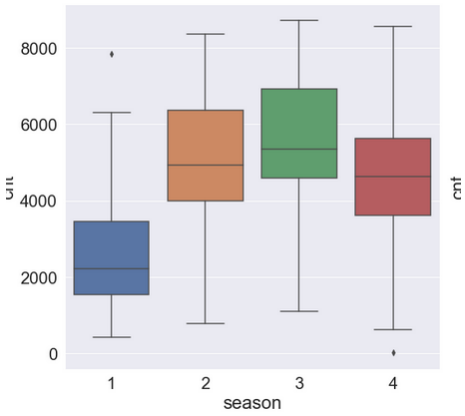
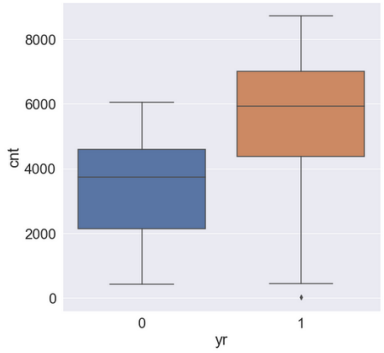
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable?

Answer:

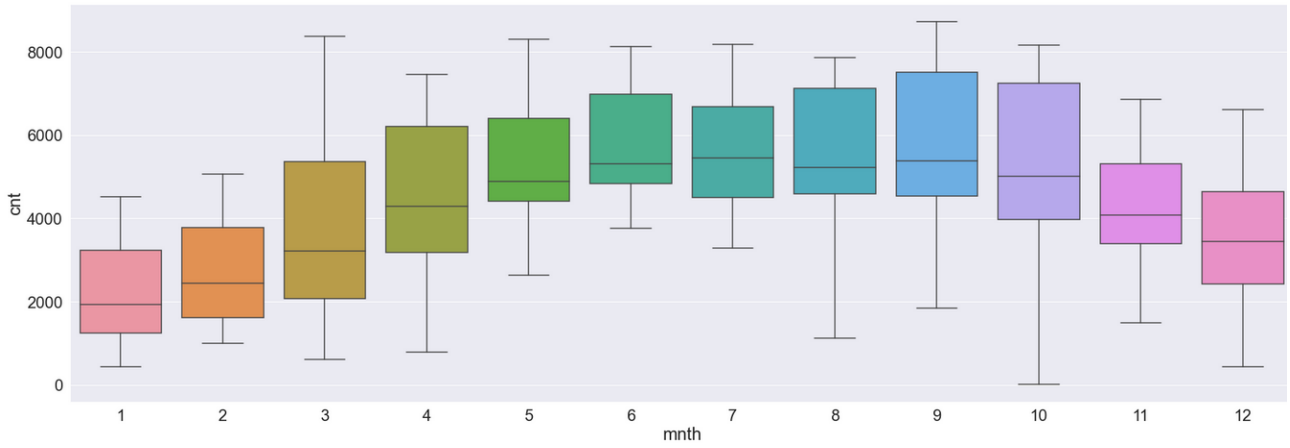
1. season: when season is 3, i.e. fall, more counts of bikes.



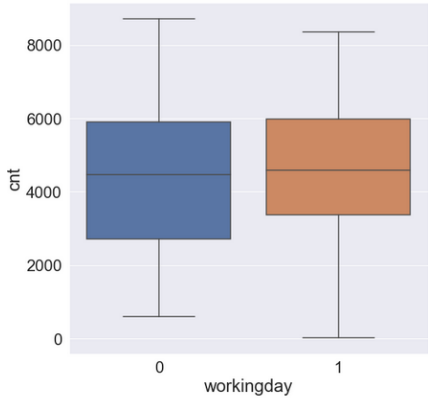
2. year: when year is 1, i.e. 2019, more counts of bikes.



