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 done analysis on categorical columns using the boxplot and bar plot. Below are the few: points we can infer from the visualization –

- Bookings appear to have increased over the fall season. And, from 2018 to 2019, the number of bookings in each season significantly grew.
- Most bookings 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 obvious that more bookings were attracted by clear weather.
- Bookings are higher on Thursday, Friday, Saturday, and Sunday than at the beginning of the week.
- Bookings appear to be less frequent when it's not a holiday, which makes sense given that during holidays, people would want to stay home and enjoy time with their families.
- Both working days and non-working days looked to have about the same amount of bookings.
- 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? (2 mark)**

   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.

Syntax -

drop_first: bool, default False, which implies whether to get k-1 dummies out of k categorical levels by removing the first level.

Let's imagine we want to build a dummy variable for a categorical column that has three different types of data. If one of the factors is neither A nor B, it is obvious that C. As a result, we don't need the third variable to find 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 is a statistical model that investigates the linear connection between a dependent variable and a set of independent variables. A linear relationship between variables means that as the value of one or more independent variables changes (increases or decreases), the value of the dependent variable changes (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.

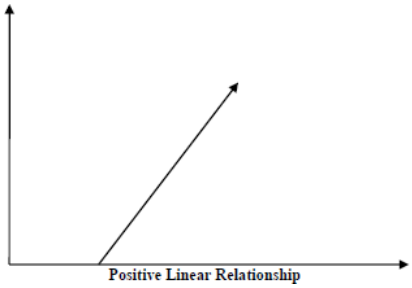
   m is the slope of the regression line which represents the effect

   X has on Y c is a constant, known as the Y-intercept. If X = 0, Y

   would be equal to c.

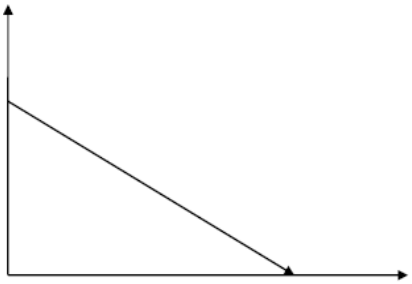
   Furthermore, the linear relationship can be positive or negative in nature as explained below–

   - Positive Linear Relationship:
     - A linear relationship is said to be positive if both the independent and dependent variables rise. The following graph can help you understand it.



Positive Linear Relationship

- Negative Linear relationship:
  - If the independent variable increases while the dependent variable decreases, the linear connection is said to be positive. The following graph can help you understand it.



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 –

✔ Multicollinearity –

- o The linear regression model assumes that the data has minimal or no multicollinearity. Multicollinearity develops when the independent variables or features are reliant on one another.

✔ Auto-correlation –

- o Another supposition The linear regression model assumes that the data has little or no auto-correlation. Auto-correlation arises when residual errors are dependent on one another.

✔ Relationship between variables –

- o The linear regression model implies a linear relationship between response and feature variables.

✔ Normality of error terms –

o Error terms should be normally distributed

✔ Homoscedasticity –

o 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, each with eleven (x, y) pairings. The most important thing to remember about these datasets is that they all use the same descriptive

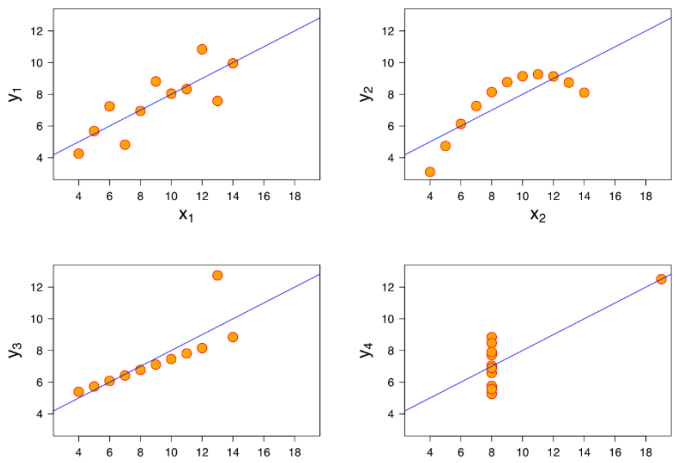
	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15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

statistics. But when things are graphed, they alter totally, and I emphasise fully. Regardless of the similarities in their summary statistics data, each graph delivers a different picture. data, each graph conveys a different tale.

