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 .Below are the few of the inferences 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 day.

2019 attracted more number of booking from the previous year, which shows good progress in terms of business.

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.

Let's say we have 3 types of values in Categorical column and we want to create dummy variable for that column. If one variable is not A and B, then It is obvious C. So we do not need 3rd variable to identify the C.

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 –

1. Normality of error terms

Error terms should be normally distributed

2.Multicollinearity check

There should be insignificant multicollinearity among variables.

3.Linear relationship validation

Linearity should be visible among variables

4.Homoscedasticity

There should be no visible pattern in residual values.

5. 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)

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

Temp

Yr

Winter

General Subjective Question

1. **Explain the linear regression algorithm in detail.**

Answer:

Linear regression is a fundamental statistical method used for modeling the relationship between a dependent variable and one or more independent variables by fitting a linear equation to observed data. It's a widely used technique in various fields, including statistics, machine learning, economics, and social sciences. Here's a detailed explanation of how linear regression works:

Problem Statement: Linear regression is used when we want to predict the value of a continuous dependent variable based on one or more independent variables. For instance, we might want to predict house prices based on features like size, number of bedrooms, location, etc.

Assumptions: Linear regression relies on several assumptions:

Linear Relationship : The relationship between independent and dependent variables is linear.

Independence: The observations are independent of each other.

Homoscedasticity: The variance of the residuals (the difference between observed and predicted values) is constant across all levels of the independent variables.

Normality : The residuals are normally distributed.

No multicollinearity : The independent variables are nt highly correlated with each other.

Linear Equation: In simple linear regression, there's one independent variable, and the relationship between the independent variable (X) and dependent variable (Y) is represented by the equation of a straight line:

Y is the dependent variable.

$Y=B0 +B1X + \varepsilon$

X is the independent variable.

B0 is the intercept (the value of Y when X=0)

B1 is the slope (the change in Y for a unit change in X)

ε is the error term representing the difference between the observed and predicted values.

Fitting the Model: The goal of linear regression is to find the best-fitting line that minimizes the sum of squared differences between the observed and predicted values. This is usually done using the method of least squares, where the parameters

and B0 and B1 are estimated to minimize the sum of the squared residuals.

Parameter Estimation: The parameters B0 and B1 are estimated using statistical techniques such as ordinary least squares (OLS) estimation. OLS finds the values of B0 and B1 that minimize the sum of the squared residuals:

$\sum i=1n(Yi-(\beta 0+\beta 1Xi))2$

Model Evaluation: Once the model is fitted, it needs to be evaluated to assess its performance and reliability. Common metrics for evaluating linear regression models include R-squared (the proportion of the variance in the dependent variable that is predictable from the independent variable(s)), adjusted R-squared, Mean Squared Error (MSE), etc.

Prediction: Once the model is validated, it can be used to make predictions. Given new values of the independent variables, the model can predict the corresponding values of the dependent variable.

Assumption Checking and Residual Analysis: After fitting the model, it's essential to check whether the assumptions of linear regression hold true. This involves analyzing the residuals (the differences between observed and predicted values) to ensure they meet the assumptions of the model.

Linear regression is a powerful and widely used technique due to its simplicity, interpretability, and effectiveness in many real-world scenarios. However, it's essential to be cautious and verify that the assumptions hold true before relying on its results. Additionally, there are variations such as multiple linear regression for more complex relationships involving multiple independent variables.

2. **Explain the Anscombe's quartet in detail.**

   Answer:

Anscombe's quartet is a set of four datasets that have nearly identical statistical properties but appear very different when graphed. This quartet was created by the statistician Francis Anscombe in 1973 to emphasize the importance of graphical data exploration and the dangers of relying solely on summary statistics.

Here's an explanation of the quartet and its significance:

The Four Datasets: Each dataset in Anscombe's quartet consists of 11 (x, y) pairs:

Dataset I: A simple linear relationship with some variance.

Dataset II: A non-linear relationship between x and y with the same regression line as Dataset I.

Dataset III: A linear relationship with one outlier that significantly affects the regression line.

Dataset IV: A perfect linear relationship except for one point that creates a misleading outlier.

