Assignment-based Subjective Questions

1. **From your analysis of the categorical variables from the dataset, what could you infer abouttheir effect on the dependent variable? (3 marks)**
   Answer:
   The analysis on categorical columns was done using the boxplot . Some fewpoints which we can infer about their effect on dependent variable are –
   a. Fall season has attracted more booking. And, in each season the booking count has increased drastically from 2018 to 2019.
   b. Most of the bookings has been done during the month of may, june, july, aug, sep and oct. Trend increased starting of the year till mid of the year and then it started decreasing as we approached the end of year.
   c. Clear weather attracted more number of bookings .
   d. Thu, Fir, Sat and Sun have more number of bookings as compared to the start of the week.
   e. On holidays, people may want to spend time at home and enjoywith family, so less number of bookings happened on holidays .
   f Booking is almost equal either on working day or non-working day.
   g In 2019 more number of booking happened as compared to the previous year, which shows a good progress in terms of business.

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

When creating dummy variables from categorical variables, setting drop_first=True is an important consideration to prevent multicollinearity and to improve the interpretability of the regression model. Here's why it's important:

Multicollinearity: When using one-hot encoding to create dummy variables, if you include all levels of a categorical variable as dummy variables, you introduce multicollinearity. Multicollinearity occurs when two or more predictor variables in a regression model are highly correlated. This can lead to issues such as unstable coefficient estimates and difficulty in interpreting the individual impact of each variable. By dropping the first level (using drop_first=True), you create independent dummy variables, reducing multicollinearity.

Interpretability: When you create dummy variables for a categorical variable with 'm' levels, you generally create 'm-1' dummy variables. This is because if you have 'm' dummy variables, they will be perfectly multicollinear, and the model won't be able to distinguish the effect of each level. By dropping one level and creating 'm-1' dummy variables, you provide a baseline level for comparison, making it easier to interpret the coefficients of the remaining dummy variables.

Avoiding the "Dummy Variable Trap": The "Dummy Variable Trap" refers to the situation when the dummy variables are highly correlated and linearly dependent, leading to multicollinearity. By dropping the first level, you avoid this trap and ensure that the dummy variables are linearly independent.

In summary, using drop_first=True during dummy variable creation is important to prevent multicollinearity, improve model interpretability, and avoid the "Dummy Variable Trap." It aligns with the assumptions of regression analysis and helps ensure the stability and accuracy of your model's results.

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 -

   a. Normality of error terms
      i. Error terms should be normally distributed
   b. Multicollinearity check
      i. There should be insignificant multicollinearity among variables.
   c. Linear relationship validation
      i. Linearity should be visible among variables
   d. Homoscedasticity
      i. There should be no visible pattern in residual values.
   e. Independence of residuals
      i. 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 −

   $Y = mX + c$

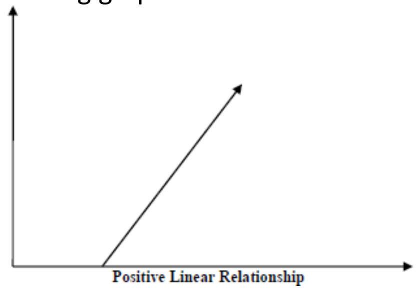
   Here, Y is the dependent variable we are trying to predict.

   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.

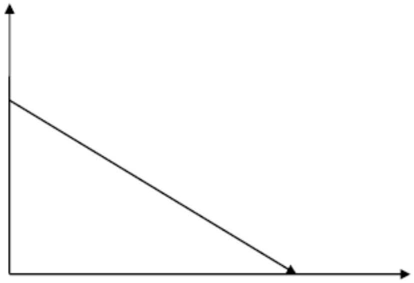
Furthermore, the linear relationship can be positive or negative in nature as explained below−

- 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 –

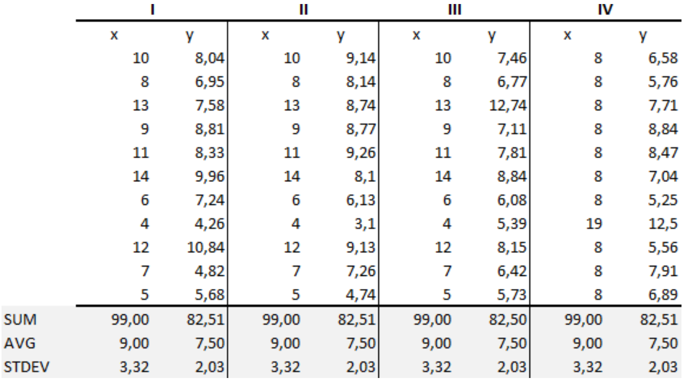
- Linear regression model assumes that the relationship between response and feature variables must be linear.

