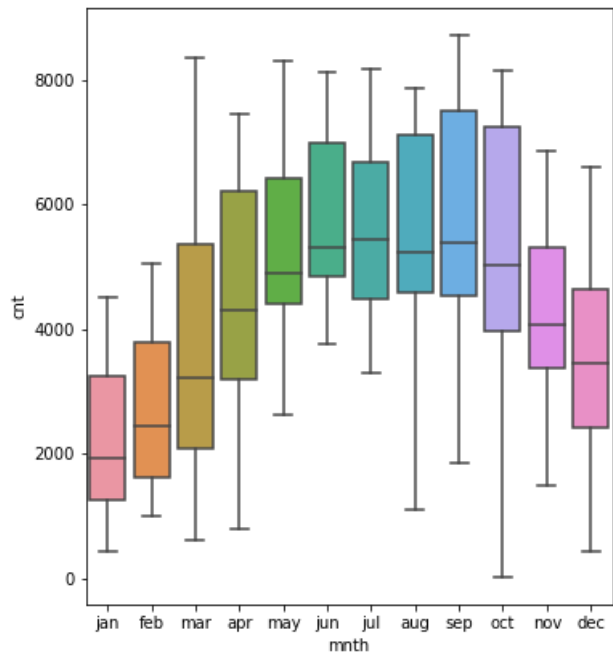
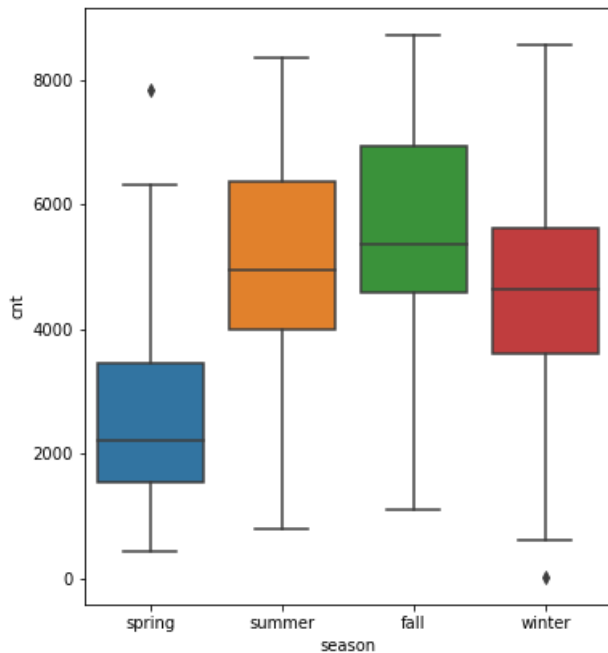


# Assignment-based Subjective Questions

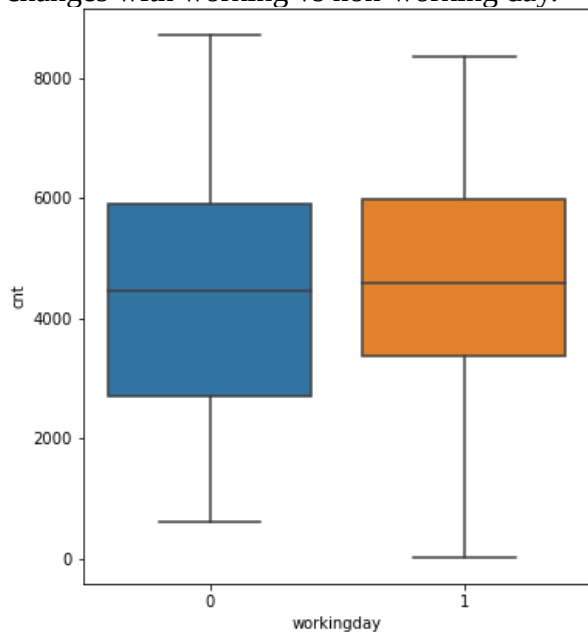
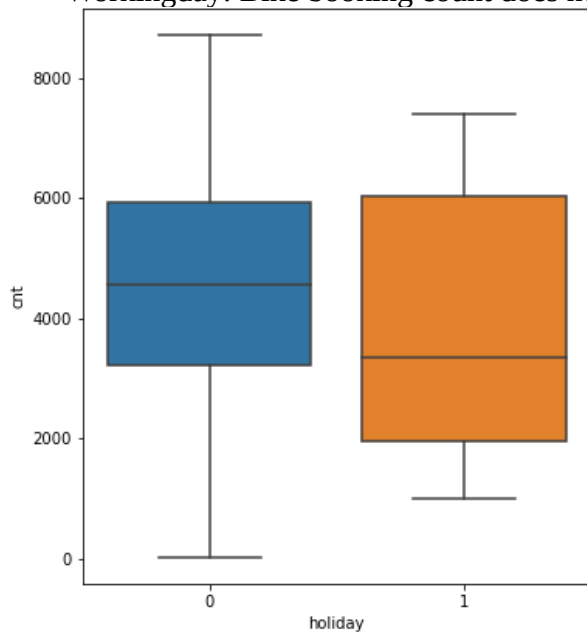
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

