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)

August, September and October months seems to have significant positive effect on the total rental counts. Summer and winter in Seasons have significant positive effect. Sunday in weekdays have positive impact. Working day also have a positive significant effect on total rental counts. Light snow and Mist in weather situation have negative impact on rentals. Holiday also have negative impact on rentals.

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)

It is necessary to avoid Dummy Variable Trap. Dummy variables are created with one hot encoding where each can take 0 or 1 value. This raise a scenario where if we don't drop one of variables then the dummy variables are highly correlated. Hence drop_first = True is used to drop one of the variables particularly first one for the convenience.

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)

Temp. Temp has a correlation of 0.63

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)

Residual distribution and Residual pattern analysis. Residual Distribution or Error terms distribution are assumed to be normally distributed (mean = 0). This can be validated by plotting dist plot of error terms. (sns.distplot(y_train_pred-y_train)). Error terms are assumed to be independent which means there should be no visible pattern of error terms across y.

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)

Top 3 positive coefficient features are Temp, yr and winter. Top 3 negative coefficient features are Light snow, windspeed and humidity.

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)

Read and Analyze the data: Preliminary analysis on the data. Pair plot, heat map, dist plot etc
Dummy Variables : Create dummy variables for Categorical variables with option drop_first = True. Drop the Categorical variable columns.
Split train-test data : Split data among train and test data before any preprocessing for Regression modelling
Rescaling : Rescaling of numerical variables is important as it will help in easy interpretation and faster conversion of gradient descent method
Training the model using RFE : This will help when there are many features and manual regression modelling is difficult. Select top 15 features with least VIF(0-5).
Significance : P values should be less than 0.05 to be significant
Test the Model for Linear model assumptions : Normal Residual distribution
Prediction on Test set : Calculate y_pred using the linear model
Analysis on test prediction : R2 value and Normal residual distribution are analyzed.

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 used to stress the importance of data visualization. This quartet comprises of four data sets that have nearly identical simple descriptive statistics, yet have very different distributions when graphed. All important features in the dataset must be visualized before implementing regression models. Anscombe's Quartet shows how regression model can be fooled .

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)

It's a measure of the strength of the association between two variables. It applies for continuous variables. Scatter plot can be used to see if the variables have linear relationship. If there is no linear relationship Correlation coefficient should not be calculated. Positive correlation indicates that both variables increase or decrease together. Negative correlation indicates that as one variable increases, so the other decreases, and vice versa

---

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 method to normalize the range of independent variables. Objective functions will not work properly without normalization in machine learning algorithms because the raw data range of different features may vary in orders significantly. Standardization brings all of the data into a standard normal distribution with mean zero and standard deviation one. Normalized scaling brings all of the data in the range of 0 and 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)

Infinite VIF value indicates that the corresponding variable may be expressed exactly by a linear combination of other variables. Infinite VIF indicates that RSquare is 1 which indirectly implies that explained variance is equal to total variance. This happens when there is a high degree of multi collinearity in the data set

---

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)

QQ plot is an intuitive visualize whether something is normally distributed. If the data adheres to the 45 degree line, its normal or close it and if it doesn’t, then its not normal. QQ plot means quantile-quantile plot. Plot compares quantiles of our data against the quantiles of the desired distribution. Quantiles are breakpoints that divide our numerically ordered data into equally proportioned buckets. In linear regression we assume that residuals are normally distributed. A Q-Q plot where the data fits closely on 45 degree line implies that the distribution is normal.
