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?

Answer:- I analyzed the categorical columns using boxplots and bar plots. Here are a few key insights we can draw from the visualizations:

- The fall season appears to have attracted more bookings, and in each season, the booking count has significantly increased from 2018 to 2019.
- Most bookings were made during the months of May, June, July, August, September, and October. The trend shows an increase from the beginning of the year until mid-year, followed by a decline as the year progressed toward the end.
- When it's not a holiday, bookings tend to be lower, which makes sense as people may prefer to stay home and spend time with their families during holidays.
- 2019 saw a higher number of bookings compared to the previous year, indicating strong business growth.
- Clear weather attracted more booking which seems obvious.
- Thursdays, Fridays, Saturdays, and Sundays have more bookings compared to the beginning of the week.
- Booking seemed to be almost equal either on working day or non-working day.

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

Answer:- The `drop_first = True` option is important because it helps reduce the number of extra columns created during dummy variable creation, which in turn minimizes correlations among the dummy variables.

Syntax:
`drop_first: bool, default False` — This parameter determines whether to generate `k-1` dummies from `k` categorical levels by removing the first level.

For example, if a categorical column has three values (A, B, and C), creating dummy variables would normally result in three columns. However, if a row is not A or B, it must be C, so we don't need a separate column for C. By using `drop_first = True`, we can omit the column for C, simplifying the dataset.

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

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?

Answer:- I have validated the assumption of Linear Regression Model based on below 5 assumptions -

- **Normality of error terms:-** The error terms should be normally distributed.
- **Multicollinearity Check:-** There should be minimal multicollinearity among the variables.
- **Linear Relationship Validation:-** A linear relationship should be evident among the variables.
- **Homoscedasticity:-** Residual values should not exhibit any visible patterns.
- **Independence of Residuals:-** There should be no autocorrelation in the residuals.

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

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

- winter
- sep
- temp

General Subjective Questions.

1. Explain the linear regression algorithm in detail.

Answer:- Linear regression is a statistical model that analyzes the linear relationship between a dependent variable and a given set of independent variables. A linear relationship means that changes in one or more independent variables will lead to corresponding changes in the dependent variable.

Mathematically, this relationship is represented by the following equation:

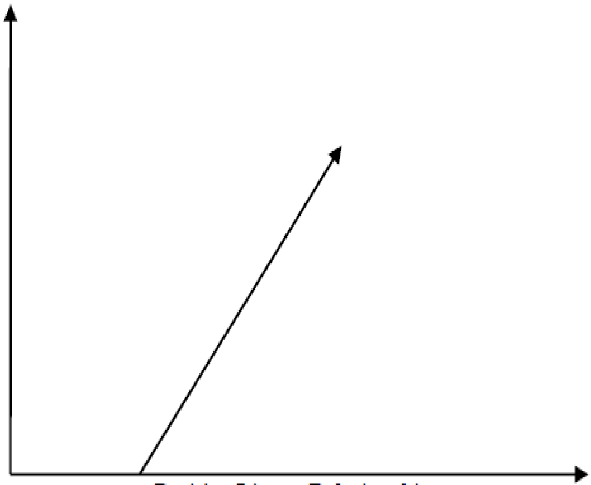
Y = mX + c

Where:

- Y is the dependent variable we aim to predict.
- X is the independent variable used for prediction.
- m is the slope of the regression line, indicating the impact of X on Y.
- c is a constant known as the Y-intercept, representing the value of Y when X is zero.

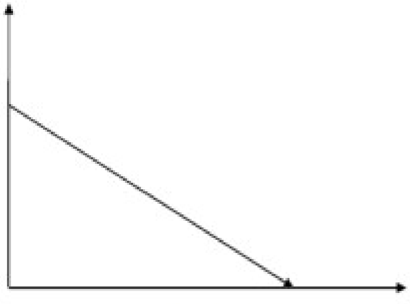
This linear relationship can be either positive or negative:

➔ **Positive Linear Relationship:-** If both the independent and dependent variables increase together, the relationship is positive. This is illustrated in the following graph:



Positive Linear Relationship

- ➔ **Negative Linear Relationship:-** If the independent variable increases while the dependent variable decreases, the relationship is negative. This is depicted in the following graph.



Negative Linear Relationship

Types of Linear Regression:- Linear regression can be categorized into two types.

