

# DATA MINING OF SHARED BIKE DATASET

Lu Han

University of Wales Trinity St. David (UNITED KINGDOM)  
1804740@student.uwtsd.ac.uk



Prifysgol Cymru  
Y Drindod Dewi Sant  
**University of Wales**  
Trinity Saint David

## Abstract

This is a two-year user's dataset about a shared cycling company in Washington. After pre-processing the dataset and change it to the corresponding appropriate format, the data is imported into Tableau and Weka for further data pre-processing and analysis.

Firstly, analyse the factors affecting the count of users, the differences between registered users and casual users at Tableau. In Weka, apply Classification algorithm and cluster algorithm to explore factors affecting users. Apply linear regression algorithm to predict the trend of user. Combine the algorithm principle and data background knowledge to obtain corresponding results of shared bicycles.

## Dataset Description

The data set is related to 2-year usage log of a bike sharing system namely Capital Bike Sharing (CBS) at Washington, D.C., USA, Including two file: day.csv/hour.csv.

Numbers of instance: 739(day.csv) + 17389(hour.csv)

Numbers of attribute: 16(day.csv)/17(hour.csv)[1]

## Attribute Information & Primary-Pre-processing

NOTE: Blue represents changed name or processed attribute which was added to the left of original attribute at Excel.

### Time:

Instant: Number of days since the 01.01.2011 (the first day in the dataset).

Dteday (MM/DD/YYYY) [date](#); hr {0-23} [hour](#); Mnth {1-12} [month](#);

Yr{0, 1}:2011/2012 [year](#)

### Time Meta-Information:

Holiday {0, 1}:1= [a holiday](#); 0= [not a holiday](#).

Working Day {0, 1}:1= [a working day](#); 0=[not a working day](#).

Season {1-4}:2= [spring](#); 3=[summer](#); 4=[fall](#); 1=[winter](#).

### Weather:

Weathersit {1-4} : The weather situation 1=['clear'](#); 2 =['mist'](#); 3 = ['light'](#); 4 = ['heavy'](#)

Temp& Atemp– normalized [0, 1] [Temperature/Feeling temperature in degrees Celsius](#)

Hum– normalized [0, 1] [Humidity: Relative humidity in percent \(0 to 100\)](#)

Windspeed – normalized [0 1] [Wind speed in km per hour](#)

### Users:

Casual; Registered; Cnt ->[total](#)

## Data in Tableau

By trying various type of data combination at different time level, the following results are obtained:

- Registered users account for the vast majority of total users and always in sync with the total user
- The users in 2012 increased significantly compared to 2011.
- The users peaked in the summer and reached the valley in January.
- On weekends, casual users are about twice as long as workdays.
- The total users has no obvious periodicity.
- Temperature is a key factor affecting the count of users.

## Data in Weka: Investigate Factors Affecting Users

### Secondary pre-processing

Remove attribute which are useless, including instant, year, casual and registered.

Apply filters

- Discretize- categorize total into discrete values
- ClassBalancer-re-weights instances so more evenly spread
- AddCluster :create a new attribute(cluster)

### Apply Classification Algorithm

Accuracy rate: J48>Naïve Bayes >Zero0

J48 is not a suitable algorithm which classification result is too complex.

### Apply K-means Cluster Algorithm

When K=4. Incorrectly clustered instance: 1.9125%.

- The newly added cluster consists of multiple data such as season, temperature, humidity and wind speed, which proves that other factors have little effect.
- Season is the most important factor affecting users.
- It is also possible that other attributes are positively related to the season.

When K=5. Incorrectly clustered instance: 28.875%.

- Summer clustering works best
- The clustering in spring is not so good

## Data in Weka: Predict Count of Users

### Secondary pre-processing

Leave only two Attribute: date& total

Add new attributes about date to create different version of dataset [2]:

- NewDate: Expression: (a1/(1000\*60\*60\*24\*365.25)-41)\*12
- Lag(delay): copy&TimeSeriesTranslate
- RemoveRange
- Classification

### Apply Linear Regression Algorithm

- Using different versions of dataset, the results are quite different.
- Adding this delay attribute gives cyclic prediction.

### Compare 1,6,12 Step-Ahead Forecast

- 1 step-ahead prediction line is the closest one to actual line.
- The more steps, the worse the prediction results.

### Periodicity

The count of total users changes in a cycle of 7 days.

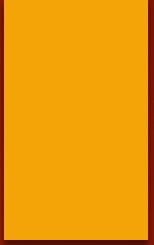
## Further Study

- Further optimization of algorithm  
Example: K-means  
(How to choose k value? /Distance calculation method?)
- Combine external sources:  
Example: Explore the impact of major events on shared bikes.
- Usage by hour:  
Example: Find peaks and valleys of users at the day.

## References

[1] Fanaee-T, Hadi, and Gama, Joao, 'Event labeling combining ensemble detectors and background knowledge', Progress in Artificial Intelligence (2013): pp. 1-15, Springer Berlin Heidelberg

[2] "University of Waikato," [Online]. Available: <https://www.cs.waikato.ac.nz/ml/weka/mooc/dataminingwithweka/>.



