Find a problem (Research target):

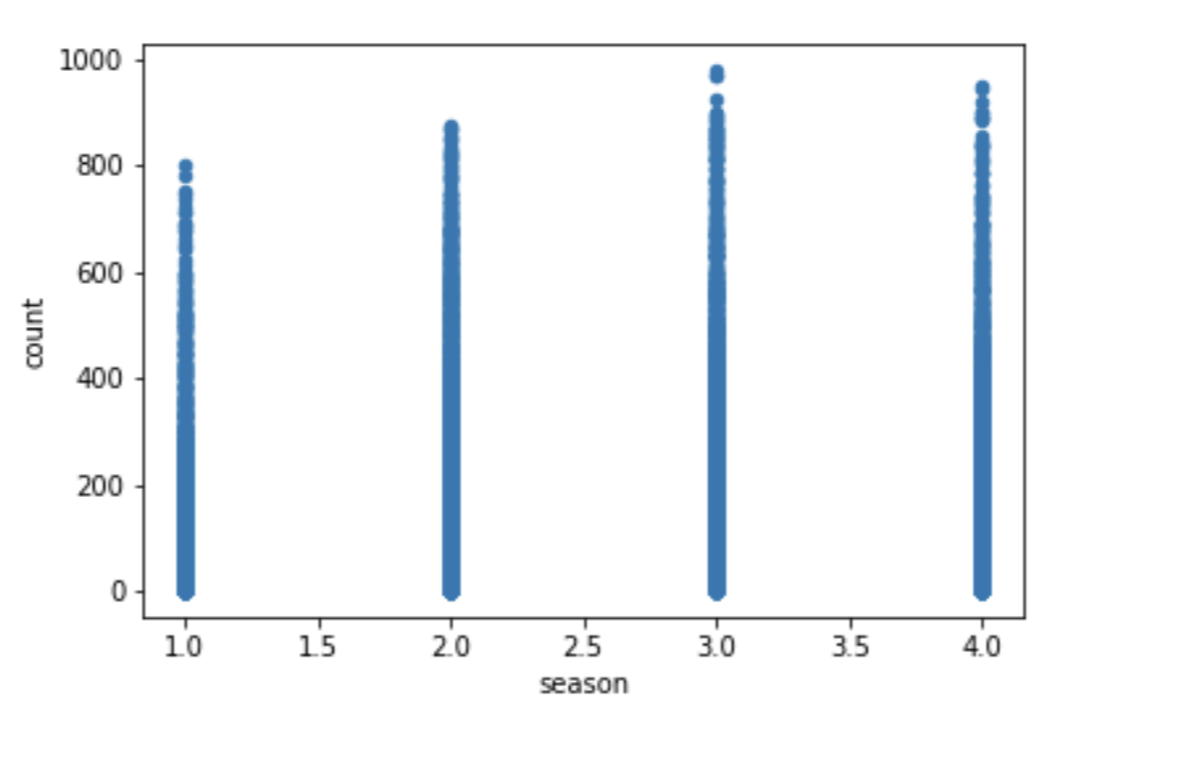
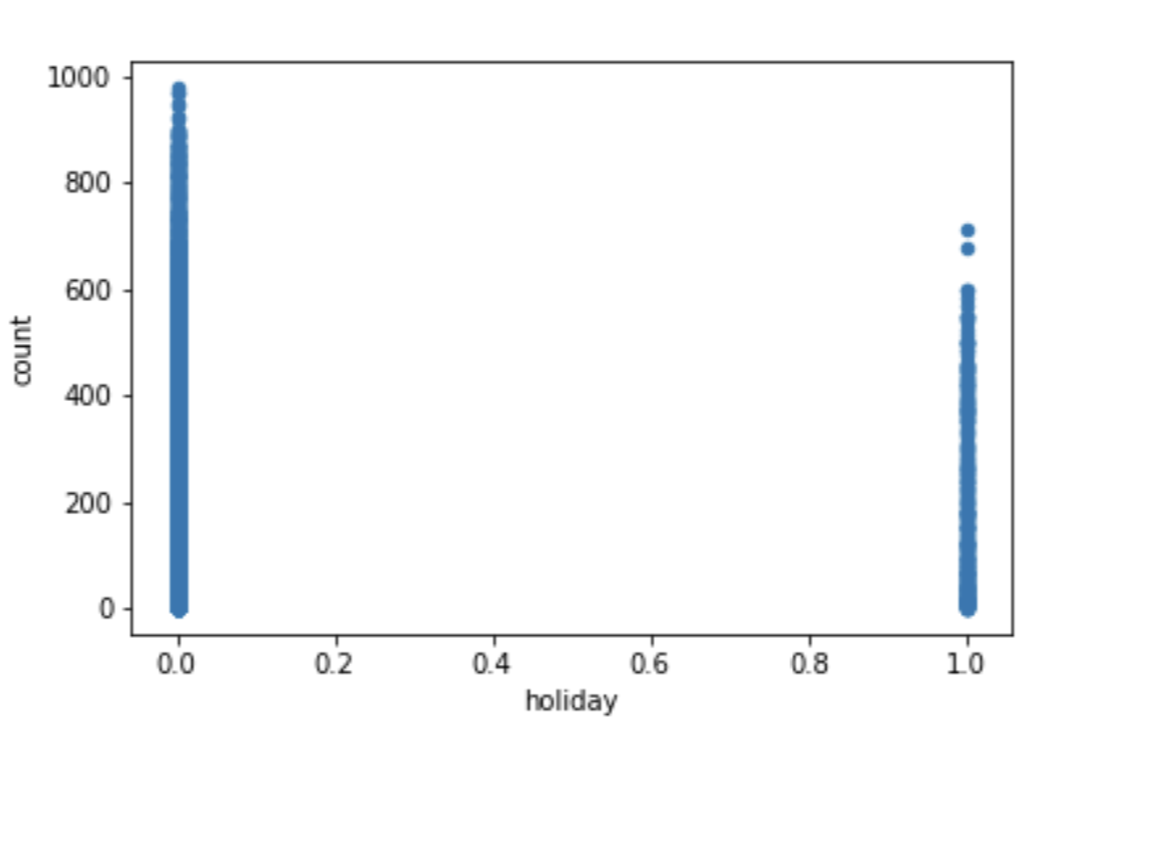
We wanted to find out what factors affect the number of bikes shared.

