Sharing Bicycle Hourly Utilization Analysis and Prediction

Dakai Zhou

1 Part 1

1.1 Data Analysis

First of all, it is import to know the data from a large scope. From figure 1, it is clear that the business grew in 2012. Figure 2 and figure 3 illustrates that people tended to use bicycle less when it is cold outside and use more when it is warm, which also support the pattern of the bicycle utilization change (goes up and goes down) though the whole year in figure 1. As is shown in figure 4, less bicycles were rented when the weather was bad. This explains the low points in figure 1 (see figure 5).

Keep going to observe more detailed information. The average amount of rented bicycles on working days are similar, but it is more than on weekends (figure 6, 7), and even much more than on holidays (figures 8).

After having insights on the data on yearly basis, seasonal basis, monthly basis and daily basis, it is time to move to hourly data. Figure 9 15 show that the utilization of sharing bicycles on working day weekends and holidays has different pattern. On working day, there are two peaks during the commuting time (8:00 and 17:00). The reason for higher count at 17:00 than at 8:00 might be the time after work is more flexible (in the morning there are risks of being late) and people might want to have some exercise after the long day work. There is also a small increment around 11:00-13:00, it may because it is more convenient to get some food for lunch by bicycle. On weekends, the peak is around 12:00-13:00, which is similar to it is on holidays.

Different seasons do not have much impact on the bicycle utilization pattern, it only influences the number of rental bicycle at each hour (figure 10 11 12 13 14).

For other factors like temperature humidity which are difficult to analyze by visualization, correlation map can be adapted. Figure 16 gives the correlations among the features. It shows higher temperatures motivates people to use more bicycle, while high humidity demotivates people. Features *temp* and *atemp* has correlation 0.99, one can be dropped when fit a model. Features *instant dteday casual* and *registered* also should be dropped. For having better feature performing, features like *season* should be split into four binary features, it can increase the accuracy of the model.

1.2 Model

Random forest was chosen as the model for this problem, the main reasons are as follows:

- Random forest works well on categorical variables and numerical variables. The given data contains both of them.
- Random forest can avoid over-fitting
- Random forest has the power to handle a large data set with higher dimensionality. The given data has many features.
- The data is not stationary, so time-series model ARIMA is not considered

The mean absolute deviations is 25.473.

2 Part 2

The top 2 features for a scalable model are:

- It should be able to deal with any amount of data, without consuming ever growing amounts of resources like memory.
- It allows fast computations for massive data-sets (parallel computing is a way to achieve this, Stochastic Gradient Descent is one algorithm example).

The scaling properties of this model is it can be done easily in a parallel way. However, it may needs a lot local RAM to build a big forest.

Online Random Forest is one solution. It can deal with massive data streams, but also massive (static) data,by running trough the data sequentially. It put data online, requires not local RAM.

The drawbacks of online random forest is that it requires a cloud platform. The algorithm is complicated.

I do not have experience on such a specific technique. But I learned parallel programming and parallel computing in the university. I also did two projects which are related to parallel programming and parallel computing.

