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 -

Key Observations from Categorical Data Analysis:

1) *Seasonal Preference*: Fall emerged as the most popular season for bookings. Interestingly, the booking count surged in each season from 2018 to 2019, indicating consistent growth.
2) *Monthly Booking Trends*: Bookings were highest from May to October, showcasing a clear upward trend from the beginning of the year until mid-year. Towards the end of the year, the booking numbers gradually declined.
3) *Weather Influence*: Unsurprisingly, clear weather conditions attracted more bookings, aligning with common expectations.
4) *Weekday Bookings*: Thursday, Friday, Saturday, and Sunday experienced higher booking numbers compared to the initial days of the week, indicating a preference for these days.
5) *Working Day Parity*: Bookings remained almost equal on both working and non-working days, showcasing a consistent customer demand regardless of the day.
6) *Business Progression*: Notably, 2019 witnessed a significant increase in bookings compared to the previous year, indicating positive progress and growth in the business.

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

Answer-

The utilization of drop_first = True holds significance as it effectively mitigates the issue of multicollinearity among dummy variables. By excluding one of the categorical levels (k-1 out of k) when creating dummy variables, we inherently define the omitted category.

For example, consider a categorical variable representing colors: Red, Blue, and Green. If we create dummy variables without dropping the first category, the model would encounter a problem of perfect multicollinearity because if it's not Red and not Blue, it's automatically Green. By dropping the first category, say Red, the model inherently understands that if a data point is not Blue, it must be Green. This prevents redundant information and enhances the model's accuracy and interpretability.

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 -

I have assessed the Linear Regression Model based on the following 5 assumptions:

1) Normality of Error Terms: The error terms should follow a normal distribution.
2) Multicollinearity Check: There should be no significant multicollinearity among the variables, ensuring independence of predictors.
3) Linear Relationship Validation: Linear relationships should be evident among the variables, indicating predictable patterns.
4) Homoscedasticity: Residuals should exhibit a consistent variance without any discernible pattern, indicating equal variance across all levels of the variables.
5) Independence of Residuals: Residuals should be independent of each other, indicating no autocorrelation or systematic pattern in the errors over time or across observations.

In essence, these checks ensure the model's assumptions are met, allowing for reliable predictions and interpretations.

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 foremost factors significantly influencing shared bike demand are:

1) *Temperature (temp)*: The temperature plays a pivotal role in shaping demand, indicating a strong correlation with bike rentals.

2) *Season (winter)*: Winter season emerges as a key factor, showcasing a distinct impact on bike usage patterns and demand.
3) *Month (September)*: Specifically, September stands out as a crucial month, demonstrating a notable influence on shared bike demand, likely due to specific events or seasonal shifts.

General Subjective Questions

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

Answer -

Linear regression is a widely used statistical method in machine learning and statistics. It's used to predict a continuous outcome variable (also called a response or dependent variable) based on one or more predictor variables (also called independent variables or features). Here's a detailed explanation of the linear regression algorithm:

1. Assumptions of Linear Regression:

Linearity: Assumes a linear relationship between predictors and the response.

Independence: Assumes that observations are independent of each other.

Homoscedasticity: Assumes that the variance of errors is constant across all levels of the predictor variable(s).

Normality: Assumes that errors are normally distributed.

2. Simple Linear Regression:

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

$$y=\beta 0 +\beta 1 \cdot x+\varepsilon$$

y is the response variable.

x is the predictor variable.

$\beta 0$ is the intercept (where the regression line crosses the y-axis).

$\beta 1$ is the slope (the change in y for a unit change in x).

ε represents the error term.

3. Multiple Linear Regression:

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

$$y=\beta 0 +\beta 1 \cdot x1 +\beta 2 \cdot x2 +\ldots+\beta n \cdot xn +\varepsilon$$

$x1 ,x2 ,\ldots,xn$ are the predictor variables.

$\beta 1 ,\beta 2 ,\ldots,\beta n$ are the coefficients for each predictor variable.

4. Fitting the Model:

The model aims to find the best-fitting line (or hyperplane in multiple dimensions) 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).

5. Interpreting Coefficients:

The intercept ($\beta 0$) represents the expected value of y when all predictor variables are 0.

Each coefficient ($\beta 1 ,\beta 2 ,\ldots$) represents the change in the response variable for a unit change in the corresponding predictor variable, holding other predictors constant.

