

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)

We can infer the below points:

- Season, weekday, month, holiday, weathersit seems to explain some variance in cnt
 - one of season and month is redundant
 - workingday, weekday does not seem to have an impact on cnt
-

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)

For n categories n columns are typically created. However one column can be represented by all the other being 0. This would result in multi collinearity. Hence to solve this problem, one column needs to be dropped. Drop_first=True drops the first and avoids this problem by dropping the redundant column.

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)

temp and atemp both have very high degree of correlation with the target variable cnt. It is clear the temp and atemp are collinear and one of them will eventually be dropped from the model.

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)

1. Residual Plot: Residuals(observed vs predicted) should be distributed around 0. Spread of residuals remains same for all fitted values. Also bell curve suggests normal distribution.
 2. Pair/Box plot between the different predictor variables and target variable
 3. VIF has been brought less than 5 to deal with multi collinearity.
 4. Durbin-watson: This value close to 2 indicates that there is likely no autocorrelation in the residuals, suggesting that the model's errors are independent.
 5. **F-statistic: 233.3 (p-value: 5.71e-181):** This very high F-statistic and an extremely low p-value imply that your overall model is statistically significant.
 6. **R-squared: 0.824:** This indicates that 82.4% of the variance in bike demand is explained by the variables in your model. Adjusted R-squared: 0.820 This is a slight reduction from the R-squared value, accounting for the number of predictors. It still suggests a good fit, with 82% of the variance explained.
 7. R-Squared for test is ~81%. Vary close to training result
-

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)

Top 3 features are:

1. Year: Highly significant. Positive impact on bike demand.
 2. Temp: Highly significant. Warmer temperatures increase bike demand.
 3. WeatherSit: Highly significant. Bad weather (snow/rain) decreases bike demand. Mist/cloudy weather reduces demand.
-

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)

<Your answer for Question 6 goes here>

Linear Regression is a supervised learning algorithm used to predict a numerical target variable Y based on one or more predictor variables X_1, X_2, \dots, X_n . The goal is to fit a linear relationship between the input variables and the target variable.

Types of Linear Regression:

1. Simple Linear Regression: Involves only one independent variable.
Formula: $Y = b_0 + b_1X + e$
2. Multiple Linear Regression: Involves more than one independent variable.
Formula: $Y = b_0 + b_1X_1 + b_2X_2 + \dots + b_nX_n + e$

Where:

Y : dependent variable

X_i = i th Independent Feature

b_0 : intercept when all $x_i = 0$

b_i : coefficient representing slope b/w x_i and y

e : Error term representing noise in data

Objective of Linear Regression:

The goal is to find the best-fitting line (or hyperplane in higher dimensions) that minimizes the error between the predicted values \hat{Y} and the actual values Y . This is typically done by minimizing the sum of squared errors (SSE):

$$SSE = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

Least Squares Method:

The most common approach to finding the best-fitting line is the Ordinary Least Squares (OLS) method. It minimizes the SSE by adjusting the values of the coefficients b_0, b_1, \dots, b_n .

Cost Function: The cost function measures how well the model fits the data. For Linear regression it is usually the MSE as shown above.

Gradient Descent (Optimization)

To minimize the cost function, we use an optimization algorithm like Gradient Descent. The idea is to adjust the coefficients b_0, b_1, \dots, b_n iteratively to reduce the cost function.

Assumptions of Linear Regression: For Linear Regression to work effectively, the following assumptions should hold.

1. **Linearity:** The relationship between the predictors and the target variable is linear.
2. **Homoscedasticity:** The variance of the errors is constant across all levels of the independent variables.
3. **Normality of Errors:** The residuals (errors) are normally distributed.
4. **No Multicollinearity:** The independent variables are not highly correlated with each other.
5. **Independence of Errors:** The residuals are independent (no autocorrelation)

Advantages of Linear Regression

1. **Simplicity:** Easy to understand and interpret.
 2. **Efficiency:** Computationally efficient and works well for smaller datasets.
 3. **Interpretability:** Coefficients provide a clear understanding of the relationship between features and the target variable.
-

Disadvantages of Linear Regression

1. **Assumption-Driven:** Requires the assumptions (linearity, normality, etc.) to be satisfied.
 2. **Sensitive to Outliers:** Outliers can have a significant impact on the model.
 3. **Limited to Linear Relationships:** Cannot model complex, non-linear relationships without transformations.
-

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)

<Your answer for Question 7 goes here>

Anscombe's Quartet is a collection of four datasets that have nearly identical statistical properties (mean, variance, correlation, regression line, etc.) but appear very different when visualized. It was created by the statistician Francis Anscombe in 1973 to highlight the importance of data visualization when analyzing data.

The quartet illustrates that relying solely on summary statistics (like mean, variance, and correlation) can be misleading and that visual inspection is crucial to understanding the true nature of the data.

Why Anscombe's Quartet is Important

1. It demonstrates that different datasets can yield similar statistical results but have entirely different distributions and relationships.
2. It emphasizes the value of plotting data to uncover underlying patterns, outliers, and nuances that summary statistics might miss.

3. It serves as a cautionary tale against blindly trusting statistical outputs without visual confirmation.

Key Observations:

- All four datasets have the **same mean** for both X and Y.
- They have the **same variance** for both X and Y.
- They have the **same correlation** between X and Y (approximately 0.82).
- They produce **nearly identical linear regression equations**.

However, despite these identical statistical properties, the datasets look **vastly different** when plotted. statistical properties, the datasets look **vastly different** when plotted.

Plotting Anscombe's Quartet

Here's what happens when you **visualize** each of the datasets:

1. Dataset I:

- A **linear relationship** with a small amount of random noise.
- The linear regression line fits well.

2. Dataset II:

- A **curved, non-linear relationship**.
- The linear regression line is not a good fit, but summary statistics do not capture this.

3. Dataset III:

- A linear relationship is **distorted by one outlier**.
- The outlier has a significant impact on the regression line, making it unreliable.

4. Dataset IV:

- Nearly all the data points have the **same X value** except for one outlier.
 - The regression line is heavily influenced by this single point, leading to a misleading fit.
-

Visual Representation

If you were to plot the four datasets, you'd see that each one has a unique distribution, despite having identical summary statistics. Visual inspection reveals the true differences among the datasets that statistical summaries obscure.

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)

<Your answer for Question 8 goes here>

Pearson's R, also known as the Pearson Correlation Coefficient, is a statistical measure that quantifies the strength and direction of a linear relationship between two variables. It is one of the most widely used correlation coefficients in statistics.

The value of Pearson's R ranges from -1 to +1:

1. +1 indicates a perfect positive linear relationship.
2. -1 indicates a perfect negative linear relationship.
3. 0 indicates no linear relationship.

Interpreting Pearson's R

Value of r	Interpretation
-1	Perfect negative linear relationship
-0.7 to -0.9	Strong negative linear relationship
-0.3 to -0.7	Moderate negative linear relationship
-0.1 to -0.3	Weak negative linear relationship
0	No linear relationship
+0.1 to +0.3	Weak positive linear relationship
+0.3 to +0.7	Moderate positive linear relationship
+0.7 to +0.9	Strong positive linear relationship
+1	Perfect positive linear relationship

A high absolute value of r does not imply causation but only a strong linear association between the variables.

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)

<Your answer for Question 9 goes here>

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)

<Your answer for Question 10 goes here>

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)

<Your answer for Question 11 goes here>
