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)

From the refined dataset (`data2`), the categorical variables and their effects on bike demand (`cnt`) include:

1. **Weather Condition (`weathersit`):**
   - `weathersit_Light Snow/Light Rain + Thunderstorm/Scattered Clouds`: Increased demand by **~2,067 rentals** compared to the reference category (`Clear or Partly Cloudy`).
   - This counterintuitive result might indicate that bikers are more resilient to light adverse weather or that specific conditions coincided with peak usage.
2. **Year (`yr`):**
   - An increase of **~1,973 rentals** was observed for 2019 compared to 2018.
   - This trend highlights the growing popularity of shared bike systems over time.
3. **Month (`mnth`):**
   - `mnth_January`: Increased demand by **~1,234 rentals**, suggesting consistent usage despite being a colder month.
   - `mnth_December`: Showed a marginally lower but still positive contribution, indicating that biking remains popular during the holiday season.
4. **Season (`season`):**
   - Summer and Fall seasons typically showed higher demand (though specific coefficients were not part of the refined results). These trends align with warmer, more conducive biking conditions.
5. **Working Day (`workingday`):**
   - Working days showed higher average demand than holidays. However, the coefficient after refinement was less pronounced, implying other variables (e.g., weather, time of year) play a stronger role.
6. **Weekday (`weekday`):**
   - The effects of weekdays (e.g., `weekday_Tuesday` or `weekday_Thursday`) were less significant in the final model. However, minor increases or decreases in demand indicate preferences for biking on specific days.
7. **Holiday (`holiday`):**
   - Holidays generally saw reduced demand compared to working days, consistent with user behavior preferring shared bikes for commuting rather than leisure.

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)

During dummy encoding of categorical variables, using `drop_first=True` avoids the **dummy variable trap** by removing one category to act as a reference. For example:

- In encoding `seasons`, if we generate dummies for fall, `spring`, `summer`, winter `drop_first=True` will exclude `fall` as the reference category.
- Same is the case for weekdays, in the current assignment, Friday was dropped basis the same.

This ensures that the model interprets the coefficients relative to the reference group, avoiding multicollinearity, and in the analysis file, this was evident as removing redundancy significantly improved model stability.

.

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)

Year (yr) had the highest positive correlation with the target variable (`cnt`), with a contribution of **~1,973 rentals**, showing a clear upward trend in bike usage.

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)

The following validations were performed using the residuals and model diagnostics:

1. **Linearity**: The residuals vs. predicted values plot displayed no discernible pattern, confirming a linear relationship.
2. **Normality**: The Q-Q plot indicated that most residuals aligned closely with the diagonal, although some outliers (e.g., values beyond **±1500**) slightly deviated.
3. **Multicollinearity**: VIF scores for the predictors were calculated, with all values belo**w 5**, ensuring no significant multicollinearity.

The model achieved an **R-squared value of 0.8495**, explaining nearly **85% of the variation in bike demand**.

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)

- The top 3 features from the final model were:

1. **Weather Condition (weathersit_Light Snow/Light Rain + Thunderstorm/Scattered Clouds):** Contributed **+2,067 rentals**.
2. **Year (yr):** Contributed **+1,973 rentals**, reflecting the growth trend in bike usage.
3. **Month (mnth_January):** Contributed **+1,234 rentals**, highlighting significant usage even in winter.

These features emerged as key determinants of bike demand in the refined analysis.

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)

Linear regression basically predicts a dependent variable Y as a linear combination of independent variables:

In this assignment, the model achieved an R^2 score of **0.8495**, indicating that **85% of the variation in bike demand** was explained by the selected predictors.

The algorithm minimizes the **sum of squared residuals** to fit the best line. For instance, in the model, minimizing residuals reduced the **mean absolute error (MAE)** to **~460 rentals**, highlighting its predictive strength.

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)

Anscombe's quartet demonstrates datasets with identical statistical properties (e.g., mean, variance, correlation) but differing distributions. Visualization in this analysis highlighted:

1. Scatterplots between predictors like `yr` and `cnt` showed linear trends, while others (e.g., `hum`) revealed weaker relationships.
2. The importance of scatterplots to detect nonlinear relationships (like temp vs. cnt).

3. How outliers (e.g., residuals exceeding **2,000 bikes**) can distort metrics without visual inspection.
4. Residual diagnostics helped detect potential outliers, ensuring reliable model inferences.

Thus, while summary statistics showed strong linearity (e.g., R^2=0.85), visual diagnostics revealed subtle deviations.

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)

Pearson's R quantifies linear correlation between two variables. For example, in this analysis:

- **Temp** and **cnt** had r=0.63 a strong positive correlation.
- **Humidity (hum)** had a weaker correlation of r=−0.4, reflecting lower demand in humid conditions.

R values closer to ±1 indicate stronger correlations.

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)

Scaling ensures features contribute equally to model training:

- In this analysis, features like `temp` (scaled to [0, 1]) and `hum` were scaled to handle large differences in magnitude.
- **Normalization**: Rescales data to [0, 1], preserving proportionality (e.g., `temp`).
- **Standardization**: Centers data around 0 with unit variance, useful for distributions like `windspeed`.

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)

An infinite VIF occurs when predictors are perfectly correlated. For example, if temp and atemp (feels-like temperature) were not dropped during preprocessing, VIF would spike due to their strong correlation (r=0.99), indicating multicollinearity.

We have observed similar high VIF values for Holiday and Sunday/Saturday variables as well

---

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)

A Q-Q plot compares residual quantiles to a normal distribution. In this analysis, the Q-Q plot showed a strong alignment, confirming normality. However, minor deviations at the extremes (residuals > **±1,500**) flagged potential outliers. Ensuring normality impacts the reliability of confidence intervals and hypothesis testing.