3 Plots

3.1 Plots from day based data

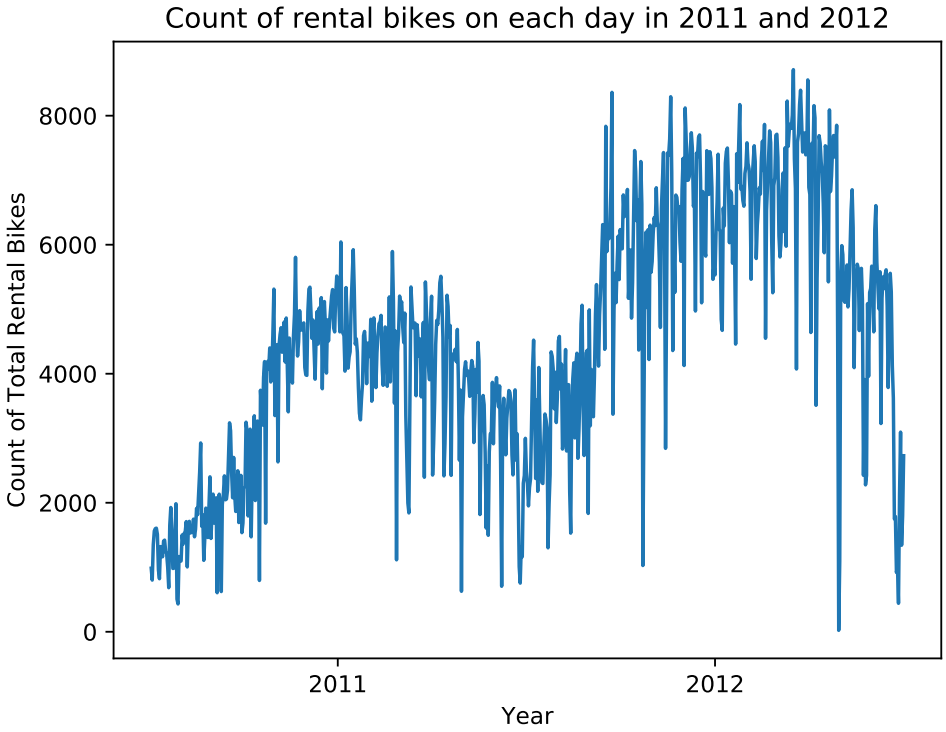

Figure 1: Average count of rental bicycle for each year

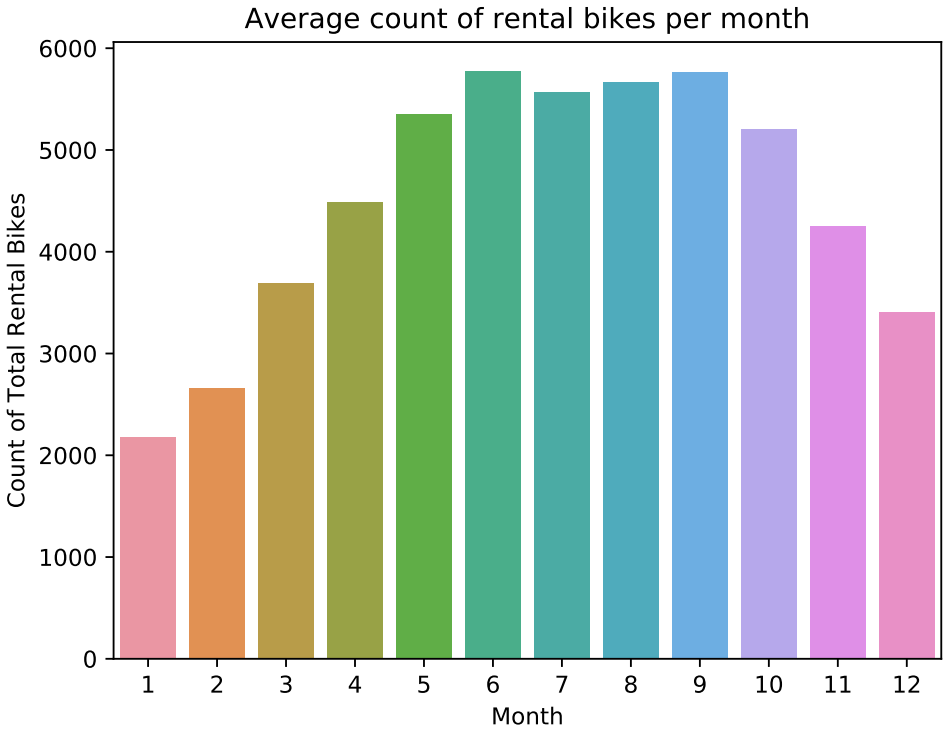

Figure 2: Average count of rental bicycle for each month

