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 :-
   - Fall season has seen an increase in bookings, with a notable rise from 2018 to 2019.
   - Most of the bookings occur from the May to Oct, with a peak during the mid-year months. Interest tends to decline toward the end of the year.
   - Clear weather positively influences booking rates, which is to be expected.
   - Thu, Fir, Sat and Sun experience higher booking numbers compared to the earlier part of the week.
   - Bookings are lower on non-holidays compared to holidays may be due to work commitments, a preference for staying home, and fewer special events.
   - Overall, bookings appear fairly balanced between working days and non-working days.
   - The year 2019 saw a significant increase in bookings compared to 2018, indicating strong business growth.
2. Why is it important to use **drop_first=True** during dummy variable creation? (2 marks)

   Answer:
   Using **drop_first=True** is important when creating dummy variables, as it helps prevent the creation of an extra column that can lead to multicollinearity among the dummy variables.

   Syntax :
   drop_first: bool, default=False, This parameter determines whether to create k−1 dummies from k categorical levels by removing the first level.

   Example :
   Let's Consider a categorical column with three values: A, B, and C. When we create Dummy variables for these values:

   - Without **drop_first=True**, we would create three columns:
     - Is_A
     - Is_B
     - Is_C
   - With **drop_first=True**, we create only two columns:
     - Is_A
     - Is_B

   This is because if a value is not A or B, it must be C, making the third column redundant. By dropping the first level, we avoid unnecessary redundancy and potential multicollinearity in our analysis.

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:

To validate the assumptions of linear regression after building the model on the training set, I performed the following checks:

- **Normality of Error Terms :** 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 show any visible pattern.
- **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? (2 marks)

Answer:

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

- Temp : **0.5480** - This feature has the highest positive impact on bike demand, indicating that as temperature increases, the demand for shared bikes also rises.
- Year : **0.2329** - This feature positively influences bike demand, suggesting that demand has generally increased over the years.
- Light_snowrain : **-0.2829** - This feature has a notable negative impact on demand, indicating that the presence of light snow or rain decreases the demand for shared bikes.

General Subjective Questions

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

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 when the value of one or more independent variables changes (increases or decreases), the value of the dependent variable also changes correspondingly.

The relationship can be represented mathematically with the following equation:
$Y=mX+c$ where:

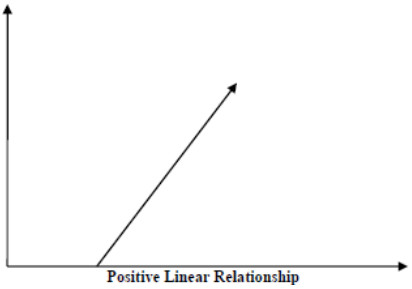
- **Y** is the dependent variable we are trying to predict.
- **X** is the independent variable used for making predictions.
- **m** is the slope of the regression line, representing the effect of X on Y.
- **c** is a constant known as the Y-intercept; when X= 0, Y equals c.

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

1. **Positive Linear Relationship :**

   A linear relationship is considered positive if both the independent and dependent variables increase together.

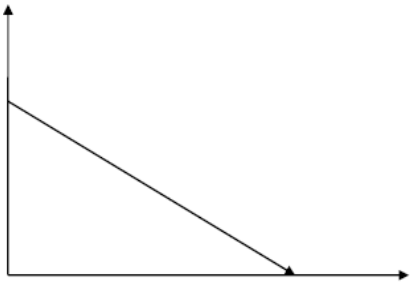
   This can be illustrated with the following graph:



Positive Linear Relationship

2. **Negative Linear relationship:**

   A linear relationship is considered negative if an increase in the independent variable corresponds to a decrease in the dependent variable. This can be illustrated with the 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.

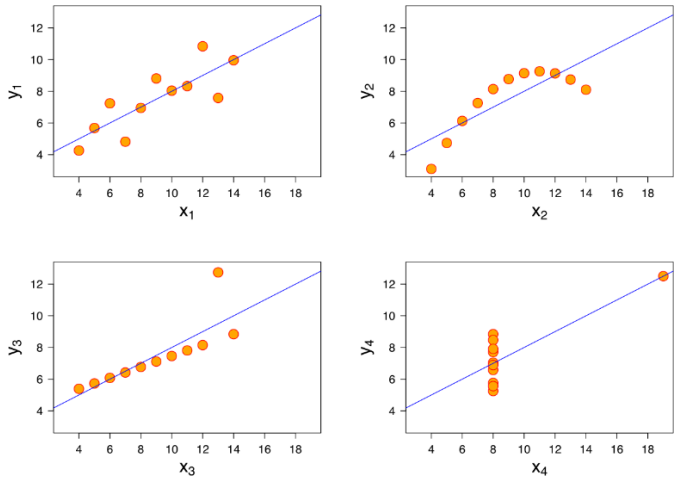
- Relationship between variables –
    - Linear regression model assumes that the relationship between response and feature variables must be linear.
- Normality of error terms –
    - Error terms should be normally distributed
- Homoscedasticity –
    - There should be no visible pattern in residual value

