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-

Here for the analysis of categorical columns I have used boxplot and bar plot. The key point we can infer from the analysis are-

i) The fall season of 2019 has the highest growth in terms of uses of Boom Bikes and in each season the booking count has increased drastically from 2018 to 2019.

ii) Most of the bookings has been done during the month of june, july, aug and sep.The trend showed a gradual increase from the beginning of the year until mid-year, after which it started to decline as the year approached its end and the number of bookings for each month seems to have increased from 2018 to 2019.

iii) As compared to the starting of the week Thu,Fri,Sat and Sun have higher number of bookings.

iv) Most of the customers prefers clear weather for booking of BoomBikes and compared to 2018 booking increased for each weather situation in 2019.

v) In 2019, there was a noticeable rise in bookings compared to 2018, 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-

When creating dummy variables for categorical data, the **drop_first** parameter is used to control the level of multicollinearity in the dataset and is particularly important in the context of regression analysis. By setting **drop_first=True**, we effectively exclude one of the levels from the categorical variable when creating dummy variables. This eliminates perfect multicollinearity because the dropped level is the reference category, and its value is inferred from the values of the other dummy variables. It reduces the number of variables in the model. When we have a large number of categories in a categorical variable, including all of them as dummy variables can lead to an overly complex and high-dimensional model. Dropping one level reduces the dimensionality and the computational complexity of the analysis.

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

Answer-

Looking at the pair-plot among the numerical variables, we can see ‘temp’ and ‘atemp’ are two numerical variables which has the highest correlation with the target variable (‘cnt’).

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Answer-

Here, I have validated the assumption of Linear Regression Model based on the 5 assumptions –

i) Normality of Residuals:- The Residuals(Error Terms) should follow Normal Distribution.

ii) Checking of Multicollinearity:- There should be no Multicollinearity present among the variables, ensuring that the predictor variables are independent.

iii) Validation of Linear Relationship:- There should be a linear relationship present among the variables.

iv) Homoscedasticity:-There should be equal variance across all levels of the variables, there should be no visible pattern among the residual values.

v) Independence of Residuals:- The residuals should be independent of each other, i.e. no Auto-Correlation should be present there.

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-

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

i) Temperature(temp)

ii) Season(winter)

iii) Month(sep)

General Subjective Questions

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

Answer-

Linear regression is a popular statistical and machine learning algorithm used for modeling the relationship between a dependent variable (target) and one or more independent variables (features). Linear regression assumes that there is a linear relationship between the independent variables (features) and the dependent variable (target). It aims to find a linear equation that best represents this relationship. Linear Regression is based on the popular equation-

"y = mx + c "

It assumes that there is a linear relationship between the dependent variable(y) and the predictor/independent variable(x).

Let's dive into the details of linear regression:

Objective: The objective in linear regression is to find the values of the coefficients (b0, b1, b2, ..., bn) that minimize the difference between the predicted values and the actual values of the target variable. This difference is often measured using a cost or loss function, with Mean Squared Error (MSE) being a common choice.

Assumptions of Linear Regression:

Linearity: The relationship between independent and dependent variables should be approximately linear.

Independence: The observations should be independent of each other.

Homoscedasticity: The variance of the residuals (the differences between actual and predicted values) should be roughly constant.

Normality: The residuals should follow a normal distribution.

Simple Linear Regression:

Simple Linear Regression involves predicting a response variable y based on a single predictor variable x. The relationship is represented as:

$$y=b0 + b1*x + \varepsilon$$

Where, y is the response variable.

x is the predictor variable.

b0 is the intercept (where the regression line crosses the y-axis).

b1 is the slope (the change in y for a unit change in x).

and ε represents the error term.

Multiple Linear Regression:

Multiple Linear Regression extends simple linear regression to predict y based on multiple predictor variables:

$$y=b0 + b1*x1 + b2*x2 +…+ bn *xn + ε$$

Where, x1 ,x2 ,…,xn are the predictor variables,

b1 ,b2 ,…,bn are the coefficients for each predictor variable,

and ε represents the error term.

Training the Model:
The model aims to find the best-fitting line that minimizes the sum of squared differences between the observed and predicted values.

This process is often done using the method of least squares, which minimizes the sum of the squared residuals (ε).

Evaluation:
Once the model is trained, we evaluate its performance using various metrics, such as R-squared (R^2), Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). These metrics help assess how well the model fits the data.