- ◆ **Simple Linear Regression:** Involves one independent variable.
- ◆ **Multiple Linear Regression:** Involves two or more independent variables.

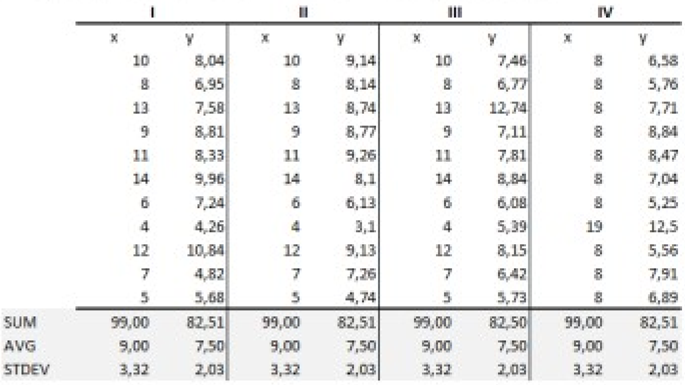
Assumptions of Linear Regression:- The linear regression model makes several assumptions about the dataset:

- **Multicollinearity:** The model assumes minimal or no multicollinearity in the data. Multicollinearity occurs when independent variables are highly dependent on each other.
- **Autocorrelation:** The model assumes minimal or no autocorrelation in the data. Autocorrelation occurs when there is a dependency between residual errors.
- **Relationship Between Variables:** The model assumes that the relationship between the response and feature variables is linear.
- **Normality of Error Terms:** The error terms should be normally distributed.
- **Homoscedasticity:** There should be no visible pattern in the residual values.

Each of these assumptions is crucial for the validity of the linear regression model, ensuring that the predictions made by the model are reliable and accurate.

2. Explain the Anscombe's quartet in detail.

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 must emphasize COMPLETELY, when they are graphed. Each graph tells a different story irrespective of their similar summary statistics.

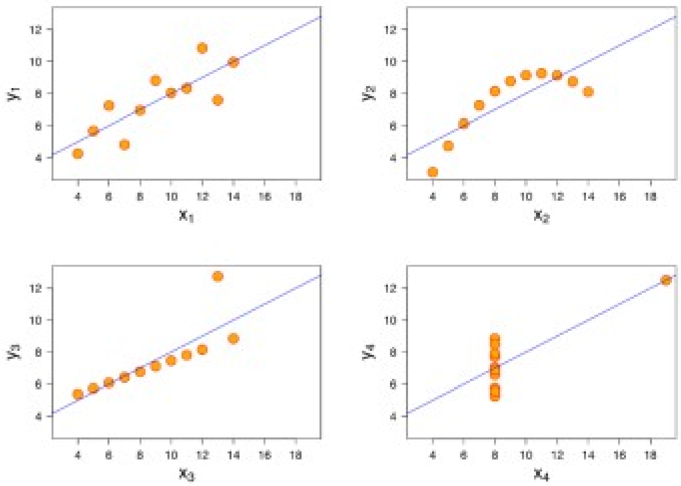


	I		II		III		IV	
	x	y	x	y	x	y	x	y
	10	8,04	10	9,14	10	7,46	8	6,58
	8	6,95	8	8,14	8	6,77	8	5,76
	13	7,58	13	8,74	13	12,74	8	7,71
	9	8,81	9	8,77	9	7,11	8	8,84
	11	8,33	11	9,26	11	7,81	8	8,47
	14	9,96	14	8,1	14	8,84	8	7,04
	6	7,24	6	6,13	6	6,08	8	5,25
	4	4,26	4	3,1	4	5,39	19	12,5
	12	10,84	12	9,13	12	8,15	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 reveal that the means and variances for both x and y are identical across all groups:

- The mean of x is 9, and the mean of y is 7.50 for each dataset.
- Similarly, the variance of x is 11, and the variance of y is 4.13 for each dataset.
- The correlation coefficient, which measures the strength of the relationship between x and y, is 0.816 for each dataset.

However, when we plot these four datasets on an x/y coordinate plane, although they display the same regression lines, each dataset tells a different story.



