Australian
National
University

COMP3430 / COMP8430
Data wrangling

Lab 1: Data Exploration using R/Rattle and Python Pandas

Objectives of this lab

- Learn about different data exploration functions that can be used on different data sets.

- Get familiar with the graphical user interface of the open source data wrangling and mining tool **Rattle**, and the **Python Pandas** library.

- Conduct data exploration using Rattle and Pandas on small example data sets.

Outline of this lab

- Data exploration using Rattle – part one (1 hour)
- Data exploration using Pandas – part two (1 hour)
- Summary

Preliminaries

- Go through the lab tutorial pdf document provided under week 3.
- Ensure you already have Rattle and the Python Pandas library installed in your machine.
- For more on R and how to download Rattle and Pandas please see the **Course Resources** link in the Course page in Wattle.
- For this lab we will be using Pandas installed as part of the open source data science package anaconda.

- In this lab you do not have a specific list of tasks to do, but rather some possibilities and questions we encourage you to investigate.

Data exploration using Rattle

- Part one of the lab basically consists of working with Rattle and load and explore a small example data set.

- Rattle is a freely available software tool that provides a graphical user interface on top of the R statistical programming language.

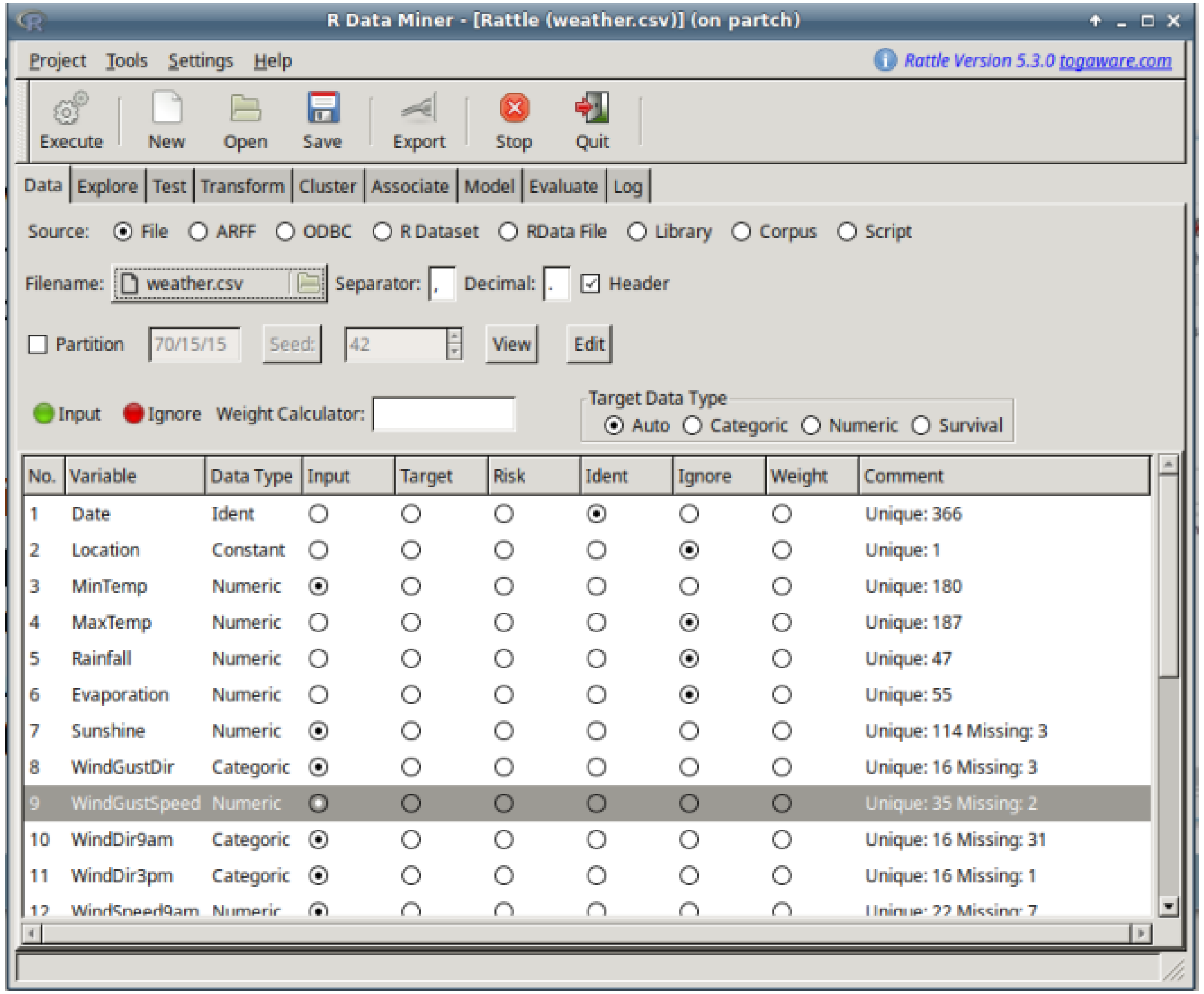
- Rattle provides access to many of the data wrangling, data mining and statistical functionalities in R.

