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? (3marks)**

   Answer:

   I have used the boxplot and bar plot to analyse categorical columns. The few conclusions we can get from the visualization are listed below. –

   I. The fall seems to have drawn additional reservations. And, from 2018 to 2019, the number of bookings in each season significantly grew.
   II. 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.
   III. It appears obvious that more bookings were lured by clear weather.
   IV. When it's not a holiday, bookings appear to be less frequent, which makes sense given that during holidays, individuals may prefer to spend time at home and enjoy time with family. Thursday, Friday, Saturday, and Sunday had more bookings than the beginning of the week.
   V. Booking seems to be much the same whether it was a working day or not.
   VI. The amount of reservations for 2019 increased over the prior year, which indicates positive business growth.

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

   Answer:

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? (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 have validated the assumption of Linear Regression Model based on below 5 assumptions -

   - **Normality of error terms**
     - **Error terms should be normally distributed**
   - **Multicollinearity check**
     - **There should be insignificant multicollinearity among variables.**
   - **Linear relationship validation**
     - **Linearity should be visible among variables**
   - **Homoscedasticity**
     - **There should be no visible pattern in residual values.**

- **Independence of residuals**
    - **No auto-correlation**

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:

   Below are the top 3 features contributing significantly towards explaining the demand of the shared bikes –

   - temp
   - winter
   - sep

General Subjective Questions

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

   Answer:

   Linear regression may be defined as the statistical model that analyses the linear relationship between a dependent variable with given set of independent variables. Linear relationship between variables means that when the value of one or more independent variables will change (increase or decrease), the value of dependent variable will also change accordingly (increase or decrease).

   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.

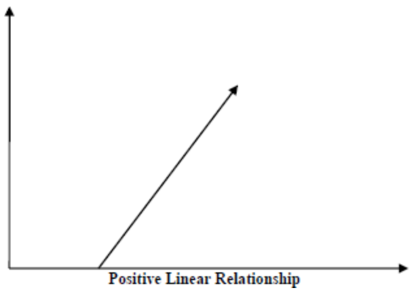
   X is the independent variable we are using to make predictions.

   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.

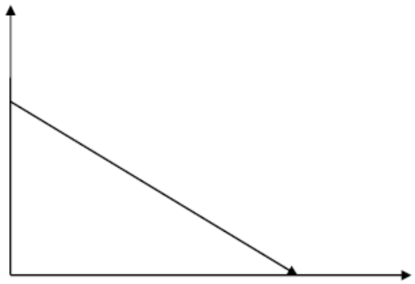
   Furthermore, the linear relationship can be positive or negative in nature as explained below−

   - 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 −



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 is of the following two types –

- Simple Linear Regression
- Multiple Linear Regression

Assumptions -

The following are some assumptions about dataset that is made by Linear Regression model –

- Multi-collinearity –
    - Linear regression model assumes that there is very little or no multi-collinearity in the data. Basically, multi-collinearity occurs when the independent variables or features have dependency in them.
- Auto-correlation –
    - Another assumption Linear regression model assumes is that there is very little or no auto-correlation in the data. Basically, auto-correlation occurs when there is dependency between residual errors.
- Relationship between variables –
    - Linear regression model assumes that the relationship between 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.

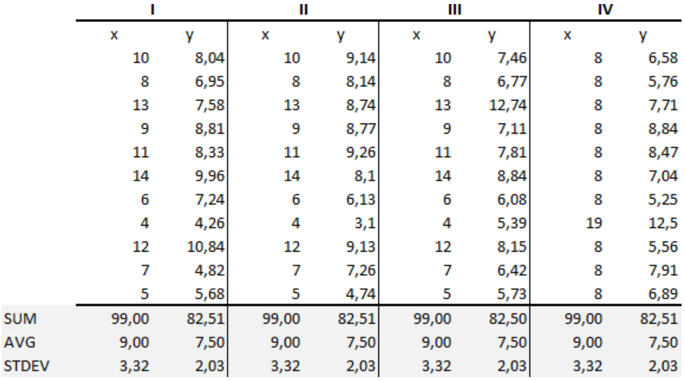
2. Explain the Anscombe's quartet in detail. (3 marks)

Answer:

Anscombe's Quartet was developed by statistician Francis Anscombe. 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.

