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:

The analysis of categorical variables using boxplots and bar plots revealed several key insights regarding booking trends. Notably, the fall season was a peak period, with bookings substantially increasing from 2018 to 2019 in all seasons. The booking pattern showed a pronounced trend, with most bookings occurring from May through October. Initially, there was an upward trend from the start of the year, peaking around mid-year, and then gradually tapering off towards the end of the year. Clear weather was another significant factor, leading to more bookings, which aligns with expectations. In terms of weekly trends, Thursday through Sunday recorded higher bookings compared to the earlier days of the week. Interestingly, non-holiday periods saw fewer bookings, possibly because people prefer to stay home and spend time with family during holidays. Despite this, the booking volume was almost consistent on both working and non-working days. A notable highlight was the year 2019, which experienced a marked increase in bookings compared to the previous year, indicating a positive trajectory for the business.

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

When dealing with categorical variables in data analysis, a common approach is to convert these variables into a format that can be easily used by algorithms, often through a process called "one-hot encoding" or "creating dummy variables." This process transforms each categorical level into a new binary column, indicating the presence or absence of that level in the original data.

In many cases, however, including all the dummy variables can lead to multicollinearity, where one variable can be predicted from the others, reducing the precision of the analysis. To avoid this, the `drop_first` option is used. When set to `True`, it removes the first level of the dummy variables.

Consider a categorical variable with three categories: A, B, and C. If we apply one-hot encoding without dropping the first level, we get three new columns: one each for A, B, and C. But in this scenario, if a data point is not A or B (as indicated by zeros in both the A and B columns), it is obviously C. Therefore, the third column (for C) is redundant, as its value can be inferred from the other two.

By setting `drop_first=True`, we eliminate this redundancy. It converts the categorical variable into two columns instead of three. If both columns are 0, it implies that the original category was the one dropped (in this case, A). This approach simplifies the model without losing any information and helps in reducing the potential for multicollinearity.

3. **Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)**

Answer:

The variable 'temp' exhibits the strongest 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 ensure the validity of the Linear Regression Model, I've verified its adherence to five key assumptions. First, the normality of error terms is crucial, requiring that these errors follow a normal distribution. Second, the model must have minimal multicollinearity, meaning the variables should not be highly correlated with each other. Third, the verification of linear relationships is important, as there should be a linear correlation evident between the variables. Fourth, the assumption of homoscedasticity is checked, which means the residuals should not show any discernible patterns, indicating consistent variance across the data. Lastly, the independence of residuals is essential, ensuring there's no autocorrelation within the residuals. These checks collectively ensure the robustness and reliability of the Linear Regression Model

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:

The three most influential features in determining the demand for shared bikes have been identified. The first key factor is 'temp,' which indicates the temperature's impact on bike-sharing demand. The second significant feature is 'winter,' highlighting the influence of the winter season on the usage of shared bikes. Finally, 'sep' emerges as the third crucial factor, representing the month of September and its specific effect on bike-sharing demand. These three elements combined play a substantial role in explaining the variations in the usage of shared bikes. The first influential factor, 'temp,' underscores the importance of temperature in determining bike-sharing demand. Temperature plays a crucial role in outdoor activities, including cycling. Warmer temperatures are generally more conducive to biking, making it a preferred mode of transport or leisure activity. The second key feature, 'winter,' highlights how the winter season impacts bike-sharing. Winter months often bring colder temperatures and harsher weather conditions, which can significantly deter outdoor activities, including biking. Factors such as snow, rain, and lower temperatures can make biking less safe and comfortable, leading to a decrease in bike-sharing usage. This seasonal effect is crucial for operators to understand and plan for lower demand periods. Lastly, 'sep,' representing September, sheds light on the temporal aspect of bike-sharing demand. September might mark a change in weather, transitioning from the heat of summer to milder temperatures, which can be ideal for biking. This month might also coincide with the start of the academic year or a return to regular work schedules after summer vacations, potentially influencing commuting patterns.

