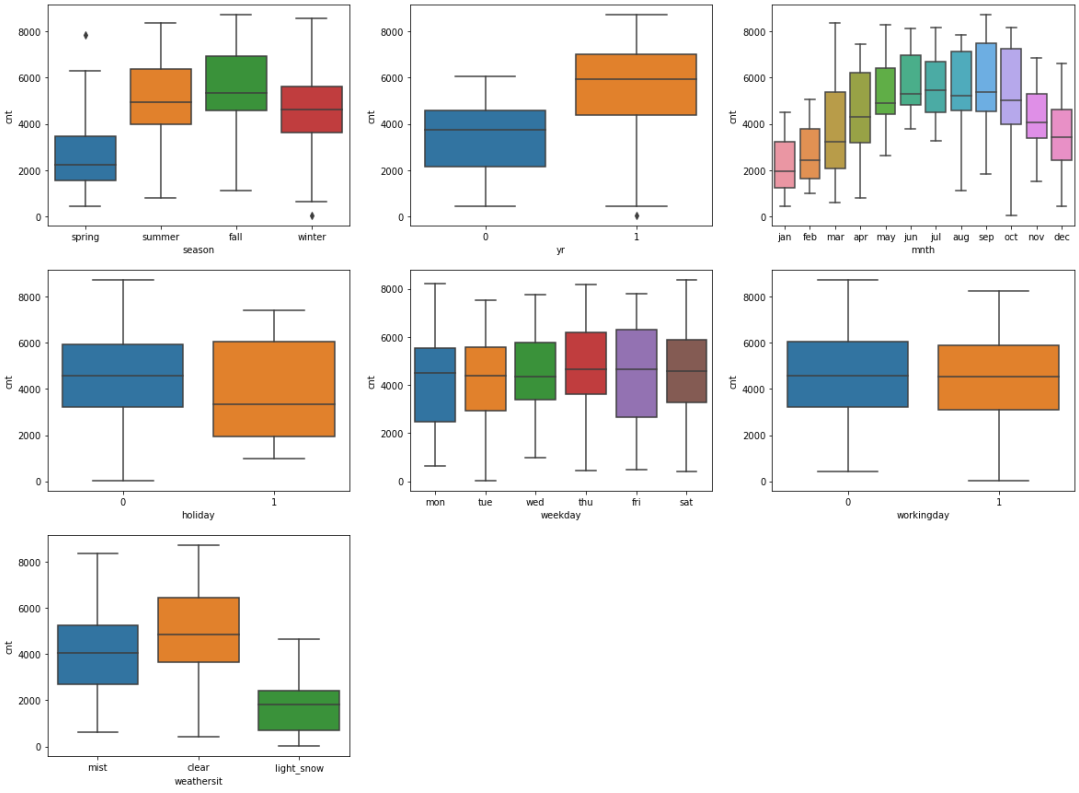
Assignment-based Subjective Questions

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

   Solution: Different categorical variables and their effect on dependent variables are as follows:

   - **season -** The demand for bike service is low in spring and high in other seasons.
   - **yr** - This is year variable (2018 is encoded as 0 and 2019 is encoded as 1). The demand for bike service is higher in 2019 than 2018.
   - **mnth** - This is the month variable. August and September have the highest and lowest on January and February.
   - **holiday** - The data showed that during non-holidays bike service demand is high.
   - **weekday -** Rentals are highest on Thursdays and Saturdays.
   - **workingday** - People mostly use bike services on working days than non-working days.
   - **weathersit** - This is the variable for weather situation. The demand for bike service is highest when the weather is clear or partly cloudy.

