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 following categorical variables:

1. season
2. yr (year)
3. month
4. holiday
5. weekday
6. weekend
7. weathersit

The dependent variable is cnt(count).

Upon doing the analysis using boxplot we can see

1. Fall season has higher count(cnt) followed by summer season.
2. There is a significant rise in count(cnt) from 2018 (0) to 1 (2019).
3. count(cnt) increases from jan till oct and then we see a drop in nov and dec months.
4. days doesn't seems to affect the cnt much (though sat and wed looks to be having higher cnt)
5. When wheather is clear then count(cnt) is higher and when its Light_Snow its drastically low.
6. Also, we don't see any outliers

Boxplot from https://github.com/kc11381/Bike_Sharing_Case_Study/blob/main/bike_sharing_predictors.ipynb

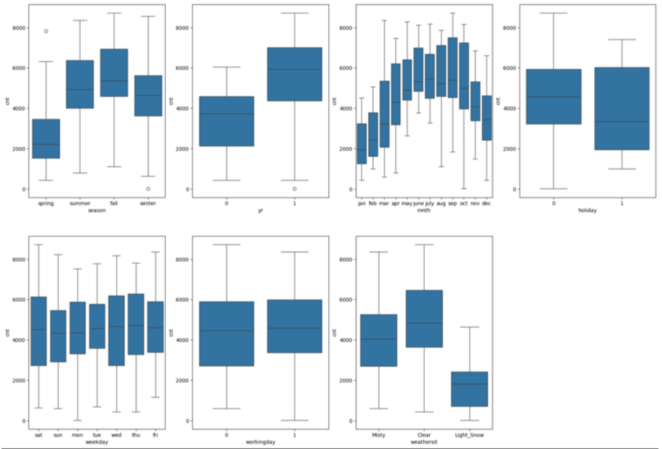


Fig. Boxplot for Cayegorical variables

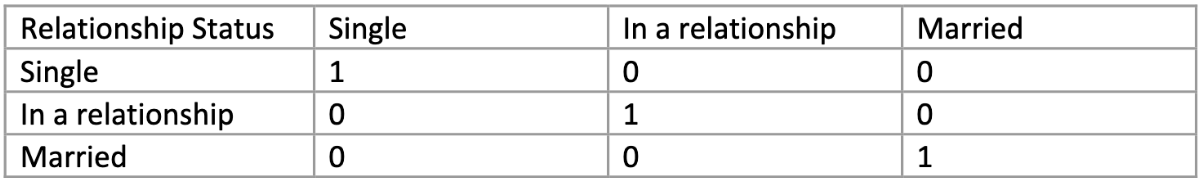
Question 2. Why is it important to use drop_first=True during dummy variable creation? (Do not edit)

Total Marks: 2 marks (Do not edit)

Answer: <Your answer for Question 2 goes below this line> (Do not edit)

We use drop_first=True to drop unnecessary columns because we just need p-1 levels for p levels of a categorical variable.

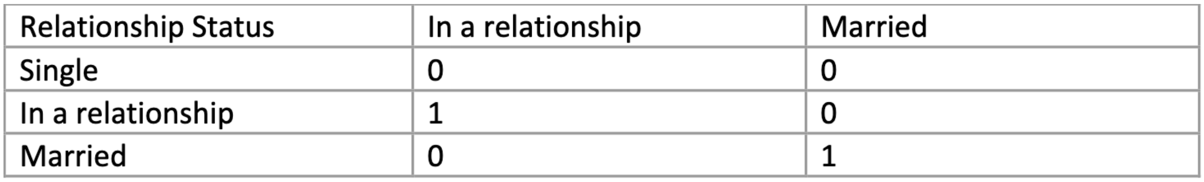
Example- For a variable say, 'Relationship' with three levels namely, 'Single', 'In relationship', and 'Married', you would create a dummy table like the following:



Relationship Status	Single	In a relationship	Married
Single	1	0	0
In a relationship	0	1	0
Married	0	0	1

But you can clearly see that there is no need of defining three different levels. If you drop a level, say ‘single’, you would still be able to explain the three levels.