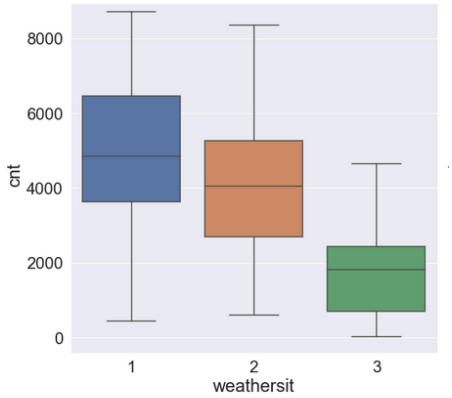
3. month: when month is from 5 to 10, i.e. may to October, more counts of bikes.



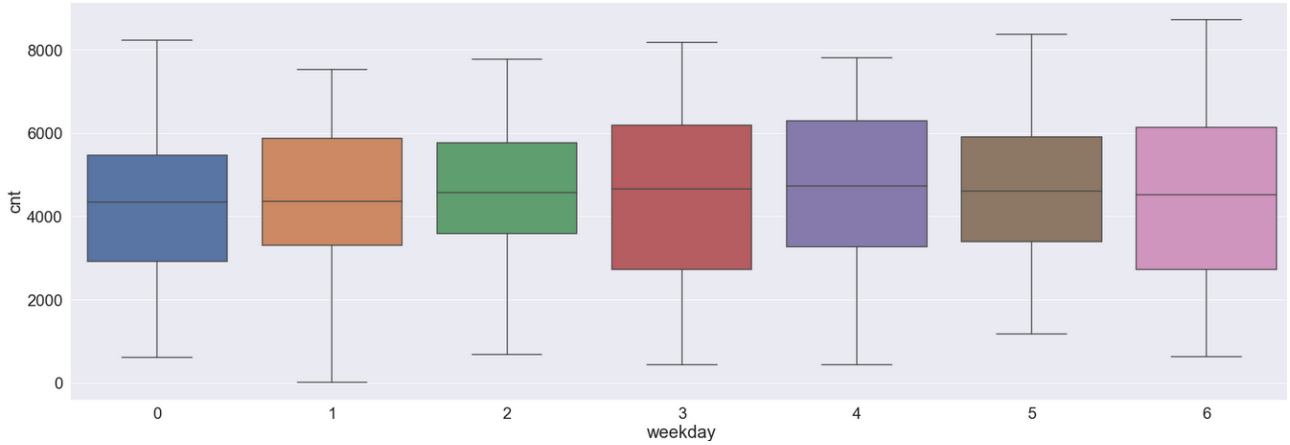
4. working day: no such effects on bike counts.



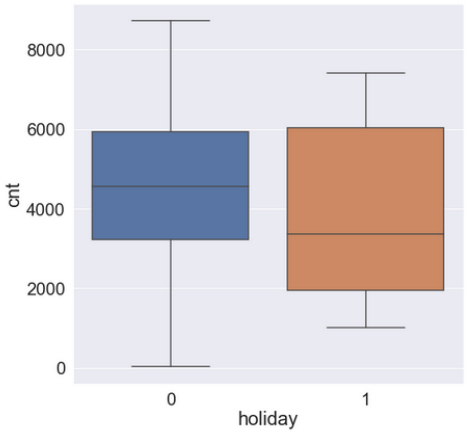
4. weather situation: more bike count at 1, i.e. clear weather.



5. weekday: no such effect on weekday on bike counts.



6. holiday: more counts when there is not a holiday.



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

Answer:
During manual one-hot encoding on a single column, we usually create dummy variables out of all the available discrete categorical values in the columns. During these process, dropping the first column signifies that when all the other dummy columns are 0 in a row, the dropped value is the value in that specific row.
It is important to drop first column to reduce the number of column as wel complexity for our model to learn.

