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)
Positive Effect:
columns: workingday,atemp.
2019 has seen more demand than 2018.
Following months will positive affect the demand: March,May, October, September
Sunday seems to have more demand compared to other days.
Negative Effect:
columns: hum,windspeed, spring season,
Light weather (Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds),
Mist weather(Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist), July month

---

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)
Default pandas function get_dummies will create k variables for k values. But as a general rule of thumb, k-1 variables are enough to define k categories. drop_first=True is the parameter to get_dummies function which will usually reduce the number of columns created by one.

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)
atemp or temp variable (depending on which column is kept) has the highest correlation with cnt target variable.

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)

First we will predict the target variable using the built model and then take residuals from the actual target variable.
These residuals should be normally distributed and have mean of zero.
Also, we can check by plotting scatter plot of residuals vs predicted dependant variable

---

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)

Positively affecting: Year 2019, atemp
Negatively affecting: Light rain
So, in summary, Year 2019, atemp and light rain are the top 3 features

---

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 defined as one dependent variable that is linearly correlated with multiple independent variables.
Linear regression algorithm works as follows:
1. Formulating the model (identify coefficients for all independent variables)
2. Define a cost function (Ordinary Least square)
3. Adjust coefficients for independent variables so that it minimizes the cost function (e.g., gradient descent)
4. Make predictions
5. Evaluate the model based on R-squared/adjusted R-squared values on train set and test set.

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>
Statistician Francis Anscombe in 1973, demonstrated that even if statistical properties of a few datasets are similar, when plotted, it can reveal very different patterns.

He demonstrated with 4 datasets having the same statistical properties.
Dataset 1: exhibited linear relationship
Dataset 2: formed parabolic curve
Dataset 3: formed straight line with one outlier
Dataset 4: All the values are identical except one outlier

So, with the help of these plots, statistical properties are not enough, visualizations are also equally important.

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 a statistical measure that quantitatively measures the linear relationship between two numerical variables.
It's range of values vary between -1 to 1.

Values vs relationship:
-1 : both variables are strongly correlated with each other negatively.
0: there is no relationship between two variables, i.e., both are independent
1: both variables are strongly correlated with each other positively

---

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 the features is a method which converts different numerical variables' values on the same scale.
Scaling is performed so that the model can accurately predict the dependent variable and algorithms can efficiently process the data.
standardized scaling will standardize the values based on mean and standard deviation. It can contain outliers.
Normalization will convert every numerical variable on the scale of 0 to 1. No outliers will be present.

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 can be infinite if R-squared for that particular independent variable is 1 indicating perfect multicollinearity.
It can happen if there are duplicate variables.
Also, in this dataset, due to dummy variables, a variable can have infinite VIF because it might be defined as a combination of other variables perfectly.
For example:Weekend = saturday+sunday

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>
The Q-Q plot is a quantile-quantile plot. The Q-Q plot is used to identify whether a particular dataset follows a particular theoretical distribution. Quantiles from the dataset are plotted on y-axis and equal quantiles are plotted on X-axis. Now, if a line can be almost perfectly fitted through the

data points, then the dataset follows the distribution to which it was compared.

In case of linear regression, Quantiles of the residuals are compared to quantiles of normal distribution so that assumption of linear regression can be verified visually.

Q-Q plot is important because it gives visual intuition and helps to identify patterns that summary statistics may miss. (Anscombe's quartet).