Summary Statistics: When one examines the summary statistics of these datasets (mean, variance, correlation coefficient, etc.), they are remarkably similar. For example, all datasets have:

Mean of x: 9.0

Mean of y: 7.5

Variance of x: 11.0

Variance of y: 4.12

Correlation coefficient: Approximately 0.816

Graphical Representation: Despite the similarities in summary statistics, the datasets look vastly different when plotted. When graphed, Dataset I shows a clear linear relationship, Dataset II reveals a non-linear pattern, Dataset III displays the impact of an outlier, and Dataset IV demonstrates how a single point can disproportionately influence regression analysis.

Significance: Anscombe's quartet underscores the importance of visualizing data before drawing conclusions. It highlights how relying solely on summary statistics can lead to overlooking important features or patterns in the data. While summary statistics provide valuable insights, they do not capture the complete picture. Visualization allows for a more comprehensive understanding of the data, enabling researchers to identify outliers, patterns, trends, and relationships that may not be apparent from summary statistics alone.

Educational Tool: Anscombe's quartet is often used as an educational tool in statistics and data analysis courses to emphasize the importance of graphical exploration and to illustrate concepts such as outliers, the influence of individual data points on regression analysis, and the limitations of summary statistics.

In summary, Anscombe's quartet serves as a powerful reminder of the need for graphical exploration alongside numerical analysis when examining data. It highlights the potential pitfalls of relying solely on summary statistics and demonstrates how visualization can reveal insights that may otherwise go unnoticed.

3. What is Pearson's R?

Answer:

Pearson's r (also known as the Pearson correlation coefficient or Pearson's r) is a statistical measure of the strength and direction of the linear relationship between two continuous variables. It quantifies the degree to which two variables are linearly related to each other.

The Pearson correlation coefficient r ranges from -1 to 1:

r=1: Perfect positive correlation. This means that as one variable increases, the other variable also increases proportionally, following a straight line.

r=−1: Perfect negative correlation. This means that as one variable increases, the other variable decreases proportionally, following a straight line, but in the opposite direction.

r=0: No correlation. There is no linear relationship between the two variables.

The formula to calculate Pearson's r for a sample is:

$r=\sum(Xi-\bar{X})2\times\sum(Yi-\bar{Y})2/\sum(Xi-\bar{X})(Yi-\bar{Y})$

Where:

Xi and Yi are individual data points.

X and Y are the means of the X and Y variables, respectively.

The numerator represents the covariance of X and Y, which measures how the two variables change together.

The denominator represents the standard deviations of X and Y, which standardizes the covariance.

Pearson's r is widely used in various fields, including psychology, economics, biology, and social sciences, to assess the strength and direction of relationships between variables.

However, it's important to note that Pearson's r measures only linear relationships and may not capture nonlinear associations between variables. Additionally, correlation does not imply causation, meaning that just because two variables are correlated does not necessarily mean that changes in one variable cause changes in the other.

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

Answer:

Scaling is a preprocessing technique used in data analysis and machine learning to transform the features of a dataset onto a similar scale. It involves adjusting the range of values of variables so that they all have similar magnitudes. Scaling is important because many machine learning

algorithms are sensitive to the scale of the input features. Failure to scale features properly can lead to biased or inefficient models.

Here's why scaling is performed and the difference between normalized scaling and standardized scaling:

Why Scaling is Performed:

Avoidance of Bias: Some machine learning algorithms, such as k-nearest neighbors (KNN) and support vector machines (SVM), are distance-based and can be sensitive to the scale of features. If features have different scales, the algorithm may give more weight to features with larger scales, potentially biasing the model.

Faster Convergence: Gradient-based optimization algorithms, such as gradient descent, converge faster when features are scaled. This is because similar step sizes in all dimensions can be achieved when the features are on the same scale, leading to faster convergence to the minimum of the optimization function.

Improved Interpretability: Scaling features to the same range can make the coefficients or importance scores of features more comparable and interpretable.

Normalized Scaling:

Normalization scales each feature to a range between 0 and 1. The formula for normalization is:

X normalized = X - X min/ Xmax −Xmin

where X is the original feature, Xmin is the minimum value of X, and Xmax is the maximum value of X.

Normalization preserves the relative relationships between values but does not handle outliers well.

Standardized Scaling:

Standardization scales each feature to have a mean of 0 and a standard deviation of 1.

The formula for standardization is:

X standardized = X−μ/σ

where

X is the original feature,

μ is the mean of X, and

σ is the standard deviation of X.

