Assignment-based Subjective Questions

Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: <Your answer for Question 1 goes below this line> (Do not edit)

The following could be inferred by visualizing the box plot for categorical columns:

1. Fall season has more booking among all the other seasons.
2. Year 2019 has more bookings than 2018.
3. Clear weather attracts more booking.
4. Months- from may to oct has more bookings done.
5. More booking has been found on Sat and Thurs.

Question 2. Why is it important to use **drop_first=True** during dummy variable creation? (Do not edit)

Total Marks: 2 marks (Do not edit)

Answer: <Your answer for Question 2 goes below this line> (Do not edit)

Drop first=True is important to remove the redundant column. Even though we drop a column
It could be easily derived by seeing the rest of the dummy values.

Example:
Male - 0
Female -1
if I drop the male column I could derived the other female column.

Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)

Total Marks: 1 mark (Do not edit)

Answer: <Your answer for Question 3 goes below this line> (Do not edit)

temp has the highest correlation.

Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: <Your answer for Question 4 goes below this line> (Do not edit)

I have validated the following:

1.Normality of error Term

For this I have plotted the histogram which shows the normal distribution.

2.Multi-collinearity

I have checked the correlation and found it is as insignificant among variables.

3.Linear Relationship:

I have plotted the scatter plot to check the linearity using component residual plot. There are patterns that are visible.

4.Homoscedasticity:

There were no visible patterns found when I plotted scatter as well as plot.

Question 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 5 goes below this line> (Do not edit)

1.Temp
2.Winter
3.Sep

General Subjective Questions

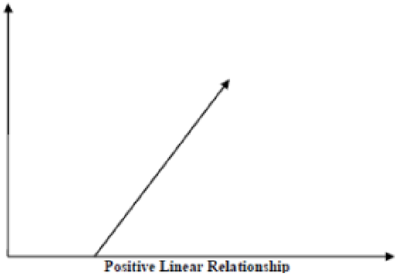
Question 6. Explain the linear regression algorithm in detail. (Do not edit)
Total Marks: 4 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Linear regression is a statistical model that examines the linear relationship between a dependent variable and a set of independent variables. A linear relationship indicates that when the value of one or more independent variables changes (either increasing or decreasing), the dependent variable will also change correspondingly.
This relationship can be mathematically expressed with the following equation:
Y=mX+cY = mX + cY=mX+c

In this equation:
Y represents the dependent variable we aim to predict.
X is the independent variable used for making predictions.
m denotes the slope of the regression line, reflecting the impact of XXX on YYY.
c is a constant known as the Y-intercept, which indicates the value of YYY when XXX equals zero.

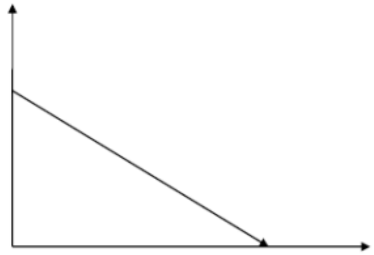
Additionally, the linear relationship can be either positive or negative.

- Positive Linear Relationship: This occurs when both the independent and dependent variables increase together. This concept can be illustrated with an appropriate graph.



Positive Linear Relationship

- Negative Linear relationship: A linear relationship will be called positive if independent increases and dependent variable decreases. It can be understood with the help of following graph –



Negative Linear Relationship

Linear regression can be categorized into two types:

