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)

I have done analysis on categorical variables using the boxplot and bar plot. Below points we can infer from the visualization:

1. Seasonal trend: Bookings were highest during the fall season.
   All seasons saw a sharp rise in bookings between 2018 and 2019, suggesting rising demand.
2. Monthly Trend: The months of May, June, July, August, September, and October saw the highest booking volumes. From the beginning to the middle of the year, the trend grew, but as the year came to a close, it started to fall.
3. Weather Impact: As anticipated, the most bookings occurred during clear weather.
4. Day of the Week: Bookings were higher on Thursdays, Fridays, Saturdays, and Sundays than at the start of the week.
5. Influence of Holidays: On holidays, fewer reservations were booked, probably because people wanted to spend more time with their families at home.
6. Working vs. Non-Working Days: On both working and non-working days, bookings were almost equal, suggesting regular usage trends.
7. Annual Growth: Bookings increased in 2019 compared to 2018, showing positive growth and business progress.

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)

Use of **drop_first = True** is crucial since it minimizes the extra column produced when creating dummy variables. As a result, it reduced the correlations that are formed between 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.

Suppose we wish to build a dummy variable for a categorical column that has three different types of values. It is evidently C if one of the variables is not A and B. Therefore, the third variable is not necessary to determine the C.

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)

By looking at pair-plot of numeric variables, ‘temp’ and ‘atemp’ variables has the highest correlation with the target variable.

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)

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

1. Linearity: The relationship between predictors and the target variable should be linear.
2. Independence of Errors: Residuals should be independent of each other.
3. Normality of Errors: Residuals should follow a normal distribution.
4. Outliers and Leverage Points: The model shouldn't be overly impacted by high leverage points or severe outliers.

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)

The top 3 features that has significant impact towards explaining the demand of the shared bikes are:

1. temp
2. fall (season)
3. sep (month)

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>

A supervised learning technique called linear regression is used to forecast a continuous target variable (y) from one or more input features (x). The line or hyperplane that minimizes the error between the actual and anticipated values is determined to be the best fit. This is a thorough explanation:

1. Objective of Linear Regression
A supervised learning technique called linear regression is used to forecast a continuous target variable (y) from one or more input features (x). The line or hyperplane that minimizes the error between the actual and anticipated values is determined to be the best fit.

Simple Linear Regression

In simple linear regression (1 predictor):
$y = \beta_o + \beta_1 x + \varepsilon$
- β_o: Intercept (value of y when x = 0).
- β_1: Slope (rate of change in y per unit change in x).
- ε: Error term (unexplained variance).

Multiple Linear Regression

In multiple linear regression (multiple predictors):
$y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + ... + \beta n x n + \varepsilon$

2. Steps in the Linear Regression Algorithm

Step 1: Hypothesis Formation

The hypothesis is that the target variable y is a linear function of the predictors. The equation is parameterized by coefficients (β_o, β_1, ..., βn).

Step 2: Loss Function (Error Measurement)

To find the best-fitting line, the algorithm minimizes the Residual Sum of Squares (RSS):
$RSS = \Sigma(y_i - \hat{y}_i)^2$
Where y_i is the actual value, $\hat{y}_i$ is the predicted value, and n is the number of observations.

Step 3: Optimization (Coefficient Estimation)

The Ordinary Least Squares (OLS) method is used to estimate the coefficients β_o, β_1, ..., βn.
For simple linear regression, the slope (β_1) and intercept (β_o) are calculated as:
$\beta_1 = \Sigma((x_i - \bar{x})(y_i - \bar{y})) / \Sigma((x_i - \bar{x})^2)$
$\beta_o = \bar{y} - \beta_1 \bar{x}$
For multiple linear regression, matrix algebra is used:
$\beta = (X^T X)^{-1} X^T y$

Step 4: Predictions

Once coefficients are estimated, predictions for new data points (x) are made using:
$\hat{y} = \beta_o + \beta_1 x_1 + \beta_2 x_2 + ... + \beta n x n$

Step 5: Model Evaluation

Evaluate the model's performance using these metrics:
1. Mean Squared Error (MSE): Average of squared differences between actual and predicted values.
   $MSE = \Sigma(y_i - \hat{y}_i)^2 / n$
2. R-squared (R^2): Proportion of variance in y explained by the predictors.
   $R^2 = 1 - (RSS / TSS)$
3. Adjusted R-squared: Adjusts R^2 for the number of predictors to avoid overfitting.

