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 analysis on categorical variables shows following information:

1. Highest bike rentals happen in fall season followed by summer season
2. This aligns with observation on the number of rentals across months in the year
3. Wednesday & Saturday account for highest rentals
4. Rental numbers shrink in non-dry weathers like cloudy and lightly snowy. No rentals during heavy rains or snowfalls
5. Median for rentals is higher on non-holidays however inter quartile range is high during holidays implying that probably causal users become active during holidays and increases overall rental numbers
6. There is significant growth in bike rentals from year 2018 to 2019

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 important to use *drop_first=True* while creating dummy variables is because *add_dummies* will create dummy variables for each value of the categorical column however this is not required. Hence, a column can be reduced and removing this would mean absence of one value implying that a particular row doesn't have either of the other defined values.

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 column has the highest correlation with 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)

Following are the elements used to validate the assumptions of linear regression after building the model on the training set:

- Residual Analysis: Depicting the error distribution
- Linearity: Depicting linear relationship on actual values vs predicted values
- p-Value: p-value of all the variables is less than 0.05
- VIF(Variance Inflation Factor): VIF for variables, ideally, should be less than 5 however for

significant variables, it can still be higher and upto 10.
- Root Mean Square Error(RMSE) should be close to 0

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)

Following are the three top contributors explaining demand of shared bikes:

1. temp
2. yr
3. Sep

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 statistical technique used in machine learning that help to assess the relation between two variables. There are independent variables, also called features/predictors and there is a target-variables, also called dependent variable. The objective to find the best fit straight line between the values of these two types of variables. This method is used in predictive analysis as well detecting relation between two or more variables.

The mathematical equation is as follows:

$y = b0 + b1*x1 + b2*x2.. + bn*xn$

b0 is the intercept i.e. value of y when x=0
b1, b2..bn shows the coefficients of all the dependent variables.

There are two types of linear regression:

1. Simple Linear Regression: It includes one variable
2. Multiple Linear Regression: It includes multiple variables

Assumptions of Linear Regression:

1. Linearity: The relationship between independent variable and target variable is linear
2. Independence: All observations are independent

3. Homoscedasticity: there is a constant variance of errors
4. Normality: Error terms are normally distributed. When plotted on histogram, it shows a bell curve.

Python libraries used in linear regression are statsmode & sklearn

---

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 refers to a set of four datasets that have similar statistical information however when data is plotted or viewed on charts/graphs, it depicts a totally different interpretation. The statistical properties of mean/median/variance, correlation, R-squares and linear regression lines for these identical data set are different when plotted on graphs.

The intent of Anscombe's quartet is to show that simply relaying of numerical values of dataset doesn't suffice data analysis and it take visual representation to identify correlations, outliers etc. that we may not be able to see by looking at just numbers.

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 is also known as Pearson correlation coefficient. It is a measure of relationship between two numerical / qualitative variables. Using Pearson's R value, we can find the strength and direction of two variables.

Person's R value ranges between -1 and 1.

-1 signifies negative relation between the variables implying that variables are inversely proportional. i.e. when one value of one variable increases, value of other variable decreases

+1 signifies positive relation between the variables showing direct relation i.e. when value of one variable increase, value of second variable also increases.

0 shows non-linear 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>

Scaling is a method to scale the quantitative variables of a dataset so that all the features contribute equally when designing a machine learning model. This is done during the preprocessing stage of data processing/analysis

Scaling is done as it adjusts the range and distribution of features of a dataset. The effect of each variable can be Equi-weighted in pro-rated manner. If not done, features with larger values may affect the model development inversely.

The two modes of scaling are normalized and standard.

Normalization scales the data such that each value falls between 0 and 1.

Standardization scales the data such that mean is 0 and standard deviation is 1 for the column.

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 (Variance Inflation Factor) is used to measure the severity of collinearity in linear regression. It is computed by (1 / (1 – R2)).
Infinite value of VIF essentially means that R2 is 1. Value of R2 can be 1 when there is a perfect correlation between two variables. As a practice, such one of such variables is dropped from the dataset to have a more rational model.

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>

Q-Q plot is a scatter plot of quantiles of two variables against each other. It gives an indication of whether distributions of two variables are similar or not.

Uses and Importance of Q-Q plot are as follows:

1. Identification of deviations: Straight line deviation in Q-Q plot can indicate deviations in normal distributions. It can depict tails or skewness.
2. Model fitment: Q-Q plots can be used to improve the model fitness by using different modelling approaches

---