- **Simple Linear Regression**
- **Multiple Linear Regression**

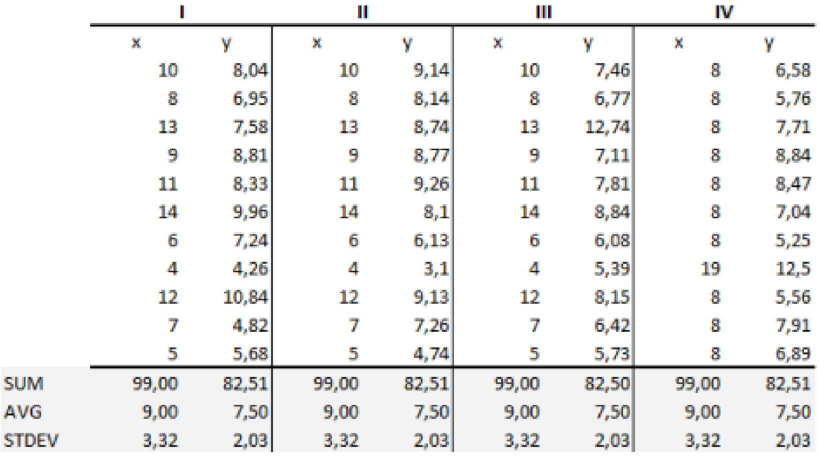
Assumptions

The following assumptions are made by the linear regression model regarding the dataset:

- **Multicollinearity**: The model assumes there is minimal or no multicollinearity, meaning the independent variables are largely independent of each other. Multicollinearity occurs when features are interdependent.
- **Autocorrelation**: The model assumes minimal or no autocorrelation in the data. Autocorrelation refers to a dependency between residual errors.
- **Linear Relationship**: The model assumes a linear relationship between the response variable and predictor variables.
- **Normality of Error Terms**: The error terms should follow a normal distribution.
- **Homoscedasticity**: Residual values should not show any specific pattern or trend.

Question 7. Explain the Anscombe's quartet in detail. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Anscombe's Quartet was created by statistician Francis Anscombe and consists of four datasets, each containing eleven pairs of (x, y) values. Remarkably, all four datasets share the same summary statistics. However, once plotted, each dataset reveals a completely different pattern, with each graph telling a unique story despite their identical statistical summaries.

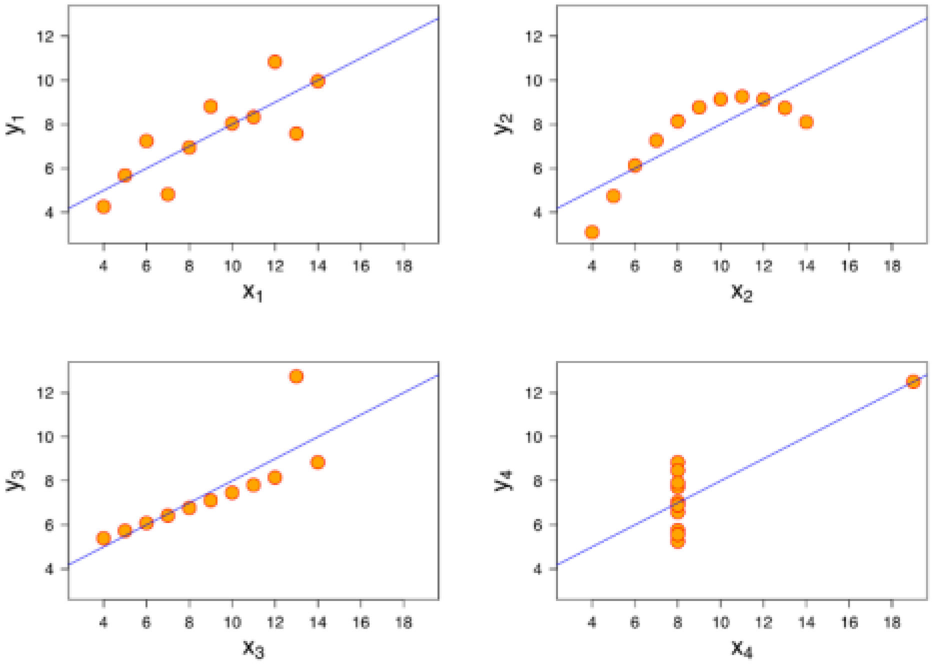


	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	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 indicate that each dataset in Anscombe's Quartet has identical means and variances for both x and y:

- The mean of x is 9, and the mean of y is 7.50 across all datasets.
- The variance of x is 11, and the variance of y is 4.13 for each dataset.
- The correlation coefficient between x and y is 0.816 across all datasets, reflecting the strength of the relationship between the two variables.

However, when we plot these four datasets on an x/y coordinate plane, we see that each dataset has the same regression line but conveys a distinctly different pattern and story.



