Linear Regression Assignment - BoomBikes Analysis
Subjective Questions and Answers

Assignment-Based Questions

Question 1: What can you infer about the effect of categorical variables on bike demand?

Looking at the categorical variables in the bike sharing data, I found some interesting patterns:

1. Season has a big impact on demand:
- Fall season is the best for bike rentals with about 5,644 bikes per day on average.
- Summer comes second with around 4,992 bikes daily.
- Winter is surprisingly decent at 4,728 bikes per day.
- Spring is the worst with only 2,604 bikes daily.
- This makes sense because spring weather can be unpredictable.

2. Weather conditions really matter:
- Clear weather drives the highest demand at 4,876 bikes per day.
- Misty or cloudy weather drops demand to 4,035 bikes daily.
- Light rain or snow hurts business with only 1,803 bikes per day.
- Heavy rain and snow kills demand at just 394 rentals daily.

3. Business is growing year over year:
- 2019 showed much higher demand at 5,634 bikes daily.
- 2018 had 3,842 bikes daily.
- That is a 47% growth rate, which is great for the company.

4. Monthly patterns follow the seasons:
- September and October are peak months with over 5,700 bikes daily.
- January and February are slowest at around 2,400 bikes daily.

Working days versus weekends dont show a huge difference, but working days are slightly better for business. All these patterns seem consistent and reliable for making business decisions.

Question 2: Why is drop_first=True important when creating dummy variables?

Using drop_first=True is really important when creating dummy variables because it prevents a problem called the dummy variable trap.

Here is what happens without it: If we have four seasons and create four dummy variables (Spring, Summer, Fall, Winter), these four variables will always add up to 1 for every row. If you know three values, you can always figure out the fourth one. For example, if Spring=0, Summer=0, and Fall=0, then Winter must equal 1.

This creates perfect correlation between the variables, which breaks the math behind linear regression. The computer cannot solve the regression equation because the correlation matrix becomes impossible to invert, which is needed for calculating the coefficients.

By using drop_first=True, we drop one category and use it as a reference point. So instead of four season dummies, we get three - Summer, Fall, and Winter. Spring becomes our reference category. Now when we interpret coefficients, Summers coefficient tells us how much higher or lower bike demand is in Summer compared to Spring.

This also makes the results much easier to interpret and prevents mathematical errors in the model calculation.

Question 3: Which numerical variable has the highest correlation with bike demand?

Temperature has the highest usable correlation with bike demand at 0.627. Although registered users shows a much higher correlation at 0.972, we cannot use that for prediction because it is part of our target variable (total = casual + registered), which would be data leakage.

Question 4: How did you validate the linear regression assumptions?

I checked the main assumptions of linear regression using plots and simple tests:

1. Linearity:

- Created a residuals vs fitted values plot to check if the relationship is linear.
- The residuals should be randomly scattered around zero with no clear pattern.
- Calculated correlation between residuals and fitted values, which was 0.002 (very close to zero).

2. Independence:

- Plotted residuals over time to check if they are independent.
- The residuals should not show any pattern over time.
- Our residuals look reasonably random.

3. Constant Variance (Homoscedasticity):

- Checked if the spread of residuals is roughly the same across all fitted values.
- Split residuals into two halves and compared their variances.
- The variance ratio was 1.58, which is acceptable (should be close to 1).

4. Normality of Residuals:

- Used histogram and Q-Q plot to check if residuals are normally distributed.
- Calculated skewness of residuals, which was -0.09 (close to 0 is good).
- Most points in Q-Q plot fell close to the diagonal line.

Overall, the basic assumptions seem reasonably satisfied, so the model results should be reliable.

Question 5: What are the top 3 features explaining bike demand?

Based on my final model using RFE feature selection, here are the top three features:

1. Year Variable (coefficient: +1974):

- Being in 2019 increases daily bike demand by about 1,974 bikes compared to 2018.
- This shows strong business growth and increasing popularity of bike sharing.

2. Temperature (coefficient: +5848):

- Each unit increase in temperature increases demand by about 5,848 bikes.
- This makes temperature the most important weather factor.
- Shows that bike sharing is very weather-dependent.

3. Light Rain/Snow Weather (coefficient: -1901):

- When there is light rain or snow compared to clear weather, demand drops by about 1,901 bikes.
- This shows how much bad weather hurts the business.

Together, these three features help explain about 84% of the variation in bike demand, making the model quite accurate for business planning.

General Questions

Question 1: Explain the linear regression algorithm in detail

Linear regression is a basic machine learning algorithm that tries to find the best straight line through data points:

1. Basic idea:

- For one variable, it finds the best line y = mx + b that fits the data.
- For multiple variables, it becomes y = b0 + b1*x1 + b2*x2 + ... + bn*xn.
- The coefficients (b1, b2, etc.) tell you how much each variable affects the outcome.