6. Evaluating the Model:

R-squared (R^2): Represents the proportion of the variance in the dependent variable that is predictable from the independent variables. Higher R^2 indicates a better fit.

Residual Analysis: Involves checking assumptions by analyzing residuals (differences between observed and predicted values) through plots and statistical tests.

7. Predictions:

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.

Linear regression is a fundamental algorithm, and its simplicity and interpretability make it a valuable tool in various fields such as economics, finance, biology, and engineering.

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

Answer -

Anscombe's quartet comprises four datasets that have nearly identical simple descriptive statistics yet have very different distributions and appear very different when graphed. It was created by the statistician Francis Anscombe in 1973 to demonstrate the importance of graphing data before analyzing it and the effect of outliers on statistical properties.

The quartet consists of the following four datasets, each containing 11 (x, y) points:

Dataset I:

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 II:

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 III:

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 IV:

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

Key Points:

Descriptive Statistics:

All four datasets have the same mean and variance for both x and y, as well as the same correlation coefficient between x and y.

These similarities in summary statistics might lead someone to believe the datasets are similar in nature.

Graphical Differences:

When graphed, Dataset I shows a clear linear relationship.

Dataset II has a more curved pattern.

Dataset III has an outlier that affects the linear correlation significantly.

Dataset IV shows no correlation except for one outlier point.

Implications:

Anscombe's quartet highlights the importance of data visualization. Relying solely on summary statistics can be misleading.

It emphasizes that statistical measures might not capture the full complexity of the dataset and that graphical exploration is crucial to understanding the underlying patterns.

In summary, Anscombe's quartet is a powerful illustration of the necessity to visualize data before drawing any conclusions, demonstrating how datasets with similar statistical properties can behave quite differently when graphed.

3. What is Pearson’s R? (3 marks)

Answer -

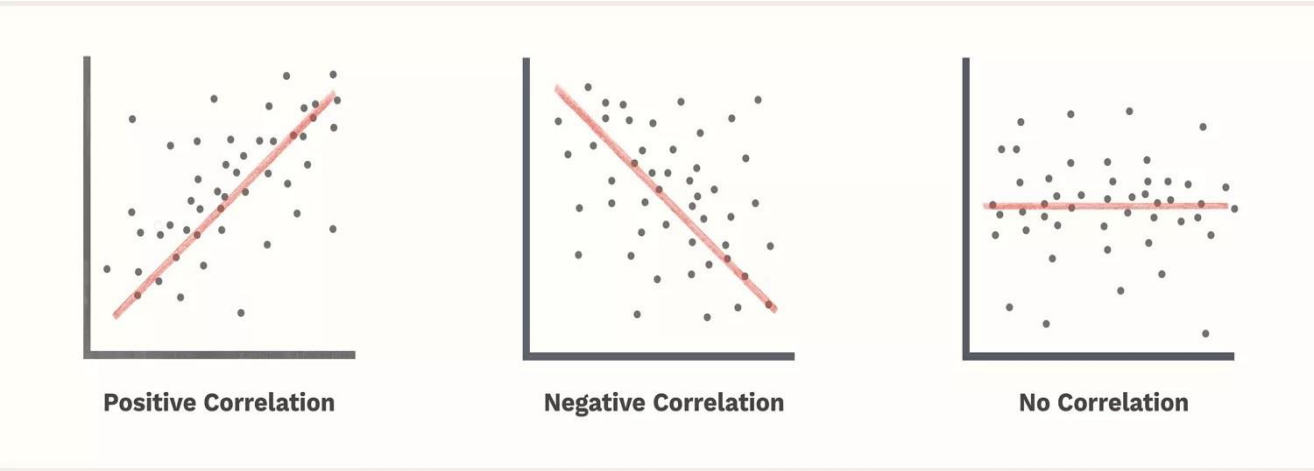
Pearson's r is a numerical measure that quantifies the strength and direction of a linear relationship between two variables. When variables tend to increase or decrease together, the correlation coefficient is positive, indicating a positive association. Conversely, if one variable tends to increase 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.

In simpler terms, if r is positive, it means as one variable goes up, the other tends to go up as well. If r is negative, it means as one variable goes up, the other tends to go down. A value near 0 indicates weak or no linear relationship between the variables.

