Capstone report

|  |  |
| --- | --- |
| Shared bikes | |
| Student names | Student numbers |
| 蒋佳巧 | 2017040370 |
| 谈文梦 | 2017040366 |
| 尤美玥 | 2017040376 |

# Idea：

## Description of data set

Dataset：《London bike sharing dataset》

Source：<https://www.kaggle.com/hmavrodiev/london-bike-sharing-dataset>

Target: Based on the features of temperature, wind speed and whether it is a holiday, to **predict the number** of shared bikes borrowed in London.

Meaning: Quantitative analysis of shared bicycle lending data can provide theoretical reference for shared bicycle companies to **dispatch shared bicycles**

Features :

|  |  |  |
| --- | --- | --- |
| **Feature name** | **meaning** | **type** |
| timestamp | date | object |
| cnt | Number of borrowed bikes | Int /constant |
| T1/t2 | temperature | Float/constant |
| hum | humidity | Float/constant |
| Wind\_speed | Wind speed | Float/constant |
| Weather\_code | different weather | Int/classified |
| Is\_holiday/is\_weekend |  | Boolean |
| season | season | Int/classified |

### Metadata:

"timestamp" - *timestamp field for grouping the data*  
"cnt" - *the count of a new bike shares*  
"t1" - *real temperature in C*  
"t2" - *temperature in C "feels like"*  
"hum" - *humidity in percentage*  
"wind\_speed" - *wind speed in km/h*  
"weather\_code" - *category of the weather*  
"is\_holiday" - *boolean field - 1 holiday / 0 non holiday*  
"is\_weekend" - *boolean field - 1 if the day is weekend*  
"season" - *category field meteorological seasons: 0-spring ; 1-summer; 2-fall; 3-winter.*

"weathe\_code" category description:

|  |  |
| --- | --- |
| *code* | *meaning* |
| *1* | *Clear ; mostly clear but have some values with haze/fog/patches of fog/ fog in vicinity* |
| *2* | *scattered clouds / few clouds* |
| *3* | *Broken clouds* |
| *4* | *Cloudy* |
| *7* | *Rain/ light Rain shower/ Light rain* |
| *10* | *rain with thunderstorm* |
| *26* | *snowfall* |
| *94* | *Freezing Fog* |

# Data visualization：

## **Hot code**

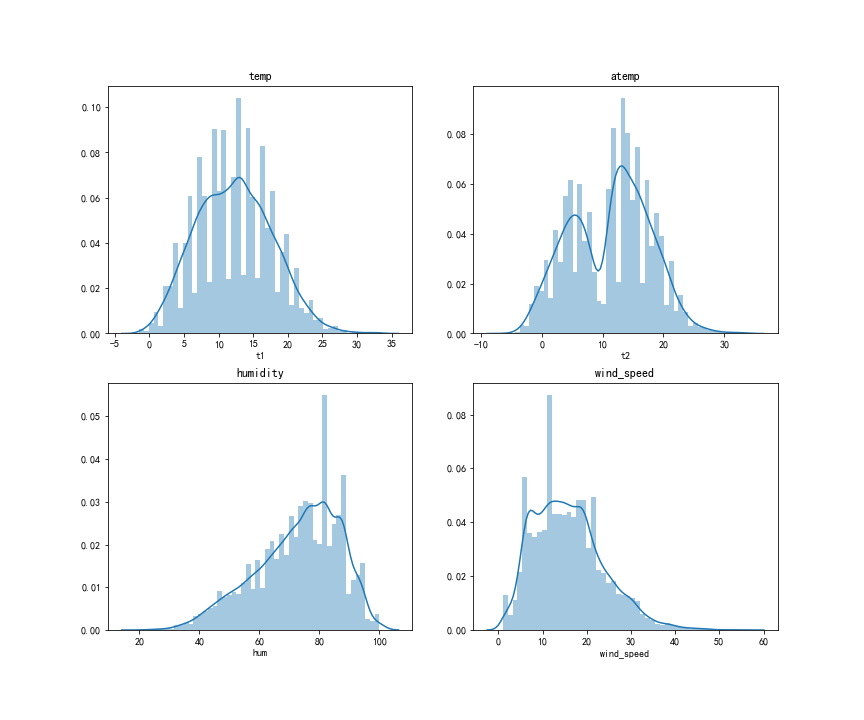
* ***Features of hot code:***

weather\_code, season. Eight kinds of weather are transformed into eight columns, each of which is a boolean variable, and the original value corresponds to one column. Four kinds of sea son homology.