- Before starting have a look at the Rattle documentation at http://datamining.togaware.com/survivor/Rattle_Data.html.

Data exploration using Rattle

- Follow the instructions in the tutorial document to start Rattle.

- We will be using the example **weather** data set that comes with Rattle in this lab, but feel free to use any other data sets later on.


R Data Miner - [Rattle (weather.csv)] (on partch)
Project Tools Settings Help
Rattle Version 5.3.0 togaware.com
Execute
New
Open
Save
Export
Stop
Quit
Data
Explore
Test
Transform
Cluster
Associate
Model
Evaluate
Log
View
Edit

Source: ● File ○ ARFF ○ ODBC ○ R Dataset ○ RData File ○ Library ○ Corpus ○ Script

Filename: weather.csv Separator: , Decimal: . ☑ Header

□ Partition 70/15/15 Seed: 42

🟢 Input 🔴 Ignore Weight Calculator:

Target Data Type: ● Auto ○ Categoric ○ Numeric ○ Survival

No.	Variable	Data Type	Input	Target	Risk	Ident	Ignore	Weight	Comment
1	Date	Ident	○	○	○	●	○	○	Unique: 366
2	Location	Constant	○	○	○	○	●	○	Unique: 1
3	MinTemp	Numeric	●	○	○	○	○	○	Unique: 180
4	MaxTemp	Numeric	○	○	○	○	●	○	Unique: 187
5	Rainfall	Numeric	○	○	○	○	●	○	Unique: 47
6	Evaporation	Numeric	○	○	○	○	●	○	Unique: 55
7	Sunshine	Numeric	●	○	○	○	○	○	Unique: 114 Missing: 3
8	WindGustDir	Categoric	●	○	○	○	○	○	Unique: 16 Missing: 3
9	WindGustSpeed	Numeric	●	○	○	○	○	○	Unique: 35 Missing: 2
10	WindDir9am	Categoric	●	○	○	○	○	○	Unique: 16 Missing: 31
11	WindDir3pm	Categoric	●	○	○	○	○	○	Unique: 16 Missing: 1
12	WindSpeed9am	Numeric	●	○	○	○	○	○	Unique: 22 Missing: 7

