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? (3 marks)**

   Answer:

   I've analyzed the categorical columns using box plots and bar plots. Here are some key observations from the visualizations:

   - The fall season appears to have garnered more bookings, with a significant increase in bookings from 2018 to 2019 across all seasons.
   - The majority of bookings occurred between May and October, showing a trend of increasing bookings from the beginning of the year until mid-year, followed by a decrease towards the year's end.
   - It's evident that clear weather conditions led to more bookings, which is understandable.
   - Thursdays, Fridays, Saturdays, and Sundays saw higher booking numbers compared to the start of the week.
   - Bookings tended to be lower on non-holiday days, which is reasonable as people may prefer to stay home and spend time with family during holidays.
   - There was a relatively equal distribution of bookings between working and non-working days.
   - 2019 saw a higher number of bookings compared to the previous year, indicating positive progress in terms of business.

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

   Answer:

   Utilizing drop_first = True is crucial as it minimizes the redundancy generated during the creation of dummy variables, thereby reducing correlations among them.

   The syntax for drop_first is: bool, default False, indicating whether to obtain k-1 dummies out of k categorical levels by excluding the first level.

   For instance, if we have three types of values in a categorical column and aim to generate dummy variables for that column, if one variable is not A and B, it is inherently C. Thus, there's no necessity for the third variable to identify C.

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

   Answer:

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

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

   Answer:

   I've assessed the validity of the assumptions for the Linear Regression Model, considering the following five criteria:

   - Normality of error terms: Ensuring that error terms follow a normal distribution.
   - Multicollinearity check: Verifying the absence of significant multicollinearity among variables.
   - Linear relationship validation: Confirming the presence of linearity among variables.
   - Homoscedasticity: Ensuring that there's no discernible pattern in residual values.
   - Independence of residuals: Verifying the absence of autocorrelation among residuals.

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

   Answer:

   The following are the top three features that significantly contribute to explaining the demand for shared bikes:

   - temp
   - winter
   - sep

General Subjective Questions

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

   Answer:

   Linear regression is a statistical model that examines the linear relationship between a dependent variable and a given set of independent variables. In this context, a linear relationship between variables implies that as the value of one or more independent variables changes (increases or decreases), the value of the dependent variable changes correspondingly (increases or decreases).

   Mathematically the relationship can be represented with the help of following equation −

   $Y = mX + c$

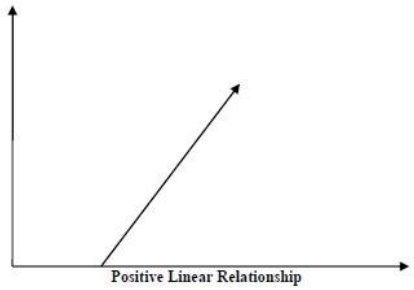
   Here, Y is the dependent variable we are trying to predict. X is the independent variable we are using to make predictions.

   m is the slope of the regression line which represents the effect X has on Y

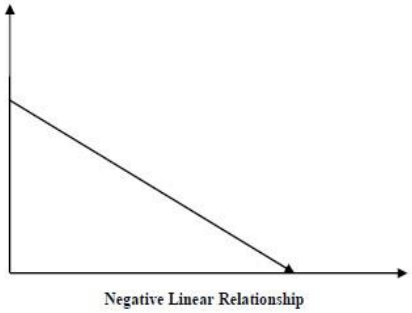
   c is a constant, known as the Y-intercept. If X = 0, Y would be equal to c.

   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. It can be understood with the help of following graph −



- 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 −



Linear regression is of the following two types −

- Simple Linear Regression
- Multiple Linear Regression

Assumptions -

Linear Regression models rely on several assumptions regarding the dataset:

1. Multi-collinearity: The model assumes minimal or no multi-collinearity in the data. Multi-collinearity arises when independent variables or features exhibit interdependence.

2. Auto -correlation: Another assumption is minimal or no auto-correlation in the data. Auto-correlation occurs when there is a correlation between residual errors, indicating a pattern in the data that the model has not captured.