Predictions:
Finally, once the model is trained, it can be used to make predictions on new, unseen data by substituting the predictor variable values into the regression equation.

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

Answer-

Anscombe's quartet is a set of four small datasets that have nearly identical simple descriptive statistics but appear very different when graphed. It was created by the statistician Francis Anscombe in 1973 to emphasize the importance of data visualization in understanding and analyzing data. The quartet is often used to illustrate the limitations of relying solely on summary statistics and the power of data visualization.

The quartet consists of the following four datasets-

Dataset 1:

x: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5

y: 8.04, 6.95, 7.58, 8.81, 8.33, 9.96, 7.24, 4.26, 10.84, 4.82, 5.68

Dataset 2:

x: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5

y: 9.14, 8.14, 8.74, 8.77, 9.26, 8.10, 6.13, 3.10, 9.13, 7.26, 4.74

Dataset 3:

x: 10, 8, 13, 9, 11, 14, 6, 4, 12, 7, 5

y: 7.46, 6.77, 12.74, 7.11, 7.81, 8.84, 6.08, 5.39, 8.15, 6.42, 5.73

Dataset 4:

x: 8, 8, 8, 8, 8, 8, 8, 19, 8, 8, 8

y: 6.58, 5.76, 7.71, 8.84, 8.47, 7.04, 5.25, 12.50, 5.56, 7.91, 6.89

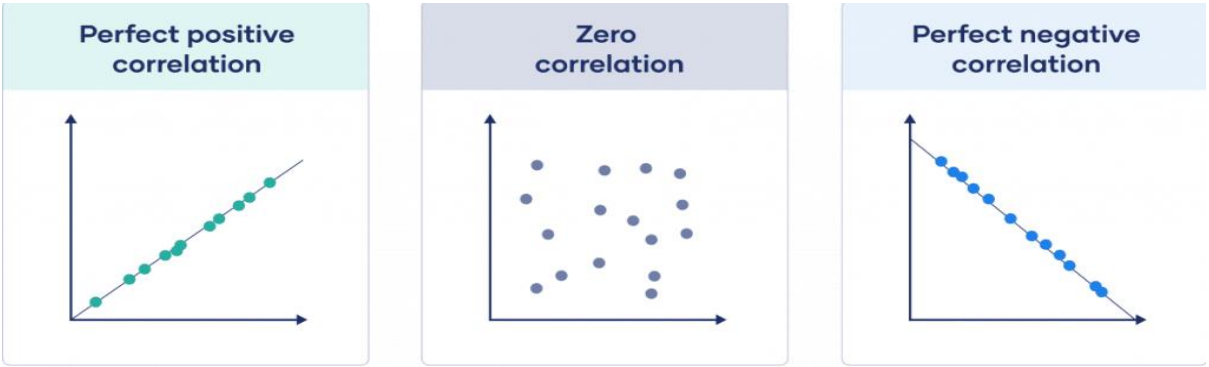
The quartet is intriguing because it shows that despite having the same summary statistics, the relationships between the variables in the datasets are vastly different. It underscores the importance of visualizing data to gain a deeper understanding of its nature and to avoid drawing conclusions based solely on statistics like means, variances, and correlations. Data visualization tools, such as scatter plots, can help reveal patterns, outliers, and relationships that might be otherwise overlooked.

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

Answer-

Pearson's correlation coefficient, often denoted as "r" or "Pearson's r," is a descriptive statistic, meaning that it summarizes the characteristics of a dataset. Specifically, it describes the strength and direction of the linear relationship between two quantitative variables. When variables tend to increase or decrease together, the correlation coefficient is positive, indicating a positive association. Conversely, if one variable tends to increase as the other decreases, the correlation coefficient is negative, indicating a negative association.

The range of Pearson's r is from +1 to -1. A value of +1 signifies a perfect positive correlation, where both variables increase or decrease perfectly together. A value of -1 indicates a perfect negative correlation, where one variable increases as the other decreases. A value of 0 suggests no linear association between the variables.



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 pre-processing refers to the process of transforming the values of variables (features) in a dataset so that they fall within a specific range or have a specific distribution. Scaling is performed to make the data more suitable for machine learning algorithms, particularly those that are sensitive to the scale of the input features.

Most of the times, collected data set contains features highly varying in magnitudes, units and range. If scaling is not done then algorithm only takes magnitude in account and not units hence incorrect modelling. To solve this issue, we have to do scaling to bring all the variables to the same level of magnitude.