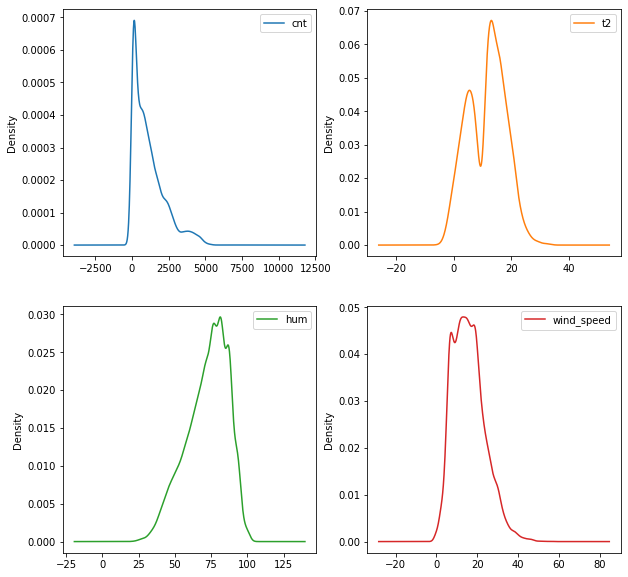
## **Visualization**

* ***Figure 1****:*

*Real temperature \ temperature "feels like" \ humidity \ wind speed ‘s distribution*

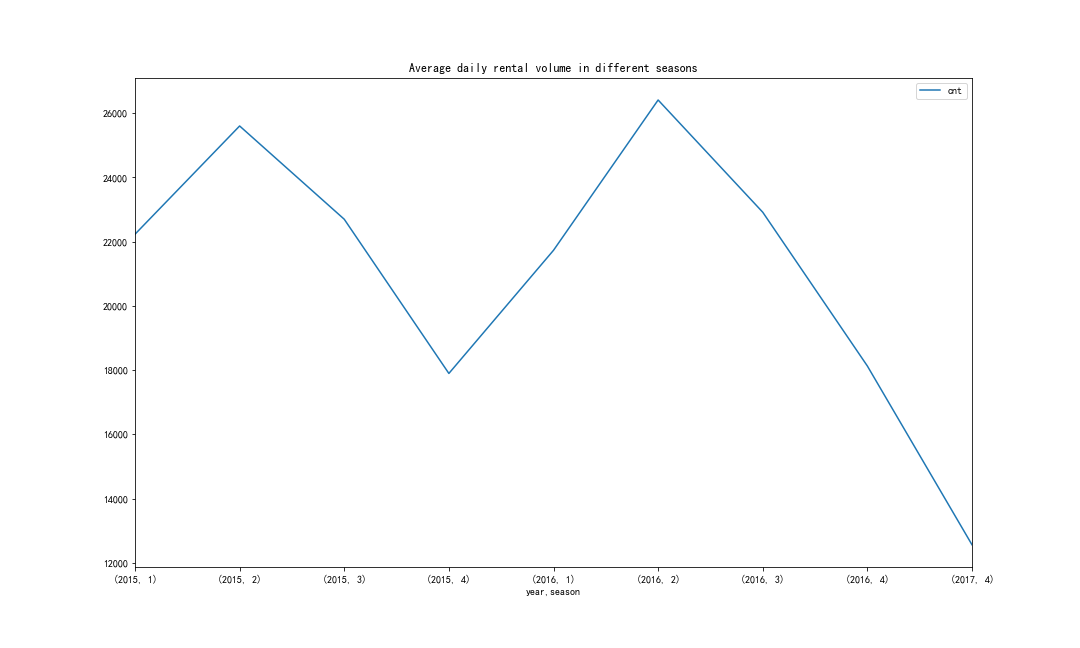
* **Figure 2**:

*The count of a new bike shares \ temperature "feels like" \ humidity \ wind speed ‘s density figure.*



