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 have used the boxplot and bar plot to analyze categorical data. The following are a few conclusions we may get from the visualization:

- The fall seems to have drawn additional reservations. And, from 2018 to 2019, the number of bookings in each season dramatically climbed.
- Most reservations were made in the months of May, June, July, August, September, and October. Beginning in January and continuing through mid-year, the trend grew before beginning to decline as the year came to a close.
- It appears evident that more bookings were lured by clear weather.
- There are more reservations on Thursday, Friday, Saturday, and Sunday than at the beginning of the week.
- Bookings appear to be lower when it's not a holiday, which makes sense given that during holidays, individuals would prefer to stay home and enjoy time with their families.
- Booking seems to be much the same whether it was a working day or not.
- The number of reservations for 2019 increased over the prior year, indicating positive company growth.

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

   Answer:

   The 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 sorts 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 C.

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

   Answer:

The target variable and ‘temp' variables are most correlated showing a correlation of 0.63

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

   Answer:

   Based on the following five assumptions, I have verified the linear regression model's underlying premise. -

   Error terms should have a regularly distributed distribution.

   Check for multicollinearity

   o Multicollinearity between variables should be negligible.

   Validating linear relationships

   o Variables' linearity should be apparent

   Homoscedasticity

   o The residual values should not exhibit any discernible pattern.

   Independent residuals

   o Absence of autocorrelation

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:

   Spring: -2344.5444

   Light rain_Light snow_Thunderstorm: -2532.5726

   yr (year): 2075.7239

   These coefficients indicate the impact of each feature on the demand for shared bikes. Spring and weather conditions like light rain, light snow, and thunderstorms have negative effects on bike demand. On the other hand, the year (2019 in this case) has a positive impact, showing an increase in demand from the previous year.

General Subjective Questions

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

   Answer:

   The statistical model known as linear regression examines the linear connection between a dependent variable and a collection of independent variables. According to the linear connection between variables, the value of the dependent variable will vary proportionally (increase or decrease) when the value of one or more independent variables changes. Mathematically the relationship can be represented with the help of the following equation −

   $$Y = mX + c$$

   Here, Y is the dependent variable we are trying to predict.

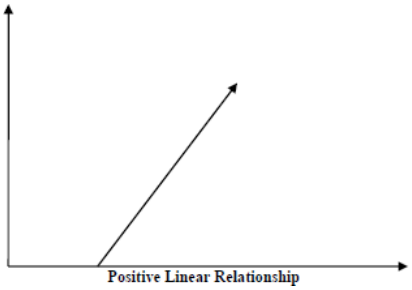
X is the independent variable we are using to make predictions.

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.

Additionally, as will be shown below, the linear connection might be either positive or negative in character.

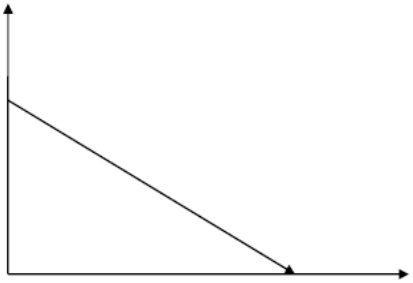
- **Positive Linear connection:**
  If both the independent and dependent variables rise, the linear connection is said to be positive. The graph below will help you understand it.



Positive Linear Relationship

- **Negative Linear connection:**

If the independent variable rises while the dependent variable falls, the connection is said to be positive. The graph below will help you understand it.



Negative Linear Relationship

Linear regression is of the following two types −

- Simple Linear Regression
- Multiple Linear Regression

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

Multi-collinearity:

The linear regression model is predicated on the premise that the data exhibit very little to no multi-collinearity. Multi-collinearity basically happens when independent variables or characteristics depend on one another.

Autocorrelation –

Another assumption of the linear regression model presupposes that the data exhibits either very little or no autocorrelation. 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 feature variables must be linear.

Normality of error terms

Error terms should be normally distributed

Homoscedasticity –There should be no visible pattern in residual values.

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

Answer:

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 anything is graphed, it completely—and I mean completely—changes. Each graph conveys a distinct message.

