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?

Utilizing a boxplot, I have conducted analysis on categorical columns. The following are a few conclusions we can get from the visualisation:
• Time of year: Fall is the busiest season for renting bikes
• Until June, demand for rental bikes is rising each month.
The peak month for demand is September, after which it starts to decline.
• There isn't a lot of difference in demands during the weekdays and working days.
Why Clear weather forecasts are in high demand.
• Demand has fallen over holidays.
• I observe a rise in demand for the next year.

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

Use of drop first = True is crucial since it aids in eliminating the excess column produced when a dummy variable is formed. As a result, it lessens the connections that dummy variables cause.
Drop first: bool, defaulting to False, indicates whether to remove the first level from the k category levels in order to obtain k-1 dummies.
Let's imagine we want to build a dummy variable for a categorical column that has three different types of data. If one factor is neither A nor B, then it is clear that C. Thus, we do not require the third variable to locate the C.

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

The desired variable and the variable 'temp' have the strongest association.

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

Using data from the Linear Regression Model, I have validated the assumption that

assuming the following five:

• Normality of error terms

Error terms ought to have a normal distribution.

• Check for multicollinearity –

There should be little multicollinearity between the variables.

There shouldn't be any discernible patterns in the residual values due to homoscedasticity.

• No auto-correlation and independence of residuals

• Verification of linear relationships – Linearity between variables should be apparent.

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

Here are the top 3 characteristics that greatly contribute to explaining the demand for shared bikes:

• temperature

• winter season

• September

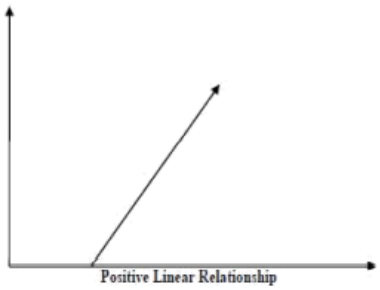
General Subjective Questions

1. Explain the linear regression algorithm in detail.

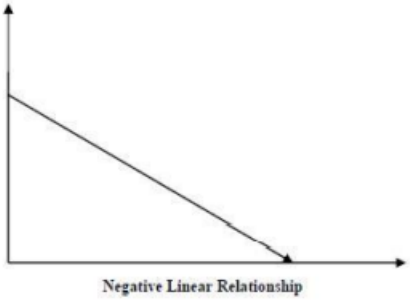
The statistical model known as linear regression examines the linear connection between a dependent variable and a set of independent variables. When there is a linear relationship between two variables, the values of the dependent and independent variables both change when one or more of the independent variables' values change (increase or decrease).

The relationship can be mathematically stated using the following equation:
$Y = mX + c$
Y is the dependent variable in this case, and our goal is to predict it.
X serves as the independent variable on which we base our forecasts.
The regression line's slope, m, represents the impact X has on Y, and the Y-intercept, c, is a constant. Y would be equal to c if X = 0.

Additionally, as will be shown below, the linear relationship might be either positive or negative in character.
.A positive linear relationship If both the independent and dependent variables rise, the linear connection is said to be positive.
The graph below will help you understand it.



• Negative Linear Relationship: If the independent variable rises and the dependent variable falls, the relationship is said to be positive.
The graph below will help you understand it.



The two types of linear regression are:
Multiple linear regression
simple linear regression

Assumptions - The linear regression model makes the following assumptions on the dataset:

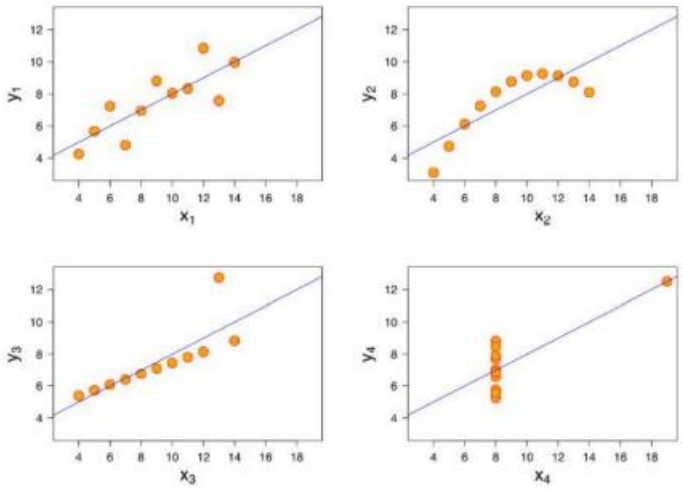
- Multi-collinearity - The linear regression model makes the assumption that the data exhibits very little to no multi-collinearity. Multi-collinearity basically happens when independent variables or features depend on one another.
- Auto-correlation -The linear regression model also makes the assumption that the data exhibits very little to no auto-correlation. In essence, auto-correlation happens when residual errors are dependent on one another.
- The linear regression model presupposes that the connection between the response and the feature variables must be linear.
- Error terms should have a regularly distributed distribution.
- Homoscedasticity - The residual values should not have any discernible pattern.

