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 analysis of categorical variables in a dataset, especially in the context of their effect on the dependent variable (in this case, bike demand or cnt), provides valuable insights. Here's how categorical variables typically influence the dependent variable and what you can infer from them in the bike-sharing context: Season (e.g., Spring, Summer, Fall, Winter):

1)Inference: Bike-sharing demand often varies by season. For instance:
Summer and Fall might have higher demand due to favorable weather conditions.
Winter might show reduced demand due to cold temperatures and snow.

2) Inference: Clear or mild weather promotes bike usage.
Adverse weather (rain, snow) typically results in lower bike demand.
Visual Evidence: Grouped bar plots of demand for different weather categories would illustrate this trend.bDay Type (e.g., Working Day vs. Weekend):

3)Inference: Demand might be higher on weekends due to leisure activities.
On working days, demand could peak during commuting hours.

4) Inference: On holidays, demand might increase due to leisure activities or decrease if people prefer other transport modes.

5) Inference: Demand might peak during morning and evening rush hours on weekdays due to commuting patterns. Demand might spread throughout the day on weekends.

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)

Using drop_first=True during dummy variable creation is important to prevent the issue of multicollinearity in the dataset, which can arise due to the inclusion of redundant information.
When creating dummy variables, a categorical variable with n categories is represented by n dummy variables, where each dummy variable corresponds to one category. If all n dummy variables are included in the regression model, it introduces a situation called the dummy variable trap.
When drop_first=True is specified:
It drops one dummy variable (typically the first category), treating it as the baseline or reference category.
The coefficients of the remaining dummy variables are interpreted relative to this baseline

---

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)

To determine which numerical variable has the highest correlation with the target variable (cnt), you would typically analyze the pair-plot and compute the correlation matrix.
The variable with the highest positive correlation to cnt will have the largest positive value.
Typically, in a bike-sharing dataset, temperature (temp) often exhibits the highest positive correlation with cnt, as warmer temperatures encourage outdoor activities and bike usage.
Other variables like humidity (hum) or windspeed might show weaker correlations, with hum potentially having a negative relationship due to discomfort caused by high humidity.

```
cnt         1.000000
temp        0.631987
atemp       0.630567
hum        -0.170997
windspeed  -0.234545
Name: cnt, dtype: float64
```

This indicates that temp has the strongest positive correlation with bike demand (cnt).

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)
Validating the assumptions of linear regression is a critical step to ensure the model is reliable and its results are interpretable. After building the model on the training set, the following assumptions are checked systematically:

1.) Linearity of Relationships
Assumption: The relationship between the independent variables and the dependent variable (cnt) should be linear.
Validation:
Residual Plot: Plot residuals (difference between actual and predicted values) against the predicted values.
Residuals should be randomly scattered around zero with no clear pattern.
Interpretation: If there is no discernible pattern in the scatterplot, the linearity assumption is likely satisfied.
2)Normality of Residuals
Assumption: Residuals should be approximately normally distributed.
Validation:
Q-Q Plot: Plot the quantiles of residuals against a theoretical normal distribution.
Histogram: Check the distribution of residuals visually.
Interpretation: If the Q-Q plot shows points close to the diagonal line and the histogram resembles a normal distribution, this assumption is satisfied.
3) Homoscedasticity (Constant Variance of Residuals)
Assumption: The variance of residuals should remain constant across all levels of predicted values.
Validation:
Residuals vs. Predicted Plot: Check if residuals are evenly spread out across all predicted values.

Interpretation: If the scatterplot shows a funnel shape (increasing or decreasing spread), the homoscedasticity assumption is violated.
4) Multicollinearity
Assumption: Independent variables should not be highly correlated with each other.
Validation:
Variance Inflation Factor (VIF): Compute VIF for each predictor.
Interpretation: A VIF > 5 or 10 indicates high multicollinearity. Drop or combine highly correlated variables if needed.

5) Independence of Errors
Assumption: Residuals should be independent of each other.
Validation:
Durbin-Watson Test: A statistic close to 2 indicates no significant autocorrelation in residuals
Interpretation: Values between 1.5 and 2.5 typically indicate acceptable independence.

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)

To identify the top 3 features contributing significantly to the bike demand (cnt) in the final model, we analyze the coefficients of the linear regression model and their statistical significance.

1) Temperature (temp): Bike demand increases with temperature (up to a comfortable range). This is typically the most strongly correlated feature with cnt.
2) Season (season dummy variables): Specific seasons (e.g., summer, fall) often see higher demand. Dummy variables representing seasons explain the seasonal variation in demand.
3)Weather Conditions: Adverse weather conditions (e.g., rain, snow) negatively impact demand. This feature helps account for daily variations due to meteorological factors.

