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)

After completing the assignment, there are following categorical variables that affects the number of bike rentals -

1. Working Day - It shows a positive but weak correlation with the number of rentals.
2. Spring (Coefficient: -0.1092, P-value: 0.000): A negative coefficient indicates that bike rentals tend to decrease albeit its a weak relation during the spring season. This effect is statistically significant.
3. Winter (Coefficient: 0.0737, P-value: 0.000): A positive coefficient suggests more bike rentals in winter & is statistically significant.
4. January (Coefficient: -0.0409, P-value: 0.020): January has a slight negative impact on bike rentals & is statistically significant.
5. November (Coefficient: -0.0613, P-value: 0.001): November also shows a decrease in rentals compared with statistical significance.
6. September (Coefficient: 0.0733, P-value: 0.000): September has a positive impact on rentals & is significant.
7. Saturday (Coefficient: 0.0715, P-value: 0.000): Saturdays show a positive impact on bike rentals.
8. Cloudy Weather (Coefficient: -0.0521, P-value: 0.000): Cloudy weather leads to fewer rentals.
9. Light Rain (Coefficient: -0.2460, P-value: 0.000): Light rain has a large negative impact on bike rentals, reducing them significantly compared.

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 during dummy variable creation to eliminate redundancy. For every categorical variable with n values, n-1 dummy variables show sufficient information. For example, instead of creating 2 dummy variables MaleGender and FemaleGender for Gender categorical variable (Male, Female), only 1 variable MaleGender having 0 and 1 denoting if it's male otherwise female is enough.

pd.get_dummies(data frame, columns='columns to drop', drop_first=True, dtype=int) uses drop_first=True to drop the first column out of n dummy columns created.

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)

Looking at the pair plot of cnt i.e. our target variable with the numerical columns, temp & atemp (feeling temp) had the highest correlation. Other than this, if we consider casual and registered users as well in the pairplot, registered which denotes the number of registered users seems to have the highest correlation followed by casual denoting casual users.

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)
To validate the assumption of linear regression after building the model on the training set:
1. I plotted the histogram of error terms or residuals which is the difference of actual value and predicted value. The plot looked like a normal distribution with mean centered around 0.
2. After that I plotted the residual on a scatter plot to see the spread of the error terms. It was Homoskedastic that meant the variance was constant.
3. I also ran the test data set on the trained model and got a R2 score of around 0.786 which is around the trained model.
4. The spread of actual vs predicted test values is also constant.

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)

September month, saturday weekday and atemp (felt temperature) are the ones with the highest coefficient in my linear regression model.

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 used to predict linear relationships. The variable to be predicted is a numerical continuous variable. It comes under supervised machine learning because we use the past data to learn and predict future data. It is divided into 2 types:

1. Simple Linear Regression where we have output variable dependent on only 1 variable.
2. Multiple Linear Regression where we have output variable dependent on more than 1 variable.

We use scatter plot to display a relation in case of simple linear regression. The equation is as follows: y = mX+ c, where y=output, m=slope and c = intercept. y here is dependent on X.

For multiple linear regression, the equation changes to y = B0 + B1X1 + … BNXN.

We find the best fitted line equation that learns our model and use it to predict the future. The residual analysis of the line or the error can be calculated using Ordinal Least Squares Method. Gradient Descent is also used for the same.

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 group of four datasets that have nearly identical simple descriptive statistics, such as the mean, variance, correlation, and linear regression line, yet they appear different when plotted on a graph. This shows us how only descriptive or quantitative statistics is not enough to infer or conclude about the data. It highlights the importance of data visualization.

1. Dataset 1 is linear and fits a regression line. The points are spread around the regression line.
2. Dataset 2 is non-linear. While the summary statistics suggest a linear relationship, a visual plot shows a curved relationship between x and y. The y-value increases , peaks and then decreases.
3. In dataset 3, most of the coordinates follow a line except one outlier.
4. In dataset 4, all the y-values are present in a single x line except one outlier far away.

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 or the Pearson correlation coefficient is used to indicate correlation between variables. It has a range from -1 to 1. -1 implies a strong negative correlation and 1 implies a strong positive correlation.
The value of one variable increases or decreases as the other value changes.
It is the most common way to measure linear correlation and is heavily used in Linear Regression.

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)

It is important to scale the features so that the coefficients obtained by fitting the regression model are not very large or very small when compared. It is advisable to standardize or normalize our dataset so that the coefficients obtained are on the same scale. There are 2 common ways to scale :

1. Min-Max Scaling
   In Min-Max scaling, we will scale the features using the formula - (x-xmin)/(xmax-xmin), where x is a series of values in a column. This will scale all the values between 0 and 1.
2. Standardization
   In this scaling method, we convert our values such that the mean is equal to 0 and standard deviation is equal to 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)

Variance Inflation Factor or VIF, gives a basic quantitative idea about how much the feature variables are correlated with each other. It is an extremely important parameter to test our linear model. The formula for calculating VIF = 1/(1 - R^2). If VIF is infinite, it means that 1-R^2 = 0 i.e. R = +1 or -1. This R is the Pearson correlation coefficient and implies that there is a high correlation whether positive or negative.

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 which is used to compare the distribution of a dataset to the normal distribution most of the time. This plot helps assess whether the data follows a specific distribution by plotting the quantiles of the sample data against the quantiles of the reference distribution.

After we have created our model using **linear regression**, we plot the error terms or residual distribution that is calculated by taking a difference of actual and predicted values. This distribution plot should be normally distributed and should have a mean around 0. A Q-Q plot is basically used to check the assumption that the residuals (errors) of the model are normally distributed. There can be various issues if the error terms are not normally distributed like non-linearity, outliers or tail ends.