Answer: We are having the following categorical features in the dataset:

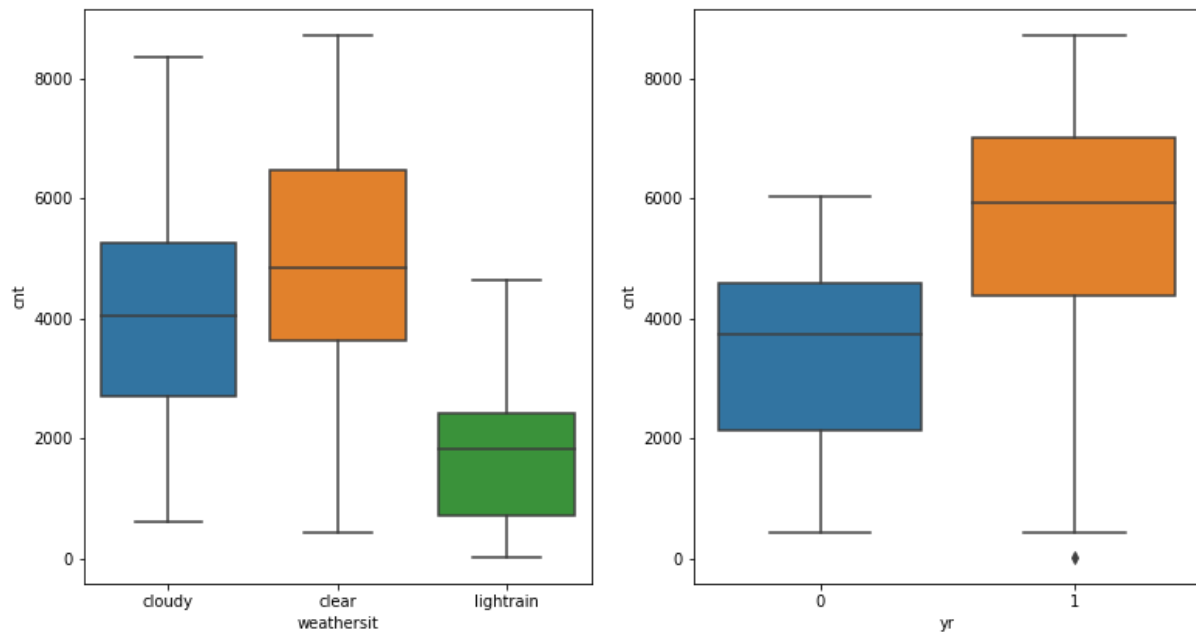
- Season: Counts of rented bikes depends on season, cnt increases in Fall and Summer season and it gets decreases with winter and spring
- mnth: Cnt varies with months as we are having higher booking with May, June, July, August, and September. And booking decreases with Winters month.



- Holiday: Cnt depends on holiday. On non-holiday we are having higher cnt value.
- Workingday: Bike booking count does not changes with working vs non-working day.



- Yr: Count of booking increases with year. We are having higher no of bike renting in 2019 year compared to 2018.



- Weathersit: Weathersit affects the bike booking. Clear weather attracts the customers and have higher booking then cloudy and then light rain.