2. Explain the Anscombe's quartet in detail.

Francis Anscombe, a statistician, created Anscombe's Quartet.
It consists of four datasets with eleven (x, y) pairings each. The fact that both datasets share the same descriptive statistics is the most important thing to keep in mind. However, when something is graphed, it completely—and I mean completely—changes. Regardless of the fact that their summary statistics are comparable, each graph has a unique narrative to tell.

	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

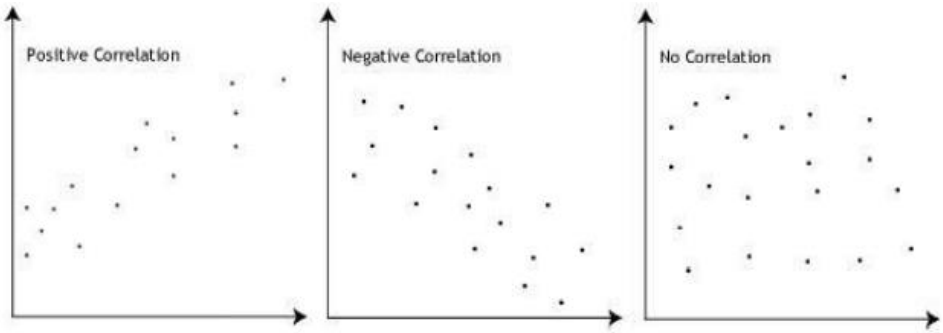
According to the summary statistics, x and y's means and variances were the same for all groups for both x and y:
For each dataset, the means of x and y are 9,7.50 respectively.
• For each dataset, the variance of x is 11 and the variance of y is 4.13.
• For each dataset, the correlation coefficient (a measure of the strength of a relationship between two variables) between x and y is 0.816.
These four datasets display the identical regression lines when we plot them on an x/y coordinate plane, but each dataset tells a different narrative.



• The linear models in Dataset I seem to be clear and well-fitting.
• Dataset II is not regularly distributed.
• Although Dataset III's distribution is linear, an outlier causes the estimated regression to be incorrect.
• Dataset IV demonstrates that a high correlation coefficient can be obtained with just one outlier.
The significance of visualisation in data analysis is emphasised by this quartet. A thorough understanding of the dataset's structure can be obtained by looking at the data.

3. What is Pearson's R?

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.



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

Feature scaling is a method for uniformly distributing the independent features in the data over a predetermined range. It is done as part of the pre-processing of the data to deal with extremely variable magnitudes, values, or units.
In the absence of feature scaling, a machine learning algorithm will typically prioritise larger values over smaller ones, regardless of the unit of measurement.

Example: If an algorithm does not use the feature scaling method, it may assume that a value of 3000 metres is greater than a value of 5 kilometers, which is not the case and causes the algorithm to produce inaccurate predictions. To solve this problem, we employ feature scaling to equalise all values' magnitudes.

S.NO.	Normalized scaling	Standardized scaling
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.

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

VIF = infinity if there is perfect correlation. A high VIF score denotes a strong connection between the variables.

The presence of multicollinearity causes the variance of the model coefficient to be exaggerated by a factor of 4 if the VIF is 4.

VIF displays a complete correlation between two independent variables when its value is infinite. If the correlation is perfect, we obtain R-squared (R2) = 1, which results in 1/ (1-R2) infinity. In order to resolve this, we must remove the variable from the dataset that is the source of this perfect multicollinearity.

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

A graphical method for assessing if two data sets originate from populations with a common distribution is the quantile-quantile (q-q) plot.

Q-Q plot application:

A q-q plot compares the quantiles of the first data set to those of the second data set. 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 plotted. The points should roughly lie along this reference line if the two sets are drawn from a population with the same distribution. The more the two data sets deviate from this reference line, the more evidence there is that they came from populations with different distributions.

Importance of the Q-Q plot:

It is frequently desirable to determine whether the presumption 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 can be gleaned from the q-q plot than from analytical techniques like the chi-square and Kolmogorov-Smirnov 2-sample tests.