3. Relationship between variables: Linear Regression assumes a linear relationship between response and feature variables. This means that the relationship between the dependent variable and independent variables should be approximately linear.

4. Normality of error terms: Error terms should follow a normal distribution.
5. Homoscedasticity: Residual values should show no discernible pattern.

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

Answer:

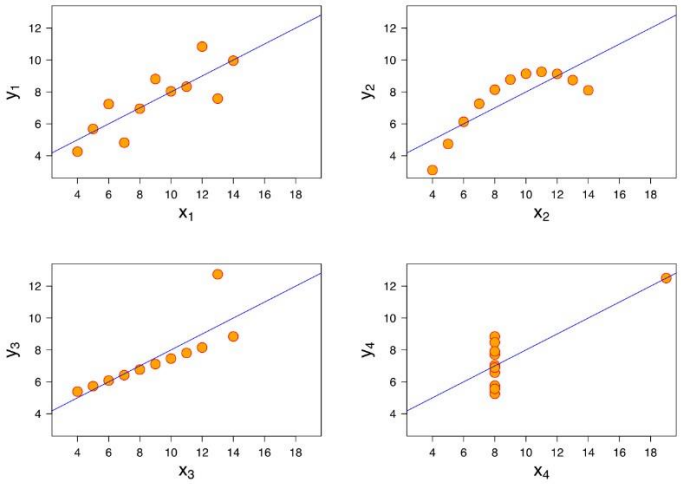
Anscombe's Quartet, created by statistician Francis Anscombe, consists of four datasets, each comprising eleven (x, y) pairs. Remarkably, despite sharing identical descriptive statistics, the interpretation drastically shifts when graphed. Each graph presents a unique narrative, completely distinct from 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.

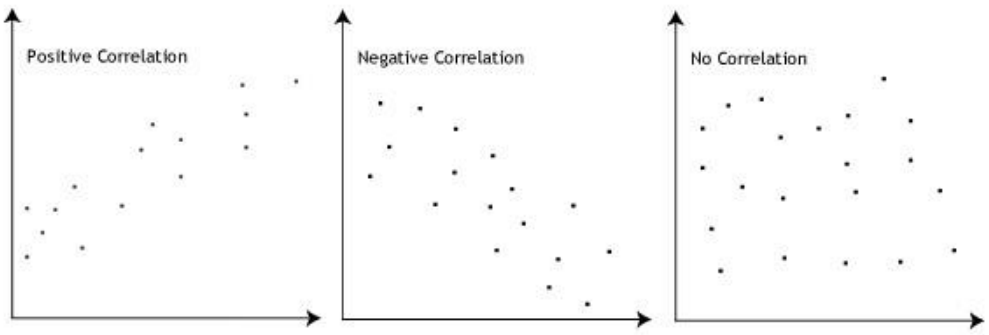
This quartet emphasizes the importance of visualization in Data Analysis. Looking at the data reveals a lot of the structure and a clear picture of the dataset.

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

Answer:

Pearson's r 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.

The Pearson correlation coefficient, r, can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables. A 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:



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

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.

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? (3 marks) Answer:**

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.

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

Answer:
The quantile-quantile (q-q) plot is a visual tool used to assess whether two datasets originate from populations with a similar distribution.

Usage of Q-Q plot:
A q-q plot compares the quantiles of one dataset against those of another. Quantiles represent the fraction or percentage of data points below a given value. For instance, the 0.3 quantile signifies the point where 30% of data falls below it. Typically, a 45-degree reference line is included. If both datasets share a common distribution, points should align closely along this line. Deviation from the reference line indicates differences in distribution between the datasets.

Significance of Q-Q plot:
When comparing two data samples, it's vital to assess whether they share a common distribution. If they do, combining them can improve estimation of common location and scale parameters. Conversely, recognizing differences in distributions is crucial for understanding data disparities. Q-Q plots offer deeper insights into these differences compared to traditional statistical tests like chi-square and Kolmogorov-Smirnov tests.