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? (3 marks)

Analysing categorical variables in a dataset can provide insights into how these variables might influence the dependent variable. Typically, this is done by examining the relationship between the categories and the dependent variable. Here's how you can approach the analysis and what you might infer:

Steps to Analyse Categorical Variables:

1. **Encoding Categorical Variables**:
   - Convert categorical variables into numerical formats (e.g., one-hot encoding, label encoding).
   - Use drop first=True in one-hot encoding to avoid the dummy variable trap (perfect multicollinearity).
2. **Summary Statistics**:
   - Calculate min & max values of the dependent variable across different categories.
   - Analyse how the average values differ between categories.
3. **Visualization**:
   - Create boxplots to visualize the distribution of the dependent variable across different categories.
   - Look for patterns, such as certain categories consistently having higher or lower values of the dependent variable.
4. **Feature Importance**:
   - After fitting the model, analyse the coefficients of the categorical variables (in linear regression) or feature importance (in tree-based models) to understand their impact.

Categorical Values Box Plot Analysis

1. Season:

- Fall having higher rental range where spring having less rental range.

2. Year:

- 2019 Rental of bike considerably increased compare to 2018

3. Weather:

- Rental having higher value during Clear weather than cloudy mist and light rain thunder.

4. WeekDay:

- Saturday and Thursday more rental than other working days.

5. HoliDay:

- - Holiday rented less than working days.

6. Month:

- September leads in monthly rental counts, with nearby months also showing strong demand. This trend follows seasonal patterns, suggesting a link between rentals and seasons.

2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)

Using `drop_first=True` during dummy variable creation is important to avoid the **dummy variable trap**, which can lead to **multicollinearity** in the model.

- **Avoiding Multicollinearity:**

  When creating dummy variables, if all categories are included, the sum of the dummy variables can perfectly predict one another, leading to multicollinearity. This can cause instability in the model coefficients and affect the interpretability of the results.

- **Ensuring a Reference Category:**

  By dropping the first dummy variable, one category is used as a reference, and the remaining categories are compared against this reference. This ensures that the model remains well-posed, with one less predictor, improving model interpretability and reducing redundancy.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

Highest correlation with cnt is seen in temp.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

❑ Linearity:

- **Assumption**:
  - The relationship between the independent variables and the dependent variable is linear.
- **Validation**:
  - **Plot Residuals vs. Fitted Values**: This plot should show no clear pattern. If the residuals are randomly scattered around zero, it suggests that the linearity assumption holds.
  - **Scatter Plots**: Plot each independent variable against the dependent variable to visually check for linear relationships.

❑ Normality :

- **Assumption:**

    - The residuals are normally distributed.
- **Validation:**

Histogram of Residuals: The histogram of the residuals should look like a bell curve (normal distribution).

❑ Multicollinearity :

- **Assumption:**
    - The independent variables are not too highly correlated with each other. Variable must be independent of each other
- **Validation:**
    - Variance Inflation Factor (VIF): Calculate the VIF for each predictor. A VIF value greater than 10 indicates high multicollinearity and suggests that the predictor may need to be removed or combined with others

❑ Homoscedasticity :

- **Assumption:**
    - **The variance of the residuals is constant across all levels of the independent** variables.
- **Validation:**
    - Plot Residuals vs. Fitted Values: In addition to checking linearity, this plot is also used to check homoscedasticity. The spread of residuals should be constant across all levels of fitted values.

❑ Autocorrelation :

    - Autocorrelation is the correlation between two of the same series. Residual Analysis of the train data

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

Our final multiple linear regression model seeks to predict bike bookings using a collection of predictor factors.

Following a thorough investigation, we determined the influence of each variable on bike bookings.

The top three variables(with co-efficient) that significantly influence bike rent are as follows:

- Yr 2036.64
- Temp 106.00
- Windspeed -35.00

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

Linear regression is a supervised learning algorithm used for predicting a continuous outcome based on one or more input features (independent variables). The goal is to establish a linear relationship between the dependent variable (target) and the independent variables by fitting a straight line (in simple linear regression) or a hyperplane (in multiple linear regression) that best represents the data.

The equation of a linear regression model can be represented as:

$$y=\beta0+\beta1x1+\beta2x2+\cdots+\beta nxn+\epsilon$$

Where:

- y is the predicted output (dependent variable).
- x1,x2,…,xn are the input features (independent variables).
- β0 is the intercept (the value of y when all xi=0).
- β1,β2,…,βn are the coefficients (slopes) that represent the change in y for a unit change in xi.
- ϵ is the error term, representing the difference between the observed and predicted values.

Training the Model

The linear regression model is trained by finding the values of β1,β2,…,βn that minimize the cost function, typically the Mean Squared Error (MSE):

$$MSE=m1i=1\sum m(yi-y\hat{}i)2$$

Where:

- m is the number of data points.
- yi is the actual value.
- y^i is the predicted value.

The minimization is usually done using optimization techniques like Gradient Descent, which iteratively updates the coefficients to reduce the error.

Prediction and Interpretation

Once trained, the model can predict the outcome for new data by substituting the values of the independent variables into the regression equation. The coefficients βi indicate the strength and direction of the relationship between each feature and the target. For example, a positive βi suggests that as xi increases, y also increases, while a negative βi suggests the opposite.

In summary, linear regression is a foundational algorithm in machine learning, providing a simple yet powerful method for predicting outcomes and understanding relationships between variables.

2. Explain the Anscombe’s quartet in detail. (3 marks)

Anscombe's quartet is a group of datasets (x, y) that have the same mean, standard deviation, and regression line, but which are qualitatively different. It is often used to illustrate the importance of looking at a set of data graphically and not only relying on basic statistic properties.

It highlights the importance of data visualization to gain a deeper understanding of the underlying patterns and relationships within the data.

3. What is Pearson's R? (3 marks)

Pearson's R, also known as the Pearson correlation coefficient, is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. It ranges from -1 to +1.

- **Value Interpretation:**
  - +1: Indicates a perfect positive linear relationship, where as one variable increases, the other also increases in a perfectly linear manner.
  - -1: Indicates a perfect negative linear relationship, where as one variable increases, the other decreases in a perfectly linear manner.
  - 0: Indicates no linear relationship between the variables.
- **Importance**:
  - In regression analysis, it helps in assessing the strength of the linear relationship between the independent and dependent variables.

4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

What is scaling

Scaling is a step of data Pre-Processing. This applied over independent variables to normalize the data within a particular range. It also helps in speeding up the calculations in an algorithm.

Why is scaling performed

In most cases, the collected data contains highly diverse characteristics in terms of scale, units and scope. If scaling is not done, the algorithm only considers the magnitude and not the units, resulting in inaccurate model. To solve this problem, we need to perform some scaling to bring all variables to the same level of magnitude.

What is the difference between normalized scaling and standardized scaling

Standardization centres data around a mean of zero and a standard deviation of one, while normalization scales data to a set range, often [0, 1], by using the minimum and maximum values. One disadvantage of normalization over standardization is that it loses some information in the data, especially about outliers.

5. You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)

This shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R2 =1, which lead to 1/(1-R2) infinity. To solve this problem we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

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

(3 marks)

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

Importance in Linear Regression: In linear regression, one of the key assumptions is that the residuals (the differences between observed and predicted values) are normally distributed. A Q-Q plot helps assess this assumption by comparing the quantiles of the residuals to the quantiles of a normal distribution. If the residuals are normally distributed, the points in the Q-Q plot will follow a straight line.