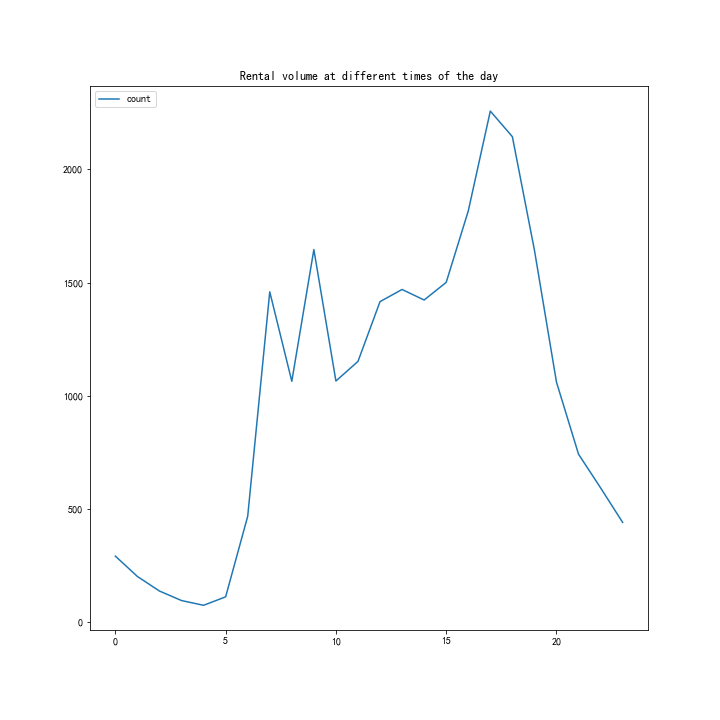
* ***Figure 3****:*

*Average daily rental volume in different seasons. It can be seen that the number of people who use shared bicycles is the most in summer and the least in winter*

**

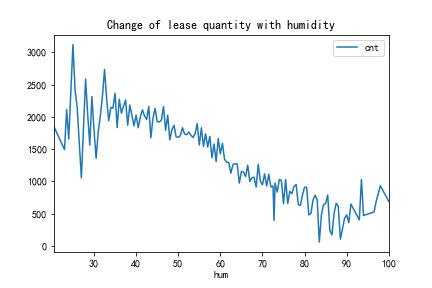
* ***Figure 4****:*

*Rental volume at different times of the day. The time periods with more number of shared bicycles are morning peak, evening peak and afternoon.*

**

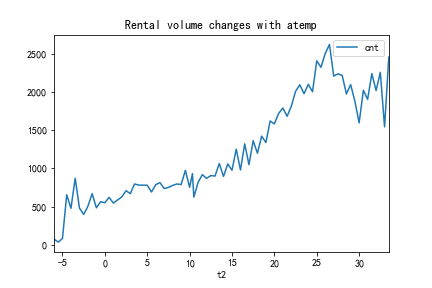
* ***Figure 5****:*

*Change of lease quantity with humidity. As shown in the figure, when the humidity is 15-50, the number of shared bikes is the most, about 2000.*

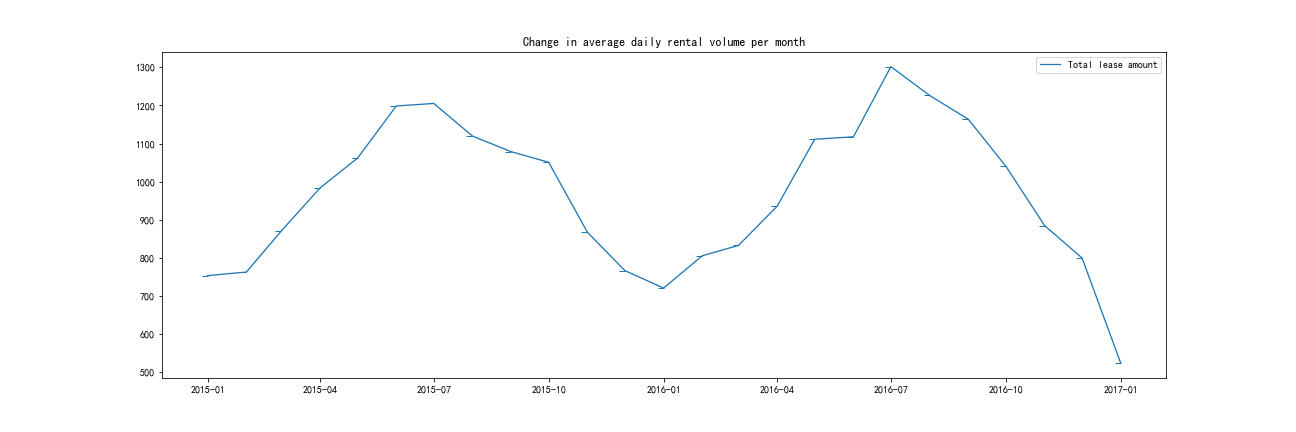
**

* ***Figure 6****:*

*Rental volume changes with “feel like” temerature. When the somatosensory temperature is about 20-30 ℃, the number of shared bikes is the highest, between 2000-2500. When the somatosensory temperature is about 0 ℃, the number of shared bikes borrowed is the least, only about 500.*

