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)

On pairplotting all categorical variables from the dataset we can see that
- the months after October have higher rent counts than the others
- Wednesday, Friday, and Saturday have higher rent counts than the other days
- Fall season has higher rent counts

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)

We need to use **drop_first=True** as that can reduce a dummy variable in our final set of inputs. Since the variable we dropped can be accounted for by the other dummy variables for this categorical variable since where this dummy variable has value **True**, all other dummy variables are **False**

also if we do not drop the column the first variable is a linear combination of all others, which breaks the assumption of linear regression that all variables are independent

---

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)

On looking at the numerical variable pairplot, we can see that both **temp** and **atemp** have the highest correlation with the target variable **cnt**

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)

I validated the assumptions of linear regression after training the model by checking the distribution of the Residual Square Errors (RSE) by subtracting y_pred from y_test and plotting the distribution, which closely matched a normal distribution which shows

that the model we obtained is a very good predictor of our dataset

We can also use the R^2 metric, which is the ratio of the variance in the model compared to the actual variance in the dataset to evaluate the model, in this case the R^2 metric for the training data is **0.76692** and for the testing data is **0.70703**

this means that the model explains around 70.7% of the variance in the dataset since the R^2 metrics for the test and training data are close together (within 10%) this means that the model has generalized farily well and is not overfitting

---

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)

Based on the final model, the 3 most significant features are
- weathersit_snowy – if the weather situation is: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
- weekday_Sat – if the day is a Saturday
- mnth_Sep – if the month is September

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 widely used statistical method for modeling the relationship between a dependent variable (target variable) and one or more independent variables (predictor variables). The goal of linear regression is to create a linear equation that best predicts the value of the dependent variable based on the values of the independent variables.

There are many types of Linear Regression
- Simple
- Multiple
- Weighted
- Regularize

We make certain assumptions in linear regression which are as follows

---

- **Linearity**: The relationship between the dependent and independent variables is linear.
- **Independence**: Each observation is independent of the others.
- **Homoscedasticity**: The variance of the error term is constant across all levels of the independent variable.
- **Normality**: The error term is normally distributed.
- **No multicollinearity**: The independent variables are not highly correlated with each other

---

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, comprising four datasets with nearly identical summary statistics, underscores the limitations of relying solely on numerical metrics.

It comprises a set of four datasets, having identical descriptive statistical properties in terms of means, variance, R-squared, correlations, and linear regression lines but having different representations when we scatter plots on a graph.

The datasets were created by the statistician Francis Anscombe in 1973 to demonstrate the importance of visualizing data and to show that summary statistics alone can be misleading.

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 Coefficient is a correlation coefficient that measures linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations; thus, it is essentially a normalized measurement of the covariance, such that the result always has a value between −1 and 1.

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 refers to adjusting the range or distribution of features in data, ensuring they

are comparable in magnitude for better model performance.

Scaling is performed as it improves model performance, prevents feature dominance, and speeds up convergence in machine learning algorithms.

Normalized scaling (Min-Max) rescales data to a fixed range (e.g., [0,1]), while standardized scaling (Z-Score) centers data to mean 0 and standard deviation 1.

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 VIF becomes infinite when there is perfect multicollinearity between independent variables. This happens when one predictor is an exact linear combination of other predictors, making the denominator of the VIF formula zero

this may happen if
- we don't drop one of the dummy variable columns
- there are columns which are linearly related, such as **temp** and **atemp**
- duplicate variables

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 (Quantile-Quantile) plot is a graphical tool used to compare the distribution of a dataset against a theoretical distribution (e.g., normal distribution). It plots the quantiles of the sample data against the quantiles of the reference distribution. If the points lie approximately along a 45-degree line, the sample follows the chosen distribution.

The uses of Q-Q plot in Linear Regression is
- checking if the residuals lie on a 45 degree line
- detecting outliers in residuals