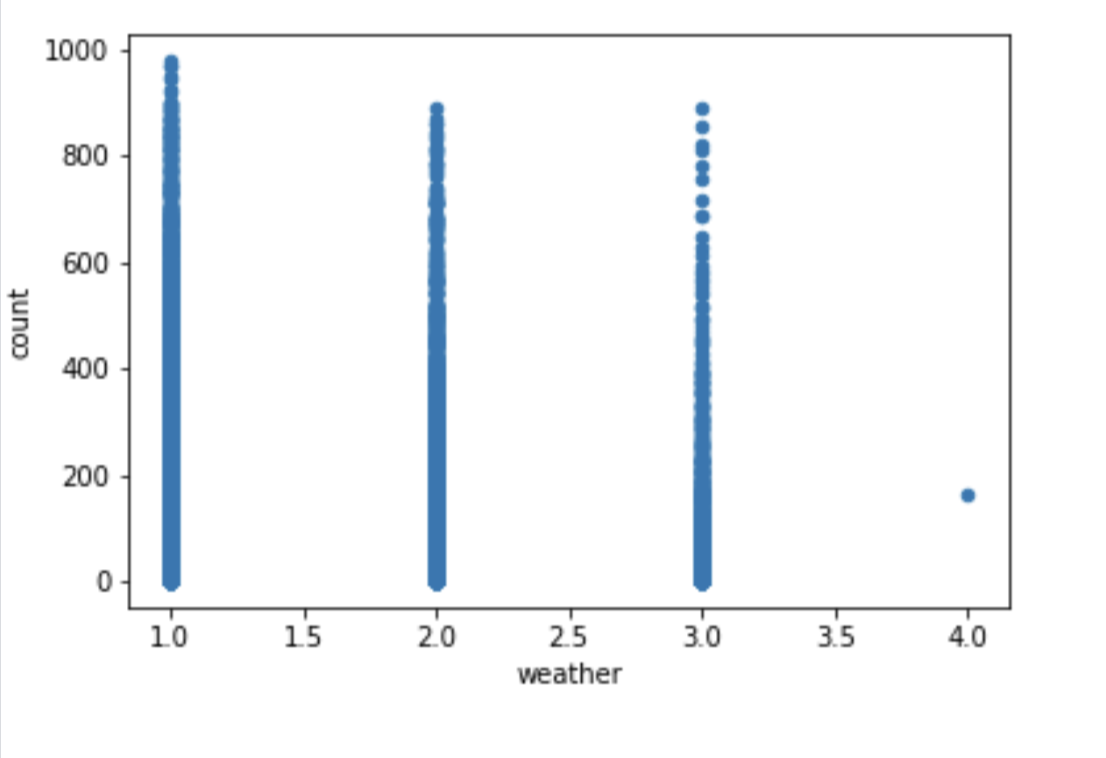
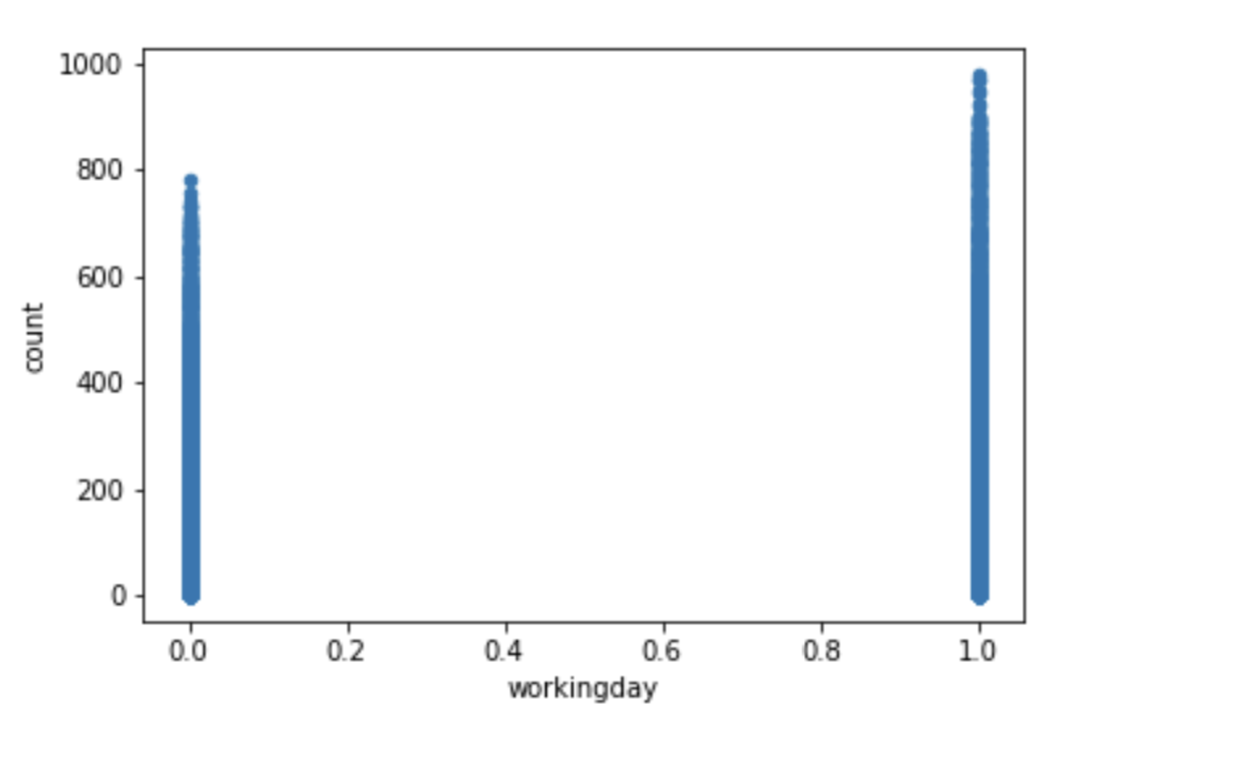
Analyze previous approaches:

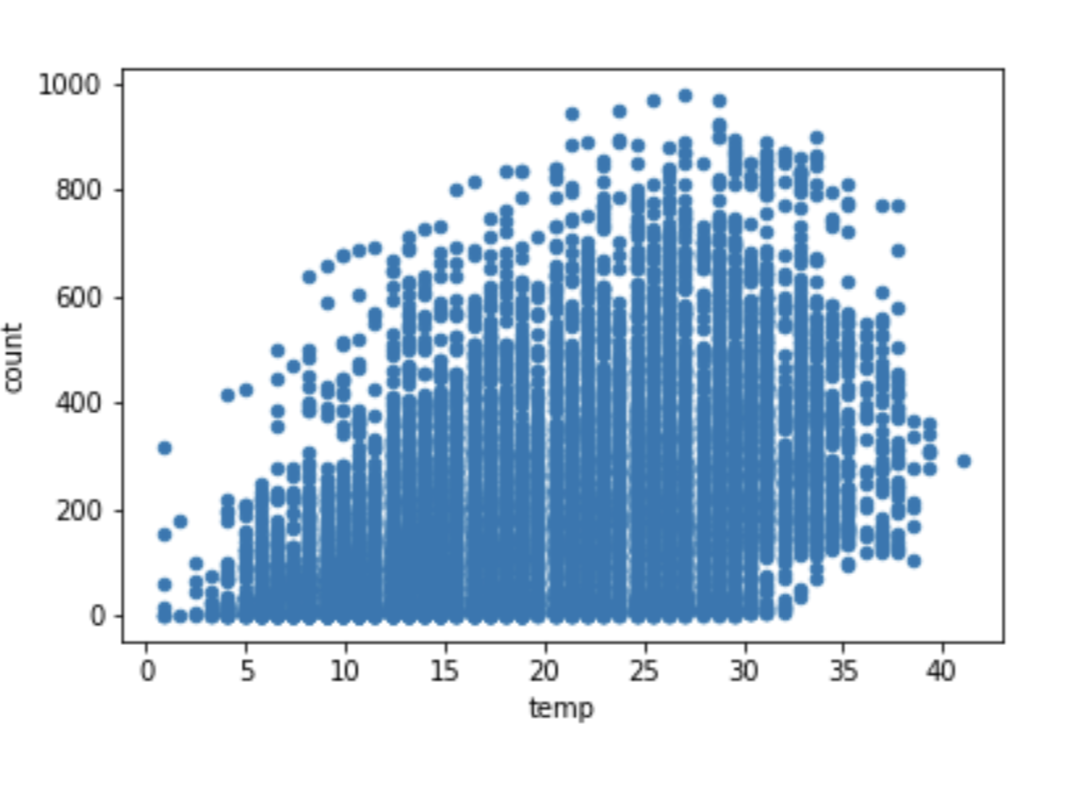
Source of code: <https://www.kaggle.com/deeprajbasu/exploringbikesharingdata>

DEEPRAJ BASU uploaded 2 years ago

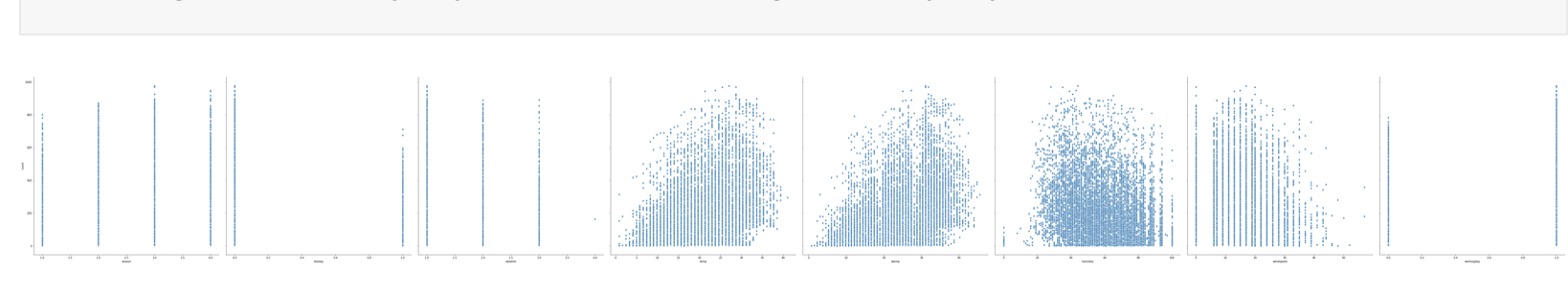
Step1: The author used separate statistics of seasons, holidays, working days, weather, windspeed and temperature to make charts. There are some examples.



