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 seems to have attracted more booking. And, in each season the booking count has increased drastically from 2018 to 2019.
- 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.
- Clear weather attracted more booking which seems obvious.
- Thu, Fir, Sat and Sun have more number of bookings as compared to the start of the week. ⎞
- When it’s not holiday, booking seems to be less in number which seems reasonable as on holidays, people may want to spend time at home and enjoy with family.
- Booking seemed to be almost equal either on working day or non-working da

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

Answer:

drop_first = True is important to use, as it helps in reducing the extra column created during dummy variable creation.

Hence it reduces the correlations created among dummy variables.

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

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:

- 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)

- temp
- winter
- sep

General Subjective Questions

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

   Answer:

   Linear regression is a statistical method used for modeling the relationship between a dependent variable (also called the response or target variable) and one or more independent variables (predictors or features). The goal of linear regression is to find the best-fit linear relationship that can be used to predict the dependent variable based on the values of the independent variables.

   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.

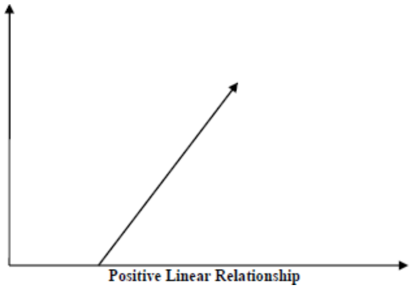
   X is the independent variable we are using to make predictions.

   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.

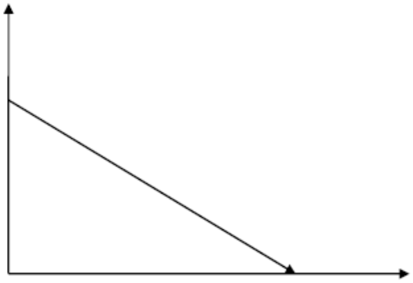
   Furthermore, the linear relationship can be positive or negative in nature as explained below−

   - 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)

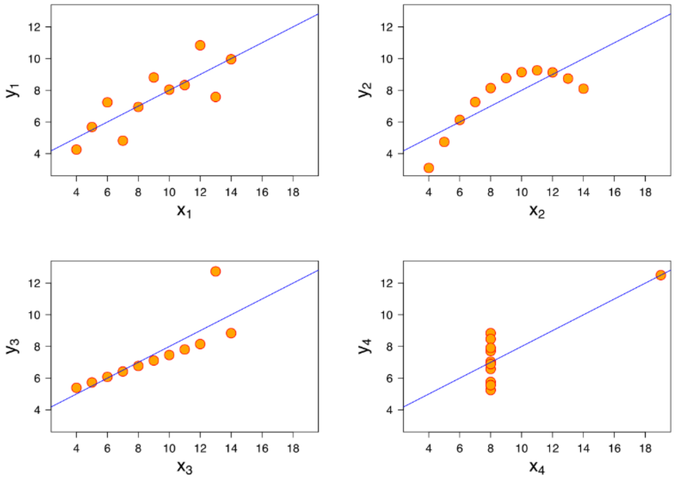
Answer:

Anscombe's quartet is a set of four datasets that have nearly identical simple descriptive statistics, yet they have very different distributions and appear quite distinct when graphed. The quartet was created by the statistician Francis Anscombe in 1973 to emphasize the importance of visualizing data and not relying solely on summary statistics. The datasets in Anscombe's quartet highlight the limitations of relying on summary statistics without visualizing the data and demonstrate the impact of outliers on statistical measures.

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) betweenx 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 showthe 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.

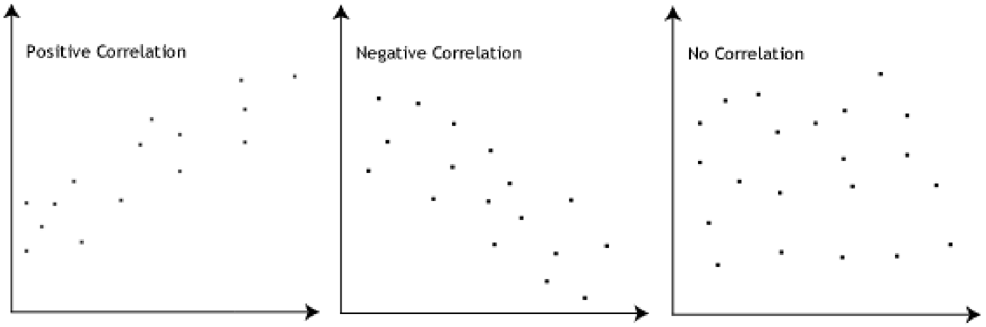
This quartet emphasizes the importance of visualization in Data Analysis. Looking at the data reveals a lot of the structure and a clear picture of the dataset.

