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)

- Season: Higher demand in summer (season 3) and fall (season 4), lower in winter (season 1)
- Year (yr): Rentals increased in the second year (2019) compared to the first year (2018).
- Month (month): Demand is highest in summer months (June–September).
- Holiday: Lower rentals on holidays.
- Weekday: No significant variation, meaning demand is stable across weekdays.
- Working Day: Higher rentals on working days compared to weekends.
- Weather Situation (weathersit): Clear weather (category 1) has the highest rentals, while worse weather reduces demand.

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)

Using drop_first=True in dummy variable creation helps prevent the dummy variable trap, which occurs when all dummy variables are used together, leading to perfect multicollinearity.

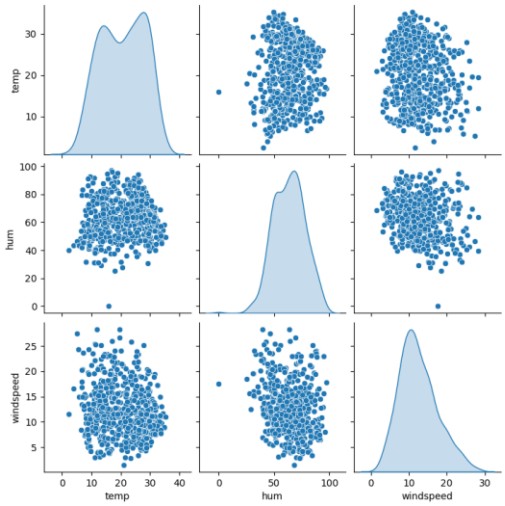
For example, if we create dummies for a categorical variable with three categories (A, B, C), we only need two dummy variables (e.g., A and B). The third category (C) is automatically inferred when both A and B are 0. Without drop_first=True, the model may include redundant information, causing instability in regression coefficients.

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)

Below is the pair-plot showing the relationship between the dependent variable and other numeric variables.



Among these, the variable “registered” exhibits the strongest correlation with the target variable. However, since the target variable is defined as the sum of casual and registered users, it’s logical to exclude both “casual” and “registered” from further analysis. Apart from these, the variables “temp” demonstrate the highest correlations with the target variable.

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)

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 equation is:
cnt = 0.1338 + 0.237 × (yr) - 0.0826 × (holiday) + 0.46 × (temp) - 0.17 × (windspeed) - 0.077 × (Spring) + 0.042 × (Summer) + 0.0705 × (Winter) + 0.091 × (clear_partlycloudy)

The top 3 influential features are “temp”, “yr”, and “clear_partlycloudy”:

- **temp**: For each unit rise in ambient temperature, the target variable goes up by 0.46 units.
- **yr**: For each unit increase in the “yr” feature, the target variable rises by 0.237 units.
- **clear_partlycloudy**: For each unit increase in this feature, the target variable increases by 0.091 units.

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 is a statistical method used to model the relationship between a dependent variable (Y) and one or more independent variables (X)

Types of Linear Regression

- Simple Linear Regression – One independent variable.
- Multiple Linear Regression – Two or more independent variables.

Assumptions of Linear Regression

- Linearity – Relationship between independent and dependent variables must be linear.
- Independence – Observations should be independent (checked using the Durbin-Watson test).
- Homoscedasticity – Residuals should have constant variance.

- Normality of Errors – Residuals should be normally distributed.
- No Multicollinearity – Independent variables should not be highly correlated (checked using VIF).

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 is a set of four datasets that appear very similar in basic statistical summaries (mean, variance, correlation) but are significantly different when visualized. It was introduced by Francis Anscombe in 1973 to highlight the importance of data visualization in statistical analysis.

The Four Datasets
Each dataset consists of 11 data points (X, Y), and all four share:

- The same mean for X and Y.
- The same variance for X and Y.
- The same correlation coefficient (~0.816).
- The same regression line.