# DATA MINING OF SHARED BIKE DATASET

Lu Han

[1804740@student.uwtsd.ac.uk](mailto:1804740@student.uwtsd.ac.uk)

# Structure of Presentation

1. Data Description
2. Pro-processing
3. Data in tableau
4. Data in Weka-Investigate Factors Affecting Users
5. Data in Weka-Prediction
6. Results
7. Further Study



# Data Description

► Dataset source:

Laboratory of Artificial Intelligence and Decision Support (LIAAD),  
University of Porto

► The data set is related to 2-year usage log of a bike sharing system  
namely Capital Bike Sharing (CBS) at Washington, D.C., USA.

|                            |               |                       |                                   |
|----------------------------|---------------|-----------------------|-----------------------------------|
| Data Set Characteristics:  | Univariate    | Number of Instances:  | 739(day.csv) +<br>17389(hour.csv) |
| Attribute Characteristics: | Integer, Real | Number of Attributes: | 16(day.csv)<br>/17(hour.csv)      |
| Associated Tasks:          | Regression    | Missing Values?       | N/A                               |

# Attribute Information

## ► Time:

1. Instant :Number of days since the 01.01.2011 (the first day in the dataset).
2. Date (MM/DD/YYYY)
3. Hour {0-23}: only contains in hour.csv
4. Month {1-12}
5. Year {0, 1} either 2011 or 2012

## ► Time Meta-Information:

1. Holiday {0, 1}: 1= a holiday; 0= not a holiday.
2. Working Day {0, 1}: 1= a working day; 0=not a working day .
3. Season {1-4} : either 2= spring; 3=summer; 4=fall; 1=winter.

## ► Weather:

1. Weathersit{1-4}: The weather situation on that day
2. Temperature – normalized [0, 1]
3. Humidity – normalized [0, 1]
4. Wind speed – normalized [0 1]
5. Feeling Temperature

## ► Users:

1. Casual
2. Registered
3. Cnt: total number

# Pre-processing

- ▶ Change unreadable attribute names to full words

Yr->year

Temp->temperature

- ▶ Converts the normalized values to their original values.

Temperature/Feeling Temperature in degrees Celsius

Relative humidity in percent (0 to 100)

- ▶ Convert numbers to corresponding words.

season number1-4 to winter-autumn

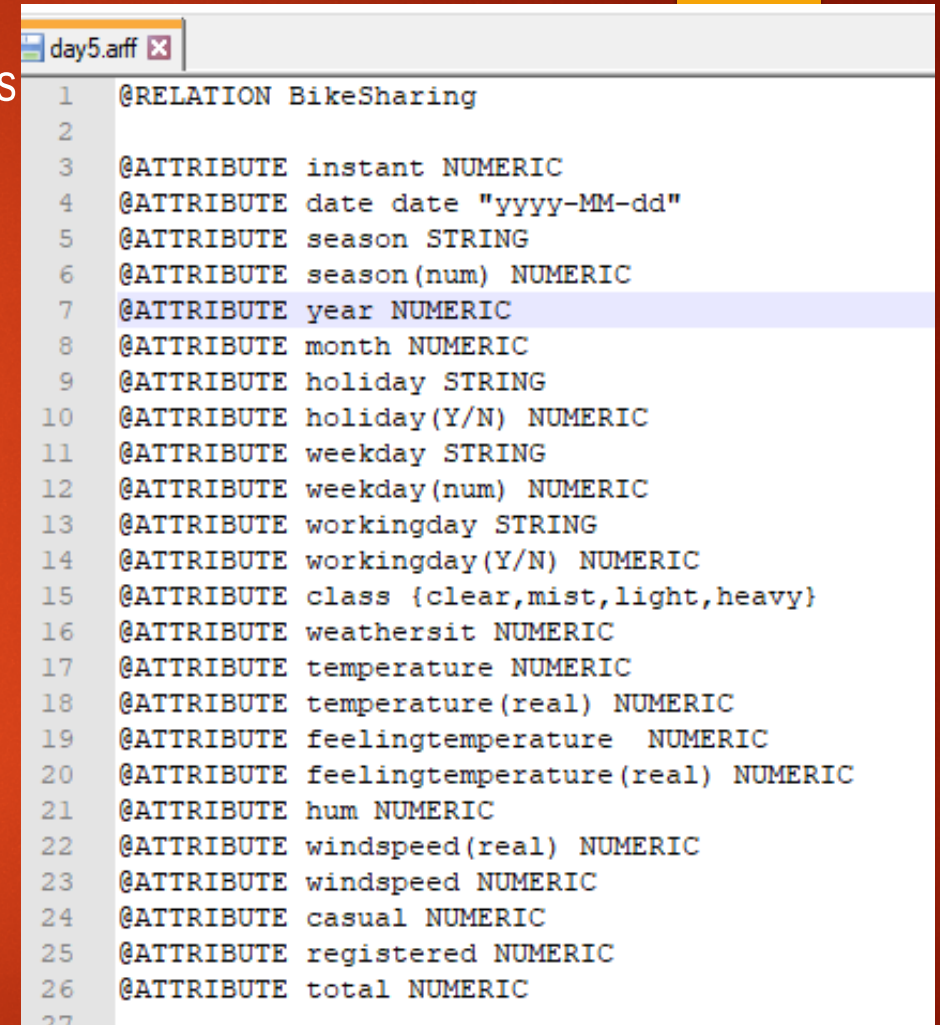
holiday 1 to 'holiday' ;0 to 'not-holiday'

weather 1 to 'clear';2 to 'mist';3 to 'light';4 to 'heavy'

