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)**

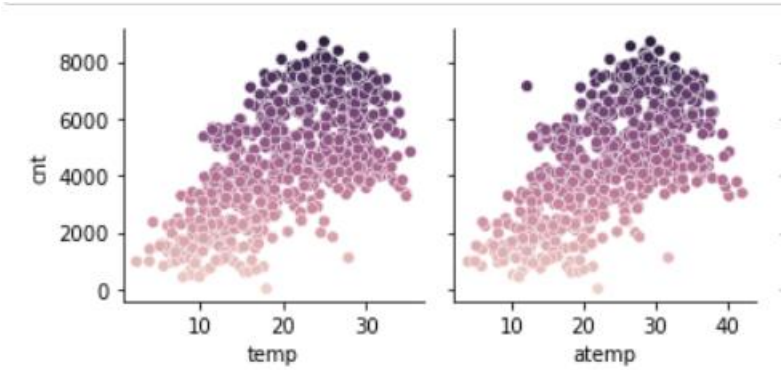
    - The season effects the LR model , we could see this from our final results , also EDA tells us Fall and summer has more bike rentals.
    - The month value also affects the LR model performance we could see that June to sept the demand for bike shared is more.
    - Weather also impacts the performance of LR model , in cloudy weather the demand for bike is more.

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

We use drop_first=True because when we create dummy variables for a particular attribute, all the categories inside it are converted to 0 and 1. For example, let's say we have a week variable ranging from 0 to 6, where 0 stands for Monday and 6 stands for Sunday. If we convert this to one-hot encoding, a response of Sunday will be represented by all 0s in all the dummy variables except the last one. This introduces redundancy, as the machine learning model can infer the missing category. Dropping the first dummy variable avoids multicollinearity and simplifies the model.

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

Temperature has the highest corelation with target variable.



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

We have followed below steps-

- Checked for all the P values to ensure that variable are significant.
- .Analysed the error term to check if the error term is normally distributed or not.
- Checked VIF to ensure that no multicollinearity is present in the data
- Also checked the Durbin Watson static score which comes around 2.047 which is close to 2 indicating there is no significant autocorrelation in the residuals of your model.

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

Atemp (Feeling Temperature): There is a positive correlation between feeling temperature (in the range of 10-30 degrees Celsius) and bike demand. As the temperature increases within this range, we observe a notable increase in bike rentals.

• **windspeed**: Lower windspeeds are associated with higher bike demand. This suggests that calm weather conditions tend to positively influence the decision to rent bikes.

• **Holiday vs. Working Day**: Bike demand is higher on holidays compared to an average working day. This difference in demand pattern highlights the influence of weekdays versus holidays on bike rental behaviour

General Subjective Questions

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

Linear regression is a fundamental statistical method used for supervised learning tasks, particularly for predicting a continuous numerical target variable (y) based on one or more independent features (x). It assumes a linear relationship between the features and the target variable.

Model: The linear regression model represents the relationship between features and the target variable as a linear equation:

y = b0 + b1*x1 + b2*x2 + ... + bn*xn (where n is the number of features)

- y: Predicted target value
- b0: Intercept (constant term)
- b1, b2, ..., bn: Coefficients (slopes) for each feature

Assumptions:

- Linear Relationship: There should be a linear relationship between the features and the target variable.
- Homoscedasticity: The variance of the errors should be constant across the entire range of features.
- Normality of Errors: The error terms (residuals) should be normally distributed with a mean of zero.
- No Multicollinearity: The features should be independent and not highly correlated with each other.

Application:

Sales Forecasting: Predicting future sales based on historical sales data and market trends.
Customer Lifetime Value: Estimating the future value of customers based on their past behaviour.

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

Anscombe's quartet is a set of four seemingly unrelated datasets, each containing 11 data points (x, y pairs). The key point is that these datasets have almost identical summary statistics:

- Mean (x)
- Mean (y)
- Standard deviation (x)
- Standard deviation (y)
- Linear regression line (slope and intercept)

However, when you plot these datasets, you discover drastically different distributions. This demonstrates the importance of data visualization and the limitations of relying solely on summary statistics.

Here's a breakdown of Anscombe's quartet: Datasets:

1. Linear: A clear linear relationship between x and y.
2. Spiral: A tight spiral pattern with no clear linear trend.
3. Cluster: Three outliers heavily influence the regression line.
4. Outlier: One outlier significantly affects the regression line.

3. What is Pearson’s R? (3 marks)

Pearson's R, also known as the Pearson correlation coefficient, is a statistical measure that reflects the linear relationship between two continuous variables. It represents the strength and direction of that association.

Key Points:

- Range: Pearson's R values range from -1 to +1.
    - -1: Perfect negative correlation (as one variable increases, the other decreases proportionally).
    - 0: No linear correlation (no predictable relationship between the variables).
    - +1: Perfect positive correlation (as one variable increases, the other increases proportionally).

Use Cases:

Pearson's R is widely used in various fields to assess the linear correlation between variables. Here are some examples:

- Machine Learning: Feature selection in linear regression models, identifying potential relationships between features and target variables.Finance.
- Analyzing the correlation between stock prices or market trends.

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

Scaling refers to the process of transforming features in a dataset to a common range.Features with larger scales can dominate the model's learning process if left unscaled. Scaling ensures all features contribute relatively equally to the model's decisions.

Normalized scaling :

Features with larger scales can dominate the model's learning process if left unscaled. Scaling ensures all features contribute relatively equally to the model's decisions.

- Standardization scales features to have a mean of 0 and a standard deviation of 1.
- Formula (Z-score): standardized_value = (original_value - mean) / standard_deviation

Use cases:

- More common in distance-based algorithms (k-NN, SVM) as it focuses on relative distances between data points.
- Useful when the distribution of features is not guaranteed to be uniform.

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

Yes this happens when features are perfectly correlated (R^2 = 1), the denominator of the VIF formula becomes zero, resulting in infinity. This indicates that the variance of the coefficient is infinitely inflated, making the coefficient unreliable and potentially causing issues in the model.

The possible cause of perfect corelation is duplicate feature or highly colinear feature. Ideally VIF<=5.

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 plot, also known as a quantile-quantile plot, is a graphical tool used to compare the quantiles (distribution) of two datasets. It helps visualize how well one data set (often your observed data) matches a theoretical distribution (like a normal distribution) or another dataset (e.g., training vs. testing data).

Linear regression assumes that the error terms (residuals) in the model follow a normal distribution with a mean of zero and constant variance. A Q-Q plot is valuable for checking these assumptions:

- Normality of Errors: If the residuals from your linear regression model form a straight line in a Q-Q plot against a normal distribution, it indicates that the error terms are likely normally distributed. This satisfies one of the key assumptions of linear regression.
- Outliers: Deviations from the line in the Q-Q plot can highlight potential outliers in your data. These outliers might need investigation or removal if they significantly affect the model.