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?

Answer: I have done analysis on categorical columns using the boxplot and bar plot. Below are the few points we can infer 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 a greater 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 a greater 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?

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?

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?

Answer: I have validated the assumption of Linear Regression Model based on below 5 assumptions:

- Normality of error terms - Error terms should be normally distributed.
- Multicollinearity check - There should be insignificant multicollinearity among variables.
- Linear relationship validation - Linearity should be visible among variables.
- Homoscedasticity - There should be no visible pattern in residual values.
- Independence of residuals - No autocorrelation

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

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

- Temp
- Winter
- Sep

General Subjective Questions

Question 1: Explain the linear regression algorithm in detail.
Answer:

Linear regression is a fundamental and widely used statistical technique for modelling the relationship between a dependent variable and one or more

independent variables. It assumes a linear relationship between the independent variables (predictors) and the dependent variable (outcome). The primary goal of linear regression is to find the best-fitting straight line that describes the relationship between these variables.

Basic Concept:

At its core, linear regression aims to find a linear equation of the form:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_n X_n + \epsilon$$

Where:

- Y is the dependent variable,
- $X_1, X_2, ..., X_n$ are the independent variables,
- $\beta_0, \beta_1, \beta_2, ..., \beta_n$ are the coefficients representing the effect of each independent variable on the dependent variable, and
- ϵ is the error term, representing the difference between the observed and predicted values.

How it Works:

1. **Objective**: The main objective of linear regression is to minimize the sum of the squared differences between the observed values and the predicted values. This technique aims to find the line that best fits the data points.
2. **Cost Function:** The most common method to achieve this is by defining a cost function, often the Mean Squared Error (MSE):

   $$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

   Where y_i are the actual values, $\hat{y}_i$ are the predicted values, and n is the number of data points.
3. **Minimization**: The goal then becomes to minimize this MSE by adjusting the coefficients:

   $\beta_0, \beta_1, ..., \beta_n$.

   This is often done using optimization algorithms like Gradient Descent.
4. **Finding Coefficients:** The coefficients are calculated using the following formulas:

   $$\beta_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

   $$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

   Where $\bar{x}$ and $\bar{y}$ are the mean values of the independent and dependent variables, respectively.

5. **Assumptions**: Linear regression also assumes certain conditions like linearity, independence of errors, homoscedasticity (constant variance of errors), and normality of errors.

Applications:

Linear regression finds applications across various fields, including economics, finance, social sciences, and natural sciences. It is used for:

- Predicting stock prices based on historical data.
- Determining the impact of advertising on sales.
- Analyzing the relationship between variables in scientific research.
- Forecasting trends in weather, sales, or population growth.

In conclusion, linear regression is a powerful statistical tool that provides a simple yet effective way to model the relationship between variables. Its simplicity, interpretability, and wide-ranging applications make it a cornerstone of statistical analysis and predictive modeling.

Question 2: Explain the Anscombe's quartet in detail.

Answer:
Anscombe's quartet is a set of four datasets that have nearly identical simple descriptive statistics, yet they exhibit very different characteristics when plotted and analyzed. This quartet was created by the statistician Francis Anscombe in 1973 to emphasize the importance of graphing data before analyzing it and to demonstrate the potential pitfalls of relying solely on summary statistics.

Description of the Quartet:

1. **Dataset 1:**

   It consists of a simple linear relationship.
   $y=3+0.5x$

   The data points are tightly clustered around a straight line.

2. **Dataset 2:**

Also, a linear relationship, but with an outlier.

One point deviates significantly from the rest.

y=3+0.5x

The outlier affects the slope and intercept of the regression line.

3. **Dataset 3:**

   Appears to have a non-linear relationship.

   y=3+0.5x+random noise

   Although it seems non-linear, it is a perfect quadratic relationship.

4. **Dataset 4:**

   Has the same

   x-values for each data point, except for one outlier.

   y=3+0.5x

The outlier drastically affects the correlation coefficient and the regression line.

Implications and Lessons:

- **Visualizing Data:**
  Anscombe's quartet highlights the importance of graphing data. Different datasets can have the same means, variances, correlations, and regression coefficients, yet look entirely different when plotted.
- **Misleading Statistics:**
  Relying solely on summary statistics can be misleading. In all four datasets, the means, variances, and correlations are the same, yet the data patterns are vastly different.
- **Outliers and Influential Points:**
  Outliers, as seen in Datasets 2 and 4, can heavily influence regression lines, affecting the interpretation of relationships.
- **Model Assumptions:**
  It underscores the need to check assumptions of statistical models. For instance, Dataset 3 shows that even though it appears non-linear, it fits a perfect quadratic relationship.