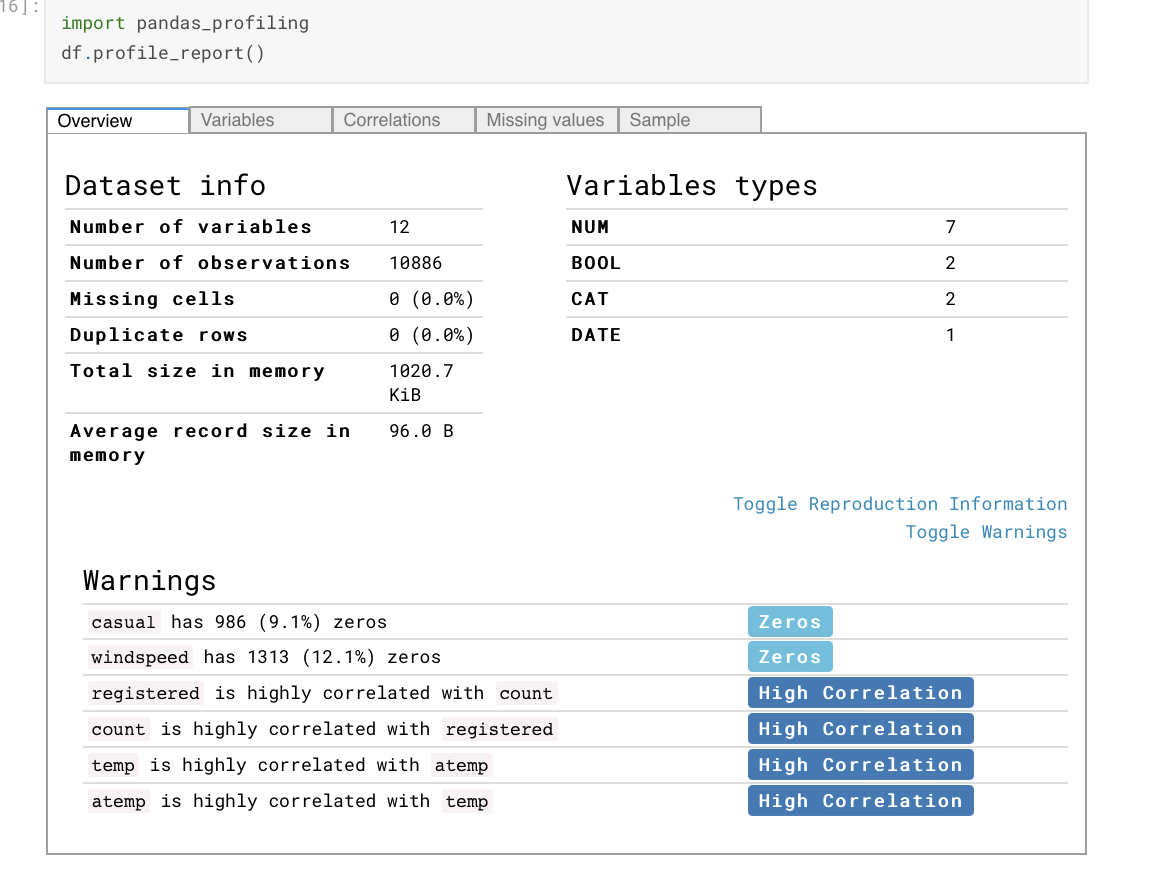




Integrated data graph result

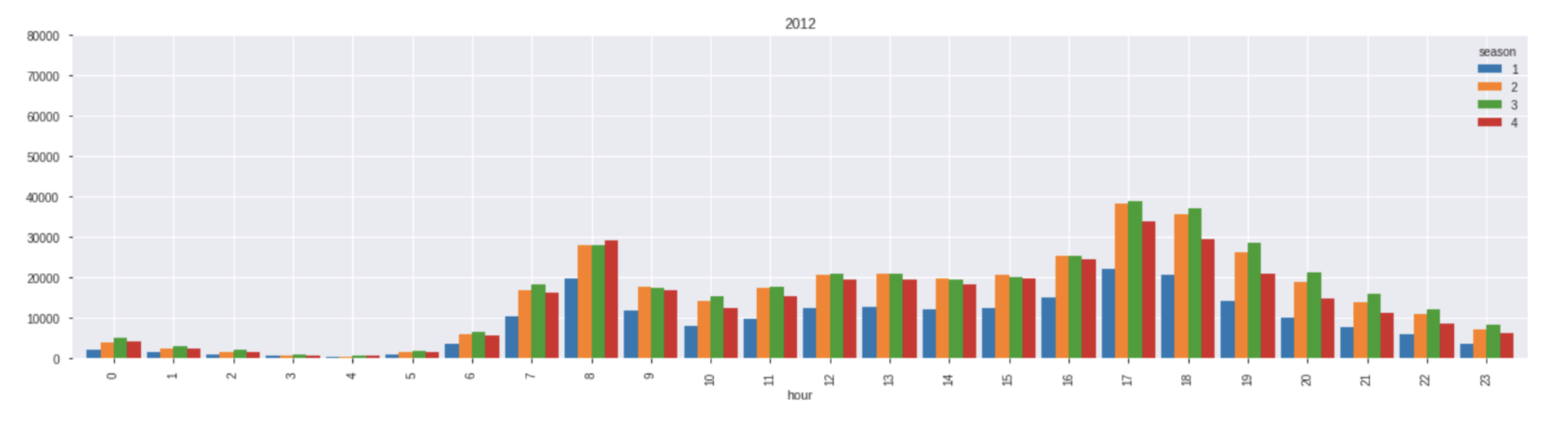


Step2: Use Pandas\_profiling to exploring data

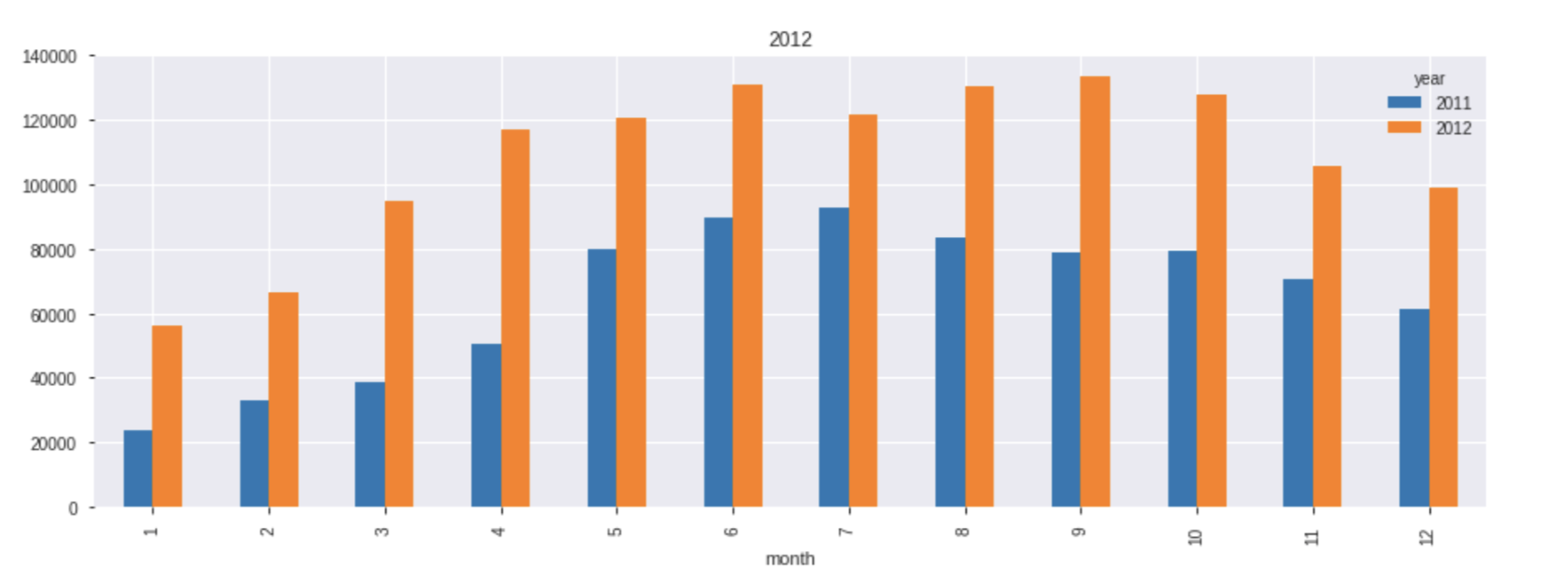


Step3: Visualizing the date time

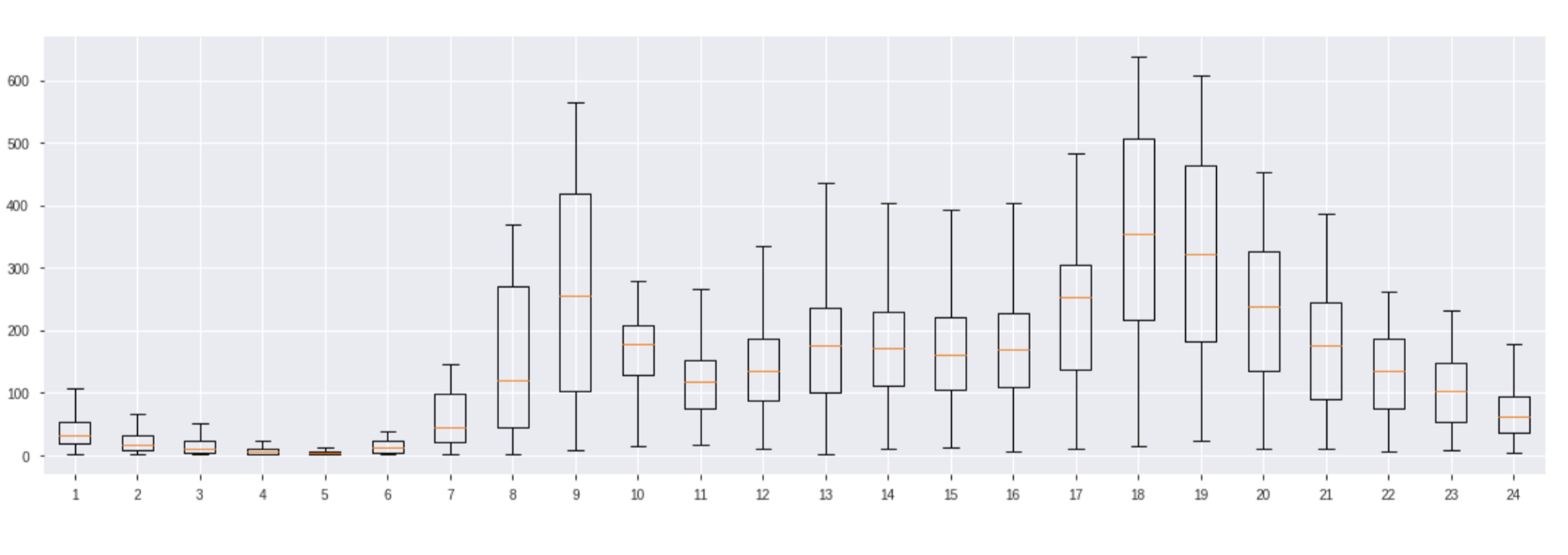
The authors used data from 2012 as a reference and hoped to get effective information from it.

The time in the date data is extracted in units of hours to judge the peak period of shared bike use.

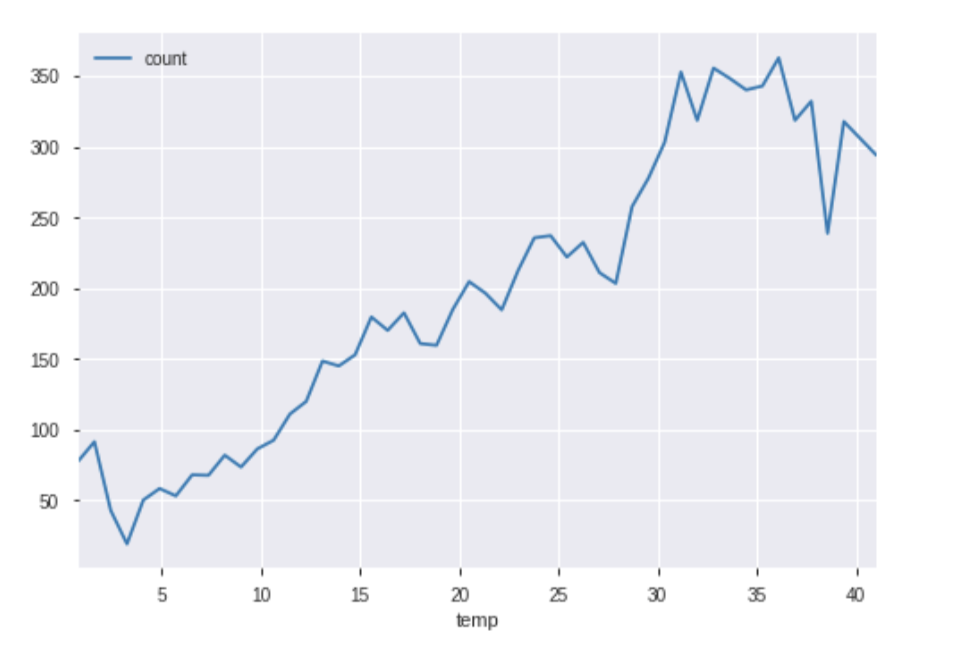
The time in the date data is extracted in monthly units to judge the peak period of shared bike use.

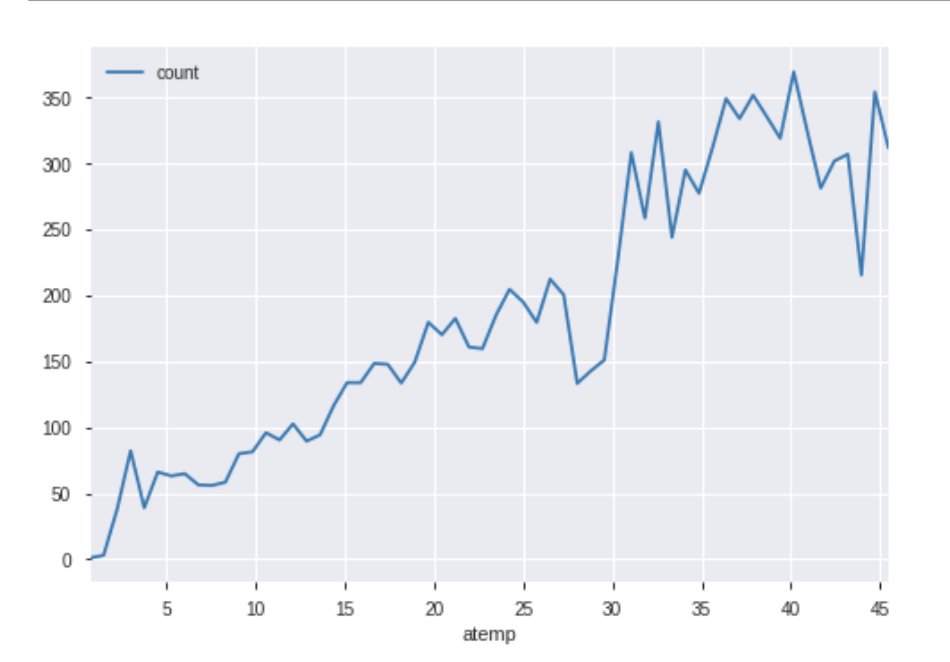


Convert data representations to box plots. The purpose of using a box plot is to see how the data is distributed, to see where the data is concentrated, what the characteristics of the distribution are, whether the data is concentrated on the side of the smaller value or the side of the larger value.



Step4：Use line plot to guess bike demand





-survey general existing approach , algorithms

Previous version: use Pandas\_profiling, generate box plot

Our team: Non-parametric test, Hypothesis test, calculate correlation coefficient，random forest algorithm

-relevance of the presented differentiation

Previous version: The first step is to chart the data separately. In the absence of a definite correlation, conclusions are unconvincing. Then the last step of bike demand prediction, which only studies the two influencing factors of temperature and body temperature.

Our team: We want to improve the following three points. The first is to check whether there are a lot of outliers. The second point is to use hypothesis testing to find out the potential influencing factors. The third point is to determine whether the correlation is strong.