2. How it works:

- It minimizes the sum of squared errors between predicted and actual values.
- Takes the difference between prediction and reality, squares it, and adds them up.
- Finds coefficients that make this total as small as possible.

3. Main assumptions:

- The relationship between variables is linear.
- Observations are independent of each other.
- Variance of errors stays constant.
- Errors are normally distributed.

4. Advantages:

- Simple to understand and interpret.
- Fast to train and make predictions.
- No parameter tuning needed.

5. Limitations:

- Can only capture linear relationships.
- Sensitive to outliers.
- Assumptions need to be checked.

Question 2: Explain Anscombes quartet in detail

Anscombes quartet is a famous statistical example that shows why you should always look at your data visually, not just rely on numbers.

What it is:

- Four different datasets, each with 11 data points.
- All four have nearly identical statistical properties.

- Same means, variances, correlation coefficients, and even the same linear regression line.

The surprising difference:
- Dataset 1: Nice linear relationship - perfect for linear regression.
- Dataset 2: Curved relationship where linear regression clearly does not fit well.
- Dataset 3: Perfect linear relationship except one outlier ruins everything.
- Dataset 4: No relationship for most points, but one outlier creates fake correlation.

The lesson:
- Identical summary statistics can hide completely different patterns.
- You could look at just the numbers and think all datasets are the same.
- Always visualize your data before analyzing it.
- Summary statistics alone can be very misleading.

This example teaches us that visualization is just as important as statistical analysis in data science.

Question 3: What is Pearsons R?

Pearsons R (correlation coefficient) measures how strong the linear relationship is between two variables.

Value range:
- Always between -1 and +1.
- +1 means perfect positive relationship (as one goes up, other goes up).
- -1 means perfect negative relationship (as one goes up, other goes down).
- 0 means no linear relationship.

Interpretation guidelines:
- Above 0.7: Strong relationship.
- Between 0.3 and 0.7: Moderate relationship.
- Below 0.3: Weak relationship.

In our project:
- Temperature had a correlation of 0.627 with bike demand.
- This indicates a strong positive relationship.
- Makes intuitive sense - warmer weather encourages more bike rentals.

Important limitations:
- Only measures linear relationships.
- Sensitive to outliers.
- Correlation does not mean causation.

Question 4: What is scaling and why is it important?

Scaling is changing your data so that all variables are on similar ranges. This is important because some variables have much bigger numbers than others.

The problem without scaling:

- Variables with bigger numbers dominate the analysis.
- Example: Age (20-80) vs Income (20,000-200,000).
- Income would dominate simply because its numbers are much larger.

Two main types:

- Min-Max Scaling: Transforms data to fall between 0 and 1.
- Standardization: Makes data have mean=0 and standard deviation=1.

Why it helps:

- All features contribute equally to the model.
- Algorithms work better and faster.
- Fair comparison of feature importance.

Best practice:

- Always fit the scaler on training data only.
- Apply the same transformation to test data.
- This prevents data leakage.

Question 5: Why might VIF be infinite?

VIF (Variance Inflation Factor) becomes infinite when you have perfect correlation between variables in your model.

Most common cause - dummy variable trap:

- When you create dummy variables for all categories.
- Example: Four season dummies always sum to exactly 1.
- If you know three values, you can perfectly predict the fourth.

Other causes:

- Including the same variable twice by mistake.
- Having variables that are mathematical combinations of others.
- One variable being a constant multiple of another.

Why this breaks everything:

- Makes the correlation matrix impossible to invert mathematically.

- Linear regression needs matrix inversion to calculate coefficients.
- Results in computational errors and unreliable results.

Simple solutions:
- Use drop_first=True for dummy variables.
- Remove duplicate or redundant variables.
- Check correlations before building the model.

Question 6: What is a Q-Q plot and its importance in linear regression?

A Q-Q plot helps you check whether your data follows a normal distribution by comparing it visually to what perfect normal data would look like.

How it works:
- Compares your actual data against perfect normal distribution.
- If data is normally distributed, points fall on a straight diagonal line.
- Deviations from the line show departures from normality.

Reading the patterns:
- Points on straight line: Data is normally distributed.
- Curve upward at ends: Distribution has heavier tails.
- Curve downward at ends: Distribution has lighter tails.
- Consistent curve: Data is skewed.

Why its important in linear regression:
- Tests if residuals are normally distributed.
- This assumption affects confidence intervals and hypothesis tests.
- If residuals are not normal, statistical inference might be wrong.

In our analysis:
- Used Q-Q plots to check our model residuals.
- Most points fell reasonably close to the diagonal line.
- This gave us confidence that our statistical tests were reliable.

Q-Q plots provide a simple visual check that complements statistical tests for assumption validation.