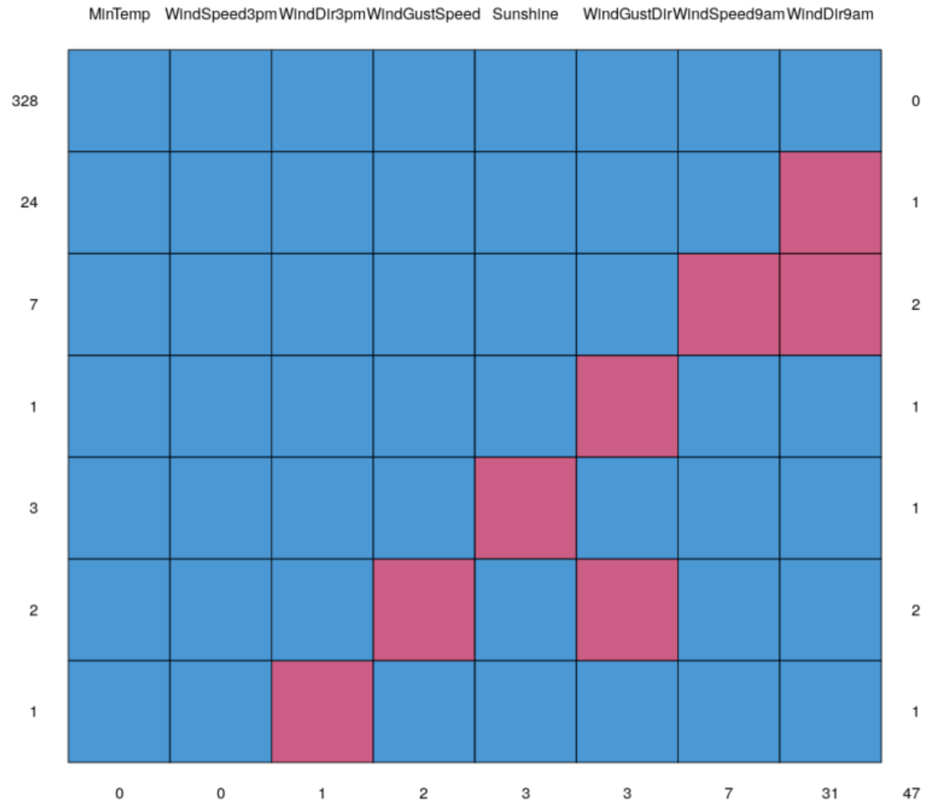
Questions to discuss

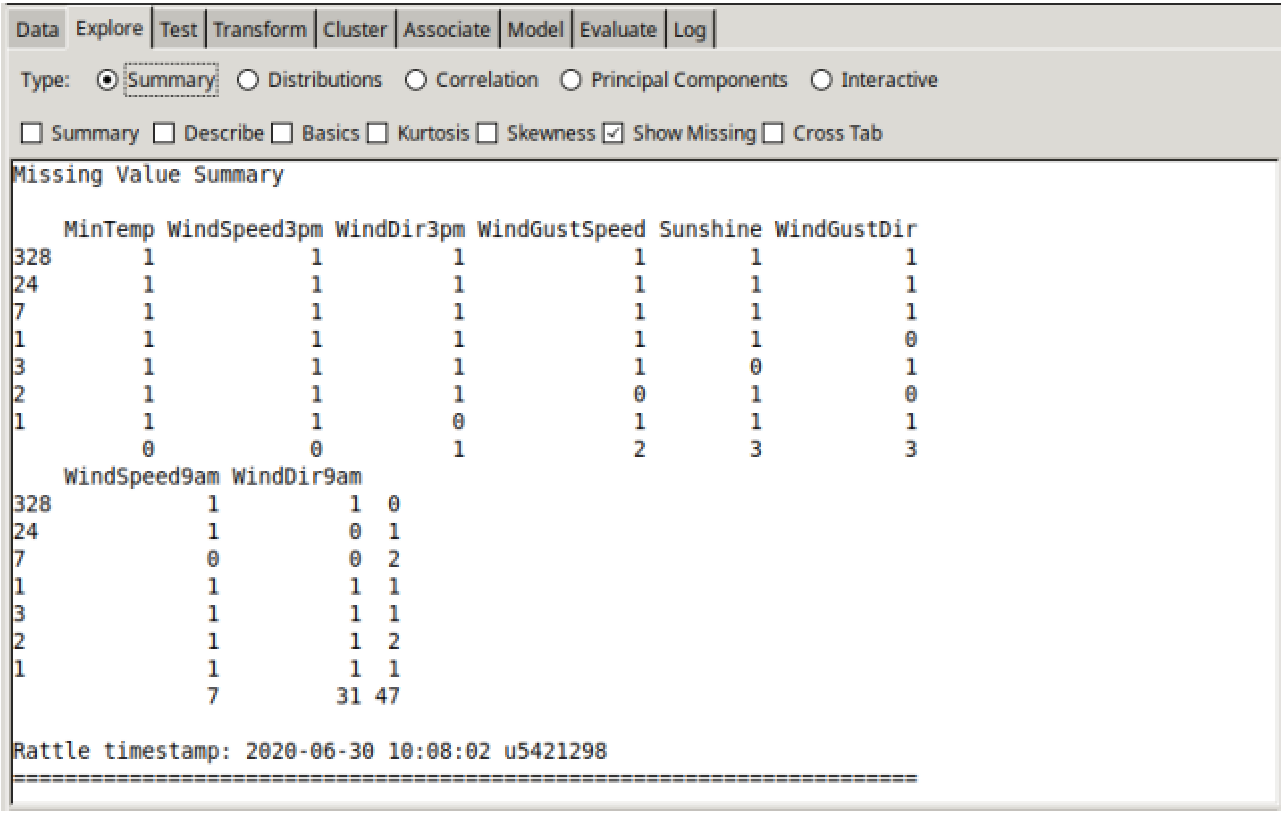
- How can you identify the correlation between attributes?
- From these correlations, can you identify the
  1. most correlated attributes
  2. least correlated attributes