**

* ***Figure 7****:*

*Change in average daily rental volume per month. From the perspective of the number of loans in different months of a year, the number of loans in July and August is the most, reaching 1200-1300 on average, while the number of loans in January is the least, only around 750. The total lending volume of a year is an unsmooth parabola. It starts to increase from the lowest in January to the highest in July / August, and then starts to decrease.*

## Discovery:

1. *Compared with the natural temperature (T1), when the body temperature (T2) is more suitable for sharing a bicycle, more people come to borrow it.*
2. *On one day, there are more people coming to borrow and share bicycles in the afternoon, morning and evening peak meetings.*
3. *In a year, in summer, the number of people who borrowed and shared bicycles was the most, and in winter the least. Judging from the body temperature, it was too cold in winter, so people were reluctant to ride bicycles.*
4. *On weekends and holidays, the number of people who borrowed shared bicycles was not as large as that on weekdays. It can be seen that there are many office workers who borrow and share bicycles on weekdays.*

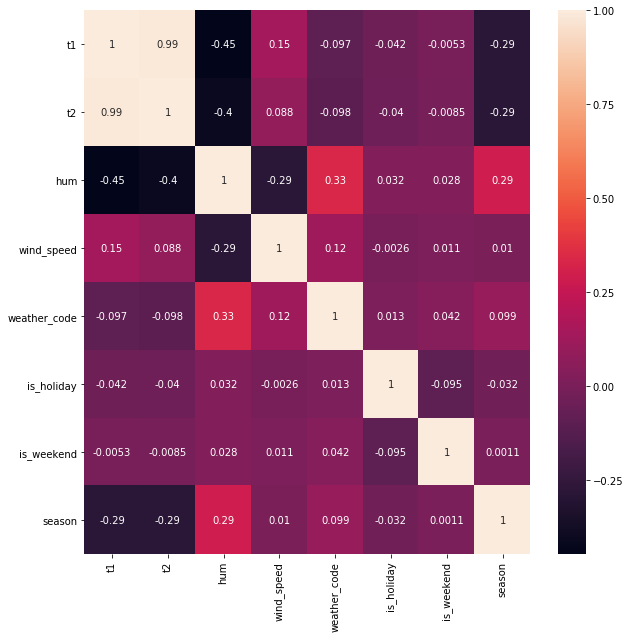
## Redundancy:

* ***Figure 8****:*

*Heat map, reflecting the correlation between features, the number is close to 1, the correlation is large, the number is close to 0, the correlation is small. It can be seen from the figure that T1 and T2 are highly correlated and need to be processed later.*

*Figure 9: Scatter diagram between two features. It can also be seen from this figure that T1 and T2 have a strong correlation and need to be processed later.*

*Figure 10: Firstly, we compare the covariance of T1 and T2, T2 is larger, which shows that T2 is more variable. By comparing the two box graphs, we find that T2 has less outliers and the range of data changes is larger than T1, so we think T2 has better research value than T1, so we delete T1 and retain T2.*



# Challenges:

1. Data preparing

For the data in the data set, it is necessary to draw the analysis chart from different seasons, different time of day, different humidity, different temperature and other conditions according to self exploration. The perspective from which thinking can be analyzed is challenging.

1. Abstract useful features

After the single hot coding, the number of feature columns has increased by 10. In the case of so many features, which features are more representative and contribute more to model training, there is no standard answer.

1. Low accurancy

After selecting the model to train the data set, the error value is large and the accuracy of prediction is low.

1. Graphical Interface

We didn't know anything about GUI programming before.

# Solution

1. Try to draw a large number of graphs from various angles, and have a deep understanding of the data set.
2. Search the Internet to learn how to choose proper features, and learn how to use Cross validation to abstract proper features.
3. Adjust the parameters of the model and try a variety of data processing methods
4. Learn from Internet.

# Improve

1. Try other ways to abstract proper features to train our model, compared with the models now to imporve accurancy.
2. Use tensorflow to train a model. Compared with the training effect of artificial selection, the training effect of non artificial selection is better.
3. In fact, the number of shared bikes borrowed actually obeys the **power-law distribution**, which is considered as a normal distribution in the study, resulting in low prediction accuracy. In order to improve the accuracy of power-law distribution, we need to study the discretization method, outlier processing method, model selection method and other related knowledge of continuous numerical value.