- ▶ Change file format

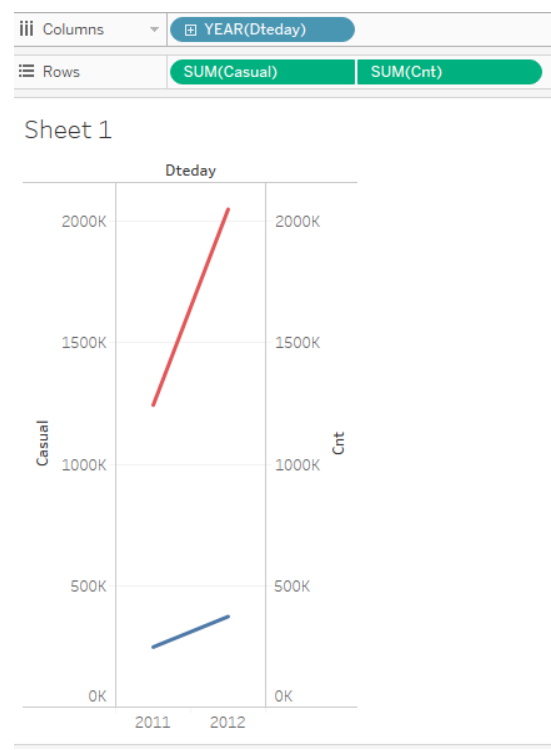
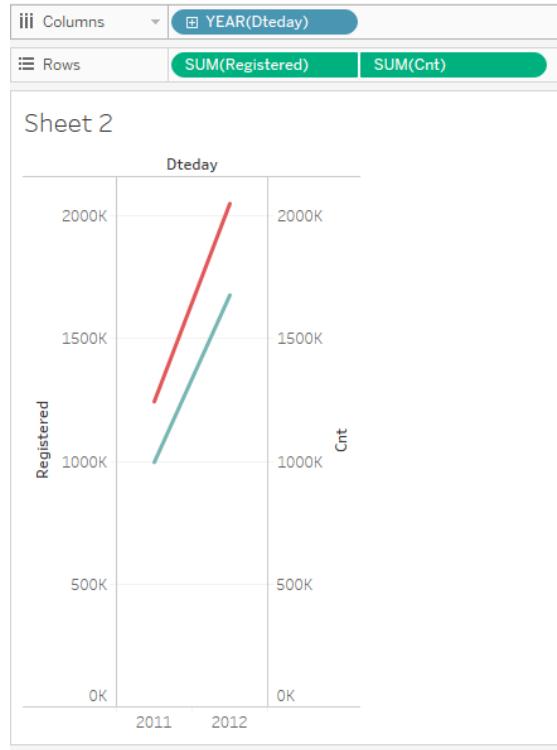
.XLSX for Tableau

.ARFF for Weka



```
day5.arff x
1 @RELATION BikeSharing
2
3 @ATTRIBUTE instant NUMERIC
4 @ATTRIBUTE date date "yyyy-MM-dd"
5 @ATTRIBUTE season STRING
6 @ATTRIBUTE season(num) NUMERIC
7 @ATTRIBUTE year NUMERIC
8 @ATTRIBUTE month NUMERIC
9 @ATTRIBUTE holiday STRING
10 @ATTRIBUTE holiday(Y/N) NUMERIC
11 @ATTRIBUTE weekday STRING
12 @ATTRIBUTE weekday(num) NUMERIC
13 @ATTRIBUTE workingday STRING
14 @ATTRIBUTE workingday(Y/N) NUMERIC
15 @ATTRIBUTE class {clear,mist,light,heavy}
16 @ATTRIBUTE weathersit NUMERIC
17 @ATTRIBUTE temperature NUMERIC
18 @ATTRIBUTE temperature(real) NUMERIC
19 @ATTRIBUTE feelingtemperature NUMERIC
20 @ATTRIBUTE feelingtemperature(real) NUMERIC
21 @ATTRIBUTE hum NUMERIC
22 @ATTRIBUTE windspeed(real) NUMERIC
23 @ATTRIBUTE windspeed NUMERIC
24 @ATTRIBUTE casual NUMERIC
25 @ATTRIBUTE registered NUMERIC
26 @ATTRIBUTE total NUMERIC
27
```