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)

Ans: Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. It provides valuable insights for prediction and data analysis. Linear regression is also a type of supervised machine-learning algorithm that learns from the labelled datasets and maps the data points with most optimized linear functions which can be used for prediction on new datasets. It computes the linear relationship between the dependent variable and one or more independent features by fitting a linear equation with observed data. It predicts the continuous output variables based on the independent input variable.

For example if we want to predict house price we consider various factor such as house age,

distance from the main road, location, area and number of room, linear regression uses all these parameter to predict house price as it consider a linear relation between all these features and price of house.

Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x)). Hence, the name is Linear Regression. In the figure above, X (input) is the work experience and Y (output) is the salary of a person. The regression line is the best-fit line for our model.

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, introduced by the statistician Francis Anscombe in 1973, is a collection of four datasets that demonstrate the importance of visualizing data before making conclusions. Despite having nearly identical statistical properties, the datasets reveal drastically different patterns when visualized.

Purpose of Anscombe's Quartet

- To emphasize that relying solely on summary statistics can be misleading.
- To highlight the importance of exploratory data analysis (EDA), particularly visualization.
- To illustrate that different datasets can yield the same statistical measures but differ significantly in their underlying structure.

Insights from Anscombe's Quartet :
Statistics Can Be Misleading: Summary statistics like mean, variance, and correlation fail to capture the nuances of datasets.
Visual inspection is crucial to understanding the data's true behavior.

Role of Outliers: Outliers can have a disproportionate influence on statistical measures.
Visualization helps identify and address such anomalies.

Limitations of Regression:
A linear regression line may not always describe the data accurately.
Checking assumptions (linearity, homoscedasticity) is vital.

Anscombe's Quartet is a powerful reminder that data analysis is more than just calculating statistics. Visual exploration helps uncover relationships, anomalies, and trends that summary measures might obscure. Always visualize your data to ensure meaningful and accurate insights!

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 R: A Measure of Correlation
Pearson's R, also known as the Pearson correlation coefficient, is a statistical measure that quantifies the linear relationship between two continuous variables. It was developed by Karl

Pearson, a prominent statistician, and is one of the most widely used methods to assess correlation.
Pearson's R measures the strength and direction of a linear relationship between two variables Pearson Correlation Coefficient: Correlation coefficients are used to measure how strong a relationship is between two variables. There are different types of formulas to get a correlation coefficient, one of the most popular is Pearson's correlation (also known as Pearson's r) which is commonly used for linear regression.

The Pearson correlation coefficient, often symbolized as (r), is a widely used metric for assessing linear relationships between two variables. It yields a value ranging from –1 to 1, indicating both the magnitude and direction of the correlation. A change in one variable is mirrored by a corresponding change in the other variable in the same direction

It is defined mathematically as:
The Pearson Correlation Coefficient, denoted as r, is a statistical measure that calculates the strength and direction of the linear relationship between two variables on a scatterplot. The value of r ranges between -1 and 1, where:

1 indicates a perfect positive linear relationship,
-1 indicates a perfect negative linear relationship, and
0 indicates no linear relationship between the variables.

Pearson's Correlation Coefficient Formula
Karl Pearson's correlation coefficient formula is the most commonly used and the most popular formula to get the statistical correlation coefficient. It is denoted with the lowercase "r". The formula for Pearson's correlation coefficient is shown below:

$$\mathbf{r = n(\sum xy) - (\sum x)(\sum y) / \sqrt{[n\sum x^2-(\sum x)^2][n\sum y^2-(\sum y)^2}}$$

The full name for Pearson's correlation coefficient formula is Pearson's Product Moment correlation (PPMC). It helps in displaying the Linear relationship between the two sets of the data.

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 a data preprocessing technique used to adjust the range and distribution of numerical features in a dataset so that they have a consistent scale. This ensures that no single feature disproportionately influences a machine learning model, especially when the model is sensitive to feature magnitudes.

Scaling is Performed because of following reasons:

i. Algorithm Sensitivity: Many machine learning algorithms, such as gradient descent-based models (e.g., logistic regression, neural networks) or distance-based models (e.g., k-NN, SVM, clustering), are sensitive to feature magnitudes.
ii. Features with larger ranges dominate those with smaller ranges, leading to biased results.
iii. Faster Convergence:

iv. Scaling improves optimization efficiency by ensuring that the gradient steps are uniform across all dimensions.
v. Improved Model Performance:
vi. Ensures that features contribute equally to the prediction, enhancing model accuracy.
vii. Facilitates Comparison: Brings all features onto a similar scale, making it easier to interpret and compare them.

Types of Scaling :

1. Normalized Scaling

- Definition: Normalization rescales the data to fit within a specific range, usually [0, 1] or [-1, 1].
- Purpose:
- Ensures all feature values lie within the same range.
- Commonly used in image processing or when features have varying units.
- Use Cases:
- Neural networks and algorithms that require features to be bounded (e.g., sigmoid activation functions).
- Key Characteristics:
- Sensitive to outliers, as they significantly affect the minimum and maximum values.

