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?

ANS:

-After conducting an analysis on the categorical columns using bar plots, the following observations can be inferred:

- The fall season seems to have attracted more bookings. Furthermore, each season has seen a substantial rise in booking counts from 2018 to 2019.

- The months of May, June, July, August, September, and October have recorded the highest number of bookings. The trend indicates a rise in bookings from the start of the year, peaking in the middle, and then declining towards the year's end.

-It is evident that clear weather conditions attract more bookings, which aligns with common expectations.

- Thursdays, Fridays, Saturdays, and Sundays exhibit a higher number of bookings compared to the earlier days of the week.

- Non-holiday periods generally see fewer bookings, which makes sense as people often choose to stay home and spend time with their families during holidays.

-The number of bookings appears to be relatively consistent between working days and non-working days.

- The year 2019 has seen a higher number of bookings compared to the previous year, indicating positive progress in terms of business growth.

- These observations offer valuable insights into booking patterns and preferences, aiding in the understanding of customer behaviour's and informing the decision-making processes for the business.

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

ANS:

-As we know we can represent a column having 'n' categories we can represent the categories by using 'n-1' dummies, pd.get_dummies will create 'n' dummies by Using `drop_first=True' It ensures that one dummy variable is dropped as a reference category, making the remaining dummy variables independent and providing clear comparisons between categories.

-during dummy variable creation is important to avoid multicollinearity issues, improve model interpretability, and reduce model complexity.

-For example, if we have a column called "students" with 3 categories: 'Class A', 'Class B', and 'Class C', using pd.get_dummies without drop_first=True would result in 3 dummy columns. However, by using drop_first=True, the first dummy column ('Class A') is dropped, and we can represent these categories with 2 dummies instead of 3. This approach simplifies the model, avoids redundancy, and facilitates clearer comparisons between categories.

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

-‘temperature’ is the variable which has the highest correlation in the target variable.

4. How did you validate the assumptions of Linear Regression after building the model on the training set?

ANS:

-I have assessed the validity of the Linear Regression Model based on the following assumptions:

1. Normality of error or Residual Analysis: -

The assumption states that the error terms (residuals) should be normally distributed. - Normality ensures that the residuals have a symmetrical distribution around zero with a constant variance.

- To verify this assumption, I examined the distribution of residuals and checked if they approximately follow a bell-shaped curve.

2. Homoscedasticity: -

Homoscedasticity assumes that the residuals exhibit consistent variance across all levels of the predictor variables. When this assumption is violated, it leads to heteroscedasticity, where the spread of residuals varies across different ranges of predictors. To evaluate homoscedasticity, I examined plots of residuals against predicted values to identify any noticeable patterns or trends.

3. Independence of residuals: -

This assumption posits that the residuals are independent of each other, meaning there is no correlation or autocorrelation present.

Autocorrelation can also be assessed using the correlation matrix

- By evaluating these assumptions, I ensured the validity and reliability of the linear regression model, allowing for accurate interpretations of the results.

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

ANS:

-The demand for shared bikes is primarily influenced by three significant features:

Temperature

Winter season

September

These three factors have been found to contribute significantly to explaining the variations in the demand for shared bikes.

General Subjective Questions:

1. Explain the linear regression algorithm in detail.

ANS:

- Linear regression is a supervised machine learning algorithm used for predicting continuous numeric values. It establishes a linear relationship between the input features (independent variables) and the target variable (dependent variable) by fitting a straight line that best represents the data. The goal of linear regression is to find the best-fitting line that minimizes the differences between the predicted values and the actual values.

-there is a way steps to perform linear regression:

1. Assumptions:

- Linearity: Assumes a linear relationship between the independent variables and the target variable.

- Independence: Assumes that the observations are independent of each other.

-Homoscedasticity: Assumes that the variance of the errors is constant across all levels of the independent variables.

- Normality: Assumes that the errors are normally distributed.

