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)

To gain insights into how categorical variables affect a dependent variable, a comprehensive analysis is essential. Here are some key points that can be deduced from such an analysis:

Relationship with the Dependent Variable:

- By examining how the dependent variable is distributed across different categories, it is possible to identify notable differences.
- For instance, certain categories may consistently correspond to higher or lower values of the dependent variable, indicating a potential influence.

Significance of Categories:

- Statistical tests such as chi-square tests (for categorical dependent variables) or ANOVA/F-tests (for continuous dependent variables) can be used to assess the significance of the relationship.
- Categories with low p-values suggest that they may have a meaningful impact on the dependent variable.

Feature Importance:

- Models like decision trees, random forests, or logistic regression can evaluate the importance of categorical variables by measuring their contribution to the predictions.

Interaction Effects:

- Some categorical variables may exert a stronger influence on the dependent variable when combined with other variables. These interactions can be explored through regression models with interaction terms or hierarchical modeling techniques.

Effect Size:

- Metrics like Cramér's V or contingency tables can quantify the strength of the relationship between categorical variables and the dependent variable.

Trends or Patterns:

- For ordinal categorical variables, it’s helpful to identify trends, such as whether the dependent variable increases or decreases across the categories.
- For nominal variables, focus on identifying specific categories that show distinct behavior.

Example Scenario: If the dependent variable is "customer churn" and the categorical variable being analyzed is "subscription plan," one might find that:

- Customers on the "Basic" plan tend to have a higher churn rate compared to those on the

"Premium" plan.

- A statistical analysis might confirm that the subscription plan is a significant predictor of churn.

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)
Setting `drop_first=True` when creating dummy variables helps avoid multicollinearity by excluding one category from each categorical variable. This prevents redundancy and eliminates the "dummy variable trap," where including all categories causes linear dependency in regression models.

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)

The numerical variable with the highest correlation to the target variable can be identified by analyzing the pair plot or correlation matrix.

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)

The assumptions of Linear Regression were validated by checking:

1. **Linearity**: Residual plots for linear relationships.
2. **Normality**: Q-Q plots or Shapiro-Wilk test for normal distribution of residuals.
3. **Homoscedasticity**: Residual vs. fitted values plot for constant variance.
4. **Multicollinearity**: Variance Inflation Factor (VIF) for independent variables.
5. **Independence**: Durbin-Watson test for autocorrelation.

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 contributing significantly to explaining the demand for shared bikes can be identified by analyzing the model's coefficients and their statistical significance (p-values). Features

with the highest absolute coefficient values and significant p-values are the most impactful.

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)

Linear Regression is a supervised learning algorithm used to predict a continuous target variable by modeling the linear relationship between independent variables (features) and the dependent variable. It aims to fit a line of the form:

$y=\beta 0+\beta 1x1+\beta 2x2+\cdots+\beta$ nxn+ϵ

Key Assumptions:

1. Linearity between predictors and the target.
2. Independence of observations.
3. Homoscedasticity (constant error variance).
4. Normal distribution of residuals.
5. No multicollinearity (low correlation among predictors).

Working:

1. **Model Formulation**: Define a linear equation with initial coefficients.
2. **Optimization**: Minimize the sum of squared errors (SSE) using Ordinary Least Squares (OLS) or Gradient Descent.
3. **Prediction**: Use optimized coefficients to estimate the target variable.

Evaluation:

Model performance is assessed using metrics like R2R^2R2, Adjusted R2R^2R2, Mean Squared Error (MSE), or Root Mean Squared Error (RMSE).

Assumption Validation:

- Linearity: Residual plots.
- Normality: Q-Q plots or histograms of residuals.
- Homoscedasticity: Residuals vs. fitted values.
- Multicollinearity: Variance Inflation Factor (VIF).
- Independence: Durbin-Watson test.

Variants like Ridge, Lasso, and Elastic Net add regularization to reduce overfitting. Linear Regression is widely used in areas like sales forecasting, housing price prediction, and experimental analysis.

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)

Anscombe's Quartet is a group of four datasets created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data before analyzing it. While these datasets have nearly identical statistical properties, their distributions differ significantly when plotted, highlighting how relying solely on summary statistics can be misleading.

1. Key Characteristics of Anscombe's Quartet

Each of the four datasets has:

- The same mean (x¯\bar{x}x¯ and y¯\bar{y}y¯).
- The same variance (σx2\sigma_x^2σx2 and σy2\sigma_y^2σy2).
- The same correlation between xxx and yyy (approximately 0.816).
- The same linear regression line (y=3+0.5xy = 3 + 0.5xy=3+0.5x).