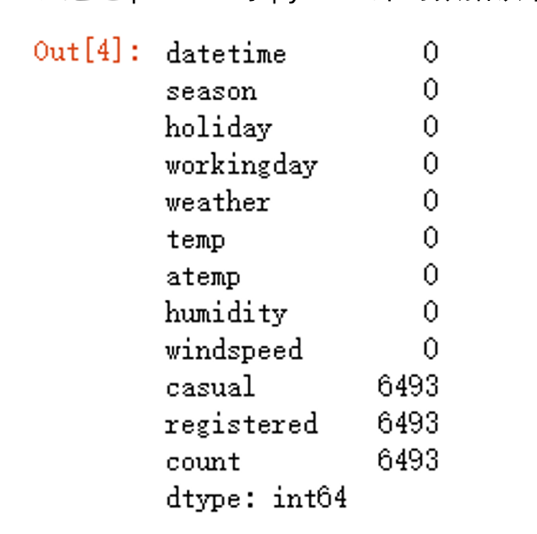
- Modeling new methodology

We added the random forest algorithm based on the last report.

The characteristics of the training model are month, day, weekday, hour, holiday, working day, weather, temp, atemp. The targets are casual, registered and count.

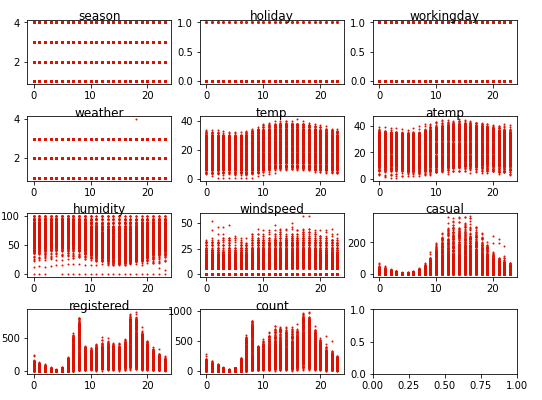
- Analyzing data

Extract the database.

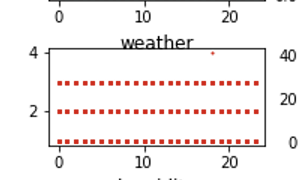


There is no null value in the visible data. The casual, registered and count in the test set need to be filled after we make predictions later.

Data observation

Scatter plots were drawn by Matplotlib and Seaborn to observe the distribution of data.

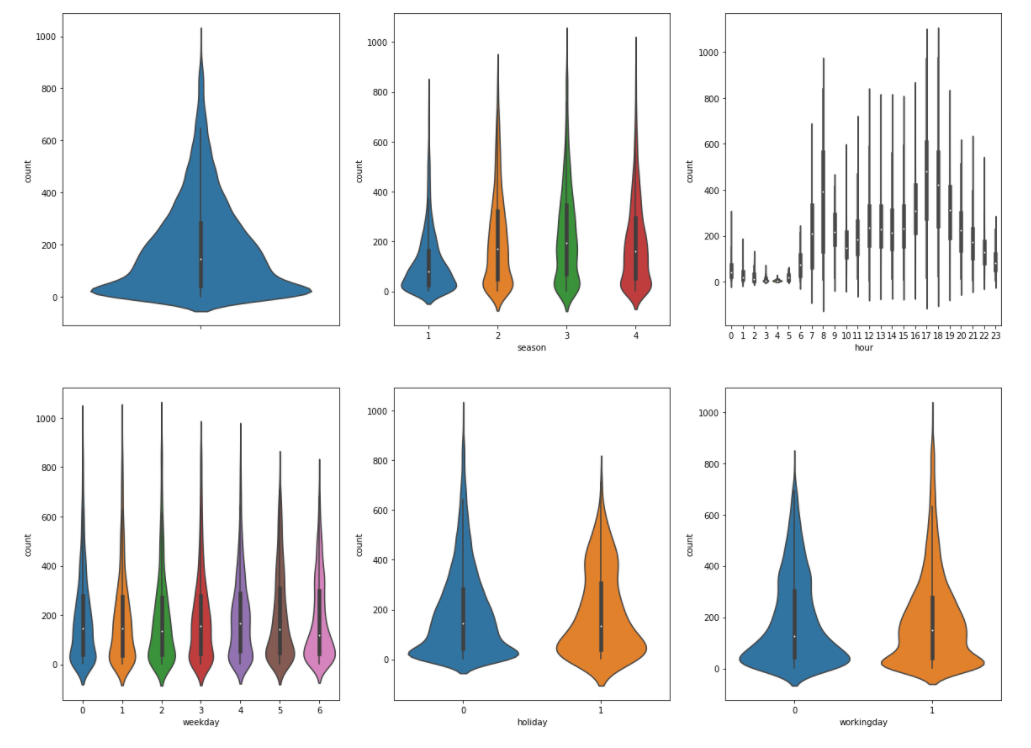
We find there is one extreme data in weather graph.





After our inspection, this data is real extreme weather.

Butterfly plots were drawn by Matplotlib and Seaborn to observe the distribution of data.



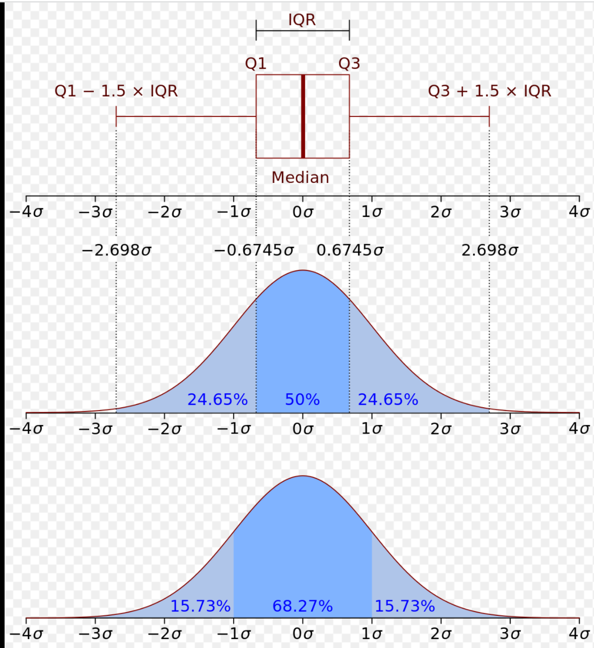
We didn't find many outliers, the data for the whole dataset is reasonable.