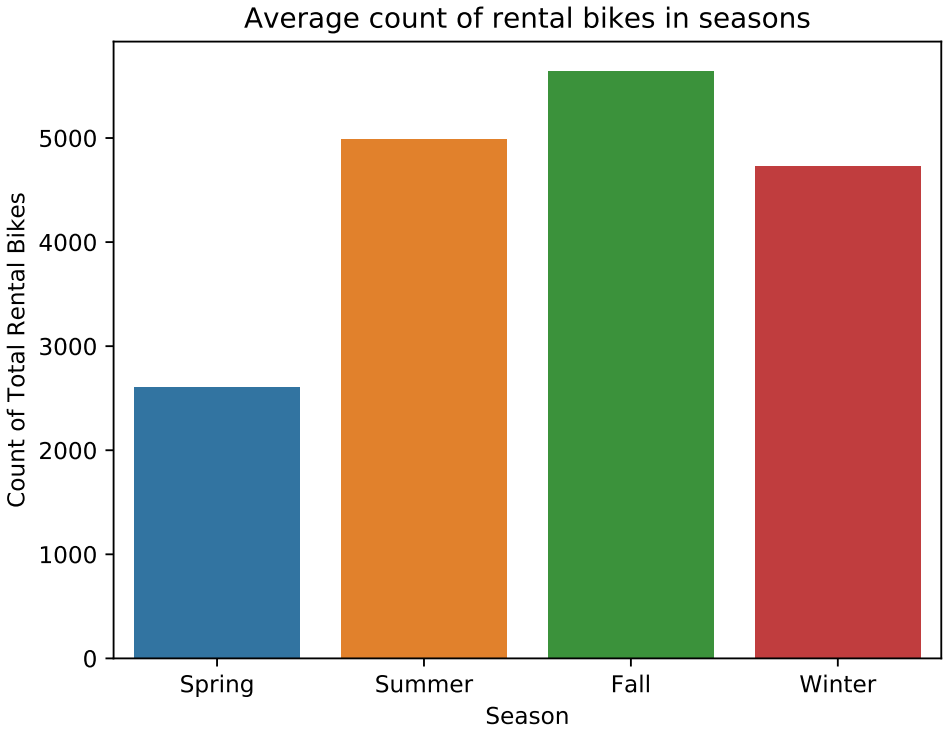

Figure 3: Average count of rental bicycle for each season

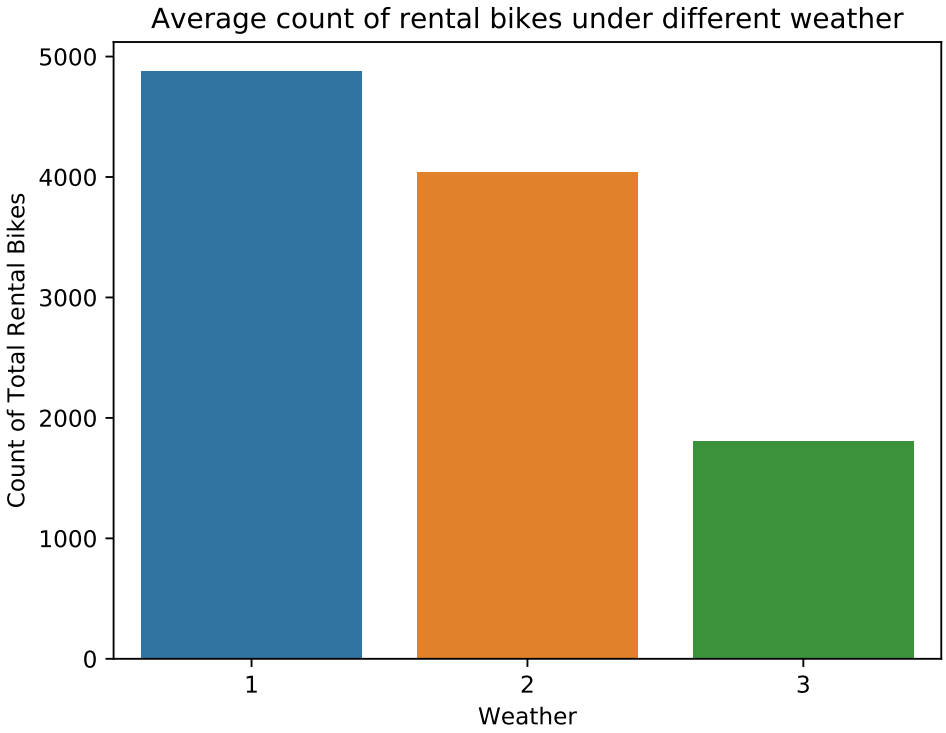

Figure 4: Average count of rental bicycle for different weather

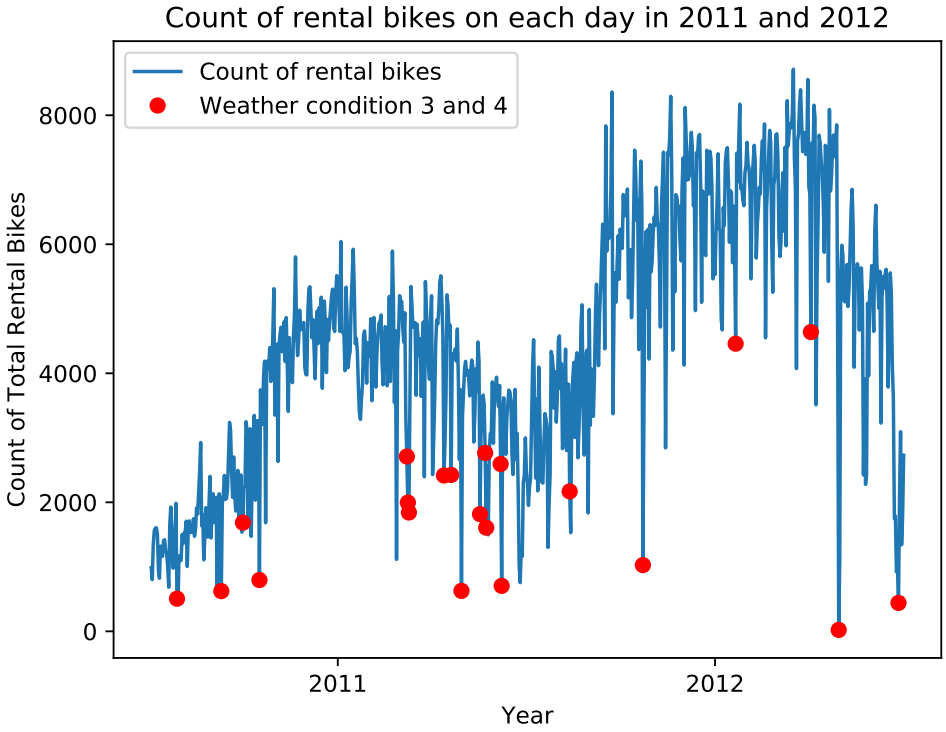