- What can you learn about the distributions of the attributes
- What can you say about the skewness of these distributions
- Can you plot these distributions in Rattle

Missing data table in Rattle

- Can you describe the missing data table in Rattle?





Type: ● Summary ○ Distributions ○ Correlation ○ Principal Components ○ Interactive

□ Summary □ Describe □ Basics □ Kurtosis □ Skewness ☑ Show Missing □ Cross Tab

```
Missing Value Summary

    MinTemp WindSpeed3pm WindDir3pm WindGustSpeed Sunshine WindGustDir
328       1            1          1             1        1           1
24        1            1          1             1        1           1
7         1            1          1             1        1           1
1         1            1          1             1        1           0
3         1            1          1             1        0           1
2         1            1          1             0        1           0
1         1            1          0             1        1           1
          0            0          1             2        3           3
    WindSpeed9am WindDir9am
328            1          1  0
24             1          0  1
7              0          0  2
1              1          1  1
3              1          1  1
2              1          1  2
1              1          1  1
               7         31 47

Rattle timestamp: 2020-06-30 10:08:02 u5421298
======================================================================
```

Data exploration using Pandas

- Pandas is an open source library providing high-performance, easy-to-use data structures and data analysis tools for Python language.
- Pandas enables you to carry out your entire data analysis workflow in Python without having to switch to a more domain specific language like R.

- Remember to **start the Anacoda distribution first**.
- Follow the instruction given in the tutorial document to start Python and import Pandas.
- Similar to Rattle, we will be using the weather data set with Pandas in this lab.

Questions to discuss

- How can you identify the correlation between attributes?
- From these correlations, can you identify the
  1. most correlated attributes
  2. least correlated attributes

- What can you learn about the distributions of the attributes?
- What can you say about the skewness of these distributions
- Can you plot these distributions?
- Extra task - Can you implement a missing data table similar to Rattle using Pandas?

Summary

- In this lab we discussed how we can use Rattle and Python Pandas to explore a data set.
- We learnt how to describe a data set using its statistics and distributions.
- Also, we learnt how we can visualise distributions in a data set.

- In the next lab we will explore how we can use Rattle and Python Pandas to clean a data set and do transformations.