Categorical variables vs target variable

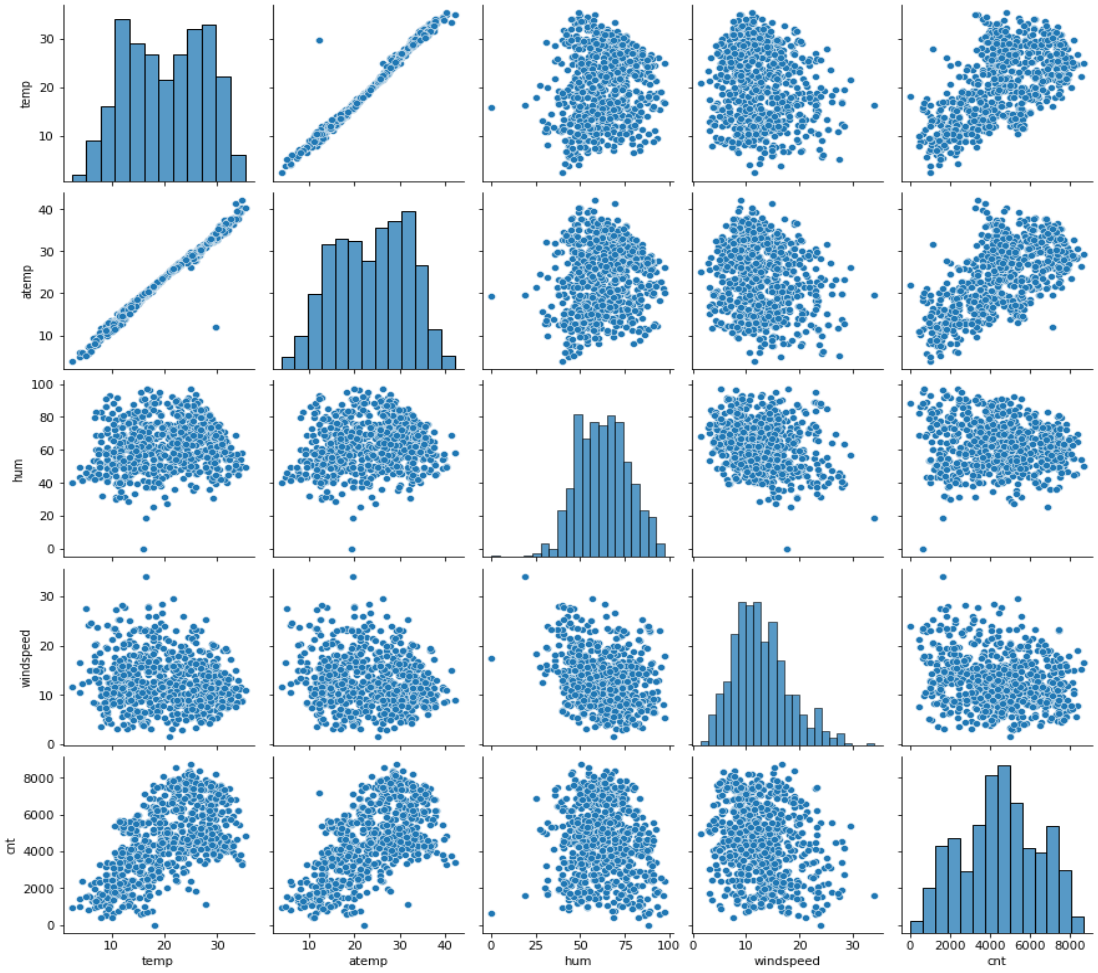


2. Why is it important to use **drop_first=True** during dummy variable creation?

Solution: It is important to use drop_first = True during variable creation because this helps in reducing correlation among the dummy variables. That variable if not deleted will create redundancy. So even after dropping that variable we will be able to explain all levels.

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

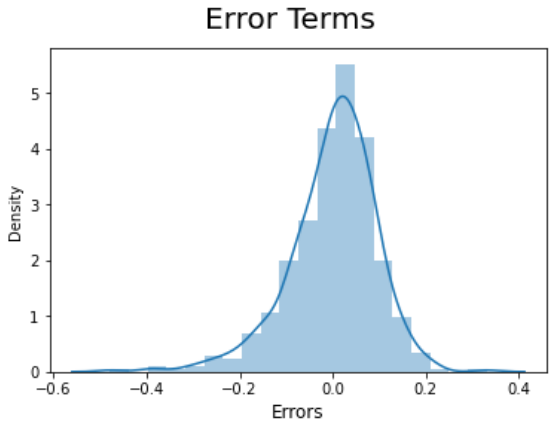
   Solution:



From above pair plot we can infer that **atemp** has the highest correlation with the target variable.

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

Solution: By checking the distribution of error terms, we found that it has normal distribution with mean at 0. The plot below shows the distribution of error terms.



5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?
   Solution: Based on the final model, the top 3 features with their coefficients contributing significantly towards explaining the demand of the shared bikes are as follows:
   - temp 0.3605
   - weathersit_light_snow -0.2859
   - yr 0.2407