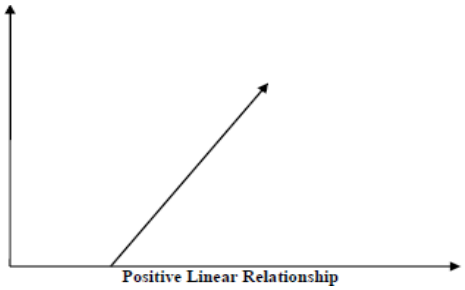
General Subjective Questions

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

Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. The core of this model is the linear association between these variables, meaning that changes in the independent variables are directly associated with changes in the dependent variable. This linear relationship is typically represented by the equation Y = mX + c, where Y is the dependent variable being predicted, X represents the independent variable used for the prediction, m is the slope of the regression line indicating the impact of X on Y, and c is the constant or Y-intercept, which is the value of Y when X is zero. Furthermore, the nature of this linear relationship can be either positive or negative. In a positive linear relationship, an increase in the independent variable leads to an increase in the dependent variable, which can be visualized in a graph where both variables rise together.
Linear regression is a fundamental algorithm in statistics and machine learning, used for predicting a quantitative response. It operates on the principle of establishing a linear relationship between a dependent variable (Y) and one or more independent variables (X). The core idea is to find a regression line, which is the best fit straight line through the data points. This line is represented by the equation Y = mX + c, where Y is the variable to be predicted, X is the predictor, m is the slope of the line, and c is the y-intercept.
The algorithm involves calculating the values of m and c that minimize the difference between the observed values and the values predicted by the model. This process, known as least squares regression, aims to minimize the sum of the squares of the differences (residuals) between the observed and predicted values.

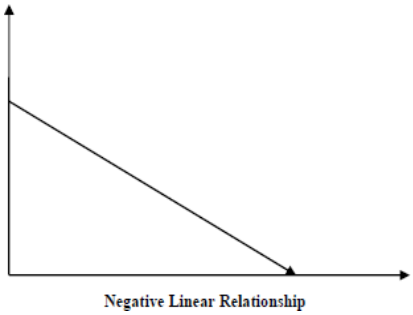
Linear regression makes several key assumptions, including linearity of the relationship between dependent and independent variables, homoscedasticity (constant variance of the residuals), and normal distribution of the residuals. It's also sensitive to outliers, which can significantly affect the slope and intercept of the regression line.

In its simplest form, linear regression is used for one independent variable (simple linear regression), but it can be extended to multiple variables (multiple linear regression). This versatility makes it a widely used tool for data analysis and predictive modeling, providing insights into relationships between variables and enabling forecasts of future observations. Beyond the basics, linear regression's effectiveness hinges on its interpretability. The coefficients (m values in the equation) quantify the strength and direction of the relationship between each independent variable and the dependent variable. This makes it invaluable for understanding which factors most influence the outcome and how they interact with each other. Additionally, linear regression can be adapted with techniques like polynomial regression for more complex relationships. In the context of machine learning, it serves as a starting point for regression analysis, often compared with more complex models to evaluate the trade-off between simplicity and predictive accuracy. Its application spans various fields, from economics to engineering, making it a cornerstone technique in both statisticals.



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 −



Linear regression is of the following two types −

1. Simple Linear Regression
2. Multiple Linear Regression

Linear regression models operate under several key assumptions about the data they analyze. One primary assumption is the absence or minimal presence of multicollinearity, which occurs when there's a dependency among the independent variables. Another assumption is that the data should exhibit little to no autocorrelation, meaning that the residual errors in the model should not be dependent on each other. Additionally, linear regression presupposes a linear relationship between the dependent and independent variables. The error terms, or residuals, are also expected to be normally distributed, adhering to the principle of normality. Lastly, the model assumes homoscedasticity, where the residuals display no discernible patterns or consistent variability across all levels of the independent variables. These assumptions are crucial for the proper functioning and accuracy of a linear regression model.