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 variableassociated 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 thediagram below:



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

Answer:

Scaling in the context of data preprocessing refers to the process of transforming variables so that they are on a similar scale. This is done to ensure that no variable has more influence than another, especially in machine learning algorithms that are sensitive to the scale of the input features. Scaling is important for various reasons, and it helps in achieving better performance and convergence in many machine learning models.

Here are the key reasons why scaling is performed:

Equal Weightage: Many machine learning algorithms use some form of distance metric to make decisions. If features are on different scales, the algorithm may give more weight to features with larger scales, leading to biased results.

Gradient Descent Convergence: In optimization algorithms like gradient descent, having variables on different scales can result in the algorithm taking a longer time to converge or, in some cases, failing to converge.

Regularization: Regularization techniques, such as L1 and L2 regularization, are sensitive to the scale of the input features. Scaling helps ensure that regularization is applied uniformly across all features.

Differences between Normalized Scaling and Standardized Scaling:

1. Range:
    - Normalized Scaling scales the data to a specific range (e.g., 0 to 1).
    - Standardized Scaling scales the data to have a mean of 0 and a standard deviation of 1.
2. Sensitivity to Outliers:
    - Normalized Scaling can be sensitive to outliers because it depends on the minimum and maximum values.
    - Standardized Scaling is more robust to outliers as it uses the mean and standard deviation.
3. Interpretability:
    - Normalized Scaling retains the original distribution of the data but brings it within a specific range.
    - Standardized Scaling transforms the data to a standard normal distribution, making it more interpretable in terms of standard deviations from the mean.

The choice between normalized and standardized scaling depends on the specific requirements of the machine learning algorithm and the characteristics of the data.

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 thisperfect multicollinearity.

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regresson. (3 marks)**

   Answer:

   A Q-Q (Quantile-Quantile) plot is a graphical tool used to assess whether a dataset follows a specific theoretical distribution, such as the normal distribution. It compares the quantiles of the observed data to the quantiles of the expected distribution.

   Use and Importance of Q-Q Plot in Linear Regression:

   Normality Assumption:

   In linear regression, one of the assumptions is that the residuals (the differences between observed and predicted values) are normally distributed. Q-Q plots are commonly used to check the normality of residuals.
   If the residuals follow a normal distribution, the Q-Q plot should exhibit a roughly straight line. Deviations from a straight line indicate departures from normality.
   Identifying Outliers:

   Q-Q plots can help identify outliers in the dataset. Outliers may cause the Q-Q plot to deviate from a straight line, particularly in the tails of the distribution.
   Outliers in the residuals can influence the assumptions of linear regression, affecting the model's accuracy and reliability.
   Model Validation:

   Q-Q plots are part of model validation and diagnostics. Checking the normality of residuals is crucial for assessing the validity of the linear regression model.
   If the Q-Q plot shows a substantial departure from a straight line, it may suggest that the model assumptions are violated, and further investigation or model refinement may be necessary.
   Interpretability:

   Q-Q plots provide a visual and intuitive way to assess the normality of residuals. If the plot deviates significantly from a straight line, it may indicate non-normality, skewness, or other distributional issues.
   The interpretability of Q-Q plots makes them accessible for both statisticians and non-statisticians in evaluating the assumptions of linear regression.
   In summary, Q-Q plots play a crucial role in linear regression by helping to validate the assumption of normality for residuals. Detecting deviations from normality is important for making reliable inferences and predictions based on the linear regression model. By using Q-Q plots, analysts can visually inspect the distribution of residuals and identify potential issues that may impact the accuracy and validity of the regression analysis.