Standardization centers the data around 0 and scales it to have a unit variance.

It is less affected by outliers compared to normalization and is often preferred when the distribution of the data is not known or is not Gaussian.

In summary, scaling is performed to ensure that features have similar magnitudes, which can improve the performance, convergence, and interpretability of machine learning models.

Normalized scaling and standardized scaling are two common scaling techniques, with normalization scaling features to a range between 0 and 1 and standardization scaling features to have a mean of 0 and a standard deviation of 1.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

Answer:

Yes, the occurrence of infinite values for the Variance Inflation Factor (VIF) is indeed possible and typically indicates a specific issue in the dataset or the model. VIF is a measure used to detect multicollinearity among predictor variables in a regression analysis. It quantifies how much the variance of the estimated regression coefficients is inflated due to multicollinearity.

The VIF for a predictor variable is calculated as:

$VIF=1/1-R2$

Where R2 is the coefficient of determination obtained by regressing the predictor variable against all other predictor variables in the model.

The presence of infinite values for VIF can occur due to the following reasons:

Perfect Multicollinearity: Infinite VIF values arise when there is perfect multicollinearity among predictor variables.

Perfect multicollinearity occurs when one or more predictor variables in the regression model are linearly dependent on each other,

meaning one variable can be exactly predicted by a linear combination of the others. In this case, the coefficient of determination

R2 becomes 1, resulting in a division by zero when calculating the VIF.

Nearly Perfect Multicollinearity: Although less common, nearly perfect multicollinearity can also lead to very high VIF values, which may appear as infinite due to computational precision. This situation occurs when there is an extremely high correlation between predictor variables, making the estimated coefficients highly unstable.

Data Issues: Infinite VIF values can sometimes be due to errors or anomalies in the dataset, such as extreme outliers, incorrect data entry, or data processing errors. These issues can lead to spurious results and should be carefully investigated.

Dealing with infinite VIF values requires identifying and addressing the underlying cause of multicollinearity in the dataset. This may involve examining the correlation matrix of predictor variables, removing redundant variables, combining or transforming variables, or re-evaluating the model specification. Additionally, it's essential to verify the data quality and integrity to ensure that no anomalies or errors are present.

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, short for Quantile-Quantile plot, is a graphical tool used to assess whether a given dataset follows a specific probability distribution or to compare the distribution of two datasets. The Q-Q plot displays the quantiles of the observed data against the quantiles of a theoretical distribution (usually a standard normal distribution), allowing for visual inspection of how closely the observed data matches the theoretical distribution.

Here's how a Q-Q plot is constructed and its importance in linear regression:

Construction of a Q-Q Plot:

First, the data points of the dataset are sorted in ascending order.

Next, the quantiles of the sorted data are calculated.

Similarly, quantiles are calculated for a theoretical distribution, typically a standard normal distribution (mean = 0, standard deviation = 1).

The quantiles of the observed data are then plotted against the quantiles of the theoretical distribution on a scatter plot.

Use and Importance in Linear Regression:

Assumption Checking: In linear regression, it is crucial to assess whether the residuals (the differences between observed and predicted values) follow a normal distribution. Q-Q plots provide a visual method for checking this assumption. If the residuals are normally distributed, the points in the Q-Q plot will approximately fall along a straight line.

Detecting Departures from Normality: Departures from normality in the residuals can indicate potential issues with the model, such as omitted variables, incorrect functional form, or heteroscedasticity. Q-Q plots allow analysts to detect deviations from normality, such as skewness or heavy tails, by observing deviations from the straight line pattern.

Model Evaluation and Diagnosis: Q-Q plots are valuable tools for evaluating the adequacy of a linear regression model. They help identify potential violations of assumptions and guide diagnostic efforts to improve model performance. By assessing the normality of residuals, analysts can make informed decisions about model validity and reliability.

Comparison of Distributions: Q-Q plots can also be used to compare the distribution of one dataset to another. This is useful for assessing differences in distributional properties between groups or comparing observed data to a theoretical distribution other than the normal distribution.

In summary, Q-Q plots are important tools in linear regression for assessing the normality of residuals, detecting departures from normality, and evaluating the adequacy of the regression model. They provide visual insights into the distributional properties of the data and help diagnose potential issues that may affect the validity and reliability of the regression analysis.