Figure 5: Average count of rental bicycle for each year

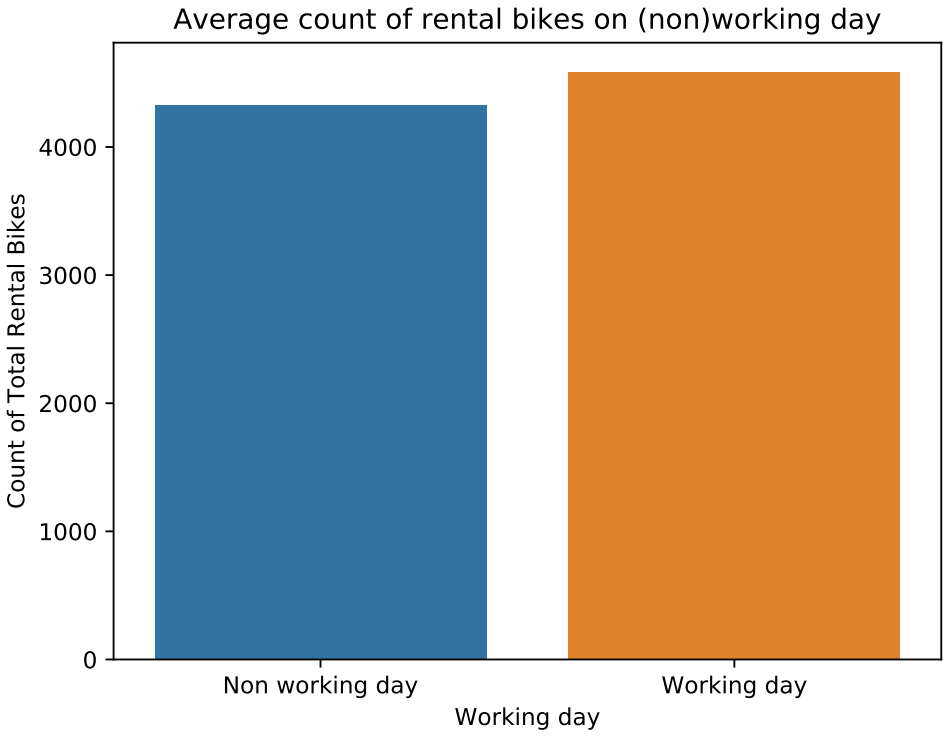

Figure 6: Average count of rental bicycle for (non)working day

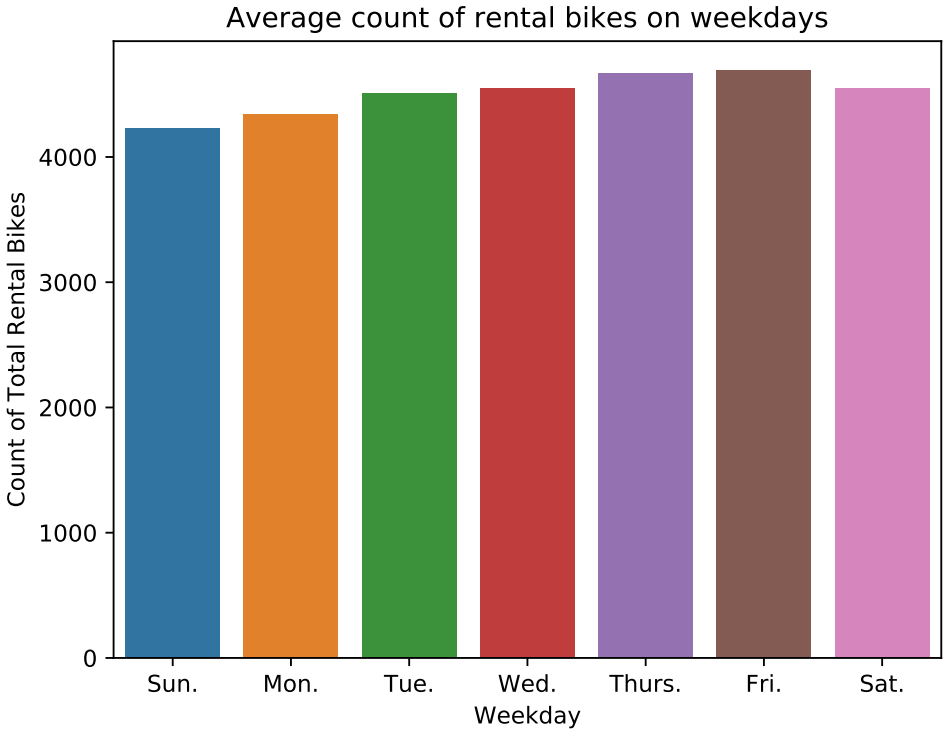

Figure 7: Average count of rental bicycle for weekdays