Let's drop the dummy variable 'Single' from the columns and see what the table looks like:



Relationship Status	In a relationship	Married
Single	0	0
In a relationship	1	0
Married	0	1

If both the dummy variables namely 'In a relationship' and 'Married' are equal to zero, that means that the person is single. If 'In a relationship' in 1 and 'Married' is 0, that means the person is in a relationship and finally, if a 'In a relationship' is 0 and 'Married' is 1, that means that the person is married.

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 and atemp has the highest co-relation with the target variable(cnt) by looking at the pairplot. From correlation heatmap value is 0.65

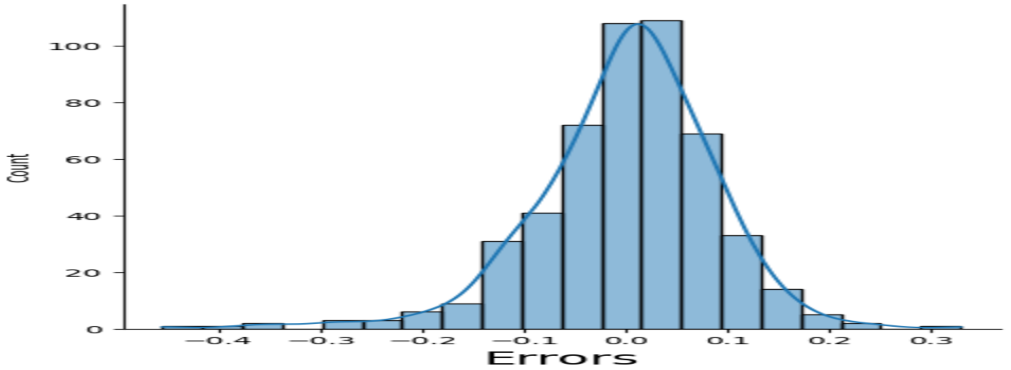
Question 4. How did you validate the assumptions of Linear Regression after building the model on the training set? (Do not edit)

Total Marks: 3 marks (Do not edit)

Answer: <Your answer for Question 4 goes below this line> (Do not edit)

We had to check 4 assumptions of the linear regression. They are tested as below

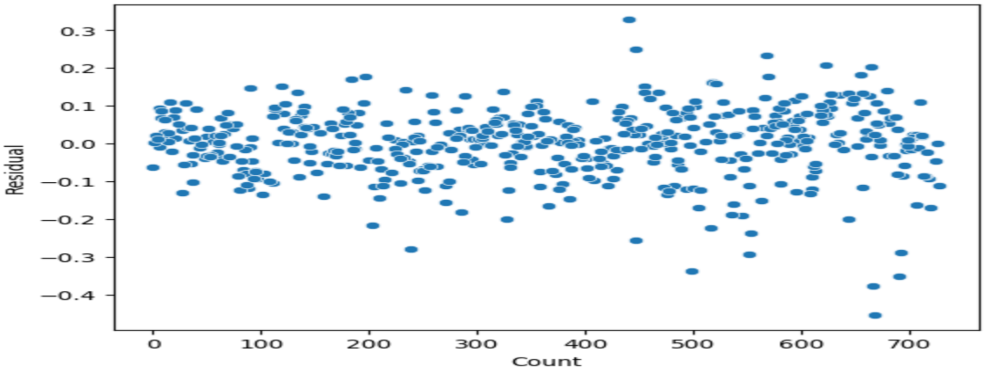
a. Histogram of the error terms and it came out to be a Normal distribution



b. VIF of the features of the final model < 5.0 (except temp as dropping that drastically decreased R-squared value)

c. Multicollinearity - Variables like temp is linearly related to cnt

d. Homoscedasticity – The variance of the residual is almost 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)

Coefficients of the final model are below

yr	**0.2348**
workingday	**0.0547**
temp	**0.4354**
windspeed	**-0.1609**
spring	**-0.0713**
summer	**0.0354**

winter	**0.0903**
dec	**-0.0467**
jan	**-0.0526**
july	**-0.0466**
nov	**-0.0447**
sep	**0.0652**
sat	**0.0670**
Light_Snow	**-0.2969**
Misty	**-0.0818**

So the top 3 features are:
a. temp
b. workingday
c. yr

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>

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>

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>

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>

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>

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>