Our conclusions are as follows

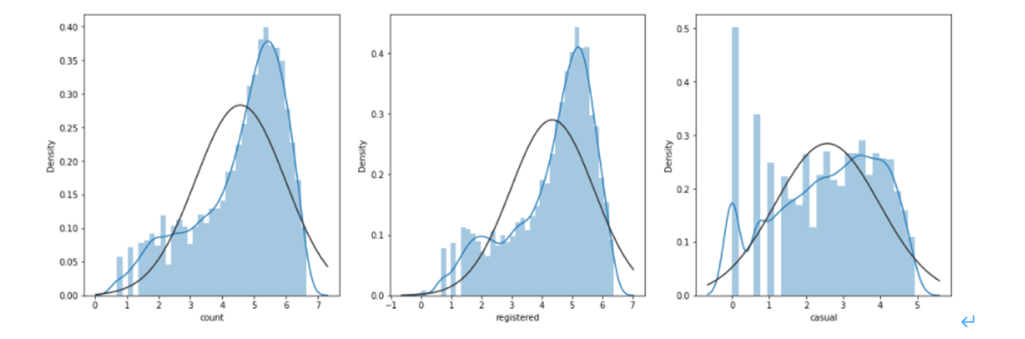
1. There are some outliers in spring and winter.
2. In the hourly chart, there are some outliers before noon.

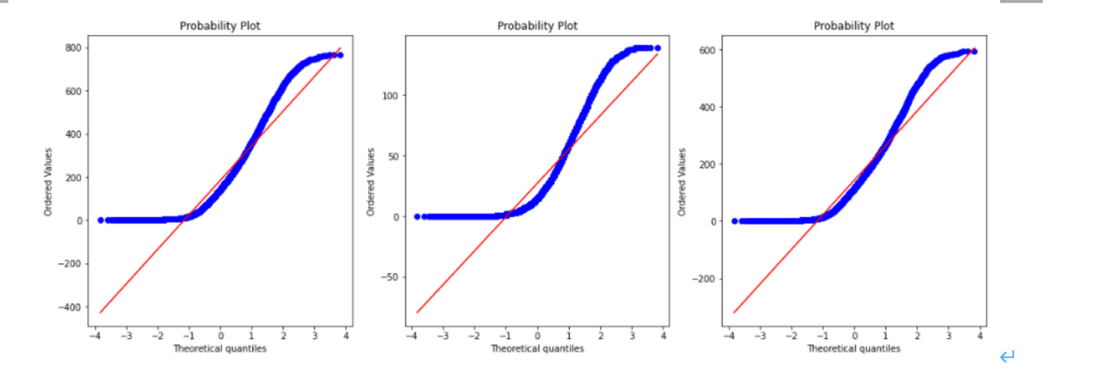
Handle outliers

Remove outliers by determining upper and lower limits. After processing, the data will be reduced by 100 to 300 entries

The upper and lower limits we set: Inter Quartile Range +/- 2iqr  


Determine if the casual, registered, count is normally distributed.



­­­­­­­

We compare the resulting graphs with the standard normal distribution graphs and find that these graphs show a short-tailed distribution. The pattern formed by the dots is curved above the line on the left and below the line on the right. This shows that the data is more concentrated closer to the mean than the standard normal distribution.

Hypothesis testing

We assume that the sample is normally distributed. If the hypothesis is true, the value should be close to 1. If the hypothesis is not met, the value is close to zero.