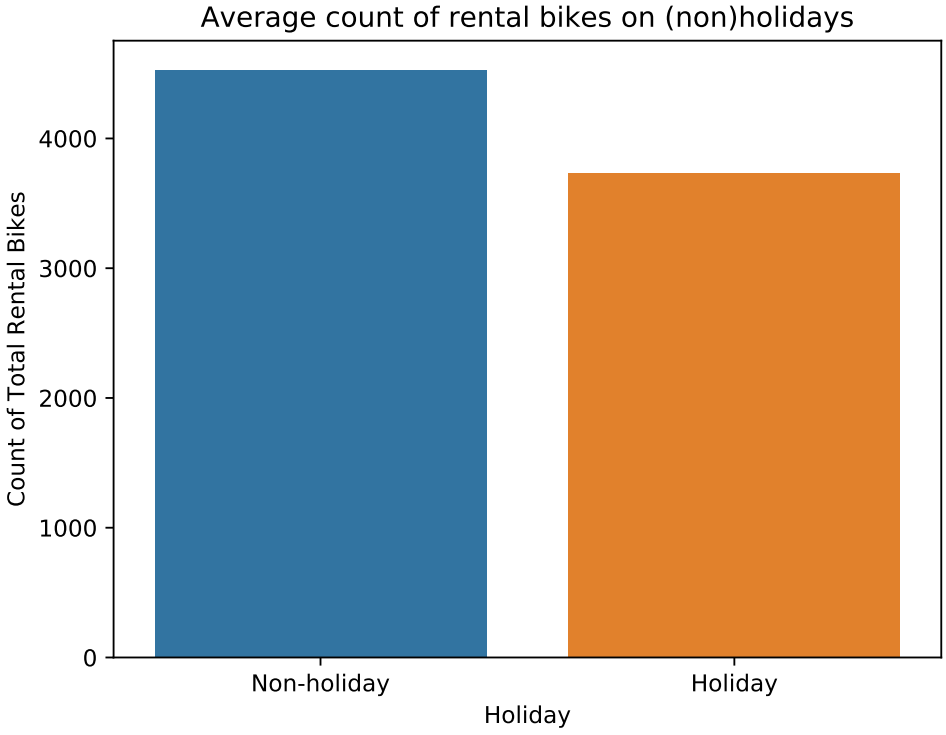

Figure 8: Average count of rental bicycle for (non)holiday

3.1.1 Plots from hourly based data

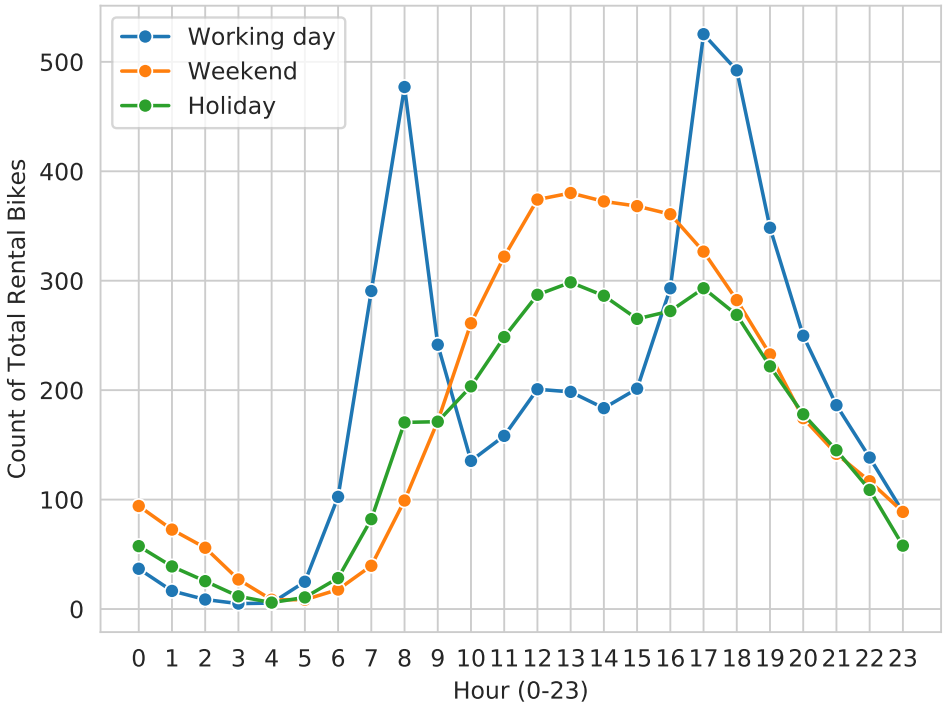

Figure 9: Average count of rental bicycle for each hour on different types of days