2. Simple Linear Regression:

Simple linear regression deals with one independent variable (X) and one dependent variable (Y). The equation of a simple linear regression can be represented as:

$Y = b0 + b1*X + \varepsilon$

- Y: Dependent variable (target variable)

- X: Independent variable (input feature)

- b0: Intercept (the value of Y when X is zero)

- b1: Slope (the change in Y for a unit change in X)

- ε: Error term (residuals)

3. Multiple Linear Regression:

Multiple linear regression extends the concept of simple linear regression to multiple independent variables. The equation can be represented as:

$Y = b0 + b1*X1 + b2*X2 + ... + bn*Xn + \varepsilon$

- X1, X2, ..., Xn: Independent variables (input features)

- b1, b2, ..., bn: Coefficients corresponding to each independent variable

- ε: Error term (residuals)

4. Model Training:

The linear regression model is trained by estimating the coefficients (b0, b1, ..., bn) that minimize the sum of squared residuals (the difference between the predicted and actual values). This process is typically done using optimization techniques like Ordinary Least Squares (OLS) or gradient descent.

5. Model Evaluation:

The trained model's performance is evaluated using various metrics such as mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), and R-squared (R2) score. These metrics assess how well the model fits the data and predicts the target variable.

6. Making Predictions: Once the model is trained and evaluated, it can be used to make predictions on new, unseen data. Given the values of the independent variables, the model calculates the predicted value of the dependent variable using the learned coefficients.

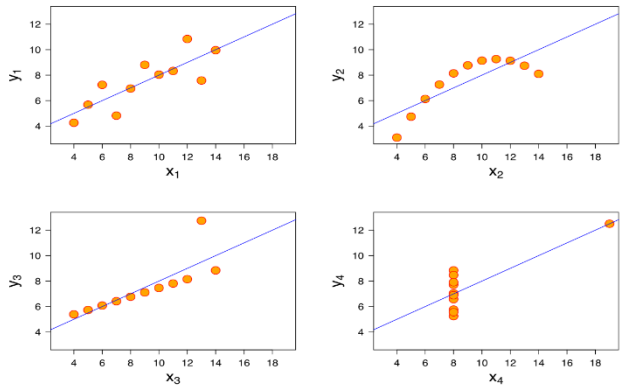
-Linear regression is a simple yet powerful algorithm widely used for tasks such as trend analysis, forecasting, and understanding the relationship between variables.

2. Explain the Anscombe's quartet in detail.

ANS:

- Anscombe's quartet is a group of datasets (x, y) that have the same mean, standard deviation, and regression line, but which are qualitatively different. It is often used to illustrate the importance of looking at a set of data graphically and not only relying on basic statistic properties.

- Anscombe's quartet comprises four datasets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (x, y) points.



- Anscombe's quartet tells us about the importance of visualizing data before applying various algorithms to build models. This suggests the data features must be plotted to see the distribution of the samples that can help you identify the various anomalies present in the data (outliers, diversity of the data, linear separability of the data, etc.). Moreover, the linear regression can only be considered a fit for the data with linear relationships and is incapable of handling any other kind of data set.

- This concept is fundamental in data analysis, emphasizing that both numerical and graphical techniques are necessary to fully understand and accurately interpret data.

3. What is Pearson's R?

ANS:

- Pearson's R, also known as the Pearson correlation coefficient, is a measure of the linear correlation between two variables. It quantifies the degree to which a relationship between two variables can be described by a straight line. The coefficient is named after Karl Pearson

- Range: The value of Pearson's R ranges from -1 to +1.

-Interpretation:

Positive Correlation: When R is positive, higher values of one variable are associated with higher values of the other variable.

Negative Correlation: When R is negative, higher values of one variable are associated with lower values of the other variable.

Magnitude: The closer the value of R to +1 or -1, the stronger the linear relationship between the two variables.

- The formula for Pearson's R is:

r = (Σ((Xi - X_mean) * (Yi - Y_mean))) / (sqrt(Σ(Xi - X_mean)^2) * sqrt(Σ(Yi - Y_mean)^2))

-Where: - Xi and Yi are the individual values of the two variables. - X_mean and Y_mean are the means of the two variables. - Σ denotes the summation symbol.