2. Explain the Anscombe's quartet in detail. (3 marks)

Answer:

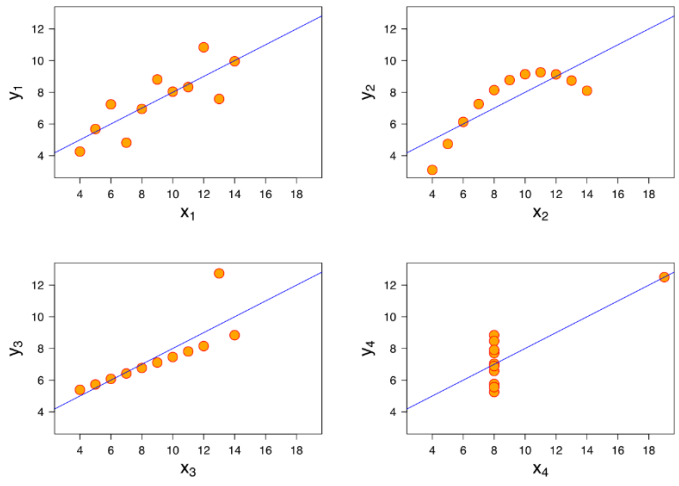
Anscombe's Quartet, created by statistician Francis Anscombe, consists of four distinct datasets, each with eleven pairs of (x, y) values. What's particularly noteworthy about these datasets is their identical descriptive statistics. However, the real divergence emerges when these datasets are visualized graphically. Despite their similar numerical summaries, each set reveals a completely different pattern or relationship when plotted, highlighting the importance of graphical analysis in statistics alongside numerical measures.

The datasets in Anscombe's Quartet present a fascinating case where, despite having identical summary statistics, they each tell a distinct story upon visualization. The mean values for x and y across all datasets are consistently 9 and 7.50, respectively. Additionally, each dataset shares the same variance values, with x having a variance of 11 and y having a variance of 4.13. Furthermore, the correlation coefficient between x and y is uniformly 0.816 across all groups, indicating a strong linear relationship. However, when these datasets are plotted on an x/y coordinate plane, the differences become starkly apparent. Despite having the same regression lines, each dataset uniquely illustrates how identical statistical summaries can mask vastly different data structures and relationship. Each dataset in the quartet has the same mean for both x and y variables (9 for x and 7.5 for y), indicating a similar central tendency across all groups. The variance, a measure of data spread, is also consistent for x (11) and y (4.13) across each dataset. Moreover, the correlation coefficient between x and y is uniformly 0.816 in all four datasets, suggesting a strong linear relationship between the variables. On the surface, these identical statistics might lead one to assume that the datasets are similar in nature.

However, the true value of Anscombe's work becomes evident when each dataset is graphically plotted on an x/y coordinate plane. The first dataset (I) appears to be a textbook example of a linear relationship, with data points closely aligned along a straight line, thereby confirming the implications of the summary statistics. Dataset II, on the other hand, reveals a clear non-linear relationship; it's more of a curve than a line, highlighting that linear correlation does not necessarily imply a linear relationship. This demonstrates the limitation of using correlation coefficients as the sole measure of association between variables.

Dataset III presents yet another scenario. While it mostly follows a linear trend like Dataset I, it includes a distinct outlier that heavily influences the regression line. If this outlier were to be removed, the relationship between x and y would appear much stronger and more linear. This dataset exemplifies how outliers can significantly impact statistical analyses, particularly regression and correlation, and it underscores the need for careful data inspection.

Dataset IV is perhaps the most striking in its divergence. It consists primarily of a constant x-value with varying y-values, except for one outlier far from the others. This outlier drives the entire correlation, suggesting a strong linear relationship where none actually exists. This example demonstrates how a single data point can disproportionately influence the overall analysis, leading to misleading conclusions.