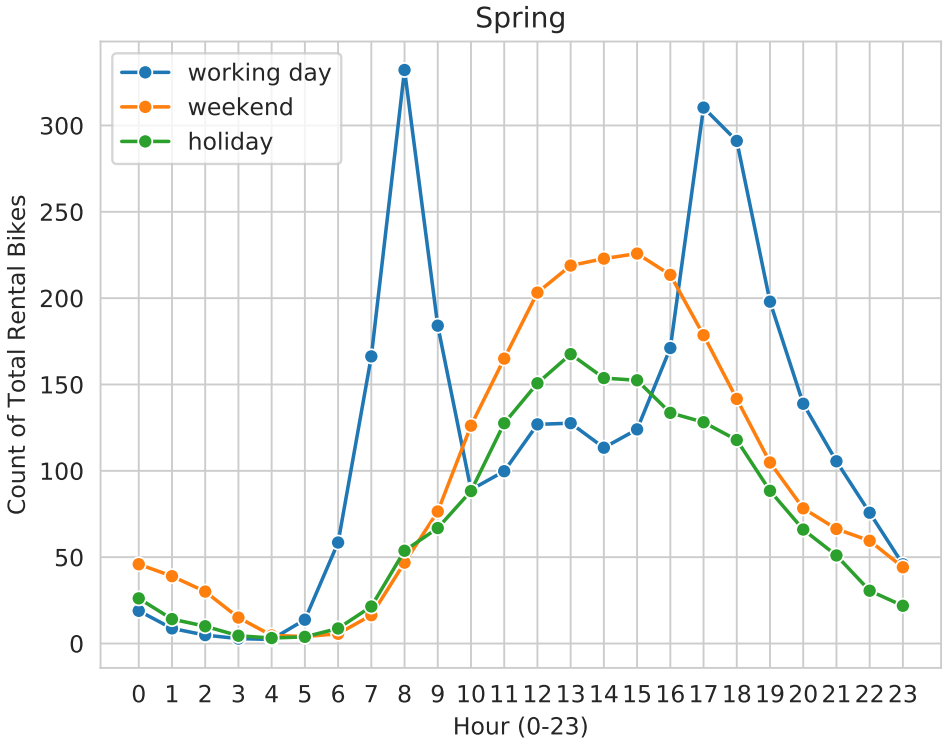

Figure 10: Average count of rental bicycle for each hour in spring

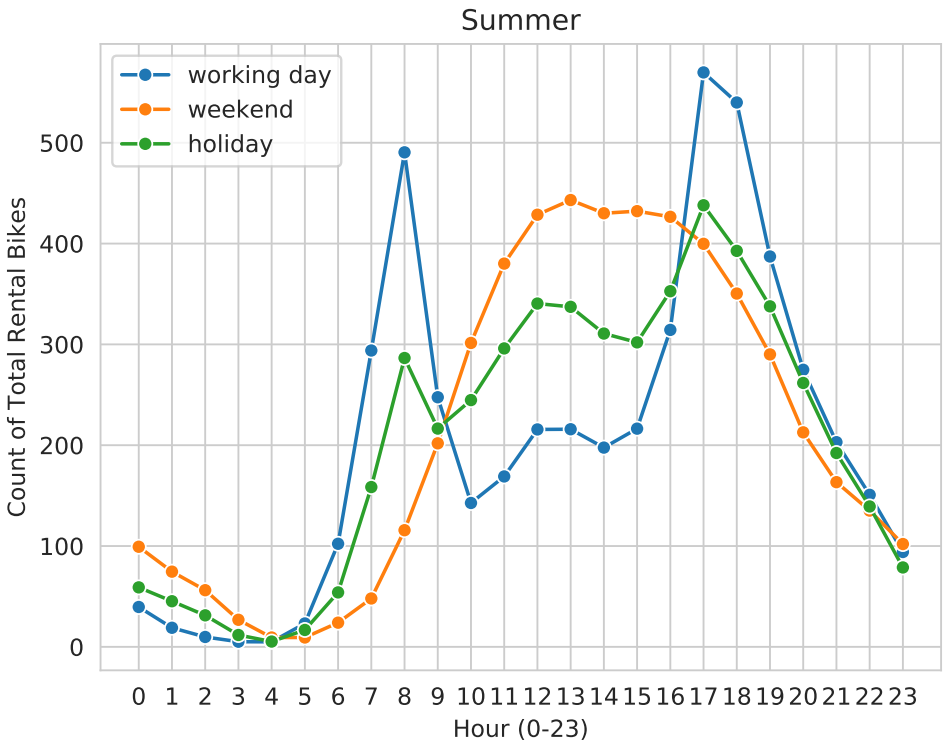

Figure 11: Average count of rental bicycle for each hour in summer

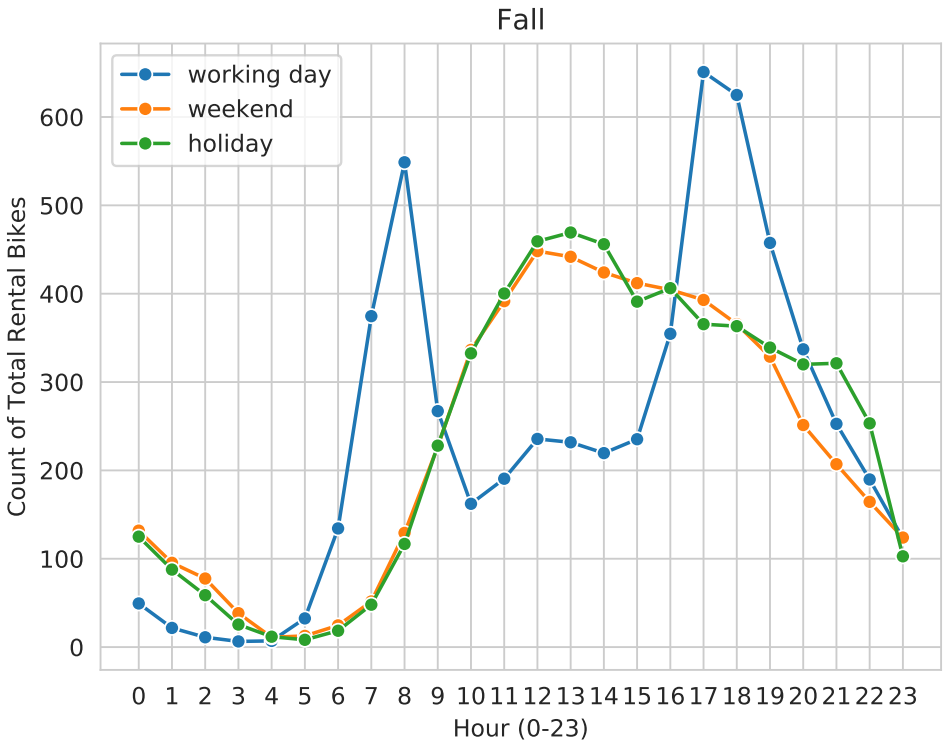