When dummy variables are created, the values given are wither 0 or 1, where 0 represents 'negative' and 1 represents 'positive'.
Thus if a column have values like A,B,C, and D. New dummy columns created are A|B|C|D.
So for the row where column values was A, in the new dummy columns dataframe, values for A,B,C,D will be 1,0,0,0 respectively.
Accordingly for B; C; ; and D, values will be 0,1,0,0 ; 0,0,1,0 : and 0,0,0,1 respectively.
e.g.
season column – fall, winter, spring, summer

	dteday	season	yr	mnth	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	casual	registered	cnt
171	21-06-2018	fall	0	6	0	2	1	2	27.914153	31.88230	77.0417	11.458675	774	4061	4835
79	21-03-2018	summer	0	3	0	1	1	2	17.647835	20.48675	73.7391	19.348461	401	1676	2077
0	01-01-2018	spring	0	1	0	6	0	2	14.110847	18.18125	80.5833	10.749882	331	654	985
265	23-09-2018	winter	0	9	0	5	1	2	24.975847	26.10625	97.2500	5.250569	258	2137	2395

creating dummy columns: 0 & 1 assigned to each column.

	fall	spring	summer	winter
171	1	0	0	0
0	0	1	0	0
79	0	0	1	0
265	0	0	0	1

dropping first column. When 1 is present at each specific column, that value is true for the row.

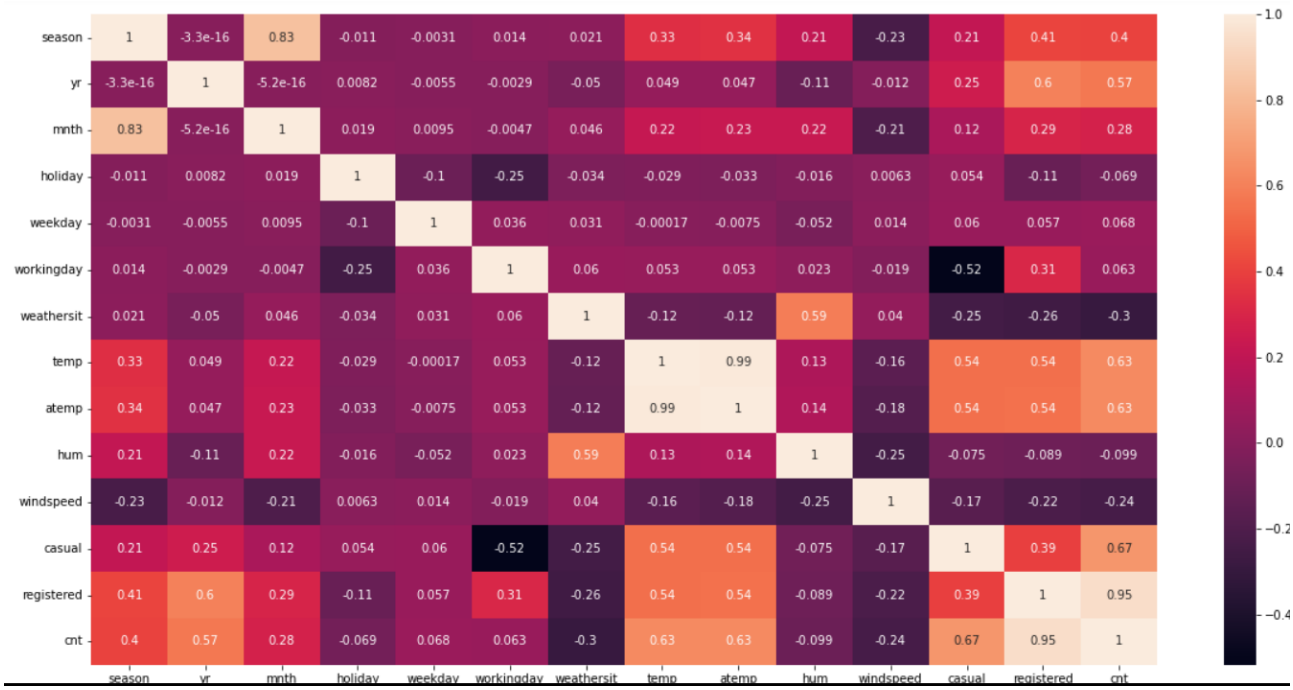
	spring	summer	winter
0	1	0	0
79	0	1	0
265	0	0	1

When all columns are 0, dropped value is true for the row.