- **Statistical Robustness:**
  Analysts should be cautious and explore their data visually before drawing conclusions. Understanding the data's underlying structure is crucial for accurate analysis and interpretation.

In summary, Anscombe's quartet is a powerful demonstration of the limitations of summary statistics and the necessity of graphical exploration in data analysis. It serves as a timeless reminder to researchers and analysts to never underestimate the value of visualizing data before drawing conclusions.

Question 3: What is Pearson's R?

Answer:

Pearson's correlation coefficient, often denoted as *r*, is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. It was developed by Karl Pearson, a pioneer in the field of statistics, and it ranges from -1 to 1.

Formula:
The formula for Pearson's correlation coefficient between two variables *X* and *Y* with *n* data points is:

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}}$$

Where:

- X_i and Y_i are the individual data points,
- $\bar{X}$ and $\bar{Y}$ are the means of X and Y respectively,
- The numerator calculates the covariance between X and Y, while the denominator normalizes this by dividing by the product of the standard deviations of X and Y.

Interpretation:
1. Strength of Relationship:

- The value of r ranges from -1 to 1.
- r=1 indicates a perfect positive linear relationship, where Y increases as X increases.

- r=−1 indicates a perfect negative linear relationship, where Y decreases as X increases.
- r=0 suggests no linear relationship between the variables.

2. Direction:

- The sign of r indicates the direction of the relationship.
- Positive r means that as X increases, Y tends to increase.
- Negative r means that as X increases, Y tends to decrease.

3. Magnitude:

- The closer r is to 1 or -1, the stronger the linear relationship.
- A value closer to 0 suggests a weaker linear relationship.

4. Example:

- For example, if
  r=0.8, it indicates a strong positive linear relationship between the variables. This means that as X increases,Y tends to increase proportionally.

Applications:

- Pearson's r is widely used in fields such as psychology, sociology, economics, and biology to assess the strength and direction of relationships between variables.
- It helps researchers understand how changes in one variable are associated with changes in another, aiding in predictive modeling and decision-making.
- In finance, it can be used to analyze the correlation between the returns of different stocks or assets, guiding portfolio diversification strategies.

Pearson's correlation coefficient is a valuable tool for understanding linear relationships between variables, providing a numerical measure of how closely related two variables are, and in what direction they tend to move together.

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

Answer:

Scaling is a preprocessing technique in data analysis and machine learning where the values of features (variables) are transformed to fit a specific scale. The purpose of scaling is to standardize the range of the independent variables so that they can be compared on a common scale. This process is crucial in situations where the variables have different units or scales of measurement, ensuring that one variable does not dominate the analysis simply because of its larger magnitude.

Why Scaling is Performed:

1. **Algorithm Sensitivity:**

Many machine learning algorithms are sensitive to the scale of the input features. Models such as k-means clustering or support vector machines (SVMs) calculate distances between data points. If the features are not on a similar scale, the algorithm might give more weight to variables with larger magnitudes, leading to biased or incorrect results.

2. **Gradient Descent Convergence:**

In algorithms like linear regression and neural networks that use gradient descent optimization, scaling helps the algorithm converge faster towards the minimum of the cost function. This is because large differences in the scales of features can cause the optimization to oscillate or take longer to converge.

3. **Interpretability:**

Scaling makes the interpretation of coefficients or feature importance easier. When variables are on the same scale, it is simpler to compare the impact of different features on the target variable.

Normalized Scaling vs. Standardized Scaling:

1. **Normalized Scaling (Min-Max Scaling):**

- Normalization scales the values of the features to a fixed range, usually between 0 and 1.
- The formula for normalized scaling is:

$$X_{\text{normalized}} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}}$$

- This method preserves the relative relationships between values but compresses the range to fit within a specific interval.

2. **Standardized Scaling (Z-score Standardization):**

- Standardization transforms the data to have a mean of 0 and a standard deviation of 1.
- The formula for standardization is:

$$X_{\text{standardized}} = \frac{X-\mu}{\sigma}$$

- Here, μ is the mean of the feature, and σ is the standard deviation.
- Standardization makes the data look like a standard normal distribution, centered around 0 with a spread of 1.

Difference:

Normalization maintains the original distribution of the data within a fixed range, suitable for algorithms that require input features to be on a bounded interval.

Standardization rescales the data to have a mean of 0 and a standard deviation of 1, preserving the shape of the distribution while making it easier to compare different features.

In summary, scaling is performed to ensure that variables are on a common scale for fair comparison and accurate model training. Normalized scaling compresses the range of data to a fixed interval, while standardized scaling centers the data around 0 with a standard deviation of 1, aiding in algorithm convergence and interpretability. The choice between the two methods depends on the specific requirements of the analysis and the characteristics of the dataset.