Figure 12: Average count of rental bicycle for each hour in fall

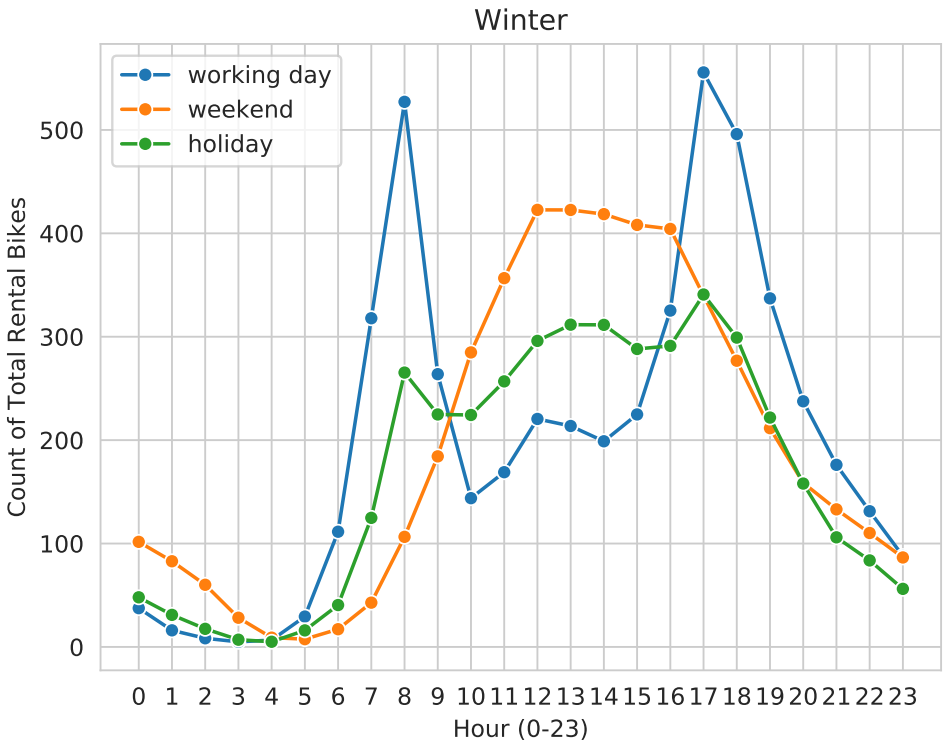

Figure 13: Average count of rental bicycle for each hour in winter

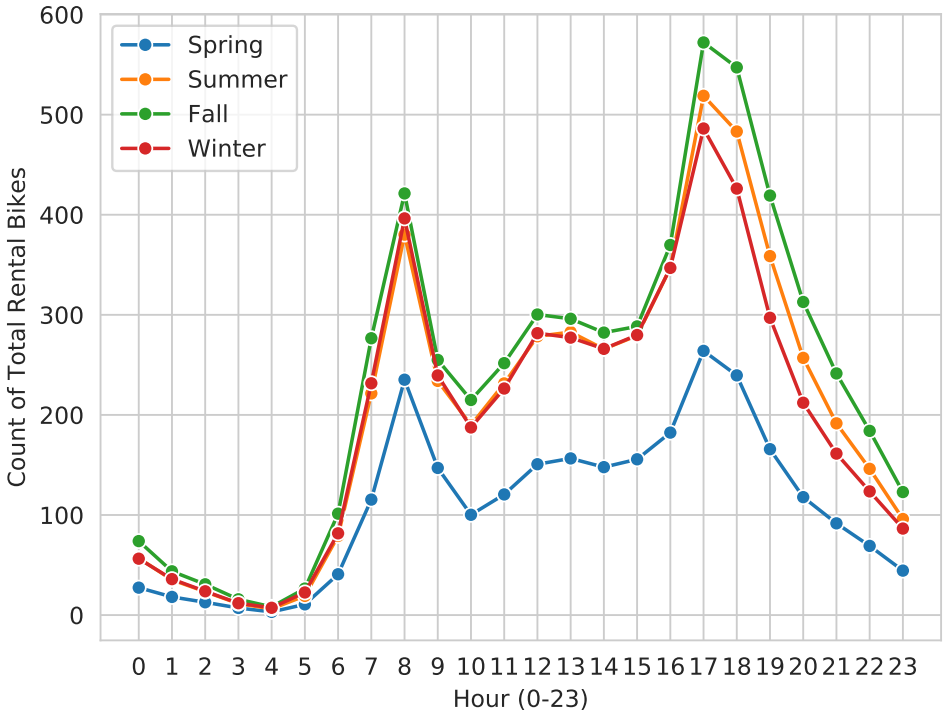