## 2. Why is it important to use drop\_first=True during dummy variable creation? (2 mark)

Answer: When we are converting the catagorical features to numeric feature. So that these can be scalarized. We break the single column into the all possible probable columns and fill the values between 0 and 1.

For example: We are having the following catagorical feature:

Season
summer
winter
fall
spring

And it will be converted to following table:

summer	Winter	Fall	spring
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1

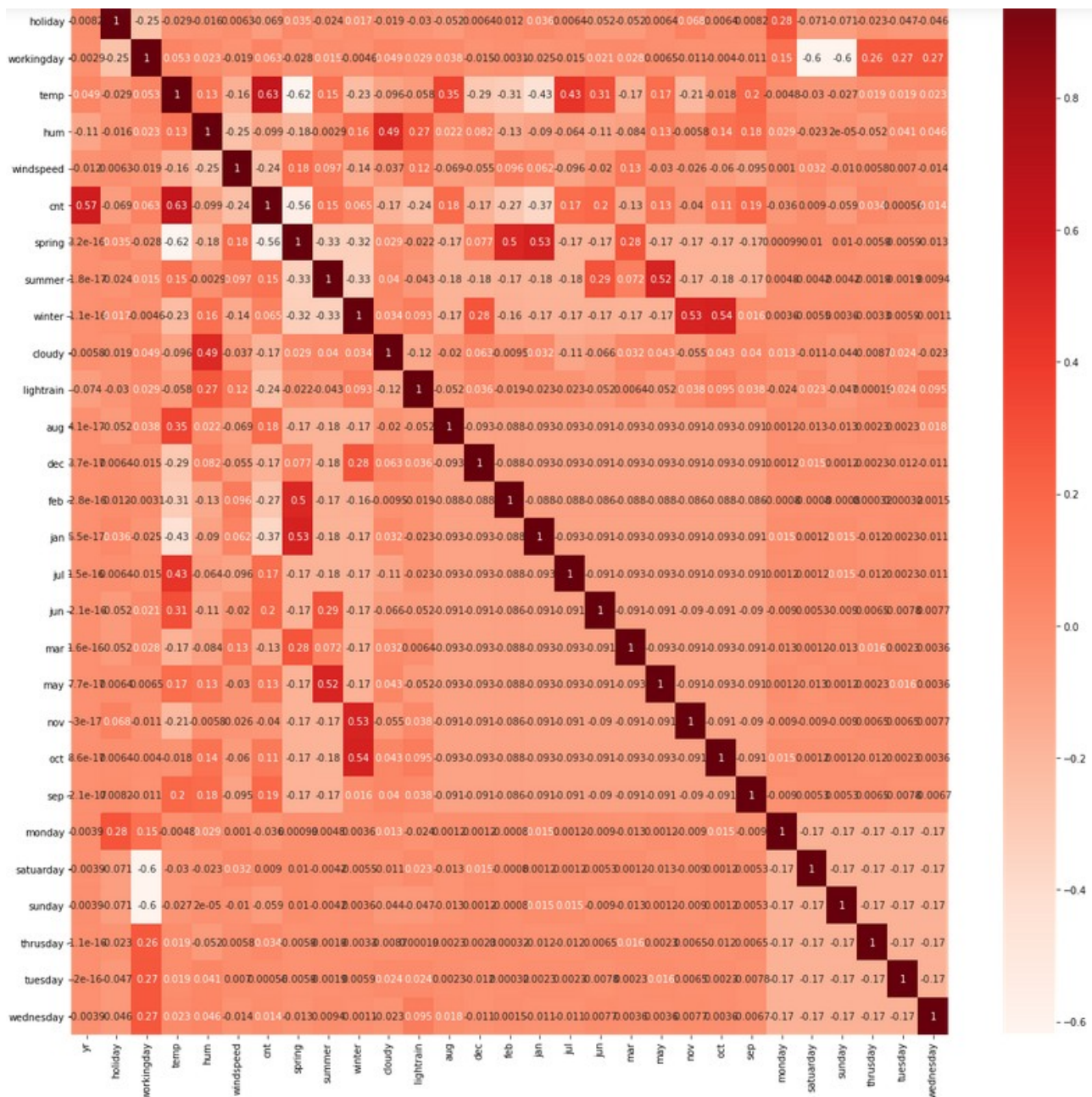
But we can extract the information with with 1 less column like if all the values are zero the we can assume it would be the final option and we can have the following table:

Winter	Fall	spring
0	0	0
1	0	0
0	1	0
0	0	1

As we can see 4 colume are reduced to 3 columns.

