Assignment-based Subjective Questions

Question 1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: <Your answer for Question 1 goes below this line> (Do not edit)

- The scatter plot reveals that bike bookings in 2019 surpassed those in 2018, indicating an increase in usage over the year.
- The data shows peak bookings during September and October, with a consistent level of high bookings from April through August, suggesting strong seasonal demand.
- The analysis demonstrates a positive correlation between temperature variables (such as temp, atemp, and humidity) and the number of bookings. This indicates that warmer and more comfortable weather conditions tend to encourage more bike rentals.
- Customers prefer to hire bikes on clear, dry days. Bookings are notably higher during the summer and winter months, reflecting a preference for biking in milder weather conditions as opposed to rainy or cloudy weather.

Question 2. Why is it important to use **drop_first=True** during dummy variable creation? (Do not edit)

Total Marks: 2 marks (Do not edit)

Answer: <Your answer for Question 2 goes below this line> (Do not edit)

- **Avoids Multicollinearity**: By dropping one dummy variable, it eliminates perfect multicollinearity among the features, crucial for the proper functioning of many statistical models that rely on the independence of predictors.
- **Reduces complexity**: Reducing the number of variables not only simplifies the model's complexity but also ensures that each parameter estimation is statistically meaningful, avoiding redundant or dependent predictors.
- **Clear Model Interpretation**: Dropping the first category sets a clear reference or baseline, making it easier to interpret how other categories compare against it in terms of their impact on the dependent variable.

Question 3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (Do not edit)

Total Marks: 1 mark (Do not edit)

Answer: <Your answer for Question 3 goes below this line> (Do not edit)

- Based on the analysis of pair plots, temperature-related factors such as actual temperature (temp), feeling temperature (atemp), and humidity (hum) show a more pronounced positive correlation with booking numbers compared to other features.

Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: <Your answer for Question 4 goes below this line> (Do not edit)

- **Distribution of Residual Errors**: A density plot of the residual errors suggested a normal distribution around zero.
- **Homoscedasticity**: A scatter plot of residuals versus fitted values did not reveal any evident patterns or increasing/decreasing variance, supporting the assumption of constant variance across the range of predictions.
- **Normality Check with Q-Q Plot**: The Q-Q plot demonstrates that residuals largely adhere to the theoretical normal distribution, confirming the normality assumption essential for the validity of the linear regression model.

Question 5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (Do not edit)
Total Marks: 2 marks (Do not edit)
Answer: <Your answer for Question 5 goes below this line> (Do not edit)

- Top 3 features contributing to the model are temp, yr, season_winter

---

General Subjective Questions

Question 6. Explain the linear regression algorithm in detail. (Do not edit)
Total Marks: 4 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 6 goes here>

Linear regression is a fundamental statistical and machine learning technique used to predict a dependent variable (outcome) based on the values of one or more independent variables (predictors). It is called "linear" because it assumes that the relationship between the dependent and independent variables can be described using a straight line.

Fundamentals of Linear Regression:

Objective: The main goal of linear regression is to find a linear relationship between the input variables (like temperature, humidity, etc.) and the output variable (like bike bookings). By establishing this relationship, you can predict future values of the dependent variable when only the independent variables are known.

How it Works:

Model Formula: In its simplest form (simple linear regression), the model predicts the outcome variable (Y) as a linear combination of the input variable (X) and two parameters, the intercept (α) and slope (β). The equation is:

$$Y = \alpha + \beta X + \varepsilon$$

Multiple Regression: In more complex scenarios, like the ones discussed in assignment, multiple regression is used where several input variables contribute to the outcome. The equation extends to:

$$Y = \alpha + \beta 1X1 + \beta 2X2 + ... + \beta nXn + \varepsilon$$

where X1, X2,...,Xn are different independent variables, and β1,β2,...,βn are the coefficients showing the influence of each variable.

Estimation of Coefficients:

Least Squares Method: The coefficients (α, β) are typically estimated using the least squares method, which finds the line (or hyperplane in multiple regression) that minimizes the sum of the squared residuals (differences between observed and predicted values).

Model Validation:

Assumptions: Linear regression relies on several assumptions:

- **Linearity**: The relationship between the independent and dependent variables should be linear.
- **Independence**: Observations are independent of each other.
- **Homoscedasticity**: The residuals (errors) should have constant variance.
- **Normality of Residuals**: The residuals should be normally distributed.

Question 7. Explain the Anscombe's quartet in detail. (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 7 goes here>

Anscombe's quartet comprises four distinct datasets that have nearly identical simple statistical properties, yet appear very different when graphed. Each dataset consists of eleven points (x, y) and was constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance and limitations of statistical graphs and the effect of outliers on statistical properties. The quartet is often used to emphasize the importance of visualizing data before analyzing it and the effect of outliers and other data anomalies on statistical interpretations.

Details of Anscombe's Quartet

Key Characteristics:

- Mean of x: The mean of the x-values is approximately the same across all four datasets.
- Variance of x: The variance is also nearly identical for the x-values in each dataset.
- Mean of y: Similarly, the mean of the y-values is about the same across the datasets.
- Variance of y: The variance of the y-values is also consistent across the four datasets.
- Correlation: Each dataset shows nearly the same correlation coefficient between x and y.
- Linear Regression Line: When a simple linear regression is performed on each dataset, the line of best fit is the same for all four datasets ($y = 3.00 + 0.500x$).

The Four Datasets

1. Dataset I: Appears to be a simple linear relationship, matching the regression line closely. This dataset is what one might expect when considering a dataset with such statistical properties.