	spring	summer	winter
171	0	0	0

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

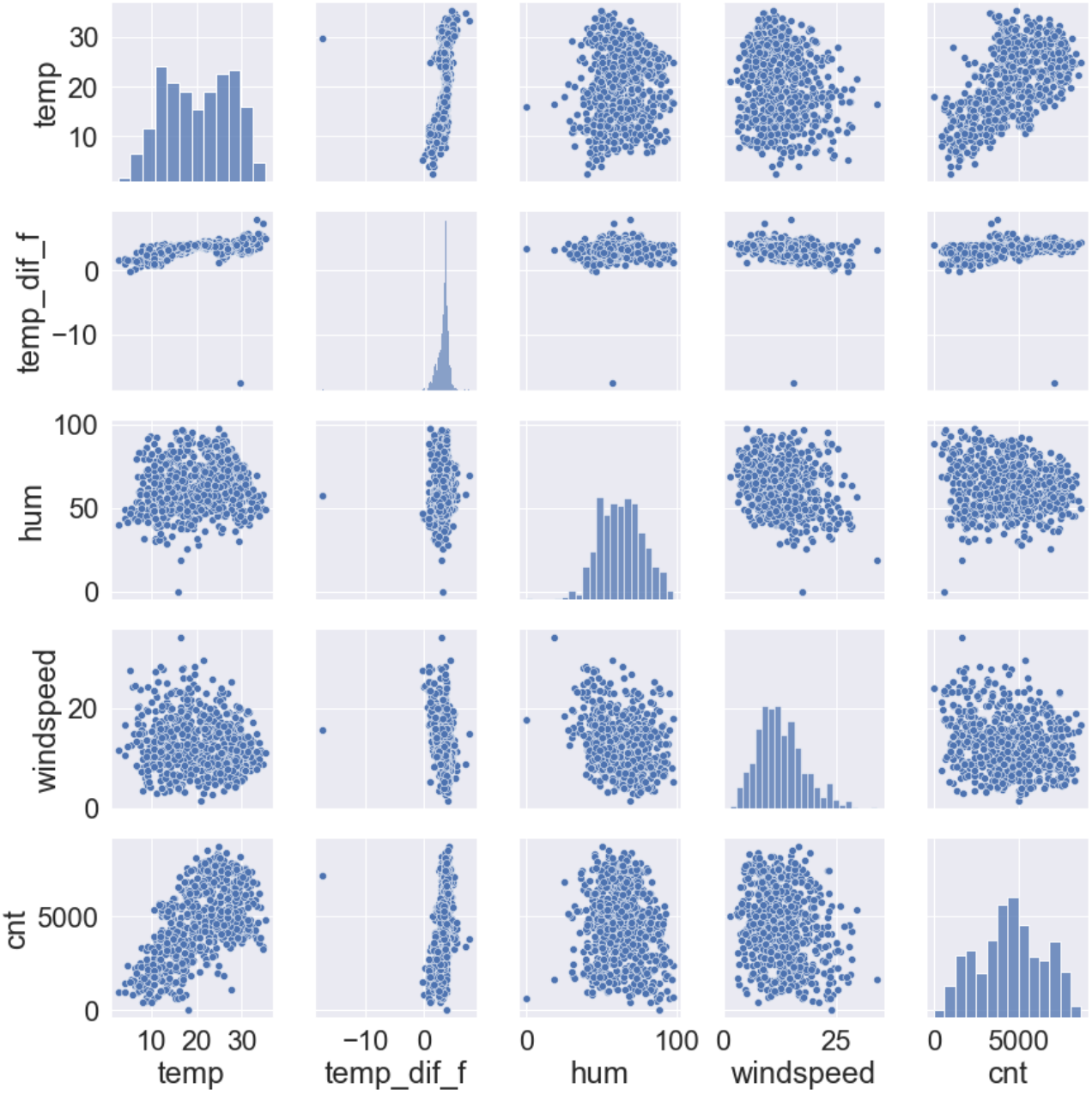
Answer:
column 'atemp' have highest correlation with the target variable 'cnt' with value of (0.6306853489531029).