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

Answer:

In statistical analysis, the Variance Inflation Factor (VIF) is a measure used to detect multicollinearity in regression models. Multicollinearity occurs when independent variables in a regression model are highly correlated with each other, which can lead to unreliable and misleading results. VIF quantifies how

much the variance of the estimated coefficients is inflated due to multicollinearity.

Understanding VIF:

1. **Formula for VIF:**

The formula for the VIF of the ith variable in a regression model is:

$$\mathrm{VIF}_i = \frac{1}{1-R_i^2}$$

Where R_i^2 is the coefficient of determination of the regression of the i-th independent variable on the remaining $p-1$ independent variables.

1. **Interpretation**:

- A VIF of 1 indicates no multicollinearity.
- Typically, a VIF greater than 10 is considered indicative of significant multicollinearity.

VIF Becoming Infinite:

Perfect Multicollinearity:

- VIF becomes infinite when there is perfect multicollinearity between variables.
- Perfect multicollinearity occurs when one or more independent variables in a regression model can be exactly predicted from a linear combination of other variables.
- In this case, the coefficient of determination R^2_i, in the VIF formula is equal to 1.

Zero Variation:

- When R^2_i=1 in the VIF formula, the denominator becomes 1- R^2_i =0.
- Division by zero results in an undefined or infinite value for the VIF.

Example:

- For instance, consider a dataset where one variable can be expressed as a perfect linear combination of other variables. In this scenario, the VIF

for that variable would be infinite.

Consequences and Remedies:

Consequences of Perfect Multicollinearity:

Perfect multicollinearity makes it impossible to estimate the coefficients of the affected variables in the regression model.

The regression coefficients become sensitive to small changes in the data, leading to unreliable results.

Remedies:

If perfect multicollinearity is detected, it is crucial to address the issue before proceeding with the analysis.

One common remedy is to remove one of the correlated variables from the model.

Another approach is to create a composite variable from the correlated variables, reducing the dimensionality of the model.

Conclusion:

In summary, the occurrence of an infinite VIF indicates perfect multicollinearity in a regression model. This situation arises when one or more independent variables can be precisely predicted from a linear combination of other variables. Perfect multicollinearity leads to unreliable coefficient estimates and can significantly affect the interpretability and validity of the regression model. Identifying and addressing multicollinearity is essential for obtaining accurate and trustworthy results in statistical analysis.

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

Answer:
A Q-Q (Quantile-Quantile) plot is a graphical tool used to assess if a set of data follows a certain theoretical distribution, such as the normal distribution. It compares the quantiles of the dataset against the quantiles of a theoretical distribution, typically a standard normal distribution (mean of 0 and standard

deviation of 1). This comparison helps to visually assess whether the data deviates significantly from the expected distribution.

How Q-Q Plot Works:

Steps to Construct:

To create a Q-Q plot, the data is first sorted in ascending order.

The sorted data points are then plotted against the corresponding quantiles from the theoretical distribution.

Interpretation:

If the points on the plot fall approximately along a straight line, it suggests that the data is normally distributed.

Deviations from the straight line indicate departures from normality.

Key Points:

Points above the line suggest that the data have larger values than expected under the theoretical distribution.

Points below the line suggest that the data have smaller values than expected.

Use and Importance in Linear Regression:

Assumption Checking:

In linear regression analysis, one of the key assumptions is that the residuals (the differences between observed and predicted values) are normally distributed.

Q-Q plots are used to check this assumption by plotting the residuals against the quantiles of a theoretical normal distribution.

Identifying Residual Patterns:

A Q-Q plot helps identify patterns or deviations in the residuals.

If the residuals follow a normal distribution, the points on the Q-Q plot will fall approximately along a straight line.

Non-linear patterns or significant deviations from the line indicate non-normality in the residuals.

Model Validity:

Ensuring that residuals are normally distributed is crucial for the validity of the linear regression model.

If the residuals do not follow a normal distribution, it could indicate that the model is missing important variables, the relationship is not linear, or there are outliers affecting the results.

Adjustments and Remedies:

Detecting departures from normality in the Q-Q plot prompts further investigation.

Remedies might include transforming variables, adding polynomial terms, or addressing outliers to improve the model's performance.

Robust Inference:

A linear regression model with normally distributed residuals allows for more robust statistical inference.

Confidence intervals, hypothesis tests, and predictions are more reliable when the assumptions of normality hold.

Conclusion:

In conclusion, a Q-Q plot is a valuable tool in linear regression analysis for checking the assumption of normally distributed residuals. It provides a visual assessment of whether the residuals follow the expected pattern of a normal distribution. By identifying deviations from normality, analysts can make informed decisions about the validity and reliability of their regression model, leading to more accurate interpretations and predictions. A well-constructed Q-Q plot ensures the robustness and integrity of the statistical analysis.