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 containing eleven (x, y) pairs. The essential point to note about these datasets is that they share the same descriptive statistics. However, the story changes completely when they are graphed; each graph tells a different story despite their similar summary statistics.

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)

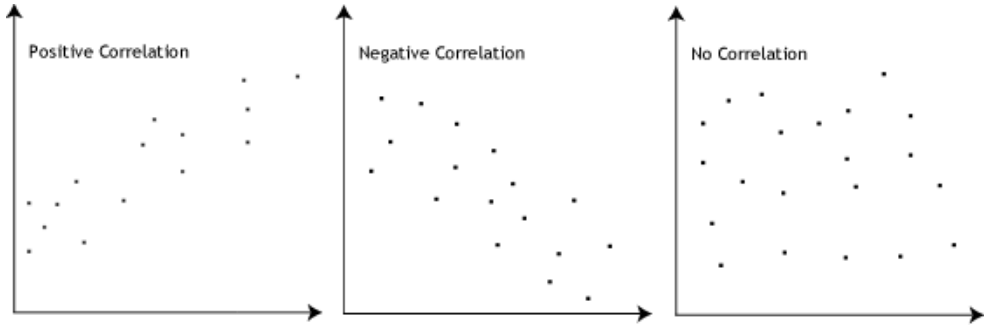
Answer:

Pearson's r is a numerical summary of the strength of the linear association between two variables. If the variables tend to increase and 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** indicates no association between the two variables.
- A value greater than **0** signifies a positive association, meaning that as one variable increases, the other variable also increases.
- A value less than **0** indicates a negative association, suggesting that as one variable increases, the other variable decreases.

This relationship is illustrated in the diagram below:



4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and sandardized scaling? (3 marks)

Answer:

Feature scaling is a technique used to standardize the independent features in a dataset to a Fixed range. It is performed during data preprocessing to handle highly varying magnitudes, values, or units. If feature scaling is not applied, machine learning algorithms may give greater weight to larger values and consider smaller values less significant, regardless of their actual unit.

Example:

If an algorithm does not use feature scaling, it might incorrectly interpret the value of 3000 meters as greater than 5 kilometers, which is not true. In this case, the algorithm would make inaccurate predictions. Therefore, feature scaling is used to bring all values to the same magnitude, addressing this issue effectively.

	Normalized scaling	Standardized scaling
1.	Uses the minimum and maximum values of features for scaling.	Uses the mean and standard deviation for scaling.
2.	Used when features are on different scales.	Used when we want to ensure zero mean and unit standard deviation.
3.	Scales values between [0, 1] or [-1, 1].	Not bounded to a specific range.
4.	Highly affected by outliers.	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 the VIF is infinite, it indicates perfect correlation between two independent variables. In this scenario, the R-squared value (R^2) equals 1, leading to the calculation of VIF as $1/(1-R^2)$, which results in infinity. A high VIF value generally suggests that the variance of the model coefficient is inflated due to multicollinearity.

For example, if VIF is 4, this means the variance is inflated by a factor of 4 because of the correlation among variables.

To resolve the issue of infinite VIF, one should drop one of the variables causing the perfect multicollinearity from the dataset.

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 technique for determining if two data sets come from populations with a common distribution.

Use of Q-Q plot:

A Q-Q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. A quantile refers to the fraction (or percent) of points below a given value. For example, the 0.3 (or 30%) quantile is the point at which 30% of the data fall below that value, and 70% fall above it. A 45-degree reference line is also plotted. If the two sets come from

populations with the same distribution, the points should fall approximately along this reference line. The greater the departure from this line, the stronger the evidence that the two data sets come from populations with different distributions.

Importance of Q-Q plot:

When comparing two data samples, it is often desirable to know if the assumption of a common distribution is justified. If so, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale. If two samples do differ, it is also useful to gain some understanding of the differences. The Q-Q plot can provide more insight into the nature of the differences than analytical methods such as the chi-square test and the Kolmogorov-Smirnov two-sample test.