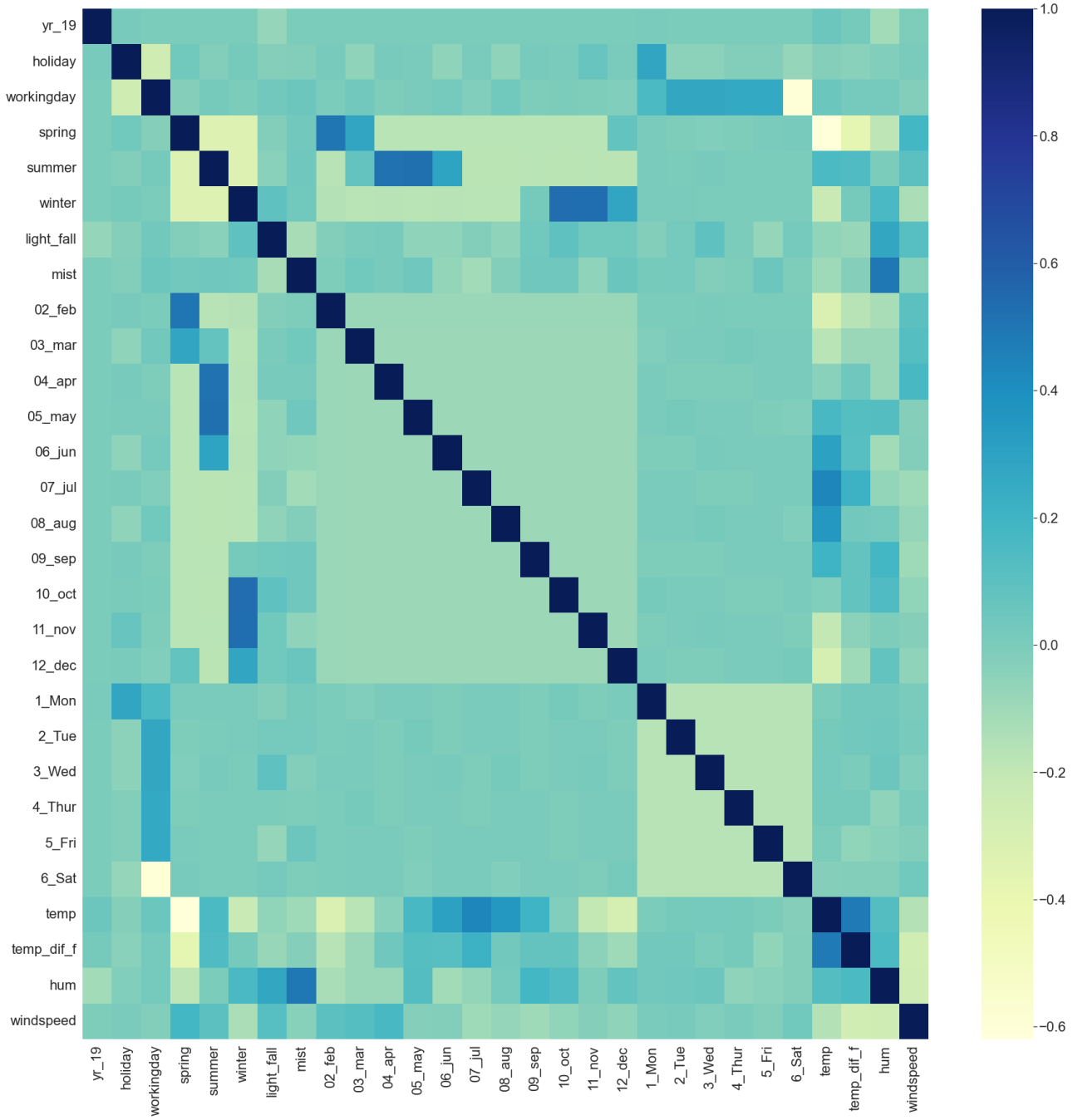
4. How did you validate the assumptions of Linear Regression after building the model on the training set?