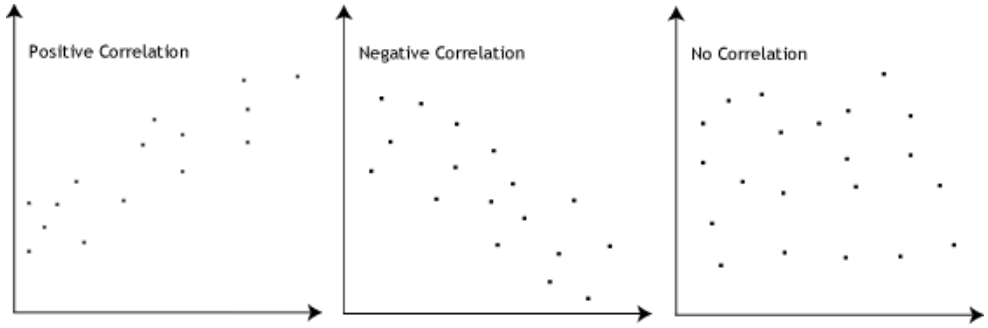
The datasets within Anscombe's Quartet each tell a unique story when visualized, underscoring the significance of graphical representation in data analysis. Dataset I displays a clear, linear pattern, making it ideal for linear modeling. Dataset II deviates from a normal distribution, indicating non-linear characteristics. The linearity in Dataset III is apparent, yet the presence of an outlier significantly skews the regression results. In Dataset IV, the influence of a single outlier is again notable, demonstrating how just one data point can drastically increase the correlation coefficient. This quartet effectively illustrates that visual examination of data can reveal crucial structural insights and provide a more comprehensive understanding of the dataset's characteristics.

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

Answer:

Pearson's r provides a numerical measure of the linear relationship strength between two variables. It is calculated in such a way that it captures the degree to which the variables move in synchronization. A positive correlation coefficient means that both variables tend to increase or decrease together. Conversely, a negative coefficient indicates that as one variable increases, the other tends to decrease, showing an inverse relationship.

The range of Pearson's r lies between +1 and -1. A zero value signifies no linear association between the variables. Values above zero denote a positive correlation, meaning both variables move in the same direction. In contrast, values below zero indicate a negative correlation, where an increase in one variable corresponds to a decrease in the other. This correlation is effectively illustrated in a scatter plot, where the nature of the relationship between the two variables becomes visually apparent.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:

Feature Scaling is an essential process in data preprocessing for machine learning. It involves adjusting independent variables within the dataset to a specific range, helping to manage variables that vary in magnitude, value, or units. The absence of feature scaling can lead to skewed results in machine learning algorithms, as they might incorrectly assign more importance to larger numbers, irrespective of their units. For example, an algorithm that doesn't employ feature scaling could mistakenly interpret 3000 meters as greater than 5 kilometers due to the numerical value, leading to inaccurate predictions. To prevent such errors and ensure uniformity in magnitude across all variables, feature scaling is used, bringing all data to a comparable scale.

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:

When the Variance Inflation Factor (VIF) reaches infinity, it indicates a perfect correlation between independent variables in a dataset. Essentially, a VIF measures how much the variance of an estimated regression coefficient increases if your predictors are correlated. If the VIF is 4, it implies that the variance of a coefficient is quadrupled due to multicollinearity. The scenario of an infinite VIF typically arises when the R-squared value in a regression analysis is 1, indicating a perfect linear relationship between variables. This perfect correlation leads to the formula 1/ (1-R2) tending towards infinity. To address this issue of perfect multicollinearity, one of the correlated variables should be removed from the dataset, ensuring a more accurate and reliable regression analysis.

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 used to assess if two datasets originate from populations with the same distribution. In a q-q plot, the quantiles of one dataset are plotted against those of another. A quantile represents the point below which a certain percentage of data falls. For example, the 30% quantile is the value below which 30% of the data lies. Typically, a 45-degree reference line is included in the plot. If the datasets share a common distribution, their points will align closely with this line. Deviations from this line suggest differences in distribution between the datasets.

The significance of a q-q plot lies in its ability to verify the assumption of a shared distribution between two samples. When this assumption holds, it's possible to combine the datasets for more robust estimates of common location and scale parameters. Conversely, if the samples differ, the q-q plot helps understand the nature of these differences, often providing deeper insights than analytical methods like the chi-square or Kolmogorov-Smirnov 2-sample tests. This makes q-q plots a valuable tool in statistical analysis, especially when comparing data distributions.