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.

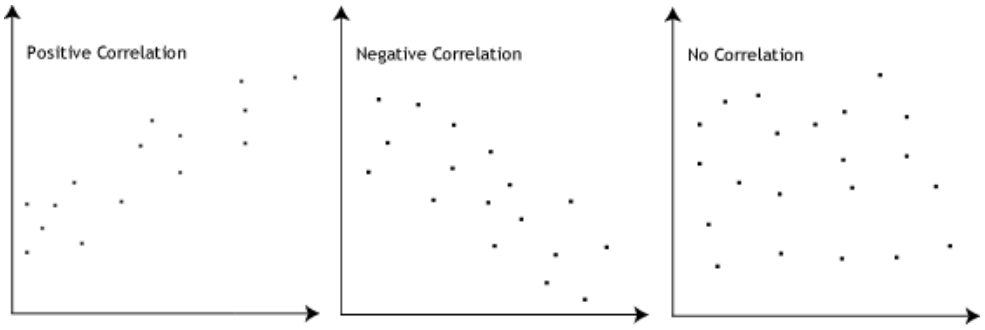
This quartet emphasizes the importance of visualization in Data Analysis. Looking at the data reveals a lot of the structure and a clear picture of the dataset.

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

Answer:

Pearson's r is a numerical representation of the strength of the linear relationship between the variables. The correlation coefficient will be positive if the variables tend to rise and fall together. The correlation coefficient will be negative if the variables tend to go up and down in opposite directions, with low values of one variable correlated with high values of the other.

The Pearson correlation coefficient, r, can range between +1 and -1. A value of 0 implies that no relationship exists between the two variables. A number greater than 0 implies a positive connection; that is, as the value of one variable rises, the value of the other variable rises as well. A number less than 0 implies a negative connection, which means that as the value of one variable rises, the value of the other variable falls. This is depicted 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 for standardizing the independent features inherent

in data within a specific range. It is used during data pre-processing to deal with drastically changing magnitudes, values, or units. If feature scaling is not performed, a machine learning algorithm will tend to weight greater values as higher and consider smaller values as lower, regardless of the unit of measurement.

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:

VIF = infinity if there is perfect correlation. A high VIF number suggests that there is a relationship between the variables. If the VIF is 4, it signifies that multicollinearity has inflated the variance of the model coefficient by a factor of four.

When the value of VIF is infinite, the correlation between two independent variables is perfect. In the event of perfect correlation, R-squared (R2) = 1, resulting in 1/ (1-R2) infinity. To address this, we must remove one of the variables from the dataset that is producing the 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 graphical tool for detecting whether two data sets are from the same population.

Use of Q-Q plot:

A q-q plot is a comparison of the first data set's quantiles to the quantiles of the second data set. A quantile is the fraction (or percentage) of points that fall below a specific number. That is, the 0.3 (or 30%) quantile is the number at which 30% of the data falls below and 70% falls above. There is also a 45-degree reference line plotted. If the two sets are drawn from the same population, the points should fall roughly along this reference line. The larger the deviation from this reference line, the stronger the evidence that the two data sets came from populations with distinct distributions.

Importance of Q-Q plot:

It is frequently desirable to determine if the assumption of a common distribution is warranted when there are two data samples. If this is the case, then location and scale estimators can combine both data sets to generate estimates of the common location and scale. If two samples differ, it is also beneficial to get an understanding of the discrepancies. Analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample tests can provide greater insight into the nature of the difference than the q-q plot.