- ✓ 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 comprises four datasets, each containing eleven (x, y) pairs. The essential thing to note about these datasets is that they share the same descriptive statistics. But things change completely, and I must emphasize COMPLETELY, when they are graphed. Each graph tells a different story irrespective of their similar summary statistics.

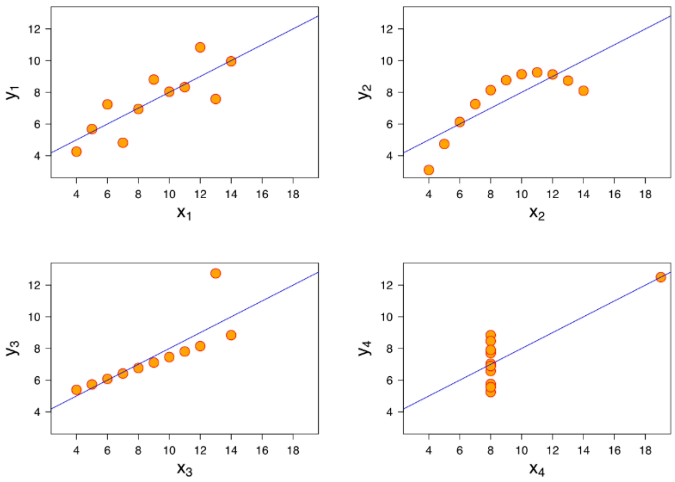


	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15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

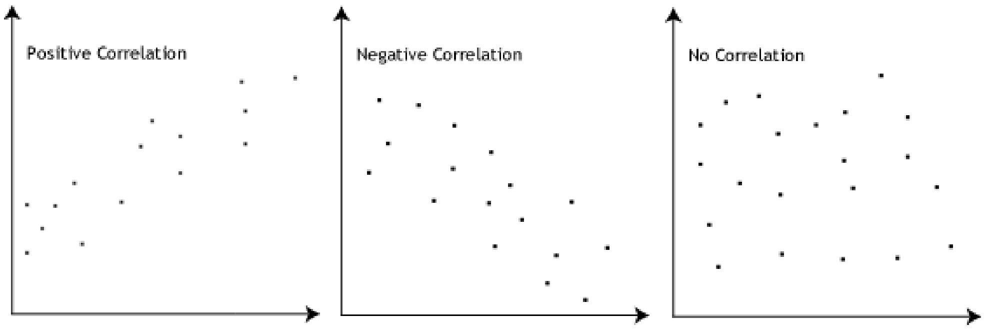
This quartet emphasizes the importance of visualization in Data Analysis. Looking at the data reveals a lot of the structure and a clear picture of the dataset.

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:

A Q-Q plot (Quantile-Quantile plot) is a graphical tool used to assess whether a given dataset follows a specific theoretical distribution, typically the normal distribution. It's a way to visually compare the quantiles of the sample data against the quantiles of a chosen theoretical distribution.
In a Q-Q plot, the x-axis represents the theoretical quantiles of the chosen distribution, while the y-axis represents the quantiles of the actual data. If the data follows the chosen distribution, the points in the Q-Q plot will approximately lie along a straight line.

Use and Importance in Linear Regression:
Assumption Checking: Q-Q plots are often used in linear regression to check the assumption of normality of residuals. Residuals are the differences between the observed values and the predicted values from the regression model. If the residuals are normally distributed, it indicates that the errors are random and unbiased, which is a key assumption of linear regression.

Detection of Departure from Normality: By comparing the points on the Q-Q plot to the straight line representing the normal distribution, you can detect deviations from normality. If the points deviate from the line in a systematic manner, it indicates that the residuals are not normally distributed. This could lead to issues with the validity of statistical tests and confidence intervals in linear regression.

Outlier Detection: Q-Q plots can also help in detecting outliers in the data. Outliers are data points that deviate significantly from the rest of the data. In a Q-Q plot, outliers will appear as points that deviate significantly from the expected straight line.

Model Validity: Ensuring that the residuals are normally distributed is crucial for the validity of the linear regression model's results. If the assumption of normality is not met, it could lead to incorrect parameter estimates, biased hypothesis tests, and inaccurate confidence intervals.

In summary, Q-Q plots are valuable tools in linear regression analysis for assessing the normality of residuals and identifying departures from this assumption. They provide a visual and quantitative way to validate the model and make necessary adjustments if the assumptions are not met.