Answer:

1. Linear relationship assumption: linear relationship present between target and feature variables(continuous).

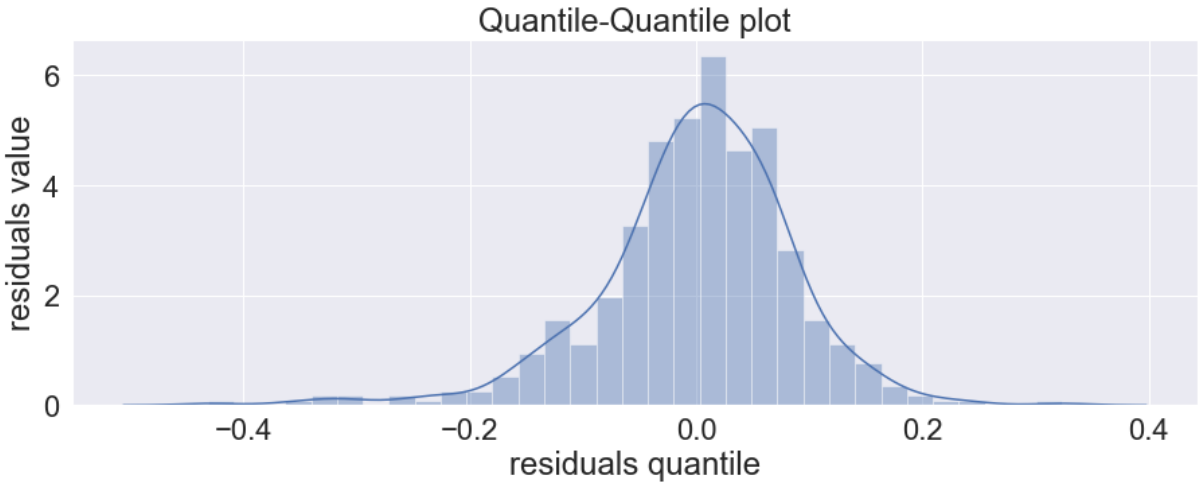


2. no correlation assumption: no collinearity observed between independent feature variables.