3. Assumptions of Linear Regression

1. Linearity: The relationship between predictors and the target variable is linear.
2. Independence: Errors (ε) are independent of each other.
3. Homoscedasticity: Errors have constant variance.
4. Normality: Errors are normally distributed.
5. No Multicollinearity: Predictors are not highly correlated.

4. Strengths of Linear Regression

- Easy to implement and interpret.
- Computationally efficient for small to medium-sized datasets.
- Useful as a baseline model for predictive tasks.
- Provides insights into relationships between variables.

5. Limitations of Linear Regression

- Sensitive to outliers and multicollinearity.
- Assumes a linear relationship between predictors and the target, which may not always hold.
- Struggles with non-linearity and complex data structures.
- Assumes constant variance of errors (homoscedasticity).

6. Extensions of Linear Regression

1. Polynomial Regression: Captures non-linear relationships by including polynomial terms (x^2, x^3, ...).
2. Regularization: Addresses overfitting by adding penalties:
   - Ridge Regression: Adds L2 penalty to shrink coefficients.
   - Lasso Regression: Adds L1 penalty to reduce some coefficients to zero, aiding feature selection.

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>

Francis Anscombe, a statistician, created Anscombe's quartet in 1973 to illustrate the value of visualizing data before studying it. Despite having virtually identical statistical characteristics, the datasets' visual representations diverge greatly, demonstrating how depending only on summary statistics can be deceptive.

Important Statistical Features of the Anscombe Quartet:
The four datasets all have the same or roughly the same statistical characteristics:

1. Mean of X and Y: For every dataset, the mean of the independent variable (x) and dependent variable (y) is the same.
2. Variance of X: Across datasets, the variance of x is the same.
3. Correlation: There is almost an equal correlation coefficient (r) between x and y.
4. Linear Regression Line: The equation for linear regression, y=mx+b, is constant across datasets.
5. Residual Sum of Squares (RSS): The amount of variation in y that the regression cannot account for is equal.

The Four Datasets
1. Dataset 1 (Typical Linear Relationship):
Appears as a standard linear relationship between x and y.
Data points are closely clustered around the regression line.

2. Dataset 2 (Non-linear Relationship):

The data follows a clear non-linear, parabolic pattern.
Applying linear regression is inappropriate since the relationship is not linear.

3. Dataset 3 (Outlier Influence):
Most points fit the linear regression line well, but a single outlier drastically affects the regression line and correlation coefficient.
This emphasizes how sensitive statistics can be to outliers.

Dataset 4 (Horizontal Cluster with One Point):
Almost all points have the same x value, except for one outlier that creates a misleading regression line.
The correlation coefficient remains high, despite the lack of a meaningful relationship.

Conclusion:
Anscombe's quartet serves as a classic illustration of the drawbacks of interpreting data solely through statistical metrics. It highlights how important it is to use visualizations in addition to numerical summaries in order to obtain precise insights and make more educated judgments.

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>

A statistical metric that measures the magnitude and direction of the linear relationship between two continuous variables is Pearson's R, sometimes referred to as the Pearson Correlation Coefficient. In statistics and data analysis, it's one of the most widely used correlation metrics.

Range:
1. −1 to 1.
2. A value of 1 indicates a perfect positive linear relationship.
3. A value of −1 indicates a perfect negative linear relationship.
4. A value of 0 indicates no linear relationship.

Interpretation:
1. r>0: Positive correlation (as one variable increases, the other tends to increase).
2. r<0: Negative correlation (as one variable increases, the other tends to decrease).
3. r=0: No linear correlation.

Assumptions:
1. Both variables should be continuous and normally distributed (for reliable interpretation).
2. Both variables should be continuous and normally distributed (for reliable interpretation).
3. Homoscedasticity: Variance of one variable should be similar across the range of the other variable.