# Data in tableau

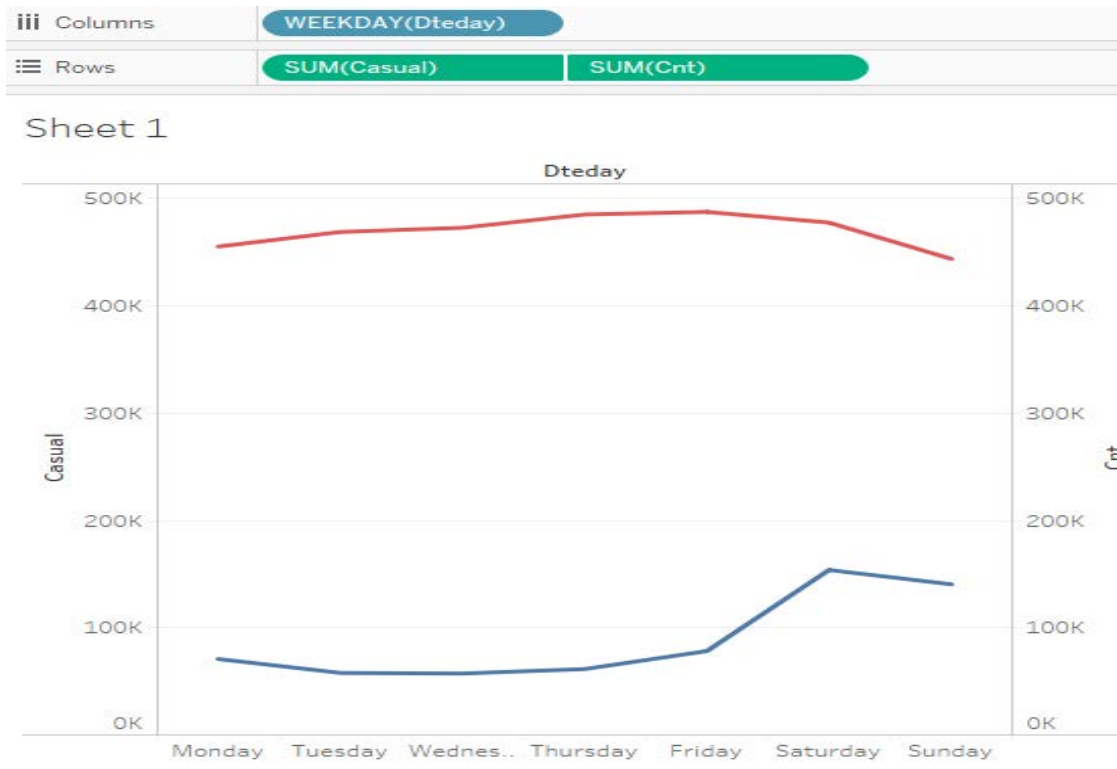


## Relationship Between Registered/Casual and Total

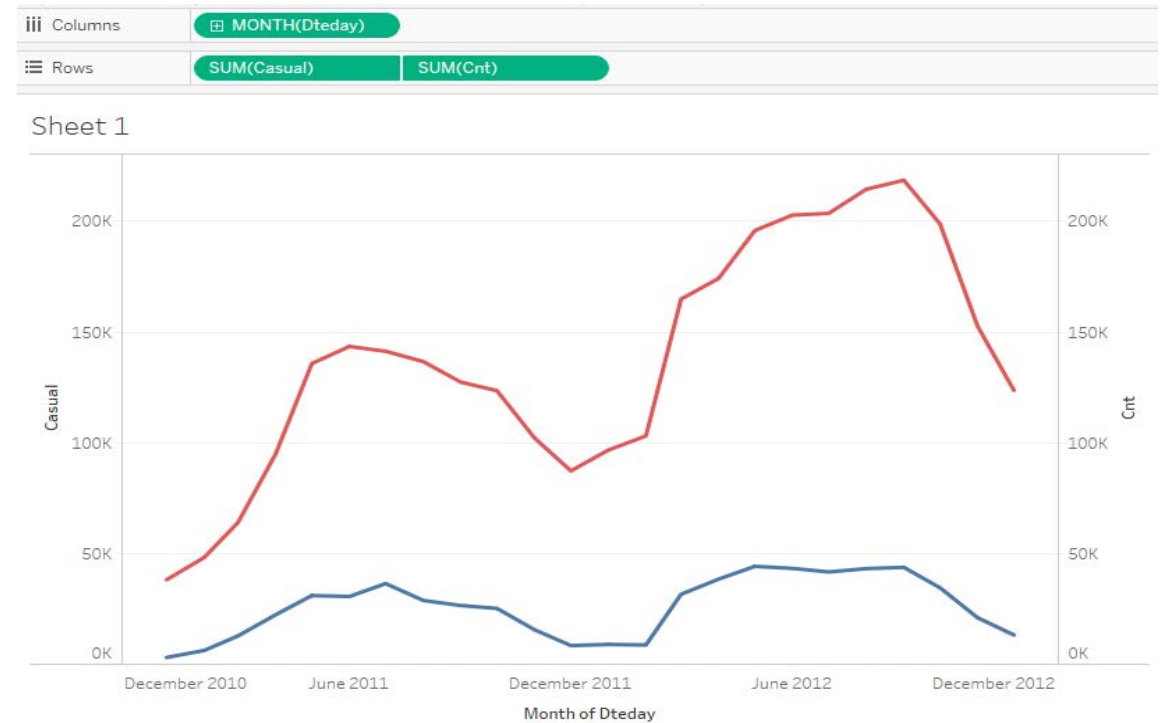
- ▶ Registered users are the vast majority, and few casual users.
- ▶ The count of shared bicycle users increased significantly from 2011 to 2012.
- ▶ Registered users are almost parallel to that of total users.

# Data in tableau

On weekends, casual users are about twice as long as workdays.