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

Answer: temp attribute is having the highest correlation with cnt (target variable) with 0.63 value.



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

**Answer:** We can validate the model on the following scenarios.

- Low p-value --> p-value should be low (less than 0.05 is acceptable)

### OLS Regression Results

```

=====
Dep. Variable:          cnt      R-squared:                0.772
Model:                  OLS      Adj. R-squared:           0.767
Method:                 Least Squares      F-statistic:           187.7
Date:                   Wed, 22 Dec 2021    Prob (F-statistic):     4.31e-154
Time:                   19:38:41           Log-Likelihood:         415.03
No. Observations:       510             AIC:                   -810.1
Df Residuals:           500             BIC:                   -767.7
Df Model:                9
Covariance Type:        nonrobust
=====

```

	coef	std err	t	P> t	[0.025	0.975]
const	0.5723	0.013	43.029	0.000	0.546	0.598
yr	0.2486	0.010	25.831	0.000	0.230	0.268
windspeed	-0.2032	0.029	-6.948	0.000	-0.261	-0.146
spring	-0.2368	0.014	-16.511	0.000	-0.265	-0.209
winter	-0.0546	0.012	-4.589	0.000	-0.078	-0.031
cloudy	-0.0901	0.010	-8.790	0.000	-0.110	-0.070
lightrain	-0.2990	0.029	-10.279	0.000	-0.356	-0.242
jan	-0.1040	0.020	-5.097	0.000	-0.144	-0.064
sep	0.0862	0.018	4.764	0.000	0.051	0.122
sunday	-0.0465	0.014	-3.383	0.001	-0.074	-0.019

```

=====
Omnibus:                47.207      Durbin-Watson:           1.996
Prob(Omnibus):           0.000      Jarque-Bera (JB):        99.456
Skew:                    -0.530      Prob(JB):                2.53e-22
Kurtosis:                 4.886      Cond. No.                 8.67
=====

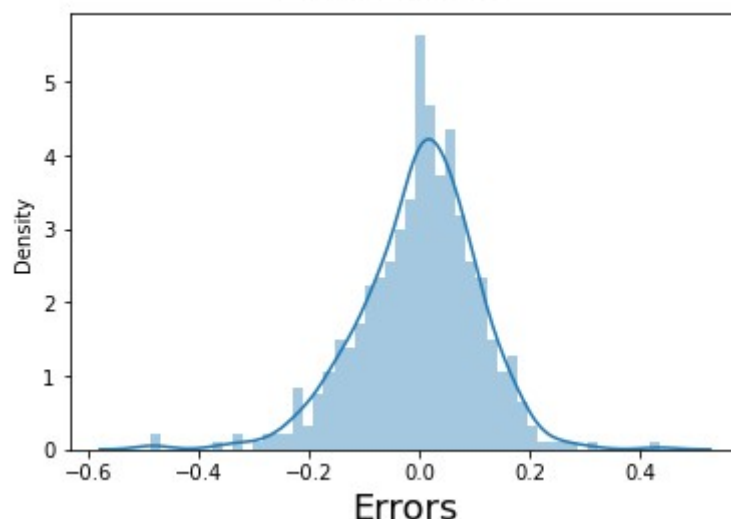
```

- b) Variance inflation factor (VIF) - VIF should be low (less than 5 is acceptable).  
as  $VIF = 1/(1-R\text{-square})$

	Features	VIF
1	windspeed	2.58
2	spring	2.12
0	yr	1.74
6	jan	1.59
4	cloudy	1.45
3	winter	1.37
8	sunday	1.15
7	sep	1.09
5	lightrain	1.08

- c) Error Rate: Normalised Error rate with zero centralized.

### Error Terms



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

**Answer:** Following features contribute the most:

- Light rain
- Yr
- Spring