The summary statistics show that the means and the variances were identical for x and y across the groups:

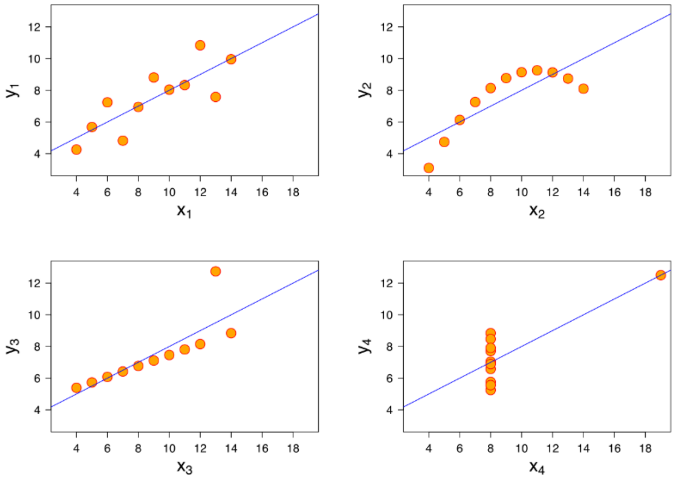
- Mean of x is 9 and mean of y is 7.50 for each dataset.



	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 correlation coefficient (how strong a relationship is between two variables) between x and y is 0.816 for each dataset
- Similarly, the variance of x is 11 and variance of y is 4.13 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 II is not distributed normally.
- In Dataset III the distribution is linear, but the calculated regression is thrown off by an outlier.
- Dataset I appears to have clean and well-fitting linear models.
- Dataset IV shows that one outlier is enough to produce a high correlation coefficient.

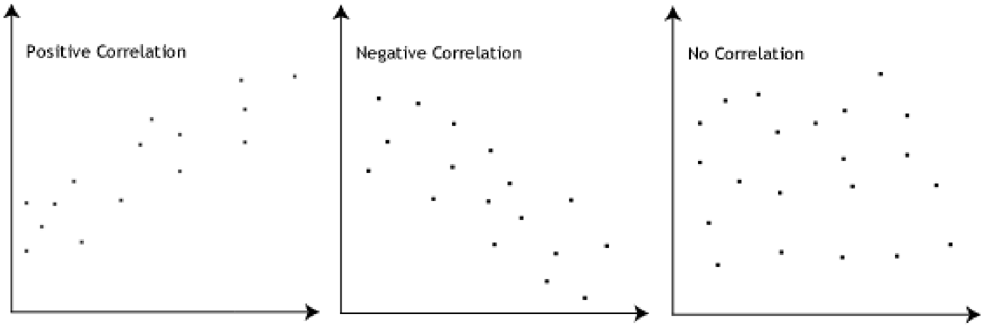
This quartet emphasizes the importance of visualization in Data Analysis. Looking at the data reveals a lot of the structure and a clear picture of the dataset.

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

Answer:

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 indicated by a value of 0. Positive associations have values greater 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 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 would often prioritise 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 metres is bigger than a value of 5 kilometres, even though this is not the case. In this scenario, the algorithm will offer incorrect 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 really affected by outliers.	It is much less affected by outliers.
3.	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
4.	Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.
5.	It is used when features are of different scales.	It is used when we want to ensure zeromean and unit standard deviation.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Answer:

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 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 common distribution is the quantile-quantile (q-q) plot.

Use of Q-Q plot:

The quantiles of the first data set are plotted against the quantiles of the second dataset in a q-q figure. 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 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 can be gleaned from the q-q plot than from analytical techniques like the chi-square and Kolmogorov-Smirnov 2-

sample tests.