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?

Total Marks: 3 marks

Answer: Via the boxplots, we can see that only some of the categorical variables are strong predictors for the target variable 'count'

- 'Season', 'yr', 'mnth' and 'weathersit': The different values of them have different impacts on the target variable 'cnt'
- 'holiday', 'weekday', 'workingday': their different values do not affect much on the target variable 'cnt'

Question 2. Why is it important to use **drop_first=True** during dummy variable creation?

Total Marks: 2 marks

Answer: If we set **drop_first=True**, the function **pd.get_dummies()** removes one (often the first) dummy variable to avoid redundancy which can cause **multicollinearity**.

So, in regression models to avoid multicollinearity issues, we should pass the parameter **drop_first=True** during dummy variable creation.

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

Total Marks: 1 mark

Answer: the variable 'temp' 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?

Total Marks: 3 marks

Answer:

- Based on **R2-squared** and **adjusted R-squared** to see the percentage of the variation in the target variable 'cnt' can be explained by the independent variables in the model.
- Step to decide with features to be selected:
    + High p-value (insignificant) - High VIF -> drop
    + High - Low:
        - High p-value - low VIF: remove these first
        - Low p-value - high VIF: remove after the above
    + Low p-value and low VIF -> keep
- For **linearity**: we can plot a scatter plot of (y_test vs y_pred)
- For **Normality of residuals:** we can plot a distplot of (y_test - y_pred) to see bell-shaped curve

---

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

Total Marks: 2 marks

Answer: To 3 features that contribute significantly towards explaining the demand of the shared bikes are:

1. ‘temp’ (temperature): 0.57
2. ‘weathersit’: -0.25*light_snow
3. ‘yr’: 0.23

General Subjective Questions

Question 6. Explain the linear regression algorithm in detail.
Total Marks: 4 marks
Answer: Please write your answer below this line.

What is Linear Regression?

A statistical method used to model the relationship between the dependent (target variable) ‘y’ and independent variables (predictors) ‘X’ by fitting a linear equation to the data.

Equation: $Y=\beta 0+\beta 1X1+\beta 2X2+...+\beta nXn+\varepsilon$

Why is it important?

- Easy to understand and explain relationships between variables
- Used for making decisions and forecasting
- To show how much a predictor contributes to the target variable.

How?

Linear Regression fits a line using the least squares method to minimize the difference between actual values (y) and predicted values ($\hat{y}$)

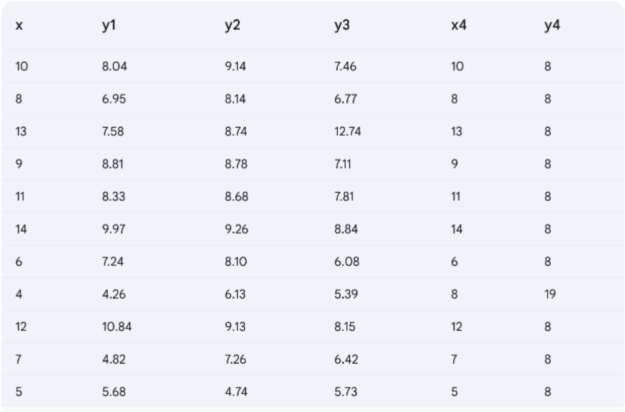
Types?

Simple (one predictor) and Multiple (2 or more predictors) Regressions.

Question 7. Explain the Anscombe’s quartet in detail.
Total Marks: 3 marks
Answer: Please write your answer below this line.

Anscombe’s Quartet is a set of four different datasets that have nearly identical statistical properties (mean, variance, correlation, and regression line) but look completely different when plotted. Therefore, statistical summaries alone are not enough in machine learning.

Examples:



x	y1	y2	y3	x4	y4
10	8.04	9.14	7.46	10	8
8	6.95	8.14	6.77	8	8
13	7.58	8.74	12.74	13	8
9	8.81	8.78	7.11	9	8
11	8.33	8.68	7.81	11	8
14	9.97	9.26	8.84	14	8
6	7.24	8.10	6.08	6	8
4	4.26	6.13	5.39	8	19
12	10.84	9.13	8.15	12	8
7	4.82	7.26	6.42	7	8
5	5.68	4.74	5.73	5	8

Statistical properties:

Property	Dataset 1	Dataset 2	Dataset 3	Dataset 4
Mean of x	9	9	9	9
Mean of y	7.5	7.5	7.5	7.5
Variance of x	11	11	11	11
Variance of y	4.12	4.12	4.12	4.12
Correlation Coefficient	0.816	0.816	0.816	0.816
Linear Regression Line	y = 3 + 0.5x	y = 3 + 0.5x	y = 3 + 0.5x	y = 3 + 0.5x

To summarize:

- Outliers can create bias in regression models.
- Linear models are sometimes wrong, so a non-linear model is needed.
- Statistical properties do not mean identical datasets.

Question 8. What is Pearson's R?
Total Marks: 3 marks
Answer: Please write your answer below this line.

Pearson's r, also called the Pearson Correlation Coefficient, quantifies the strength and direction of the linear relationship between two continuous variables.

- It ranges between -1 and 1
- Values: closer to the 2 ends (-1 or +1): tell a strong correlation and near 0 imply a weak or no correlation
- Positive (+): the variables move in the same direction,
- Negative (-): The two variables move in opposite directions

Question 9. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling?
Total Marks: 3 marks
Answer: Please write your answer below this line.

What?

Scaling is the process of transforming numerical data into a specific range or distribution to ensure that all features contribute equally to a model.

Why?

Scaling is needed because real-world data values may vary in degrees of magnitude, range, and units. Therefore, in order for machine learning models to interpret these features on the same scale (maybe for faster speed), we need to perform feature scaling.

Normalization vs. standardization?

In normalization (Min-Max Scaling), the data range is [0,1] or [-1,1] and data does not follow a normal distribution.

In standardization (Z-score Scaling), Mean euqals to 0 and std Dev. equals to 1. Data follows normal distribution (bell-shaped).

Question 10. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

Total Marks: 3 marks
Answer: Please write your answer below this line.

Variance Inflation Factor (VIF) measures multicollinearity between independent variables in a regression model. The higher the VIF, the higher the possibility of multicollinearity.

When VIF becomes **infinite**, it indicates **perfect multicollinearity** in the dataset. Think of dropping one of the correlated features.

Question 11. What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

Total Marks: 3 marks
Answer: Please write your answer below this line.

A Q-Q plot (Quantile-Quantile plot) a visual way to compare the quantiles of your data to the quantiles of the theoretical distribution.

In linear regression, one key assumption is that the residuals (errors) should be normally distributed. A Q-Q plot is used to verify this assumption. If residuals are normally distributed, the model is valid. Otherwise, reject the model.