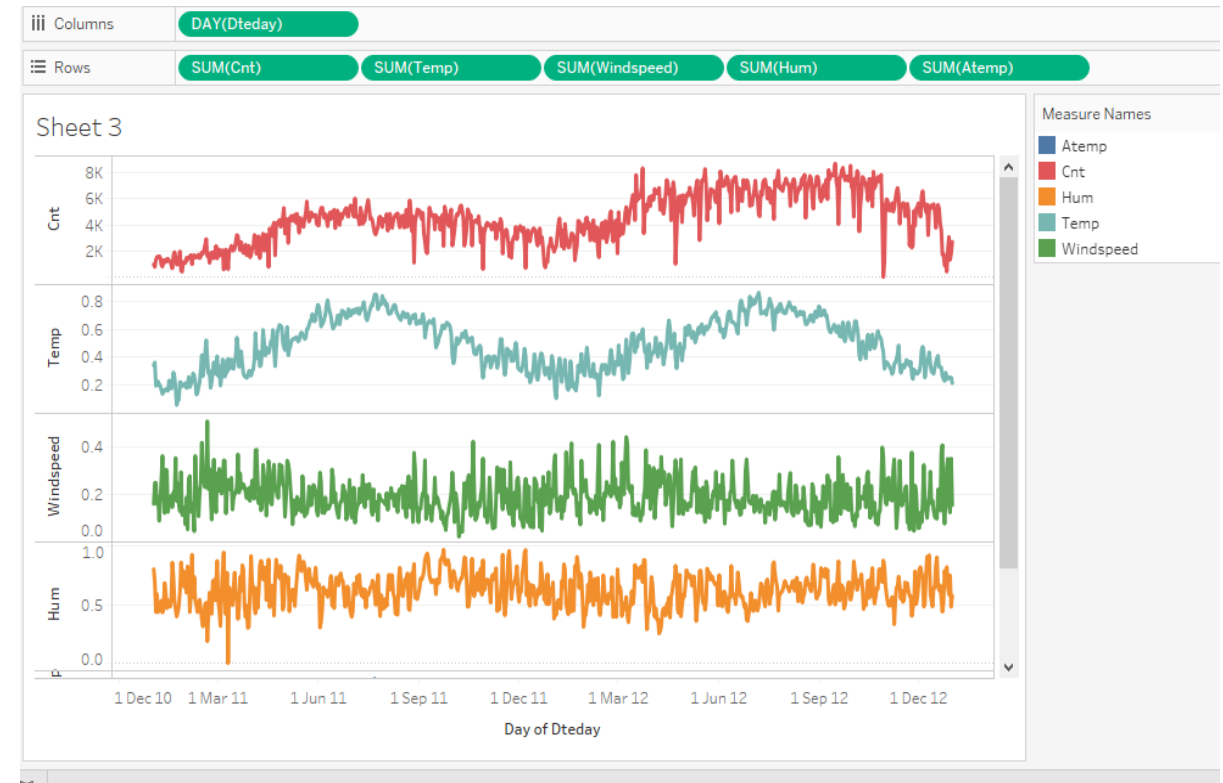
The users peaked in the summer and reached the valley in winter.



# Data in tableau

- Compare temperature/ wind speed/ humidity with users:

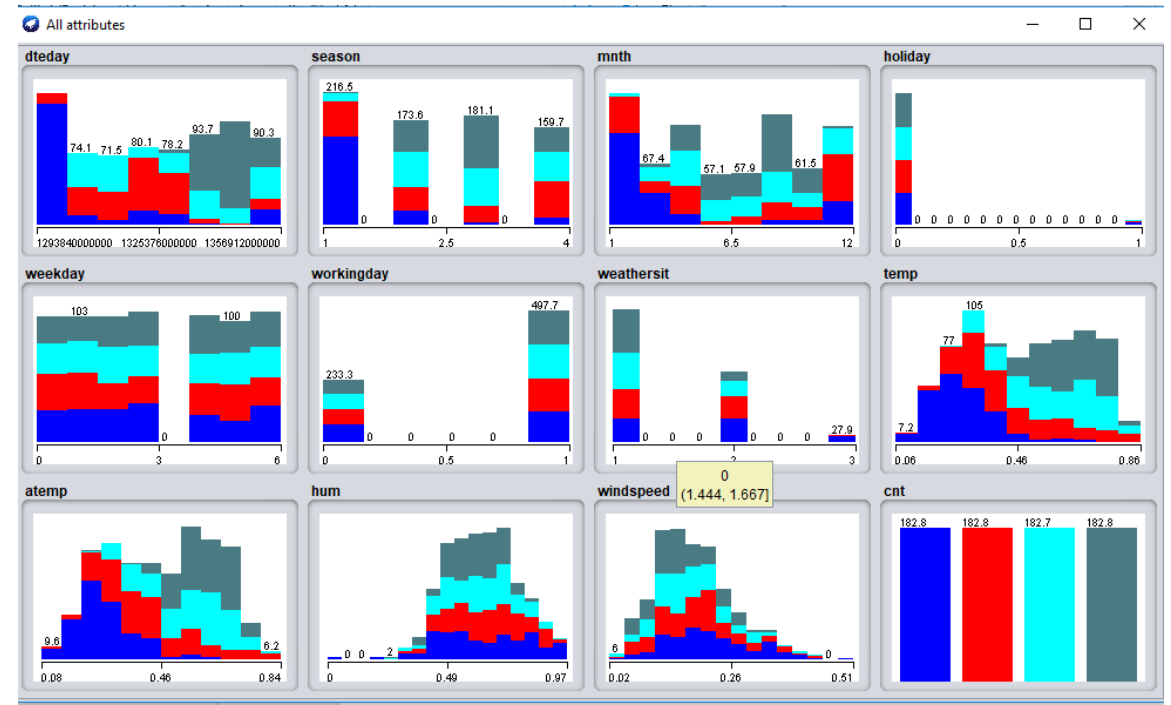
Temperature is a key factor affecting the number of users.