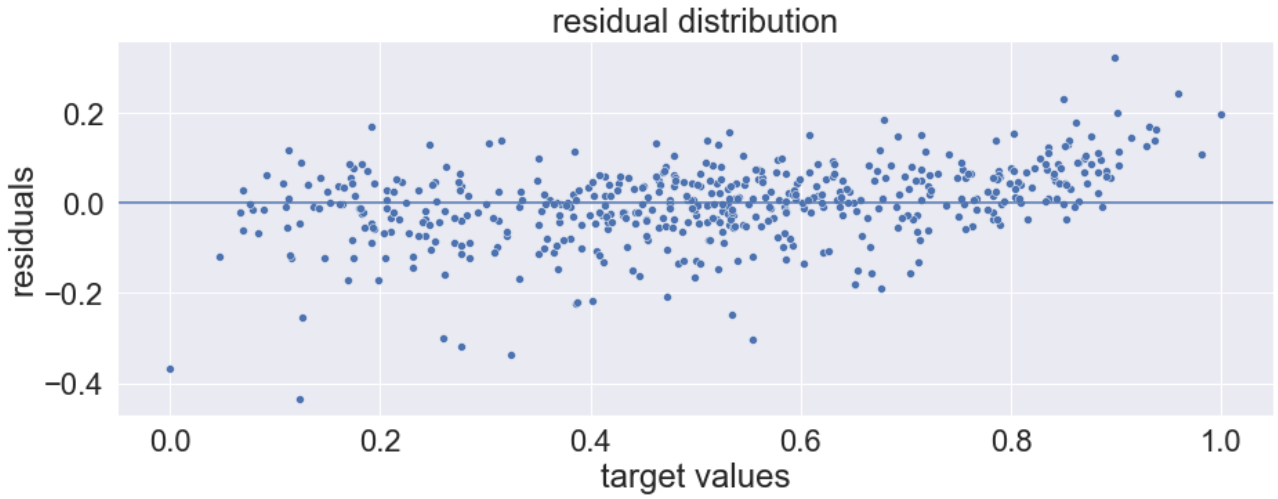


3. error terms

a. normal distribution of error terms and zero mean of errors



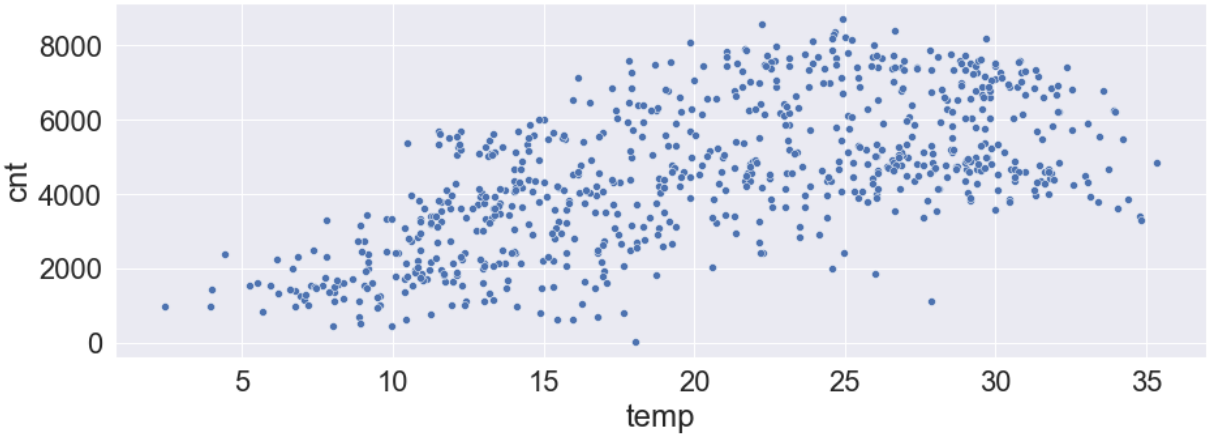
b. residuals are homoscedasticity in nature, i.e. have constant variance.



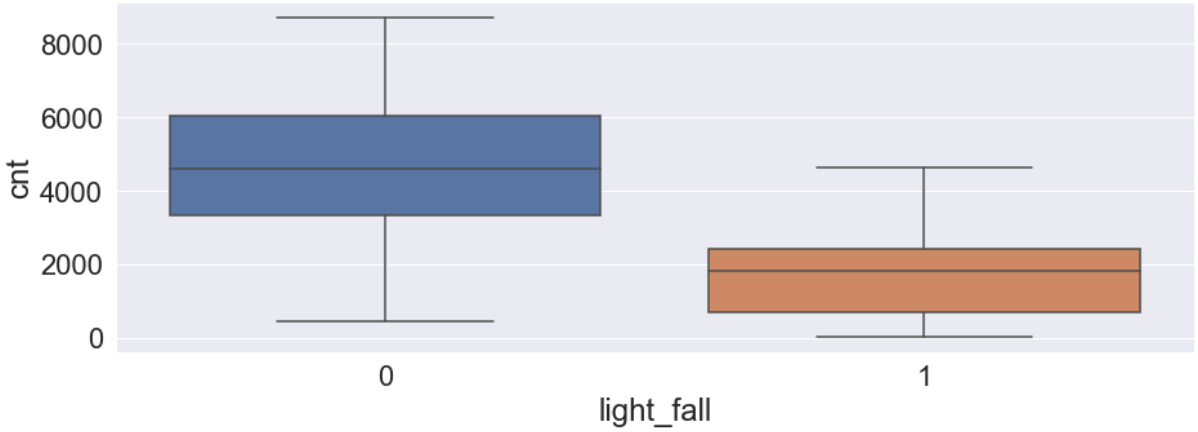
5. Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes?

Answer:
features having very high coefficients absolute value are more contributing to the shared bikes demand.

- temp: temperature: 0.434234
  - bike demand increases with temperature.



- light_fall: light fall weather situation: -0.254504
  - more bike demand when there is no light fall.



- yr_19: year 2019: 0.232232
  - more demand for year 2019