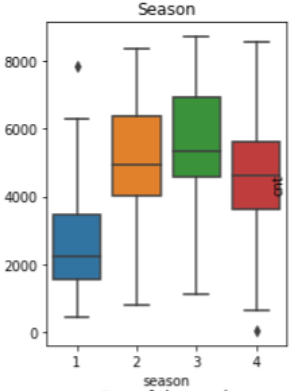
## Assignment-based Subjective Questions

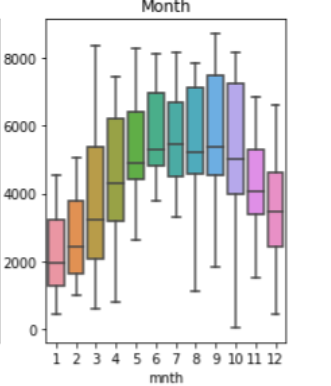
1. From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)

**Answer:** The categorical variables from the dataset are: **season, mnth, weekday, weathersit, holiday and workingday**.

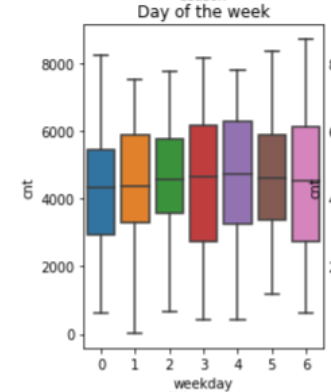
* **Season**: Most of the bike booking from BoomBikes happen in rainy season with a Median above 5000 count. The second highest number of bookings happen in Summer close to 5000 count. Spring has least number of bookings. Season is a good predictor variable.



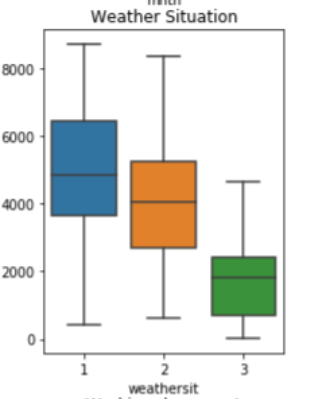
* **mnth**: Most of the booking happen from May to September close to 5000 booking per month. April, October and November months have a booking of more than 4000.



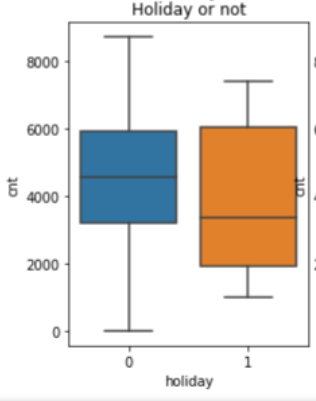
* **weekday (Day of the week):** 0 indicates Tuesday and 6 indicates Monday here. All the days have same median around 4300 to 5000 approximately. The model shows Monday would be better day.



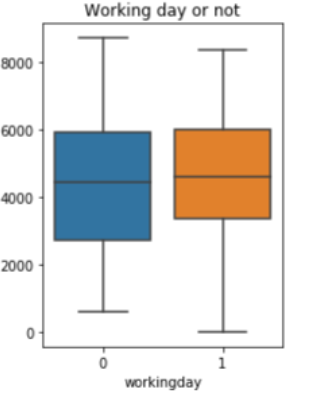
* **weathersit**: When the weather situation is cloudy or Mist (value = 1), there are a greater number of bookings happening on the BoomBikes. The median is close to 5000 bookings



* **holiday**: The plot shows median more than 4500 counts, when it is not a holiday.



* **workingday**: Workingday has value 1 for a working day. Working day have higher number of bookings.



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

**Answer:** Dummy variables are created for categorical variables. If there are ‘k’ number of values in a categorical variable, then we need to represent it in (k-1) dummy variables and then drop the original column. The drop\_first=True will achieve in creating (k-1) dummy variables and mapping the values accordingly.

E.g., Four seasons are indicated by 3 dummy variables:

Dummy variables for Season column

|  |  |  |  |
| --- | --- | --- | --- |
| **Season** | **Season\_summer** | **Season\_rain** | **Season\_winter** |
| 1: spring | 0 | 0 | 0 |
| 2: summer | 1 | 0 | 0 |
| 3: fall | 0 | 1 | 0 |
| 4: winter | 0 | 0 | 1 |

1. Season (1:spring) when Season\_summer=0, Season\_rain=0, Season\_winter=0
2. Season (2:summer) when Season\_summer=1, Season\_rain=0, Season\_winter=0
3. Season (3:fall) when Season\_summer=0, Season\_rain=1, Season\_winter=0 indicated by Season\_rain
4. Season (4:winter) when Season\_summer = 0, Season\_rain=0, Season\_winter=1 indicated by Season\_winter

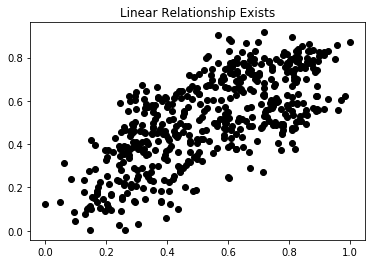
3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