Applications:
1. Identifying the strength and direction of linear relationships in data.

2. Used in feature selection for machine learning, as a strong correlation may indicate redundancy.

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:
A data preparation method called scaling is used to modify numerical feature values so they fall inside a given range or distribution. It guarantees that no feature's size will cause it to control the learning method.

Application of Scaling:
1. Improves Model Performance: Many machine learning algorithms (e.g., gradient descent-based models like logistic regression or neural networks) are sensitive to the scale of input data, and scaling helps in faster convergence.
2. Prevents Feature Domination: Features with larger ranges may dominate others, leading to biased results.
3. Better Interpretation: Scaling makes feature values easier to interpret when visualizing or analyzing data.
4. Ensures Metric Compatibility: In distance-based models like k-NN or clustering, unscaled features can distort distance calculations.

Types of Scaling:
1. Normalized Scaling: Normalization transforms data values to fall within a specific range, typically [0, 1] or [-1, 1].
2. Standardized Scaling: Standardization scales data to have a mean of 0 and a standard deviation of 1 (unit variance).

Key Differences Between Normalization and Standardization:

Aspect	**Normalization**	**Standardization**
Range	Scales to a fixed range (e.g., [0, 1] or [-1, 1]).	No fixed range: values centered around 0 with unit variance.
Use Case	Uniformly distributed data or bounded features.	Data with Gaussian distribution or outliers.
Effect on Outliers	Sensitive to outliers.	Less sensitive to outliers.

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 Variance Inflation Factor (VIF), which measures how much the variance of a regression coefficient is inflated by the existence of other correlated predictors, is a useful tool for quantifying multicollinearity in datasets. One predictor variable can be perfectly explained as a linear combination of other predictor variables when the VIF value is infinite, a phenomenon known as perfect multicollinearity.

To handle infinite VIF follow below steps:

1. Eliminate Redundant Features: Choose a feature that is exactly connected, then eliminate it.
2. Drop One Dummy Variable: To escape the dummy variable trap, always drop one dummy variable while encoding categorical data.
3. Feature Selection: To cut down on the number of features, employ strategies like Lasso regression, forward selection, and backward elimination.
4. Regularization: Multicollinearity is lessened by penalties introduced by algorithms such as Ridge regression.
5. Variance Decomposition: Use the correlation matrix to examine feature correlations and eliminate variables that lead to singularity.

By addressing perfect multicollinearity, the VIF values become finite, improving model interpretability and stability.

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>

A graphical tool called a Q-Q (Quantile-Quantile) plot is used to compare a dataset's distribution to a theoretical probability distribution, usually the normal distribution. The quantiles of the theoretical distribution are plotted against the quantiles of the observed data.

- If the data follows the theoretical distribution**, the points will align closely along a 45-degree line.**
- **Deviations from the 45-degree line** indicate departures from the theoretical distribution.

Use of a Q-Q Plot in Linear Regression ->
In linear regression, several assumptions need to be met for the model to perform optimally. A Q-Q plot is commonly used to check the normality assumption:

1. Normality of Residuals:
   - Linear regression assumes that the residuals (differences between observed and predicted values) are normally distributed.
   - By plotting a Q-Q plot of the residuals against a normal distribution, we can visually assess whether this assumption is satisfied.
2. Detecting Outliers and Skewness:

- Large deviations in the tails of the Q-Q plot indicate outliers.
- Patterns like curves or bends may suggest skewness or other non-normal features.

Importance of a Q-Q Plot in Linear Regression ->

1. Validating Model Assumptions: Verifies that the residuals of the model meet the normality assumption, which is essential for confidence intervals and hypothesis testing.
2. Enhancements to the Guiding Model: The Q-Q plot's deviations from normalcy may indicate the need for adjustments (such as square root or logarithmic) to enhance model performance.
3. Evaluating Model Fit: This indirectly validates the suitability of the linear model by assisting in determining if the residuals behave as anticipated.
4. Robustness to Deviations: The Q-Q plot aids in determining when deviations from normalcy are serious enough to require attention, even though minor deviations may not have a substantial effect on model performance.