However, when plotted, they show very different distributions:

- Dataset 1 – A typical linear relationship.
- Dataset 2 – A nonlinear, quadratic-like curve.
- Dataset 3 – A linear pattern with an outlier affecting the regression.
- Dataset 4 – A vertical line with one extreme outlier, misleading correlation.

Key Takeaways

- Summary statistics alone are not enough to understand data.
- Visualizing data (scatter plots, histograms, etc.) is crucial.
- Outliers can significantly affect regression models and correlations.

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, also known as Pearson's correlation coefficient, is a statistical measure that quantifies the strength and direction of a linear relationship between two continuous variables. It's widely used in statistics to assess how well the relationship between two variables can be described by a straight line.

Range: Pearson's R ranges from -1 to +1.
+1: Perfect positive linear relationship (as one variable increases, the other increases proportionally).
0: No linear relationship (the variables don't move together in a predictable way).
-1: Perfect negative linear relationship (as one variable increases, the other decreases proportionally).

Interpretation: The closer the value is to 1 or -1, the stronger the linear relationship. Values near 0 suggest little to no linear correlation (though there could still be a non-linear relationship).
It's worth noting that Pearson's R only measures linear relationships and assumes the data is normally distributed and measured on an interval or ratio scale.

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 is a preprocessing technique used in data analysis and machine learning to transform the values of numerical variables into a specific range or distribution. The goal is to ensure that features (variables) are on a comparable scale, preventing those with larger ranges or units from disproportionately influencing algorithms that rely on distance, magnitude, or gradient calculations (e.g., regression, neural networks, or k-means clustering).

Scaling is performed for several reasons:

- **Comparability**: Features often have different units or ranges (e.g., height in centimeters vs. weight in kilograms). Without scaling, a feature with a larger range could dominate the results of an algorithm, even if it's not inherently more important.
- **Algorithm Performance:** Many machine learning algorithms (like gradient descent-based models, SVMs, or KNN) assume features are on similar scales. Scaling improves convergence speed and accuracy.
- **Numerical Stability:** Large, unscaled values can lead to computational issues like overflow or rounding errors.
- **Interpretability:** Scaling can make coefficients or feature importance easier to compare in some models.

Key Differences

Aspect	Normalized Scaling (Min-Max)	Standardized Scaling (Z-Score)
Range	Fixed (e.g., [0, 1])	Unbounded (depends on data)
Mean	Not necessarily 0	Always 0
Std Deviation	Varies	Always 1
Outlier Sensitivity	High (affected by min/max)	Lower (based on mean/std)
Use Case	Bounded inputs, non-normal data	Normal-like data, robust models

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)

The Variance Inflation Factor (VIF) is a measure used in regression analysis to detect multicollinearity—when independent variables in a model are highly correlated with each other. A VIF value of infinity indicates a severe issue with the data, and happens due to:

- **Duplicate Variables**: If two variables are identical
- **Linear Dependencies**: If a variable is a perfect linear combination of others
- **Over-specified Model**: Including more variables than observations (or linearly dependent variables) can collapse the regression
- **Data Errors**: Mistakes like accidentally including the same variable twice under different names can trigger this.

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 (Quantile-Quantile plot)** is a graphical tool used to compare the distribution of a dataset with a theoretical distribution, typically the **normal distribution**. It helps determine whether the residuals (errors) in a regression model follow a normal distribution.

Use of Q-Q Plot in Linear Regression

- Checking Normality of Residuals – One key assumption of linear regression is that residuals should be normally distributed. If the Q-Q plot shows significant deviations from the diagonal line, this assumption is violated.
- Detecting Skewness – If points curve upward or downward, it indicates skewed data.
- Identifying Outliers – Points that deviate significantly from the diagonal line suggest potential outliers.
- Improving Model Validity – If residuals are not normally distributed, transformations (e.g., log transformation) or alternative models (e.g., non-linear regression) may be needed.

Importance of Q-Q Plot in Linear Regression

- Ensures the validity of hypothesis testing (e.g., p-values in regression models).
- Helps in choosing the right model for prediction.
- Prevents biased parameter estimates caused by non-normal residuals.