
Bike Sharing Assignment

By Anshaj Singh

Assignment-Based Subjective Questions

- ❖ Q1. From your analysis of the categorical variables, what can you infer about their effect on the dependent variable?
- ❖ Ans. : From the analysis, seasonal and weather-related categorical variables showed a clear impact on bike demand. For example, summer and winter tended to have higher rental counts, while spring showed noticeably lower demand. Weather categories also had strong effects: clear weather days had the highest rentals, cloudy days reduced demand slightly, and light snow / rain sharply reduced the number of rides. This suggests that seasonal patterns and weather conditions are important drivers of bike rental demand.

Assignment-Based Subjective Questions

- ❖ Q2. Why is it important to use `drop_first=True` during dummy variable creation?
- ❖ Ans. : Using `drop_first=True` helps prevent the dummy variable trap, which occurs when dummy variables add perfect multicollinearity to the model. By dropping one category from each variable, we set a “baseline category,” making the model stable and ensuring that coefficients are interpretable. Without dropping the first dummy, linear regression can fail due to redundant columns.

Assignment-Based Subjective Questions

- ❖ Q3. Which numerical variable had the highest correlation with the target (cnt)?
- ❖ Ans. : Based on the pair-plot and correlation matrix, the variable “atemp” (feels-like temperature) had the highest positive correlation with the target variable cnt.

Assignment-Based Subjective Questions

- ❖ Q4. How did you validate the assumptions of Linear Regression?
- ❖ Ans. : I validated the assumptions in the following ways:
 - Linearity:
Scatter plots between predictions and residuals showed no major non-linear pattern.
 - Normality of residuals:
A Q–Q plot and a histogram of residuals were examined; residuals appeared roughly normal.
 - Homoscedasticity:
Plotting residuals vs predicted values showed roughly constant spread.
 - Multicollinearity:
Calculated VIF values and removed highly correlated variables like temp, weekday dummies, and month dummies.

Assignment-Based Subjective Questions

- ❖ Q5. Based on the final model, what are the top 3 features contributing significantly to bike demand?
- ❖ Ans. : From the final model, the top three positive contributors were:
 1. yr – Bike rentals increased significantly in 2019 compared to 2018.
 2. atemp – Warmer “feels-like” temperatures increased bike demand.
 3. season_Winter – Winter months surprisingly showed higher rental counts compared to the baseline season.

General Subjective Questions

- ❖ Q1. Explain the linear regression algorithm in detail.
- ❖ Ans. : Linear regression models the relationship between a dependent variable (Y) and one or more independent variables (X) by fitting a straight line that best describes the data. It works by minimizing the sum of squared residuals, which is done using Ordinary Least Squares (OLS). The model assumes linear relationships, constant variance of errors, normally distributed residuals, and no multicollinearity. Once trained, it provides coefficients that show how a one-unit change in each predictor affects the target variable.

General Subjective Questions

- ❖ Q2. Explain Anscombe's quartet.
- ❖ Ans. : Anscombe's quartet is a famous set of four datasets that have identical statistical summaries (mean, variance, correlation, regression line) but look completely different when plotted. It illustrates the importance of visualization because relying only on summary statistics can be misleading. Each dataset behaves differently (non-linear patterns, outliers, unusual spreads), showing why graphs are essential in data analysis.

General Subjective Questions

- ❖ Q3. What is Pearson's R?
- ❖ Ans. : Pearson's correlation coefficient (r) measures the strength and direction of a linear relationship between two numerical variables. Its values range from -1 to $+1$ (from perfect -ve correlation to perfect +ve correlation). It helps understand how strongly two variables move together in a straight-line pattern.

General Subjective Questions

- ❖ Q4. What is scaling? Why is it performed? Difference between normalization & standardization.
- ❖ Ans. : Scaling is the process of transforming numerical features to a consistent range so that models treat them fairly.
Scaling is performed as many models (including linear regression) perform better when features have similar scales, especially when using regularization or gradient-based methods.
- ❖ Normalization (Min–Max Scaling):
Rescales data to 0–1 range.
Sensitive to outliers.
- ❖ Standardization (Z-score Scaling):
Centers data at mean = 0, std = 1.
Less affected by outliers and preferred for regression.

General Subjective Questions

- ❖ Q5. Why do VIF values sometimes appear infinite?
- ❖ Ans. : VIF becomes infinite when a predictor is perfectly collinear with another variable (or a combination of variables). In other words, when one feature can be predicted exactly from others, the denominator in the VIF formula becomes zero, causing the value to blow up to infinity. This indicates extreme multicollinearity, and the variable should be removed.

General Subjective Questions

- ❖ Q6. What is a Q-Q plot and why is it used in linear regression?
- ❖ Ans. : A Q-Q plot (Quantile–Quantile plot) compares the distribution of residuals to a theoretical normal distribution. If the residuals follow the diagonal line closely, it means they are approximately normal. Normality of residuals is one of the assumptions of linear regression and helps ensure valid hypothesis tests, confidence intervals, and model reliability.