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)
Please find below the analysis of categorical variables from the dataset:

1. Season: Average Bike Rentals decrease during the spring season. This might be imparted to unpredictable or unfavorable weather conditions in that particular reason for which the dataset belongs to. Might be an off or less busy season. It is the highest in fall maybe due to comfortable temperatures, low humidity and fall has lesser rainy days. The other seasons, winter and summer have decent demand for bike rentals.
2. The number of bike rentals shows an increasing trend from 2018 to 2019. And this might be considered an important factor for the future, as bikes are becoming trendier, the demand may surge.
3. Bike demand shows a seasonal trend and increases from January till May, fairly consistent around the same demand from May till October (with a small dip), and decreases during November and December. This may be due to the holiday season, people going out for vacations and celebrations at home and the extreme winter.
4. Lesser demand during the holidays as people might prefer celebrating or relaxing compared to the non-holidays.
5. Weathersit is an important parameter, if the situation is clear or partly cloudy, more is the demand for bike rentals compared to snowy, or rainy situations.

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)
The use of drop_first = True during dummy variable is important due to the following reasons:

1. Say we have three categories in a column, dummy variable creation will initially lead to three columns with binary value for each category. Now, mathematically if the first created column is 1, the other two variables will have 0 as value. Therefore, necessarily, with just two columns we can still have the information of the additional category.
2. Due to dropping one column, we are reducing the model complexity without loss of any important information as the other dummy columns still give information of this dropped column.

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)
There are two parts to the solution:

1. Variables highly correlated but dropped from the model: Casual and Registered are the two variables which are highly correlated with the target variable, however, these two are removed because:
    a. At a given point in time, we might not have a total idea of registered vehicles or the casual customers coming in to book a bike. Therefore, it might be better to drop one or both of the variables. As a machine learning objective, we should not keep the

same/similar variable as predictor variables and it is given that the sum of registered and casual is the total vehicle rented (target variable).

b. Variables comparatively highly correlated: Temperature and Experienced Temperature also show some extent of correlation with the target variable and highly correlated with each other. In the final model, therefore, decided to keep only one variable out of these two.

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)

I validated the assumptions of Linear Regression after building the model in the following ways:

1. Correlation/Multicollinearity Check:
    a. Correlation matrix is plotted at the initial stage to refrain keeping highly correlated variables
    b. VIF (Variance Inflation Factor) - VIF is a measure of multicollinearity, usually a VIF of greater than 5 might lead to having variables with high correlation. Therefore, along with a p-value check to understand the importance of variables, did a VIF to drop the features leading to multicollinearity.
    Multicollinearity should not be present during Linear Regression model training because, when we talk about individual weights (betas) assigned to a variable which indicates how much change in y happens due to unit change in a variable x, considering the remaining variables (independent) remains constant would otherwise be proven false with the presence of multicollinearity.
    c. Residual Analysis: Plot the residuals to understand if they are normally distributed. Two methods can be leveraged, one via plotting the distribution directly to gauge it or by plotting a QQ Plot, if it comes in a straight line approximately, then it is normally distributed

---

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)

The top three features based on the final model are:

1. Atemp: This refers to feeling temperature and also should some linear behavior with the target variable based on the pair plot. Thus this variable (kept even though VIF was higher due to business understanding), comes as an important variable.
2. weathersit3 - This variable refers to light snow, light rain, thunderstorm and has a coefficient value of -0.29 indicating that if this feature value becomes 1, then the rental bike count for that day will decrease
3. season_spring - As discussed in the first questions, the spring season led to lesser demand for rental bikes and thus have a coefficient of -0.14

Apart from this Year comes as an important variable due to the reason that the demand shows an uptrend with passage of years

