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)

We have 7 Categorical Variables namely Season, year, Month, Holiday, Weekday, WorkingDay and Weathersit. I have used bar plots to understand the patterns in them.

The inferences for each categorical variable is as follows.

1. **Season**: **Fall** has the highest demand for bikes followed by **Summer** and **Winter** which leaves **spring** to have the least demand.
2. **Year**: We have the stats of year 2018 and 2019 and there is a rise in demand from 2018 to 2019.
3. **Month**: The months June, July, August, September have the highest demand for bikes. December has the lowest bike rental. January, February and December has less demand for bikes.
4. **Holiday**: The holiday days have more demand compared to non-holiday days. Could be because people having time available for biking.
5. **Weekday**: Fridays and Saturdays have a very little spike in demand, if not all the days have similar demand.
6. **Weathersit**: Bike demand as expected is good in clear and manageable weather when compared harsh weather like snow, rain etc. These play a important role in determining the demand of the bikes.

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** because

1. **Multicollinearity**:
   When you make dummy variables, each category becomes a separate column, so having all of them can create overlap. This overlap messes with the model's understanding of each variable's individual impact. Dropping the first category helps avoid this issue, keeping things clearer for the model.
2. **Reducing complexity**:
   Here we're dropping one category to act as a baseline. This removes extra info that doesn't actually add value, so the model doesn't affect by unnecessary columns. It keeps everything clean and straightforward.
3. **Model Interpretation**:
   By dropping one category, the model can compare the remaining categories to that baseline, so the coefficients in the output are easy to interpret. Each coefficient just shows how much each category differs from the baseline, making the results more intuitive.

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)
Based on the heat map, **temp** and **atemp** has a 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)
I have validated using

1. **Normality of Residual errors**:
   I plotted a histogram of the residuals and they are normally distributed which is represented by the bell curve.
2. **Multicollinearity**:
   I calculated the VIF scores. I aimed for VIF score being less than 5 which means collinearity would not be an issue
3. **Homoscedasticity**:
   This basically a check for constant variance. For this I plotted a graph between predicted values and residuals and there wasn't any constant pattern, so they have constant variance.

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 3 features contributing significantly towards explaining the demand of the shared bikes are

1. **temp (Temperature)**: Higher the temperature, higher is the demand. Hence directly proportional.
2. **Yr (Year)**: The year 2019 is a strong predictor, it indicates that demand over the years is increasing
3. **Light snowy Rain Weather**: Adverse weather conditions effect the demand for bikes negatively.

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 an algorithm used for predicting a continuous target variable based on one or more input features. The main concept is to find out the best fitting line that represents the relationship between the input variable/variables and output variable.

1. **Equation**

Here we use the equation

Y = mx+c

Where,

Y is the target variable

X is the feature variable(input)

m is the slope (gives relation between x and y)

c is the intercept

for multiple features

y=b0+b1x1+b2x2+...+bnxn

where,

each b is coefficient showing the weights of each x.

2. **Best Fit**

Here the aim is to find the values of the coefficients, basically the values of m, c or all the b values so that line fits as closely as possible. This is done by reducing the residuals. It is least squares technique to minimize the residuals. Residual is basically the difference between actual and predicted values.

3.**Cost Function**

It is used measure how well the line is fitting. Here MSE (Mena squared method) is used. It calculates the average of the squared differences between actual and predicted values. MSE and model fit are inversely proportional. Lower the MSE, better the model fits.

4.**Making Predictions**

Once the best fit is determined, we can use it to make predictions on new data by providing the values of features(x values) into the line equation. The y value(output) will be the predicted value

---

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 this set of four small datasets created by a statistician named Francis Anscombe in 1973. The point of these four datasets is to show that just looking at summary statistics (like mean, variance, correlation) can totally mislead you about what's actually happening with the data.

1. **Same Stats, Different Stories**

Each of the four datasets in Anscombe's quartet has nearly identical summary statistics—same mean, same variance, same correlation between xxx and yyy, and the same linear regression line. But here's the thing: if you actually plot these datasets on a graph, they look completely different from each other. They tell totally different "stories."

2. **Why Looking at Graphs is Key**

The big takeaway from Anscombe's quartet is that you can't rely only on numbers to understand data. You really need to visualize it, like with scatter plots or other charts, to get a true picture. For example, in one of the datasets, the points are arranged in a perfect straight line, while in another, they're a curve. Another set has an outlier that throws off the line entirely. So yeah, numbers alone can be super misleading.

3. **What This Means for Data Analysis**

Anscombe's quartet is a reminder that visualizing data can reveal patterns or problems that summary stats won't show. It basically teaches us that data analysis isn't just about numbers; it's about context. Without graphs, you'd miss out on all the insights these datasets are really showing.

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, or Pearson's correlation coefficient, is a measure that tells you how strong the relationship is between two variables. It gives you a number between -1 and 1. If you get 1, it means there's a perfect positive relationship (when one goes up, the other goes up too). If it's -1, then there's a perfect negative relationship (one goes up, the other goes down). And if it's around 0, that means there's pretty much no relationship at all.

Advantages

Pearson's R is super helpful because it quantifies the relationship, so we do not just guess by looking at the graph. It actually gives number to it, which makes comparisons easier.

Disadvantages

It only captures linear relationships. If the variables are related in a curvy or non-linear way, Pearson's R might not show that. Hence, it's a good measure, but it's not perfect for every situation.

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 basically adjusting the values of your data so they fit within a specific range or have a certain distribution. It's done to make sure that different features in a dataset are on the same scale, especially when they are of different units (like height in cm and income in dollars). Models can get confused if some features have way bigger numbers than others.

We scale because certain machine learning algorithms are sensitive to the scale of the data. For instance, algorithms like k-nearest neighbors, support vector machines, and neural networks perform better when features are similar in range. If they're not, features with larger scales can end up dominating, even when they're not necessarily more important.

- **Normalized Scaling**: Normalization adjusts values to a specific range, usually between 0 and 1. It's great when you want all features to be directly comparable within a specific interval. It's often used when you have data that doesn't necessarily follow a bell curve.

- **Standardized Scaling**: Standardization, on the other hand, adjusts data so it has a mean of 0 and a standard deviation of 1. It's useful when data is normally distributed (bell curve) or when you're working with algorithms that assume a normal distribution.

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>

When we see an infinite VIF (Variance Inflation Factor), it usually means there's perfect multicollinearity in your data. In simpler terms, one of your predictor variables can be perfectly predicted by a combination of other predictor variables.

VIF measures how much a variable's variance is inflated due to multicollinearity. If one variable is an exact linear combination of others, it throws VIF to infinity because the model can't separate out its unique impact. It's like the model's way of telling , "I don't know what this variable adds because it's basically the same as the others."

When you hit an infinite VIF, it's a sign to double-check for redundant variables. You might need to drop one of the predictors or combine similar ones so the model can better understand the independent impact of each feature.

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 (Quantile-Quantile) plot is a graph that helps you check if your data follows a particular distribution, usually a normal distribution. It's basically a plot where the quantiles of your data are compared to the quantiles of a theoretical distribution. If the points on the plot form a straight line, your data is pretty close to that distribution (like normal).

In linear regression, we assume that the residuals (errors) are normally distributed. A Q-Q plot helps you visually check this assumption. If your residuals are normally distributed, you should see a roughly straight line in the Q-Q plot. If they curve or form some other pattern, it suggests the residuals might not be normal.

This matters because if the residuals aren't normal, it can mess with the validity of your regression results, especially things like confidence intervals and hypothesis tests. A Q-Q plot is a quick, visual way to check this and make sure the model's assumptions hold up.