2. Standardized Scaling

- Definition: Standardization (or Z-score normalization) transforms the data to have a mean of 0 and a standard deviation of 1.
- Centers the data and removes unit dependency.
- Suitable for algorithms that assume normality or are sensitive to variance (e.g., PCA, linear regression).
- Key Characteristics: Robust to outliers compared to normalization but still affected if extreme values dominate the dataset.
- Differences Between Normalization and Standardization:

Aspects	Normalization	Standardization
1. Definition	Rescales data to a specific range (e.g., [0, 1]).	Centers data to mean 0 and standard deviation 1.
2. Formula	(X−min)/(max−min)(X - \text{min}) / (\text{max} - \text{min})(X−min)/(max−min)	(X−μ)/σ(X - \mu) / \sigma(X−μ)/σ
3. Output Range	Bounded (e.g., [0, 1] or [-1, 1]).	Mean = 0, SD = 1 (no fixed range).
4. Effect on Distribution	Retains original distribution shape.	Transforms to a standardized distribution.
5. Sensitivity to Outliers	Highly sensitive.	Less sensitive, but still affected.
6. Use Cases	Neural networks, image processing.	PCA, regression, distance-based models.

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)

The Variance Inflation Factor (VIF) measures the degree of multicollinearity among the independent variables in a regression model. A VIF value of infinity occurs when there is perfect multicollinearity, meaning one or more independent variables can be expressed as an exact linear combination of the others.

VIF Formula: The VIF for a variable XiX_iXi is calculated as:

VIF(Xi)=11−Ri2\text{VIF}(X_i) = \frac{1}{1 - R^2_i}VIF(Xi)=1−Ri21

Where:

- Ri2R^2_iRi2: The coefficient of determination when XiX_iXi is regressed on all other independent variables.

Why VIF Becomes Infinite:

- When Ri2=1R^2_i = 1Ri2=1, the denominator (1−Ri2)(1 - R^2_i)(1−Ri2) becomes zero, making the VIF infinite.
- Ri2=1R^2_i = 1Ri2=1 implies that XiX_iXi is perfectly predictable using the other independent variables, indicating perfect multicollinearity.

An infinite VIF indicates perfect multicollinearity. This issue needs to be addressed by removing redundant variables, avoiding dummy variable traps, or using advanced regularization techniques to ensure the regression model is stable and interpretable.

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)

Q-Q plot (Quantile-Quantile plot) is a graphical tool used to compare the distribution of a dataset to a theoretical distribution (usually the normal distribution). It plots the quantiles of the dataset against the quantiles of the reference distribution.

If the points on the Q-Q plot lie approximately on a straight line, it indicates that the data follows the reference distribution. Deviations from the straight line indicate departures from the reference distribution.

Components of a Q-Q Plot

1. X-axis:
    - Theoretical quantiles from the reference distribution (e.g., normal distribution).
2. Y-axis:

- Quantiles from the observed data

➢ **Use and Importance of a Q-Q Plot in Linear Regression:**

In linear regression, the Q-Q plot is primarily used to validate the assumption of normality of residuals, which is critical for reliable statistical inference.

1. Validate the Assumption of Normality:
   - Linear regression assumes that the residuals (differences between predicted and actual values) are normally distributed. A Q-Q plot helps check if this assumption holds.
   - Residuals following a normal distribution will appear as points forming a straight diagonal line in the Q-Q plot.
2. Detect Deviations from Normality:
   - If the points deviate significantly from the straight line:
     - S-shaped curve: Indicates heavy tails (e.g., leptokurtic distribution).
     - Inverted S-shaped curve: Indicates light tails (e.g., platykurtic distribution).
     - Outliers: Points far away from the line indicate outliers in the data.
3. Impact of Non-Normal Residuals:
   - Non-normality can lead to biased confidence intervals and hypothesis tests in regression, making it harder to trust model interpretations.
4. Helps in Model Diagnostics:
   - If residuals deviate from normality, remedial measures can be taken, such as:
     - Applying data transformations (e.g., log, square root).
     - Using robust regression techniques.

How to Interpret a Q-Q Plot;

- Straight Line:
  - Residuals are normally distributed.
- Curved Line:
  - Indicates skewness in the residuals.
  - Upward curve: Positive skewness.
  - Downward curve: Negative skewness.
- Extreme Points:
  - Suggest the presence of outliers.

When to Use a Q-Q Plot in Linear Regression:

1. After fitting a regression model:
   - To validate the assumption of normality for residuals.
2. When residuals deviate significantly:
   - To diagnose issues with the model or data (e.g., outliers, non-linearity).

Conclusion

A Q-Q plot is an essential diagnostic tool in linear regression to assess the normality of residuals. By visually inspecting deviations from a straight line, analysts can identify potential issues with the data or model, allowing them to take corrective actions to ensure robust model performance.