The main reasons for scaling are-

i) Regularization: Regularization techniques (e.g., L1 and L2 regularization) assume that all features are on a similar scale. Scaling helps ensure that regularization is applied uniformly to all features.

ii) Faster Convergence: Gradient-based optimization algorithms used in machine learning, like gradient descent, converge faster when working with scaled features. This can speed up training time.

Difference between normalized scaling and standardized scaling-

Normalized Scaling (Min-Max Scaling):

Range: Transforms data to a specific range, typically [0, 1].

Formula: $X_{normalized} = (X - X_{min})/(X_{max} - X_{min})$

It's useful when we want to constrain the values within a specific range or when the data follows a uniform distribution.

It preserves the relationships between data points and can be suitable for algorithms that rely on distances.

Standardized Scaling (Z-score Scaling):

Range: Scales the values to have a mean of 0 and a standard deviation of 1.

Formula: $X_{standardized} = (X - \mu)/\sigma$

It's useful when the data doesn't follow a normal distribution, and we want to make it approximately normally distributed.

Standardization centers the data around zero, which can help when working with models that assume a mean-centered data distribution.

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

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

Answer-

The Variance Inflation Factor (VIF) is a statistical measure used to assess multicollinearity in a regression analysis. A VIF value of infinity (or approaching infinity) typically occurs when there is perfect multicollinearity among the predictor variables. Perfect multicollinearity is a situation where one or more independent variables can be perfectly predicted from other independent variables in the model. In other words, one variable is a perfect linear combination of the others.

In mathematical terms, if the determinant of the correlation matrix of the predictor variables is equal to zero, it indicates perfect multicollinearity. When calculating the VIF using the formula:

$VIF = 1/(1-R2)$

If $R2=1$ due to perfect multicollinearity, the denominator becomes zero, leading to $VIF=\infty$.

A VIF value of infinity (or approaching infinity) typically occurs for one of the following reasons:

Linear Dependence: When one variable is a perfect linear combination of the others, meaning one can be expressed as a constant multiple of the other(s).

Small Sample Size: In some cases, particularly with very small sample sizes, VIF values may become unstable and tend towards infinity because of limited data available to estimate the coefficients precisely.

It's crucial to address multicollinearity appropriately, as ignoring it can lead to unreliable model results and misleading interpretations. To address multicollinearity, we can consider the following strategies:

i) Feature Selection
ii) Combine Variables
iii) Regularization Techniques
iv) Principal Component Analysis(PCA)

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” or "Quantile-Quantile plot," is a graphical tool used in statistics to assess whether a dataset follows a specific theoretical distribution. It compares the quantiles of the dataset to the quantiles of a known probability distribution, typically a normal distribution (the most common use of Q-Q plots). The Q-Q plot is a valuable diagnostic tool for understanding the distribution of data and identifying deviations from a theoretical distribution.

The Q-Q plot works in the following orders-

First, The data is sorted in ascending order, and the quantiles are calculated for the dataset. Then, the quantiles are also calculated for the chosen theoretical distribution, typically a normal distribution, which serves as a reference. Then, The quantiles from the dataset are plotted on the x-axis, and the quantiles from the theoretical distribution are plotted on the y-axis.

Use and importance of a Q-Q plot in linear regression:

Validation of Distribution: Q-Q plots are used to assess the distribution of the residuals in linear regression. Checking the normality of residuals is an important assumption in linear regression. A Q-Q plot helps to determine if the residuals follow a normal distribution. Deviations from a straight line in the Q-Q plot can indicate departures from normality.

Outlier Detection: Outliers in the dataset can significantly impact regression results. A Q-Q plot can reveal outliers as points that deviate from the straight line, helping to identify observations that might be problematic.

Model Adequacy: Q-Q plots are part of the model diagnostic process. When building regression models, it's essential to assess whether the model assumptions are met. A Q-Q plot is a graphical means to check if the assumptions of normality and linearity hold.

Comparing Distributions: Q-Q plots can be used to compare the distribution of a dataset to various theoretical distributions, not just normal. This can be helpful for choosing appropriate statistical tests and models.

In summary, a Q-Q plot is a valuable tool for assessing the distribution of data, particularly in the context of linear regression. It helps ensure that the assumptions underlying the regression model are met and provides insights into the quality and reliability of the model's results.