- Assumptions:

Linearity: Pearson's R measures the strength and direction of a linear relationship. If the relationship is not linear, Pearson's R may not be an appropriate measure.

Homoscedasticity: The variability of one variable should be similar across all values of the other variable.

Normality: The variables should be approximately normally distributed, especially if performing significance testing.

- Pearson's R is a powerful and widely-used statistic for quantifying the linear relationship between two continuous variables. It is simple to calculate and interpret, making it a fundamental tool in statistics and data analysis.

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

- Scaling is a preprocessing step in machine learning that involves adjusting the features of a dataset to a uniform scale. This process ensures that all features have comparable ranges, which is crucial for some machine learning algorithms and data analysis methods. The purpose of scaling is to handle problems associated with the magnitude and units of the features, and to prevent features with larger values from overshadowing the learning process.
- The main reasons for performing scaling are:
- Comparison of Features
- Gradient Descent Optimization

- .Regularization Techniques: in this technique there is 2 type of scaling:
- Normalized Scaling (Min-Max Scaling):
- Normalized scaling, also known as Min-Max scaling, rescales the features to a fixed range, typically between 0 and 1.
- The formula for normalized scaling is:
X_scaled = (X - X_min) / (X_max - X_min)
- X is the original feature value, X_min and X_max are the minimum and maximum values of the feature, respectively.
- Normalized scaling preserves the original distribution of the data but scales it to a specific range.

- Standardized Scaling (Z-score normalization):

- Standardized scaling, also known as Z-score normalization, transforms the features to have zero mean and unit variance.

- The formula for standardized scaling is:

X_scaled = (X - X_mean) / X_std

- X is the original feature value, X_mean is the mean of the feature, and X_std is the standard deviation of the feature.

- Standardized scaling centers the data around zero and scales it based on the spread of the data.

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

ANS:

- The occurrence of an infinite value of Variance Inflation Factor (VIF) is known as "perfect multicollinearity." It happens when there is an exact linear relationship between one or more independent variables in a regression model. Perfect multicollinearity leads to unstable parameter estimates, making it impossible to compute accurate VIF values for the affected variables.

- Duplicate or Redundant Variables:

-When two or more variables in the dataset are identical or perfectly correlated, it leads to redundancy in the information they provide. For example, having both "height in inches" and "height in centimeters" as independent variables would introduce perfect multicollinearity.

- Data Transformation Issues:

- Applying inappropriate data transformations can also result in perfect multicollinearity. For instance, if you convert a continuous variable into categorical bins and include all the bins as independent variables, it can lead to perfect multicollinearity.

- Creation of Derived Variables:

- When new variables are created from existing variables using mathematical operations, it's essential to avoid introducing perfect multicollinearity inadvertently. For example, if you

create a new variable by summing two existing variables, and those two variables are perfectly correlated, it will result in perfect multicollinearity.

- To address the issue of perfect multicollinearity and infinite VIF values, it is necessary to identify and remove the redundant or perfectly correlated variables from the model.

- It is important to note that while infinite VIF values indicate the presence of perfect multicollinearity, high VIF values (but not infinite) suggest strong multicollinearity between variables. In such cases, it is advisable to assess the impact of multicollinearity on the model's performance and consider addressing it by removing or transforming the correlated variables.

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

ANS:

- A Q-Q plot, short for quantile-quantile plot, is a graphical tool used to assess the distributional similarity between a given dataset and a theoretical distribution, such as the normal distribution. It plots the quantiles of the observed data against the quantiles of the theoretical distribution, allowing visual comparison and identification of deviations from the expected distribution.

- The use and importance of a Q-Q plot:

- Assessing Normality Assumption.

- Detecting Skewness and Outliers.

- Model Evaluation and Assumption Checking.

- Comparison of Distributions.

- In summary, the Q-Q plot is a valuable tool in linear regression for assessing the normality assumption, detecting skewness and outliers, evaluating model adequacy, and comparing distributions. It helps in understanding the distributional characteristics of the residuals and provides insights into potential issues that may affect the validity and reliability of the linear regression analysis.

Name:-Akshar patel