# Data in Weka

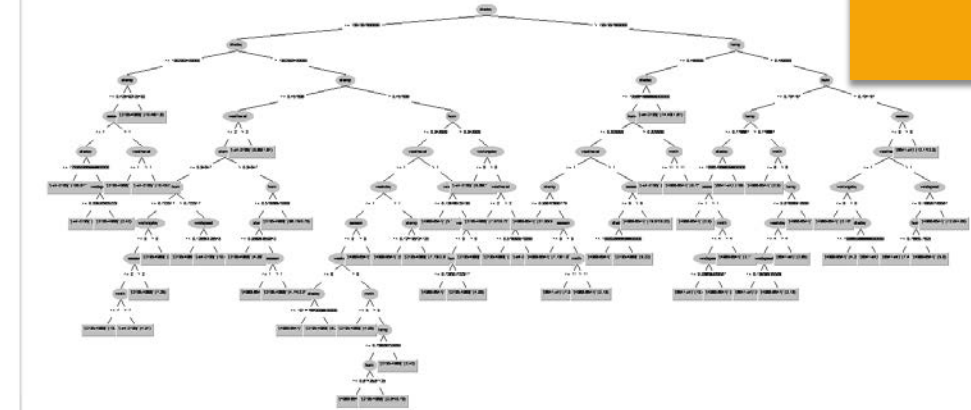
## -Investigate Factors Affecting Users

- ▶ Remove attribute which are useless, including instant, year, casual and registered.
- ▶ Apply filters
  - ▶ Discretize- categorize total into discrete values
  - ▶ ClassBalancer-re-weights instances so more evenly spread



# Apply J48 Classify Algorithm

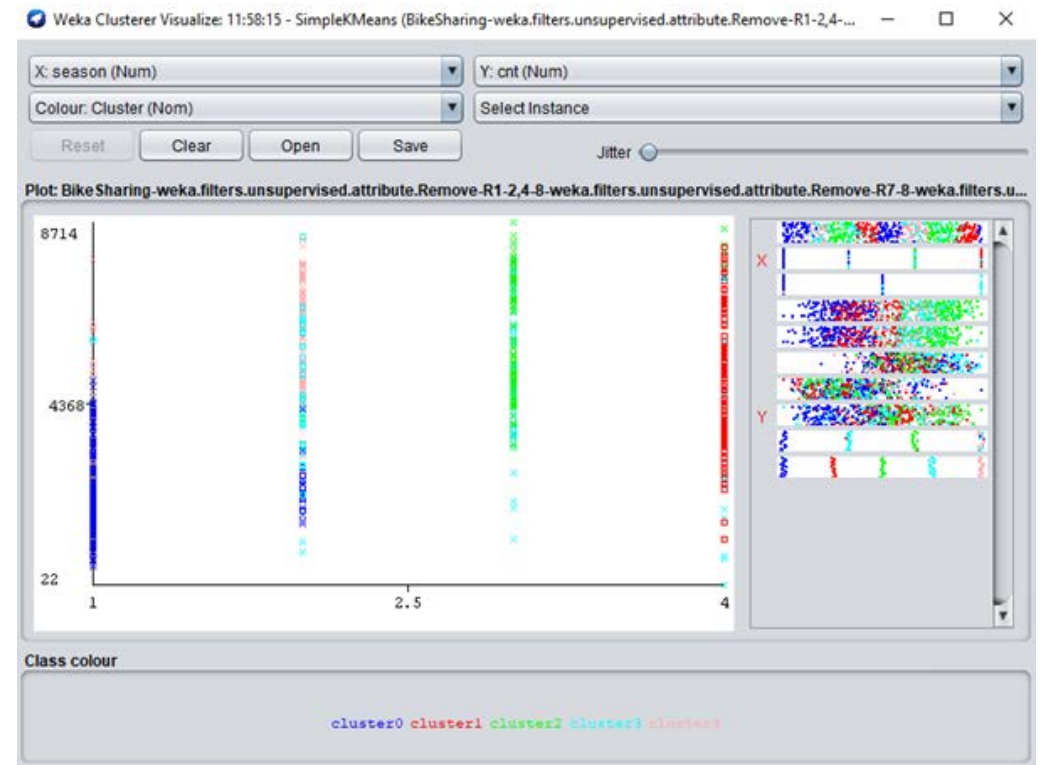
- ▶ It is not a good classification result which is too complex.
- ▶ Performance on the first and fourth class is better than on the second and third class.
- ▶ The accuracy rate is 74.6%, not bad.
- ▶ Further exploration
  - ▶ Trying to delete the attribute does not simplify the J48 tree
  - ▶ Zero0(25%): worst
  - ▶ Naive Bayes



| Class         | Precision | Recall | F-Measure |
|---------------|-----------|--------|-----------|
| '(-inf-2195]' | 0.842     | 0.851  | 0.803     |
| '(2195-4368]' | 0.652     | 0.662  | 0.551     |
| '(4368-6541]' | 0.622     | 0.655  | 0.638     |
| '(6541-inf)'  | 0.828     | 0.826  | 0.827     |
| Weighted Avg  | 0.746     | .0.744 | 0.744     |

