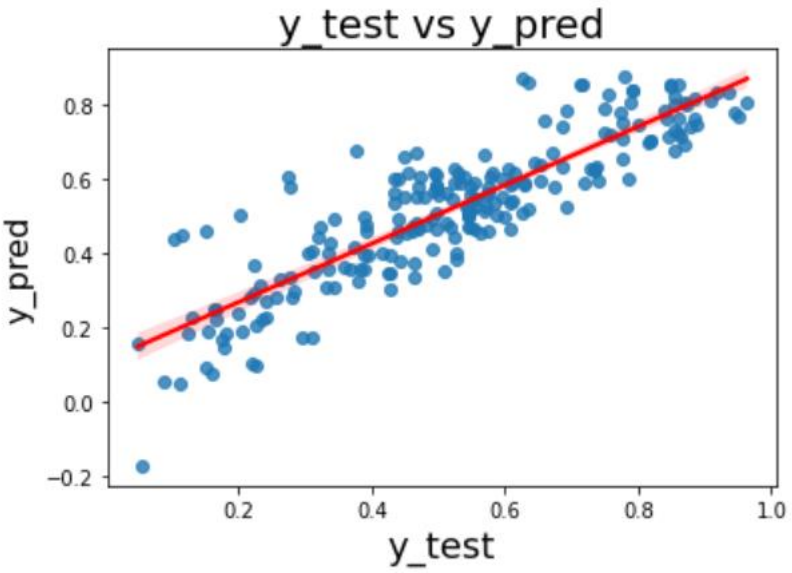
Answer: Linear Regression is a supervised machine learning algorithm where the predicted output is continuous and has a constant slope. It's used to predict values within a continuous range, (e.g. sales, price) rather than trying to classify them into categories (e.g. red, green).

Mathematically the relationship can be represented with the help of following equation:
Y = mX + c

Here, Y is the dependent variable we are trying to predict.
X is the independent variable we are using to make predictions.
m is the slope of the regression line which represents the effect X has on Y
c is a constant, known as the Y-intercept. If X = 0, Y would be equal to c.

Linear regression models can be classified into two types depending upon the number of independent variables:

i) Simple linear regression: When the number of independent variables is 1
ii) Multiple linear regression: When the number of independent variables is more than 1



Further the relationship can be Positive or Negative.

- ➔ A linear relationship will be called **positive** if both independent and dependent variable increases.
- ➔ A linear relationship will be called **negative** if independent increases and dependent variable decreases.

Ques 2: Explain the Anscombe's quartet in detail.

Answer: It was developed by statistician Francis Anscombe to illustrate the importance of plotting the graphs before analysing and model building, and the effect of other observations on statistical properties. It comprises 4 datasets, each contains eleven (x, y) pairs.

Anscombe's Quartet can be defined as a group of four data sets which are nearly identical in simple descriptive statistics, but there are some peculiarities in the dataset that fools the regression model if built. They have very different distributions and appear differently when plotted on scatter plots.

Anscombe's Data										
Observation	x1	y1		x2	y2		x3	y3	x4	y4
1	10	8.04		10	9.14		10	7.46	8	6.58
2	8	6.95		8	8.14		8	6.77	8	5.76
3	13	7.58		13	8.74		13	12.74	8	7.71
4	9	8.81		9	8.77		9	7.11	8	8.84
5	11	8.33		11	9.26		11	7.81	8	8.47
6	14	9.96		14	8.1		14	8.84	8	7.04
7	6	7.24		6	6.13		6	6.08	8	5.25
8	4	4.26		4	3.1		4	5.39	19	12.5
9	12	10.84		12	9.13		12	8.15	8	5.56
10	7	4.82		7	7.26		7	6.42	8	7.91
11	5	5.68		5	4.74		5	5.73	8	6.89
				Summary Statistics						
N	11	11		11	11		11	11	11	11
mean	9.00	7.50		9.00	7.500909		9.00	7.50	9.00	7.50
SD	3.16	1.94		3.16	1.94		3.16	1.94	3.16	1.94
r	0.82			0.82			0.82		0.82	

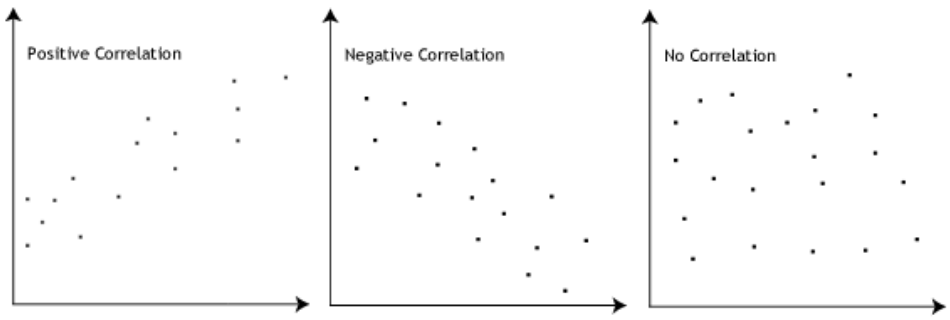
The four datasets can be described as:

- Dataset 1: this fits the linear regression model pretty well.
- Dataset 2: this could not fit linear regression model on the data quite well as the data is non-linear.
- Dataset 3: shows the outliers involved in the dataset which cannot be handled by linear regression model
- Dataset 4: shows the outliers involved in the dataset which cannot be handled by linear regression model

Ques 3: What is Pearson's R?

Answer: Pearson correlation coefficient is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

It is a statistic that measures the linear correlation between two variables. Like all correlations, it also has a numerical value that lies between -1.0 and +1.0. Value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as the value of one variable increases, the value of the other variable decreases. This is shown in the diagram below:



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

Answer: 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.

Example: If an algorithm is not using feature scaling method then it can consider the value 3000 meter to be greater than 5 km but that's actually not true and, in this case, the algorithm will give wrong predictions. So, we use Feature Scaling to bring all values to same magnitudes and thus, tackle this issue.

Normalized scaling	Standardized scaling
Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
Really affected by outliers	Much less affected by outliers
It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.

Ques 5: You might have observed that sometimes the value of VIF is infinite. Why does this happen?
Answer: In vase if there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity.

When the value of VIF is infinite it shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R-squared (R2) =1, which lead to 1/ (1-R2) infinity. To solve this, we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

Ques 6: What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
Answer: Quantile- Quantile plot (Q-Q plot) is a graphical technique for determining if two data sets come from populations with a common distribution.

Use of Q-Q plot:
A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second dataset. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value. A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.

The advantages of the Q-Q plot are:

1) The sample sizes do not need to be equal.

2) Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot. For example, if the two data sets come from populations whose distributions differ only by a shift in location, the points should lie along a straight line that is displaced either up or down from the 45-degree reference line.