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)

The categorical variables that I have used in my model are yr , month , holiday , weekday, workingday, summer , Light snow/rain , Mist and clouds .

Year, weekday, working day, and summer have statistically significant positive effects on rental counts.

Holiday, light snow/rain, and mist/clouds significantly reduce rental counts.

Month has a minor positive effect but remains statistically significant

---

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 , prevents our model from the dummy variable trap , it occurs when all the categories of a categorical variable are included in the dataset , hence it increases multicollinearity so in order to prevent this we drop the first category out of all the other categories since that first category can be defined using the other categories , if all the other categories are 0 then it means it belongs to first category.

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)

Registered column has a very high correlation value (0.95)

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 Analysis : Here by calculating the residuals first , formula for residuals is residuals = y_true – y_predicted , then we will plot the residuals on a histogram and see if the residuals are normally distributed or not , it should be aligned to mean 0 forming a bell shaped normal distribution curve .

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 my model the top three features are :

atemp– the coefficient is 0.5861and the p value is 0.00 making it a highly significant column.

yr– the coefficient is 0.2324 and p value is 0.00 , making it highly significant.

mnth– the coefficient is 0.0058 and p value is 0.0060, making it highly significant.

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)

<Your answer for Question 6 goes here>
Linear regression is a supervised ml algorithm useful for the prediction of y (dependent variable) with the help of x (independent variables) , for one variable the equation is y = mx+c here c is the value of y for x = 0 and m is the slope i.e. the change in y for a unit change in x . we calculate the residuals using y actual and y predicted (residuals = y_actual – y_predicted) we tend to minimize it for getting the best fit line. With the help of gradient descent we adjust the weights (w) and bias(b) referring to the equation y = wx + b . It is useful for data that follow a linear relationship between the dependent and the independent variables.

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)

<Your answer for Question 7 goes here>
Anscombe's quartet is useful to illustrate the importance of plotting before we analyze the data and build a model over it. The 4 datasets here have the same statistical information that is the mean and standard deviation for all are same but if we plot them the scatter plots are entirely different for all four datasets. Hence we get a clear picture of the data and proceed if it is fit to be used for linear regression algorithm. For ex the dataset 1 will fit the linear regression model pretty well , dataset 2 can't fit the linear regression model because the data is not linear in nature it follows non linear trend , dataset 3 shows the outliers involved in the dataset which can not be handled by the linear regression , dataset 4 shows totally a non linear structure with outliers which is not fit for Linear Regression model.

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)

<Your answer for Question 8 goes here>
Pearson's R also known as Pearson Correlation Coefficient is the statistical measure that describes the strength and direction of a linear relationship between two variables , it ranges from -1 to +1 , +1 indicates a perfect positive linear relationship , -1 indicates a perfect negative linear relationship and 0 means no relationship.

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)

<Your answer for Question 9 goes here>

The process of changing the feature values of the dataset to fall into the particular range of distribution is called as feature scaling. It guarantees that all the features contribute equally to the model , many ml algorithms are sensitive to the magnitude of the features , with the help of scaling we get the convergence faster (gradient descent) hence model training becomes faster.

Difference between Normalization and Standardization is :

Normalization or MinMax Scaling scaled down the values to [0,1] , it is best for bounded data where range is necessary , formula for normalization is x – xmin / (xmax – xmin) it is sensitive to outliers.

Standardization refers to transforming the data to have 0 mean and standard deviation 1 , formula for standard deviation is (x – mean)/standard deviation , it is not sensitive to outliers .

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)

<Your answer for Question 10 goes here>

VIF Infinite refers to high multicollinearity it exists when we can define the predictor variable as the linear combination of the other variables , when VIF is infinite it is very difficult for the model to distinguish between the contribution of each independent variable since they are contributing as a whole or we can say as a combination. VIF = 1 / 1-R^2 , if R^2 is equal to 1 i.e. High Multicollinearity this implies 1/1-1 = 1/0 = infinite value.

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)

<Your answer for Question 11 goes here>

A Q-Q Plot is a graphical technique for determining if a dataset follows a specific theoretical distribution, such as a normal distribution. The quantiles of the dataset are shown against the quantiles of the theoretical distribution. If the data closely fits the theoretical distribution, the points in the Q-Q plot will align along a straight diagonal line. However, deviations from this line indicate that the presumed distribution has changed. For instance, heavy or light tails are shown by deviations at the plot's extremities, whereas skewness may be indicated by a curved pattern. Q-Q plots are often used to assess the normality of regression model residuals or to find outliers in a dataset.