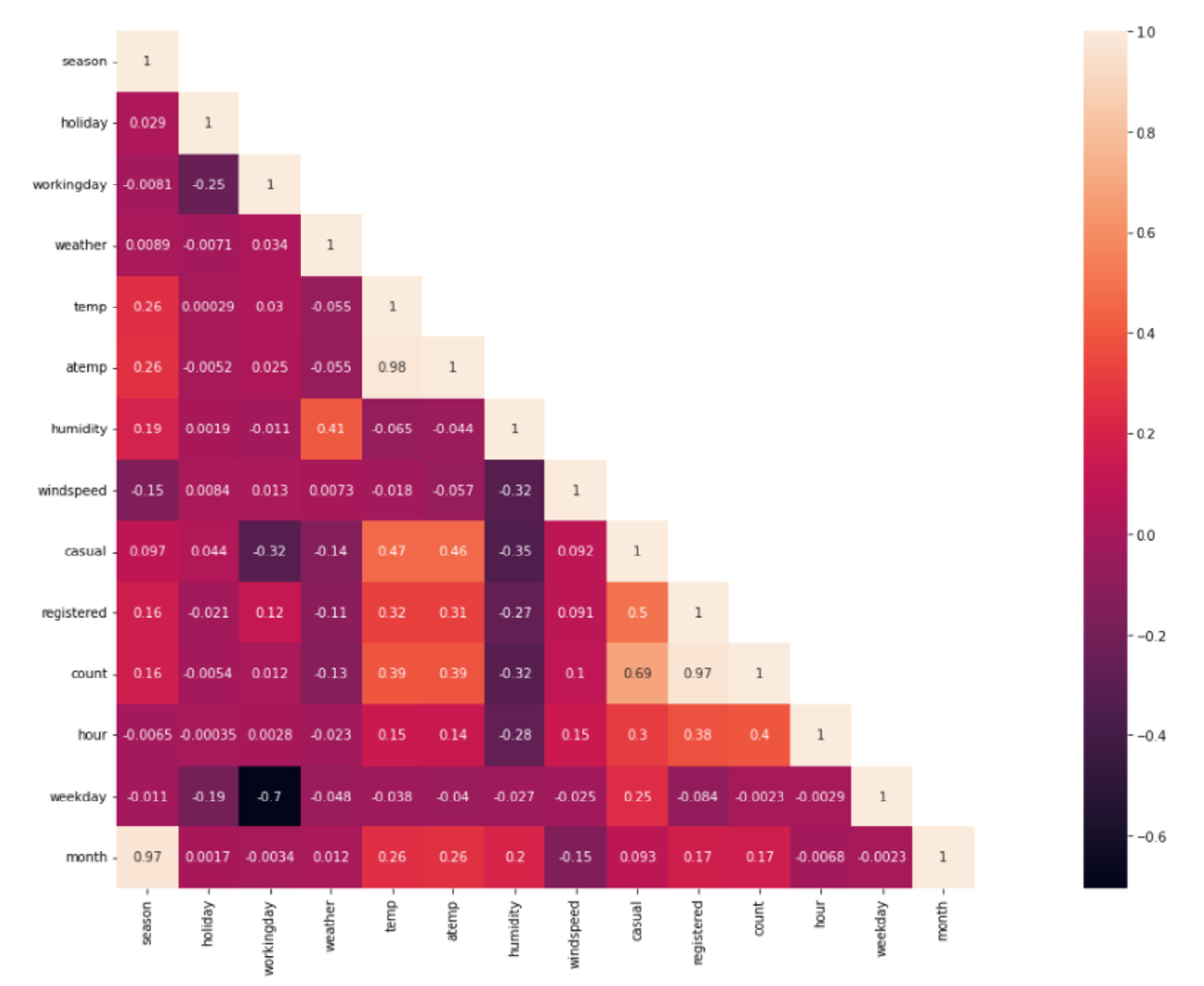
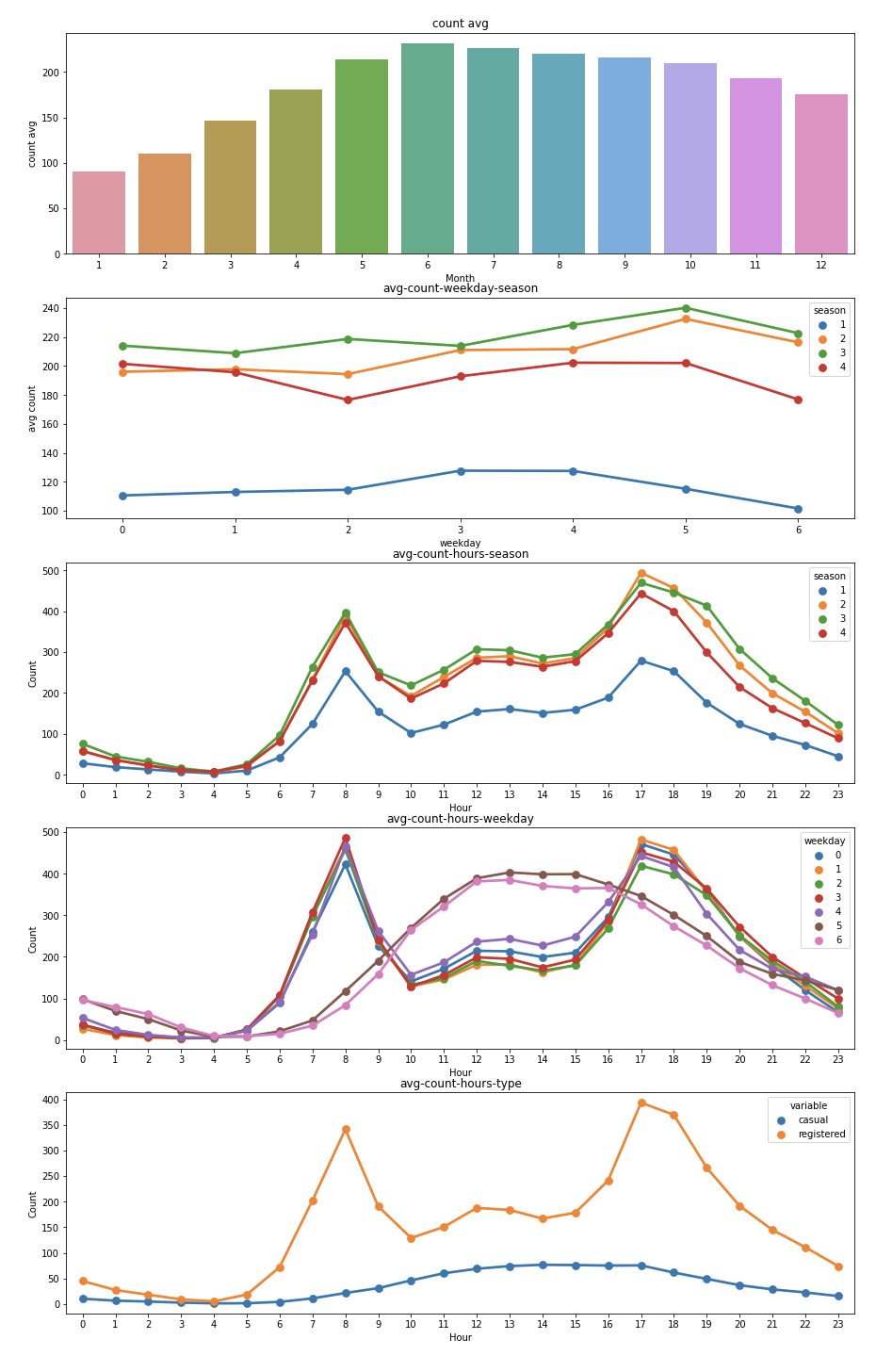


Chart analysis



Conclusion:

The number of users

Average monthly users: 6, 7, 8

The number of weekly users in four seasons is the highest in summer and autumn, the lowest in spring, and the lowest on Wednesday

Most people use it during rush hour every day

Correlation between the number of users and climate characteristics:

The body temperature (atemp), non-registered users (casual) and registered users (registered) had a strong positive correlation, while the negative correlation with humidity was obvious, and had no obvious relationship with wind speed

Modeling with random forest algorithm

Model explanation

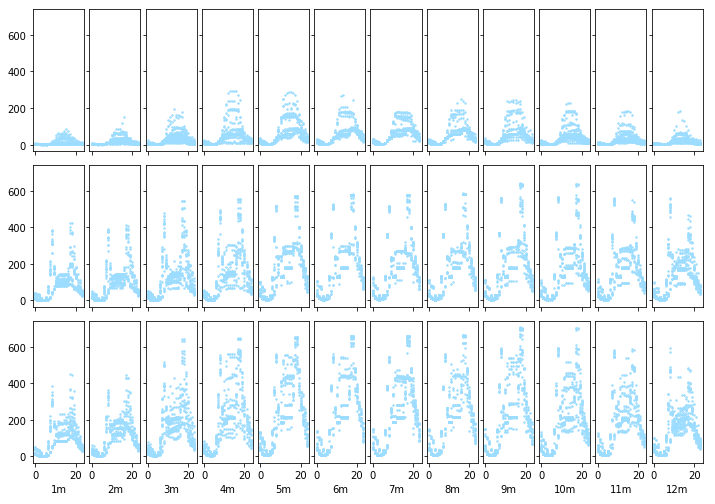
model\_casual: Random Forest models that use meteorological characteristics, time training to predict casual values.

model\_registered: Random Forest models that use meteorological characteristics, time training to predict registered values.

model\_count: Random Forest models that use meteorological characteristics, time training to predict count values.

Result explanation

The trained model was used to fill the three fields casual, registered and count in the test data set, and the final scatter chart was drawn as follows. The final result is a 3-row, 12-column diagram. The first line shows a predicted scatter chart for casual results for all months, the second line shows a predicted scatter chart for registered results and the third line shows a predicted scatter chart for count results. We found that the predicted results were consistent with the distribution characteristics of the training data set



collaborative evaluation

We finally decided this topic after consulting relevant materials and data sets in the early stage. 어즉민 is responsible for the previous approaches analysis and reports, 황자신is responsible for data analysis and prediction of random forest algorithm, and이건희 took part in coding, algorithm recommendation and preparation for final publication. Although there were some difficulties in decision-making during the project, we communicated with each other with delay and finally completed the project

Github Link：https://github.com/yuzemin07/BigData\_team13