2. Dataset II: Shows a clear curvature (quadratic relationship), which indicates that a simple linear regression is not appropriate despite having the same regression line as dataset I.
3. Dataset III: Appears linear but with one outlier affecting both the slope of the regression line and the correlation coefficient. Without this outlier, the correlation would be much higher and the line of best fit would look different.
4. Dataset IV: Consists mainly of a cluster of points with one significant outlier far from the cluster, heavily influencing the regression line. Without this outlier, the x-values would have almost no variation, and the correlation would be near zero.

Implications and Lessons

Anscombe's quartet is a compelling demonstration of why it's critical to graph data before analyzing it. Here are some lessons it teaches:

- Visual Inspection: It highlights the need to look at a graphical representation of the data before analyzing it using summary statistics or regression models.
- Impact of Outliers: The datasets illustrate how outliers can significantly affect the results of statistical analyses.
- Limitations of Statistics: The quartet serves as a cautionary tale about the limits of relying solely on summary statistics. Statistical measures can be the same but hide vastly different distributions and relationships.
- Model Assumptions: It emphasizes checking the assumptions underlying a statistical model. For linear regression, verifying the linearity of the data's relationship is crucial, as non-linear relationships (like those in datasets II and III) can lead to incorrect conclusions if treated linearly.

Question 8. What is Pearson's R? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 8 goes here>

Pearson's R, also known as the **Pearson correlation coefficient** is a measure of the linear correlation between two variables, denoted as X and Y. It quantifies how well the relationship between these variables can be described by a linear equation.

Definition and Calculation

The Pearson correlation coefficient is mathematically defined as the covariance of the two variables divided by the product of their standard deviations:

$r = \text{cov}(X, Y) / (\sigma X * \sigma Y)$

Where:

- $\text{cov}(X, Y)$ is the covariance between variables X and Y.
- σX and σY are the standard deviations of X and Y, respectively.

Properties

- **Value Range**: Pearson's R values range from -1 to 1. A correlation of -1 indicates a perfect negative linear relationship, +1 indicates a perfect positive linear relationship, and 0 indicates no linear relationship.
- **Symmetry**: The correlation between X and Y is the same as the correlation between Y and X.
- **Unitless**: Pearson's R is dimensionless, facilitating the comparison of correlations across different studies or datasets.

Interpretation

- **Positive Correlation**: If R is greater than 0, it indicates a positive relationship; as one variable increases, the other also increases.
- **Negative Correlation**: If R is less than 0, it indicates a negative relationship; as one variable increases, the other decreases.
- **Strength**: The closer the absolute value of R is to 1, the stronger the linear relationship.

Question 9. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 9 goes here>

Scaling is a method of transforming data to alter the range of variable values, typically to bring them into a consistent range. This process is crucial in data preprocessing, especially in statistical modeling, where different features might measure using different units and scales (e.g., kilometers, kilograms, or counts).

Why is Scaling Performed?

- **Equal Importance**: Many machine learning algorithms, such as those involving distance calculations assume that all features have the same scale for them to treat all features with equal importance during training.
- **Avoiding Bias**: Without scaling, features with larger ranges could dominate the decision process in some models, introducing a bias to those features with naturally larger values or greater variance.

Types of Scaling

Normalized Scaling: Normalization adjusts the data values so that they fall within a specified range, typically 0 to 1, or -1 to 1.

Standardized Scaling: Standardization transforms data to have a mean of zero and a standard deviation of one. It does not bind values to a specific range, which might be useful for features with outliers or when there is no specific range requirement for the data.

Question 10. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 10 goes here>

The condition where the Variance Inflation Factor (VIF) becomes infinite typically occurs when there is perfect multicollinearity among the independent variables in a regression model. VIF is a measure used to detect the level of multicollinearity in a set of multiple regression variables. It quantifies how much the variance of an estimated regression coefficient increases if your predictors are correlated.

When analyzing the formula for VIF, $VIF_i = 1 / (1 - R^2_i)$, it becomes clear how the metric can reach infinity. The denominator, $1 - R^2_i$, becomes zero when R^2_i equals 1. This scenario occurs when the independent variable i can be perfectly predicted from the other independent variables in the model, resulting in perfect multicollinearity. Essentially, $R^2_i = 1$ signifies that 100% of the variance in variable i is explained by the other variables, leaving no residual variance. When this happens, the denominator of the VIF formula becomes zero, and dividing by zero yields an infinite result.

Question 11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.
(Do not edit)
Total Marks: 3 marks (Do not edit)
Answer: Please write your answer below this line. (Do not edit)

<Your answer for Question 11 goes here>

In the context of linear regression, Q-Q plots are primarily used to analyze the normality of residuals. Residuals in linear regression are the differences between the observed values of the dependent variable and those predicted by the model. The assumptions of linear regression include the idea that these residuals are normally distributed.

Importance of Q-Q Plots in Linear Regression

Assumption Check: One of the key assumptions of linear regression is that the residuals are normally distributed, particularly when the model is used for hypothesis testing regarding the regression coefficients. A Q-Q plot provides a visual means to inspect this assumption. If the residuals are normally distributed, the points on the Q-Q plot will align closely with a straight line.

Spotting Anomalies: The Q-Q plot can also help in identifying outliers. Points that deviate significantly from the straight line in the plot indicate anomalies in the data. These outliers can potentially skew the regression results and might necessitate further investigation or remedial measures such as transformations or robust regression techniques.