**Answer:** Temp variable has highest correlation

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

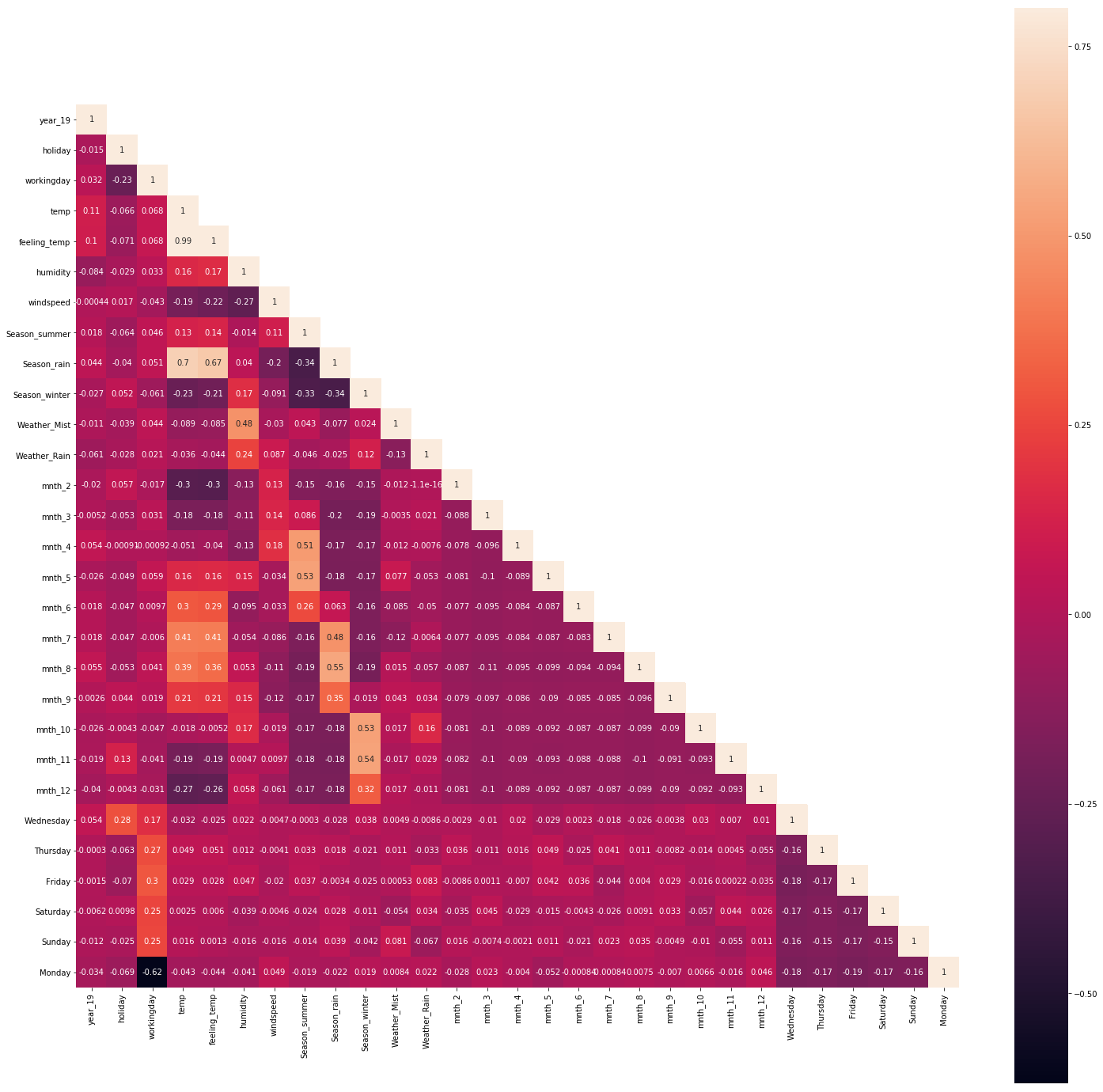
1. **Linear Relationship between the features and the target**

- Linear relationship exists with 'temp' feature



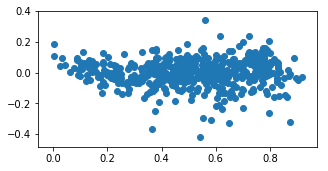
1. **Little or no Multicollinearity between the features**

- We can note that there are no multi collinearity



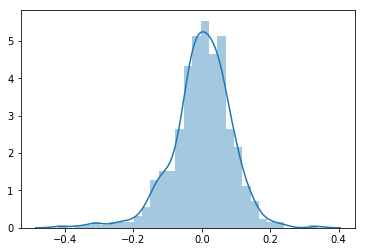
### Homoscedasticity Assumption

We are unable to findout any pattern with the noise, hence it is a Homoscedasticity



1. **Error terms are normally distributed with mean zero**

* We can notice almost normal distribution with residuals



1. **Variables follow a Normal Distribution**

* We can notice almost normal distribution on the q-q plot

### 

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

* **temp**: The bike demand in BoomBikes increases by 0.5309 times when there is raise in temperature.
* **year\_19**: In the year 2019, there has been an increase in demany in BoomBikes by a ratio of 0.2292 when compared to 2018. As the years increases, the demand will also increase
* **Season\_winter**: The demand for Bike sharing is high in winter by 0.1349 times

## General Subjective Questions

1. Explain the linear regression algorithm in detail. (4 marks)

**Answer:** Linear Regression algorithm is based on Supervised Learning methods. It is used to predict the target variable. There are two types of linear regression:

* Simple linear regression
* Multiple linear regression

1. Simple linear regression: This is basic type of regression model. This explains relationship between dependent variable and one independent variable and is depicted by a straight line. Usually a scatter plot is plotted and straight line is obtained from the data points.

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)