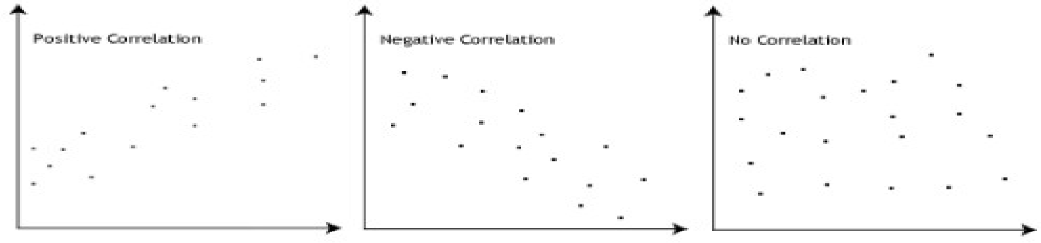
- Dataset I shows a clean, well-fitting linear model.
- Dataset II is not normally distributed.
- Dataset III has a linear distribution, but an outlier significantly affects the calculated regression.
- Dataset IV illustrates that a single outlier can result in a high correlation coefficient.

This quartet highlights the crucial role of visualization in data analysis, as examining the data visually can reveal the underlying structure and provide a clearer understanding of the dataset.

3. What is Pearson's R?

Answer:- Pearson's r provides a numerical summary of the strength of the linear relationship between two variables. If the variables tend to increase or decrease together, the correlation coefficient will be positive. Conversely, if one variable tends to increase while the other decreases, the correlation coefficient will be negative.

The Pearson correlation coefficient, r, ranges from +1 to -1. A value of 0 signifies no association between the variables. A positive value indicates a positive association, meaning that as one variable increases, the other does as well. A negative value indicates a negative association, where an increase in one variable corresponds to a decrease in the other. This relationship is illustrated in the diagram below:



4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?

Answer:- Feature scaling is a technique used to standardize the independent features in a dataset to a fixed range. This process is a crucial part of data preprocessing, especially when dealing with features that have varying magnitudes, values, or units. Without feature scaling, a machine learning algorithm might incorrectly assign higher importance to features with larger values and less importance to those with smaller values, regardless of their actual significance.

For example, without feature scaling, an algorithm might mistakenly interpret 3000 meters as being greater than 5 kilometers, even though they represent the same distance. This could lead to inaccurate predictions. By applying feature scaling, all values are brought to the same magnitude, which helps prevent such issues and ensures the algorithm treats features appropriately.

Normalized Scaling:-

1. Minimum and maximum values of features are used for scaling.
2. It is used when features are of different scales.
3. Scales values between [0, 1] or [-1, 1].
4. It is highly affected by outliers.
5. Scikit-Learn provides a transformer called MinMaxScaler for normalization.

Standardized Scaling:-

1. Mean and standard deviation are used for scaling.
2. It is used when we want to ensure a zero mean and unit standard deviation.
3. It is not bounded to a certain range.
4. It is much less affected by outliers.
5. 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?

Answer:- If there is a perfect correlation between variables, the Variance Inflation Factor (VIF) becomes infinite. A high VIF value indicates the presence of multicollinearity, which means there is a correlation between the variables. For example, if the VIF is 4, it means that the variance of the model coefficient is inflated by a factor of 4 due to multicollinearity.

When VIF reaches infinity, it indicates a perfect correlation between two independent variables. In such cases, the R-squared (R^2) value equals 1, leading to $1 / (1-R^2)$ being infinite. To address this issue, one of the variables causing the perfect multicollinearity needs to be removed from the dataset.

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

Answer:- A quantile-quantile (Q-Q) plot is a graphical technique used to determine whether two data sets come from populations with a common distribution.

Use of Q-Q Plot:

A Q-Q plot compares the quantiles of one data set against the quantiles of another. A quantile represents the percentage of data points that fall below a certain value. For example, the 0.3 (or 30%) quantile is the value below which 30% of the data falls, with 70% above it. A 45-degree reference line is typically included in the plot. If the data sets share a common distribution, the points on the Q-Q plot should align closely with this reference line. The more the points deviate from this line, the stronger the evidence that the data sets come from populations with different distributions.

Importance of Q-Q Plot:

When comparing two data samples, it's important to assess whether they share a common distribution. If they do, location and scale estimators can combine both data sets to provide better estimates of the common location and scale. If the samples differ, understanding the nature of these differences is valuable. The Q-Q plot offers more detailed insight into these differences than analytical methods like the chi-square test and the Kolmogorov-Smirnov two-sample test.

- ✔ **Completed By Abhishek Singh Chauhan.**