The graphs are -



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

Answer -

Definition: Scaling in the context of data analysis refers to the process of transforming numerical variables to a standard scale, making it easier to compare different variables on equal footing. It ensures that variables with different units and magnitudes are comparable and do not introduce bias in the analysis.

Why Scaling is Performed:

Magnitude Normalization: Variables often have different units (e.g., height in meters vs. income in dollars) and scales, which can skew the analysis. Scaling brings them to a similar scale.

Algorithm Performance: Many machine learning algorithms are sensitive to the scale of the input features. Scaling ensures fair treatment among variables in algorithms like k-means clustering, support vector machines, etc.

Interpretability: Scaling makes it easier to interpret the coefficients in linear models. Without scaling, coefficients represent changes in the target variable for a one-unit change in the predictor, which might not make sense for all predictors.

Difference Between Normalized Scaling and Standardized Scaling:

Normalized Scaling (Min-Max Scaling):

Range: Scales the values to a specific range, usually [0, 1].

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

Advantage: Preserves the relationships between the original data points.

Use Case: When the distribution of data does not follow a normal distribution.

Standardized Scaling (Z-Score Normalization):

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

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

Advantage: Works well for data following a normal distribution. Preserves the shape of the original distribution.

Use Case: When dealing with algorithms that assume normality, like linear regression, or when the data follows a normal distribution.

In summary, scaling is essential to ensure fair comparisons and accurate modeling in data analysis. The choice between normalized and standardized scaling depends on the nature of the data and the requirements of the specific analysis or algorithm being used.

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

Answer -

A situation where the Variance Inflation Factor (VIF) is calculated as infinite typically occurs due to perfect multicollinearity among the predictor variables. Perfect multicollinearity happens when one or more independent variables in a regression model can be exactly predicted from other variables, leading to an infinite correlation between them.

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=∞.

Perfect multicollinearity can create issues in regression analysis. It makes it impossible to estimate unique coefficients for the correlated variables because their changes are perfectly predictable from one another. This situation typically arises in the following cases:

a)Dummy Variable Trap: When dummy variables are created for categorical variables, and one of the dummy variables can be exactly predicted from the others. For example, in a binary category represented by two dummy variables, if one variable is 1, the other must be 0, leading to perfect correlation.

B)Linear Dependence: When two or more variables are perfectly linearly dependent, meaning one can be expressed as a constant multiple of the other(s).

C)Data Errors: Sometimes, data errors or coding mistakes can create artificial perfect correlations between variables.

To handle this issue, it's crucial to identify and resolve multicollinearity problems, often by dropping one of the correlated variables or using techniques like Principal Component Analysis (PCA) to transform the variables and remove multicollinearity.

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 whether a dataset follows a particular theoretical distribution, like the normal distribution. It compares the quantiles of the dataset against the quantiles of the theoretical distribution.

How it Works:

Sorting: First, the data points are sorted in ascending order.

Theoretical Quantiles: For each data point, a corresponding theoretical quantile from the specified distribution is calculated.

Plotting: The data quantiles are plotted against the theoretical quantiles. If the points fall along a straight line, the data closely follows the theoretical distribution.

Use and Importance in Linear Regression:

Normality Check: In linear regression, it's essential to ensure that the residuals (the differences between observed and predicted values) are normally distributed. Deviation from normality can affect the validity of statistical inferences.

Identifying Outliers: Q-Q plots help in identifying outliers and skewed distributions. Outliers can significantly impact the regression model, leading to biased results.

Assumptions Validation: Linear regression assumes that the residuals are normally distributed. Q-Q plots provide a visual assessment of this assumption. If the residuals follow a straight line in the Q-Q plot, it indicates normality.

Robustness Check: Q-Q plots are a robust way to check normality because they are not influenced by small sample sizes.

Importance:

Statistical Validity: Ensuring that the residuals are normally distributed is crucial for making valid statistical inferences, constructing confidence intervals, and performing hypothesis tests in linear regression.

Model Accuracy: A linear regression model with normally distributed residuals tends to make more accurate predictions and provides reliable estimates of model parameters.

In summary, Q-Q plots are vital tools in linear regression analysis, helping to validate assumptions, detect outliers, and ensure the statistical validity and accuracy of the model.