Figure 14: Average count of rental bicycle for each hour each season

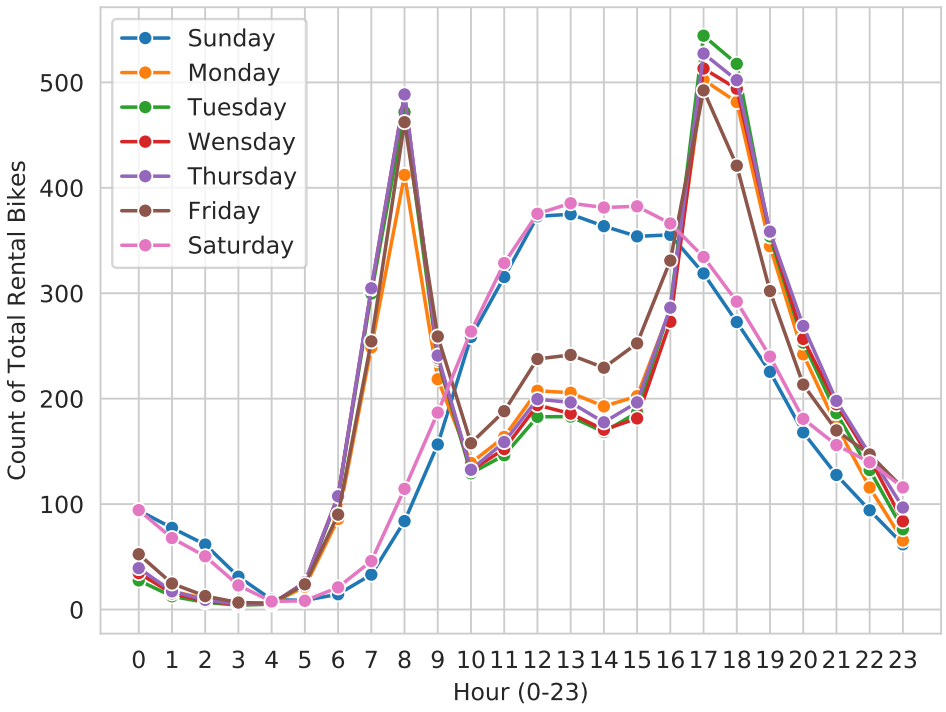

Figure 15: Average count of rental bicycle for each hour each weekday

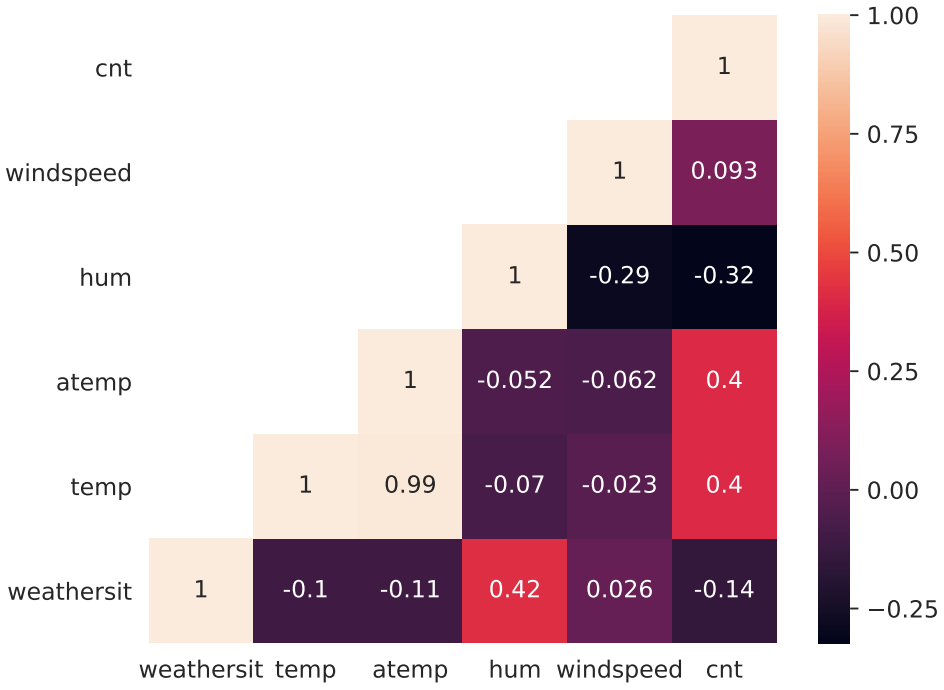

Figure 16: Correlation map among the features