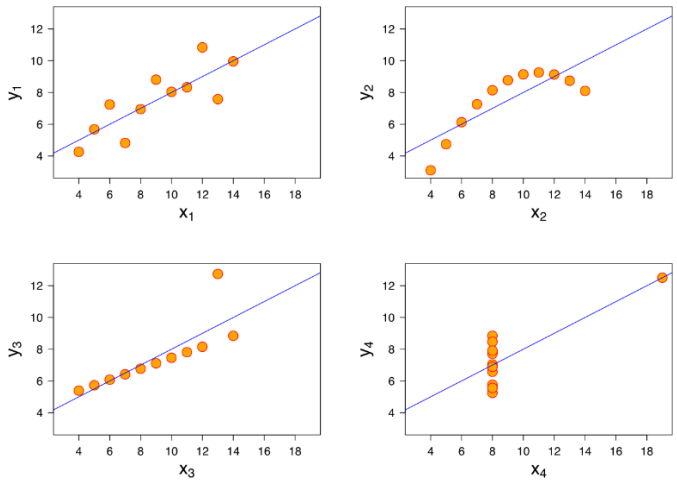
	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	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, 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 link between two variables) between x and y is 0.816.



The four datasets display the same regression lines when we plot them on an x/y coordinate plane, but each one has a different interpretation:

• 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 may be obtained with just one outlier.

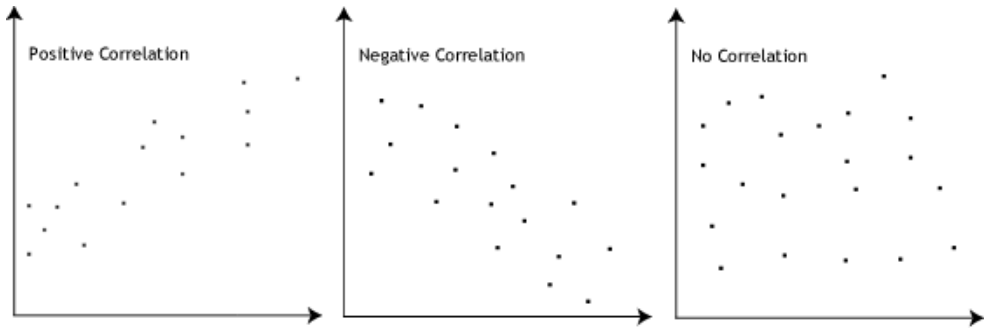
The significance of visualization in data analysis is emphasized by this quartet. A thorough understanding of the dataset's structure may be obtained by looking at the data.

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

Answer:

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. Positive associations have values larger than 0, meaning that if one variable's value rises, so does the value of the other. 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. The illustration below demonstrates this:



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 method for uniformly distributing the independent characteristics 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 would often prioritize larger values over smaller ones, regardless of the unit of measurement.

Example: If an algorithm does not use feature scaling, it may assume that a value of 3000 meters is bigger than a value of 5 kilometers, even if this is not the case. In this scenario, the algorithm will offer incorrect forecasts. To solve this problem, we employ feature scaling to equalize all values' magnitudes.

S.NO.	Normalized scaling	Standardized scaling
1.	The minimum and maximum values offeatures are used for scaling	Mean and the standard deviation is used forscaling.
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:

VIF = infinite if there is a perfect correlation. A high VIF score denotes a strong connection between the variables. The existence 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 have R-squared (R2) = 1, which results in 1/ (1-R2) infinite. To fix this, we must remove the variable from the dataset that is the exact multicollinearity's cause.

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

(3 marks)

Answer:

A graphical method for assessing if two data sets originate from populations with a similar 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 drawn. The points should roughly lie along this reference line if the two sets are drawn from a population with the same distribution. The more this reference line deviates, the stronger the evidence. The two data sets have originated from populations with various distributions, leading to the conclusion that.

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 may be gleaned from the q-q plot than from analytical techniques like the chi-square and Kolmogorov-Smirnov 2-sample tests.