Assignment-based Subjective Questions

1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

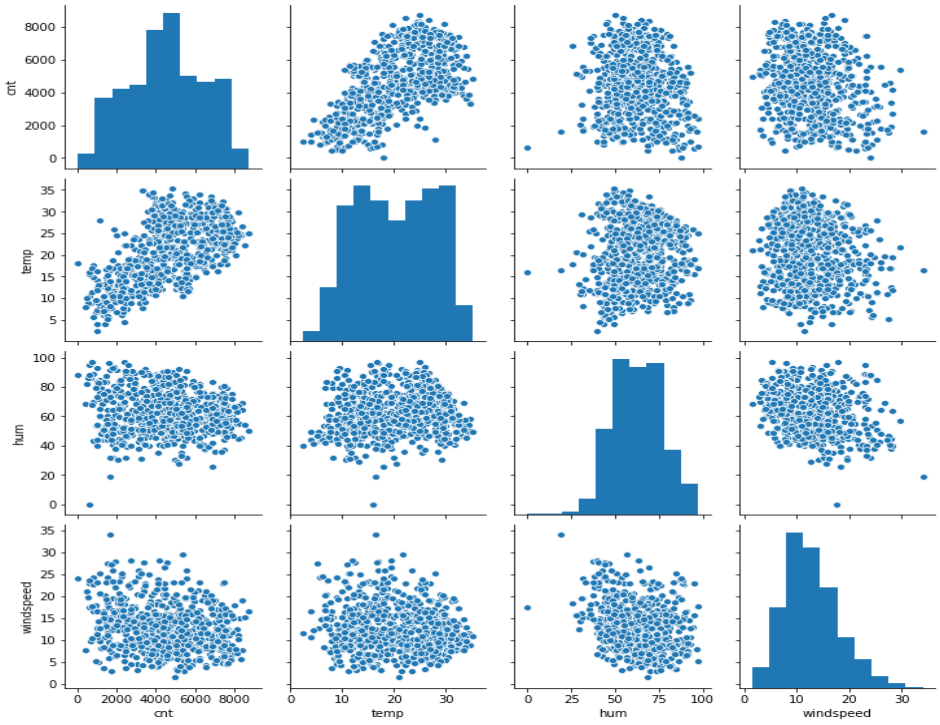
Ans 1. The categorical variables are seasons, weathersit, holiday, mnth, yr, weekday.

- **Spring has the lowest median, 25th and 75th quartile in comparison to other seasons, when it comes to total count of bikes.**
- **When it comes to month January has the lowest median, 25th and 75th quintile while july has the highest median, and 75th quartile.**
- **There were less rentals during holidays.**
- **There were less rentals during light show and rain which seems reasonable.**
- **2019 had more rentals than 2018.**

2. Why is it important to use drop_first=True during dummy variable creation? (2 mark)

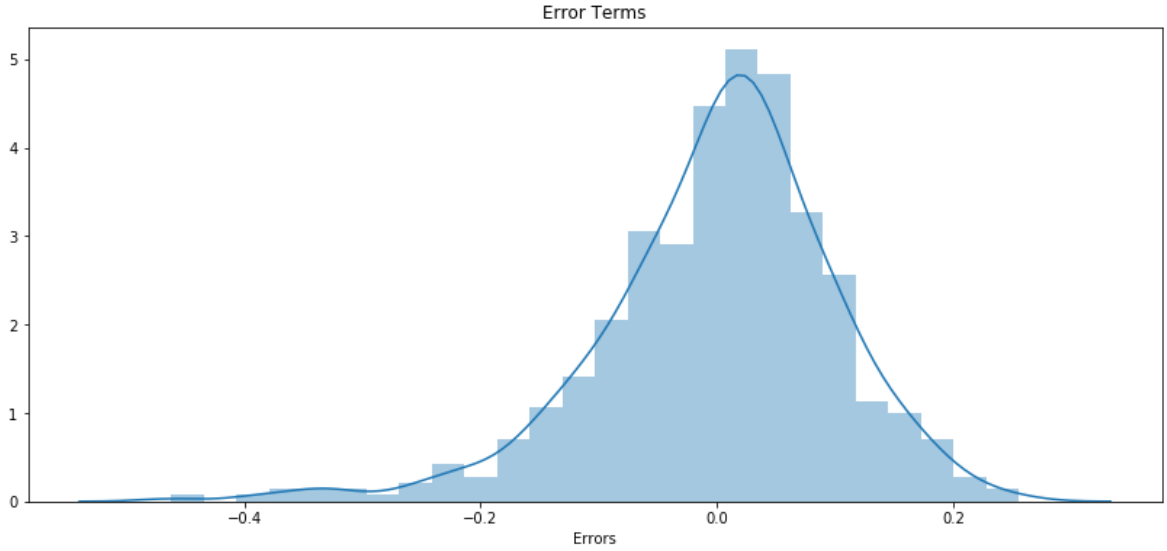
- **It reduces correlations among dummy variables.**
- **And multicolinearity among the dummy variables.**
- **It also reduces one column , which also means that if all things in that category are 0 then the dropped item is 1.**

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)



Variables Temp and atemp have highest correlation with cnt variable.

4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3marks)



- **Using residual sum of squares, which Is the difference between true and predicted value from training. Which tells us if errors are distributed around 0 or not.**
- **Using VIF(variance inflation factor) i.e if VIF are less than 5, then model meets the assumptions.(HETEROSKIDASTICITY), thus making sure that data does not have a lot of variance.**

5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)

- **(year *0.2403)***
- ***(temp* 0.4380)**
- ***(Light Snow & Rain * -0.2027)***

General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)
   -
2. Explain the Anscombe's quartet in detail. (3 marks)
3. What is Pearson's R? (3 marks)
4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)
5. You might have observed that sometimes the value of VIF is infinite. Why does this happen?

(3 marks) 6.

What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression.

(3 marks)