Despite these similarities, the datasets are very different in their structure and behavior, as revealed through visualization.

Anscombe's Quartet consists of four datasets with identical statistical properties (mean, variance, correlation, and regression line) but different distributions. It emphasizes the importance of visualizing data, as:

1. **Dataset 1**: A simple linear relationship with well-spread points around the regression line.
2. **Dataset 2**: A non-linear relationship where the regression line is not appropriate.
3. **Dataset 3**: Mostly constant yyy-values with one influential outlier, highlighting the effect of outliers.
4. **Dataset 4**: Identical xxx-values with one extreme outlier that skews the regression line.

The Quartet teaches that summary statistics can be misleading, and visualization is crucial for accurate analysis.

Anscombe's Quartet is a group of four datasets created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data before analyzing it. While these datasets have nearly identical statistical properties, their distributions differ significantly when plotted, highlighting how relying solely on summary statistics can be misleading.

3. Lessons from Anscombe's Quartet

1. **Importance of Visualization**:
    - Summary statistics like mean, variance, and correlation can hide the true nature of the data.
    - Visualizing data through scatter plots or other graphs reveals patterns, relationships, and anomalies.
2. **Limitations of Statistical Measures**:
    - Correlation and regression alone may not capture non-linear relationships, outliers, or other complex structures.

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)

Pearson's RRR is a measure of the linear correlation between two variables, ranging from -1 to 1. A value of 1 indicates a perfect positive linear relationship, -1 indicates a perfect negative linear relationship, and 0 indicates no 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)

Scaling is the process of adjusting data to fit within a particular range or distribution, ensuring that all features are on a comparable scale, which is important for many machine learning algorithms.

Why Scaling is Important:

- **Improved Model Performance**: Algorithms like gradient descent, k-nearest neighbors, and SVMs can perform poorly if features have different scales.
- **Faster Convergence**: Proper scaling can help algorithms converge more quickly.
- **Equal Contribution**: It ensures that each feature contributes equally to the model, avoiding domination by features with larger ranges.

Difference Between Normalization and Standardization:

- **Normalization**: Scales the data to a fixed range, typically [0, 1], using the formula:
  Normalized=x−min(x)max(x)−min(x)\text{Normalized} = \frac{x - \min(x)}{\max(x) - \min(x)}Normalized=max(x)−min(x)x−min(x)
  It is useful when features have different units or need to be within a specific range.
- **Standardization**: Transforms data to have a mean of 0 and a standard deviation of 1, using the formula:
  Standardized=x−μσ\text{Standardized} = \frac{x - \mu}{\sigma}Standardized=σx−μ
  This is appropriate for algorithms that assume a Gaussian distribution or require similar variance across features.

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)

A **VIF (Variance Inflation Factor)** becomes infinite when perfect multicollinearity exists in the dataset. This happens when one predictor variable is an exact linear function of another, meaning the variables are perfectly correlated (correlation coefficient is 1 or -1).

Why This Occurs:

- **Exact Linear Relationship**: If one predictor can be precisely predicted from another, the variance of the coefficient estimate for that variable becomes infinite.
- **Matrix Inversion Issue**: Since VIF is calculated using the inverse of the correlation matrix, perfect collinearity (e.g., x1=2x2x_1 = 2x_2x1=2x2) makes the matrix non-invertible, causing the VIF to be infinite.

To address this, it's important to remove or combine the highly correlated predictors to eliminate multicollinearity.

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)

A Q-Q (Quantile-Quantile) plot is a graphical tool used to compare the distribution of a dataset to a theoretical distribution (typically the normal distribution). It plots the quantiles of the observed data against the quantiles of the reference distribution. If the data follows the reference distribution, the points will fall approximately along a straight line.

In linear regression, one of the key assumptions is that the residuals (errors) are normally distributed. A Q-Q plot can be used to visually check this assumption by plotting the residuals against a normal distribution.
Importance of a Q-Q Plot in Linear Regression:
Assess Normality of Residuals: A Q-Q plot helps to determine if the residuals follow a normal distribution, which is crucial for making reliable inferences (e.g., hypothesis testing) and calculating confidence intervals.
Detect Non-Normality: If the points deviate significantly from the straight line, it indicates that the residuals may not be normally distributed, which could affect the validity of the regression model.
Model Diagnostics: It is an important diagnostic tool to assess the goodness of fit and ensure that the assumptions of linear regression are met.