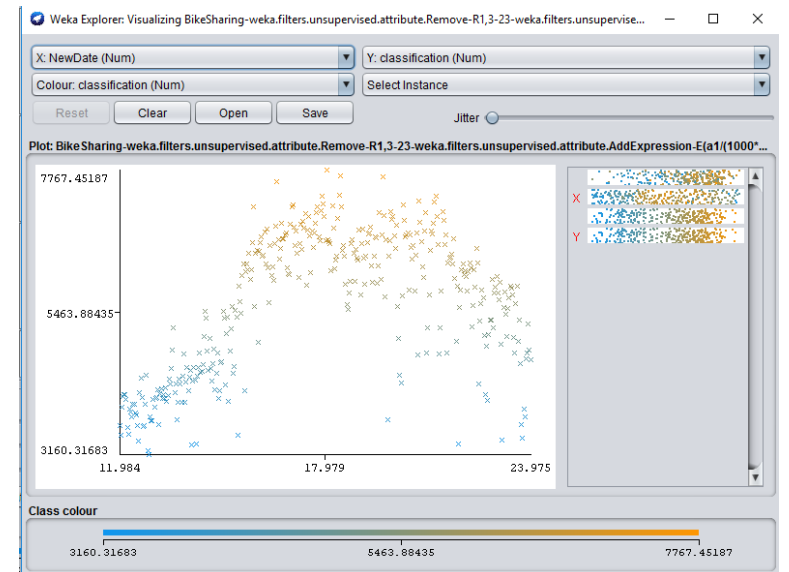
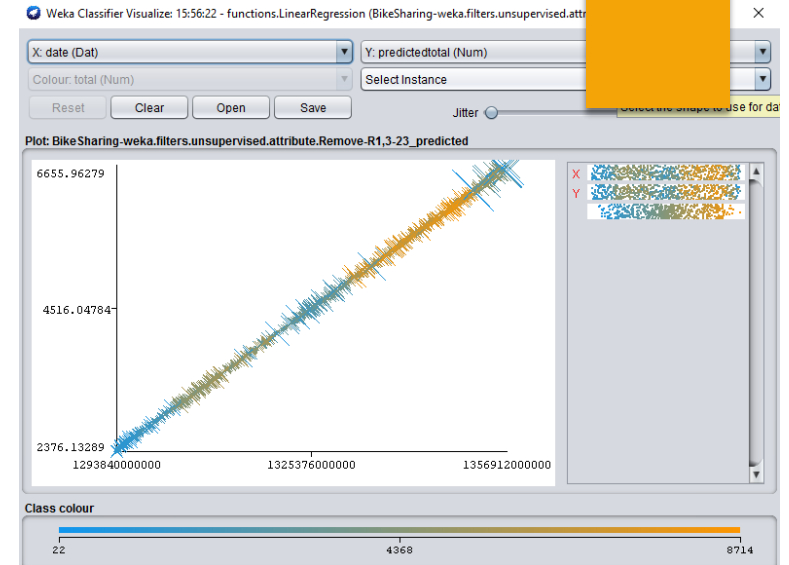
# Apply K-means Cluster Algorithm

- ▶ Use the AddCluster to create a new attribute: cluster
- ▶ When K=4. Incorrectly clustered instance: 1.9125%.
  - ▶ The newly added cluster consists of multiple data such as season, temperature, humidity and wind speed, which proves that other factors have little effect.
  - ▶ Season is the most important factor affecting users.
  - ▶ It is also possible that other attributes are positively related to the season.
- ▶ When K=5. Incorrectly clustered instance: 28.875%..
  - ▶ Summer clustering works best
  - ▶ The clustering in spring is not so good



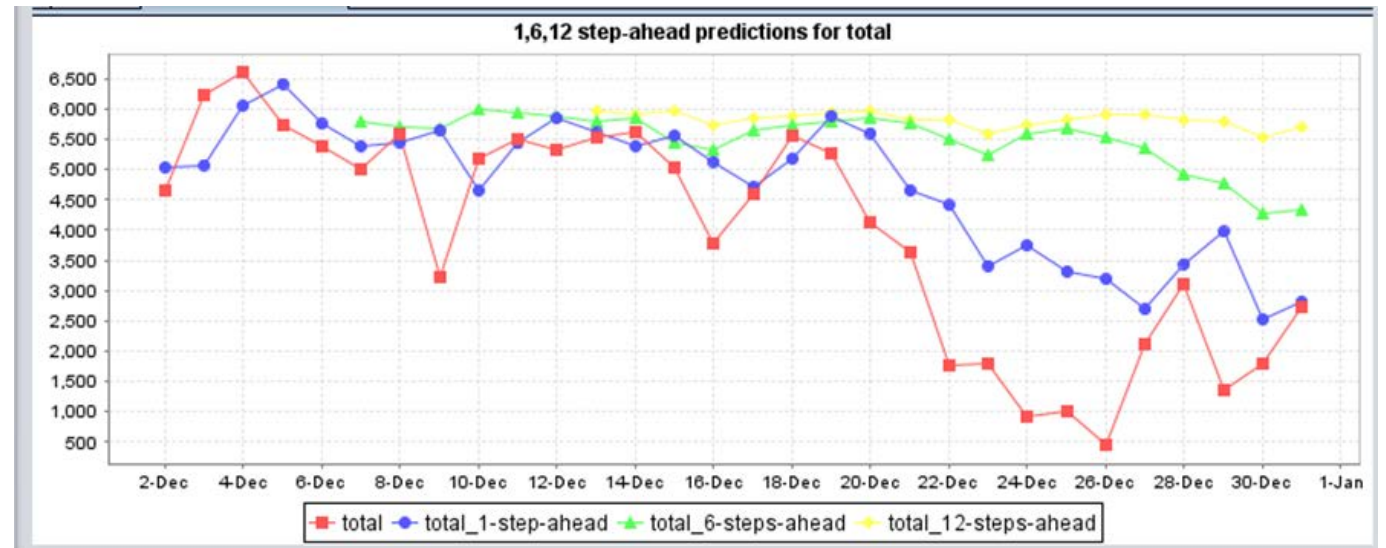
# Data in Weka- Prediction

- ▶ Leave only two Attribute : date& total
- ▶ Algorithm: Linear regression
- ▶ Create five Version of dataset:
  1. Original
  2. NewDate: Expression:  $(a1/(1000*60*60*24*365.25)-41)*12$
  3. Lag(delay): copy&TimeSeriesTranslate
  4. RemoveRange
  5. Classification
- ▶ The size of these crosses indicates the size of the error at this point.
- ▶ Using different versions of data, the predictions are quite different.
- ▶ Adding delay attribute gives cyclic prediction.