- Dataset I shows a clear and well-fitting linear pattern.
- Dataset II has a non-normal distribution.
- In Dataset III, the distribution is mostly linear, but an outlier skews the regression line.
- Dataset IV illustrates how a single outlier can result in a high correlation coefficient.

Anscombe's Quartet highlights the critical role of visualization in data analysis. Visualizing the data provides insight into its structure, revealing patterns that summary statistics alone cannot capture.

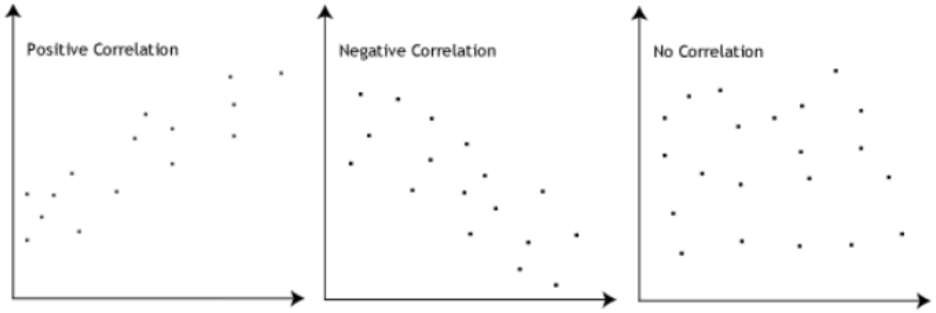
Question 8. What is Pearson's R? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Pearson's correlation coefficient (r) provides a numerical measure of the strength and direction of the linear relationship between two variables. If both variables increase or decrease together, rrr will be positive. Conversely, if one variable tends to increase as the other decreases, rrr will be negative.

The Pearson correlation coefficient ranges from +1 to -1:

- +1 indicates a perfect positive linear relationship.
- -1 indicates a perfect negative linear relationship.
- 0 means there is no linear relationship between the variables.

An r value greater than 0 suggests a positive association—when one variable increases, so does the other. An r value less than 0 suggests a negative association—when one variable increases, the other decreases. This relationship is illustrated in the diagram below:



Question 9. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

Feature scaling is a method used to standardize the independent features in a dataset within a specific range. This process is crucial during data preprocessing, especially when dealing with features that have varying magnitudes, values, or units. Without feature scaling, machine learning algorithms may disproportionately prioritize larger values over smaller ones, leading to inaccurate interpretations and predictions.

For instance, if an algorithm does not implement feature scaling, it might mistakenly consider 3000 meters as greater than 5 kilometers, which is incorrect. By applying feature scaling, we ensure that all values are brought to a similar magnitude, helping to avoid such miscalculations and improving the overall accuracy of predictions.

S.NO.	Normalized Scaling	Standardized Scaling
1	Uses the minimum and maximum values of features for scaling.	Uses the mean and standard deviation for scaling.
2	Applied when features have different scales.	Applied to achieve a zero mean and unit standard deviation.
3	Scales values to a range of [0, 1] or [-1, 1].	Not limited to a specific range.
4	Highly affected by outliers.	Much less impacted by outliers.
5	Scikit-Learn has a transformer called MinMaxScaler for normalization.	Scikit-Learn has a transformer called StandardScaler for standardization.

Question 10. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

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.

Question 11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

The quantile-quantile (Q-Q) plot is a graphical tool used to assess whether two datasets likely originate from populations with a similar distribution.

Purpose of a Q-Q Plot:

A Q-Q plot compares the quantiles of one dataset against those of a second dataset. A quantile represents the proportion (or percentage) of points below a specific value. For instance, the 0.3 (or 30%) quantile is the value below which 30% of the data falls, with the remaining 70% above it. A 45-degree reference line is also included in the plot. If the datasets share a common distribution, the points will generally align with this reference line. The more the points deviate from this line, the stronger the indication that the datasets come from populations with different distributions.

Importance of the Q-Q Plot:

When comparing two data samples, it's often useful to determine if they share a common distribution. If they do, estimates for location and scale can be combined to improve accuracy. If they differ, the Q-Q plot can reveal how they differ, providing more detailed insights than purely analytical methods like the chi-square or Kolmogorov-Smirnov two-sample tests.