---

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 as the name suggests, is a statistical approach to understand and identify the linear relationship between a target variable and one or more independent variables. The types of Linear Regression are: a. Simple LR, b. Multiple LR, c. Polynomial LR. T**he idea behind Linear Regression is to find the Best Fit Line, which means finding a line (in terms of Simple Linear Regression) or finding a hyperplane (in terms of Multiple Linear Regression).** The idea is to find a line/ hyperplane which makes the least/minimum error from the data points i.e. **the hyperplane that goes by the closest from the data points with the least error on an assumed to be linear data.** In mathematical terms, the whole idea is to find the value of slope (coefficients) and the intercept which will give the best fit line or the hyperplane. We have to define the loss function which here is the squared difference between actual value and the predicted value and then we apply partial derivatives to calculate the value of slope and the intercepts. However, there is another formula using Gradient Descent as well which is a non-closed form solution. This is computationally less expensive and starts with random values (initialized) for slope and intercepts and then takes gradual steps in the right direction towards finding the minima of the error function.

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 consists of four datasets, which are nearly identical in terms of the simple descriptive statistics such as mean etc. However, these are very different distribution and appear very different when plotted in a graph.
This was brought up to demonstrate the importance of graphing a data while analyzing it and also how outliers can impact or influence the basic descriptive statistics.
Key Pointers:

1. Visualization is essential and summary statistics alone might not provide a complete understanding of the data
2. Statistics sometimes can be misleading, for example if we have three data points and two columns (-5, 0, 5) and (-10, 0, 10), the mean comes to be zero but the distribution is not similar in terms of spread. Similarly, there will be datasets with similar variance or basic descriptive statistics but the distribution may be very different
3. Outliers can disproportionately impact statistical measures

---

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's Correlation Coefficient measures how two variables are linearly related. It quantifies the strength and direction of the linear relationship between two continuous

variables.
The beauty of this statistical measure is that it quantifies the value between -1 to +1 thereby giving us an understanding of how highly two variables are correlated irrespective of scale of the variable.
If the value is + 1, it indicates perfectly linear relationship
If the value is 0, there is no linear relationship
If the value is - 1, then it indicates perfectly opposite (negative) 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 may be defined as a process to change or adjust the range of independent variables to a desired range from its original range without losing on information from the data.
Scaling is performed due to the following reasons:

1. When dealing with Machine Learning models like Linear Regression, it focuses on giving weights to the variables; now, two very different scales for independent features may hamper the model prediction i.e. it prevents features with larger ranges from disproportionately influencing the model
2. The interpretation of the coefficients becomes difficult
3. Algorithms like gradient descent etc will converge faster when scaled as the calculation and gradient movement from all directions will become easier.
4. For models which are based on distance calculation such as K-Means or K Nearest Neighbors, which are backed by distance calculation, it ensures all features are equally important when calculating the distance

Difference between Normalization and Standardization:

Sl No	Normalization (Min Max Scaling)	Standardization (Z-score scaling)
1	It scales the data to a fixed range i.e. 0 and 1	It scales the data to have a mean of 0 and a standard deviation of 1. However, the distribution is not always within a fixed range
2	X' = (X-Xmin)/(Xmax-Xmin)	X' = (X-mu)/stdev

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>
The equation for VIF is as follows:

$$VIF = 1 / (1-R^2)$$

Here, R^2 indicates the variance of the one variable as explained by the other variable.
From the equation, VIF can be infinity only when R squared becomes 1, i.e. one variable fully explains the variance in another variable. This can only happen if we have the same feature twice or if there is perfect multicollinearity between variables.
One example from the dataset, can be temp and atemp, which are very highly correlated with correlation coefficient of 0.99. If these two are used to calculate VIF then, VIF value will come to be almost close to infinity.

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 or Quantile Quantile Plot is used to compare the distribution of a dataset with a theoretical distribution. The idea is that if the Q-Q plot is approximately along a straight line, the dataset matches the theoretical distribution.
Here, the X-axis shows theoretical quantiles from the referencing distribution and Y-axis refers to sample quantiles from the dataset
Use and Importance of Q-Q Plot:
This particular plot helps in identifying if a distribution matches a theoretical distribution if the data is coming to be a straight line.
This is leveraged in checking one of the assumptions of linear regression which states that the residuals should be normally distributed.
Here, the underlying theoretical distribution taken is normal and if the residuals fall in a straight line (approx), then we can say that the residuals are normally distributed.