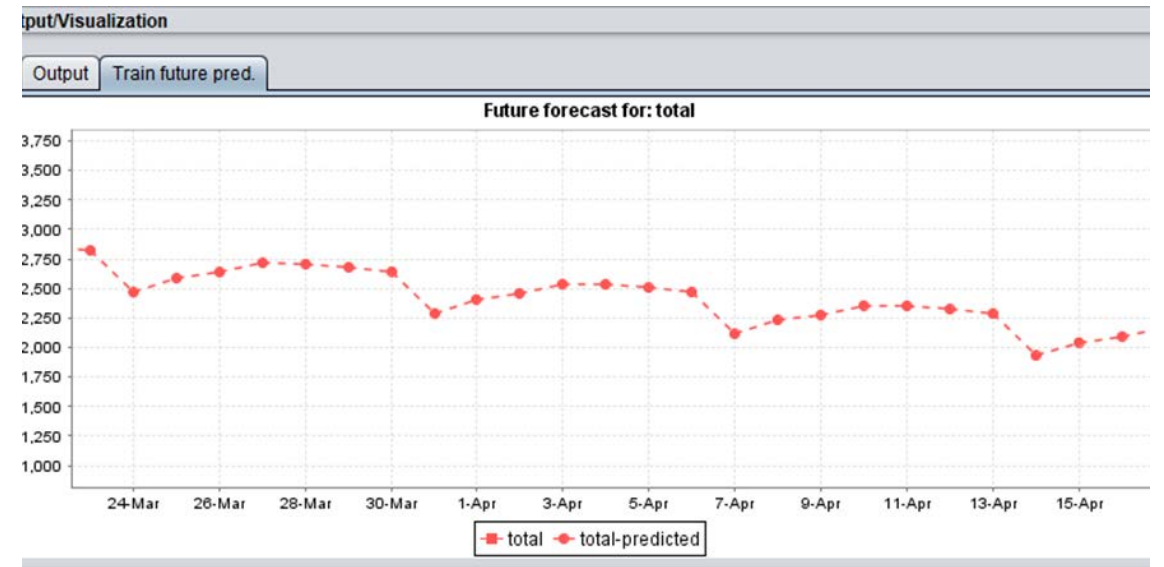
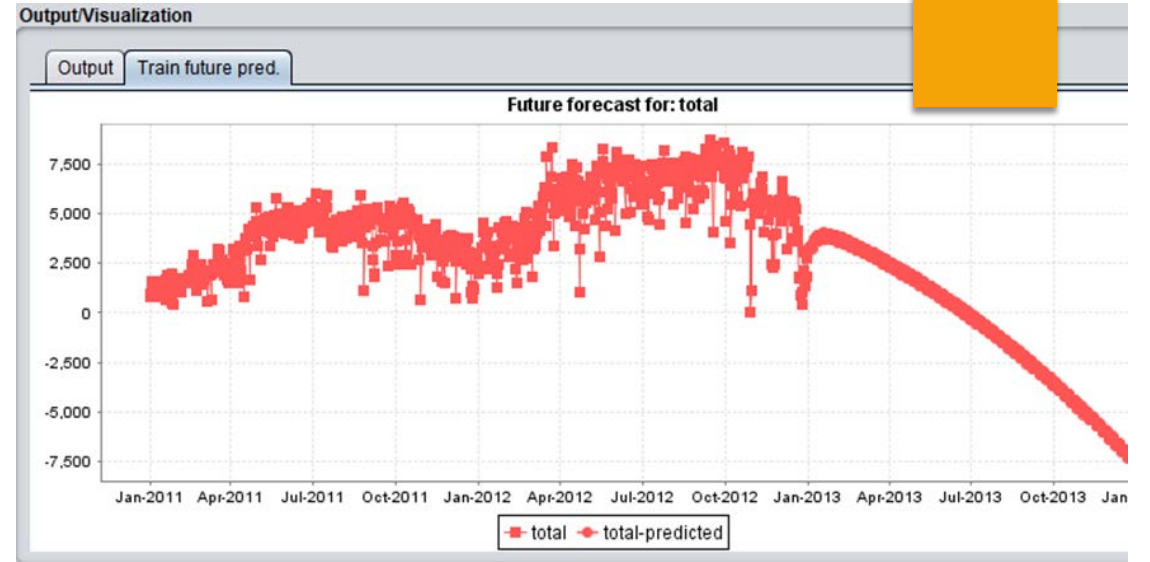
# Compare 1,6,12 Step-Ahead Forecast

- ▶ Advanced configuration at Forecast panel
  - ▶ Base learner: choose filter classifier
  - ▶ Classify: linear regression
  - ▶ Filter: remove: 1,4,16 invert
- ▶ 1 step-ahead prediction line is the closest one to actual line.
- ▶ The more steps, the worse the prediction effect.



# Periodicity

- The count of total users changes in a cycle of 7 days.



# Results

- ▶ Registered users account for the vast majority of total users and always in sync with the total user
- ▶ The shared bicycle users in 2012 increased significantly compared to 2011.
- ▶ The number of users peaked in the summer and reached the valley in winter.
- ▶ On weekends, casual users are about twice as long as workdays.
- ▶ Temperature is a key factor affecting the count of users.
- ▶ The count of users changes in a cycle of 7 days.

# Further study



## Further optimization of algorithm

Example: K-means(How to choose k value?/Distance calculation method?)



## External sources:

Example: The impact of major events on shared bikes.



## Usage by hour

Example: Find peaks and valleys of users at the day.



Thank you!