Australian National University

COMP3430 / COMP8430
Data wrangling

Lab 2: Data cleaning and transformation using R/Rattle and Python Pandas

Objectives of this lab

- Get familiar with the functionalities available in Rattle and the Python Pandas library that can be used to conduct data cleaning and data transformation on smaller example data sets.

- In this lab we will work on imputing missing values in attributes and learn how to perform data transformation tasks on attribute values using Rattle and Pandas.

Outline of this lab

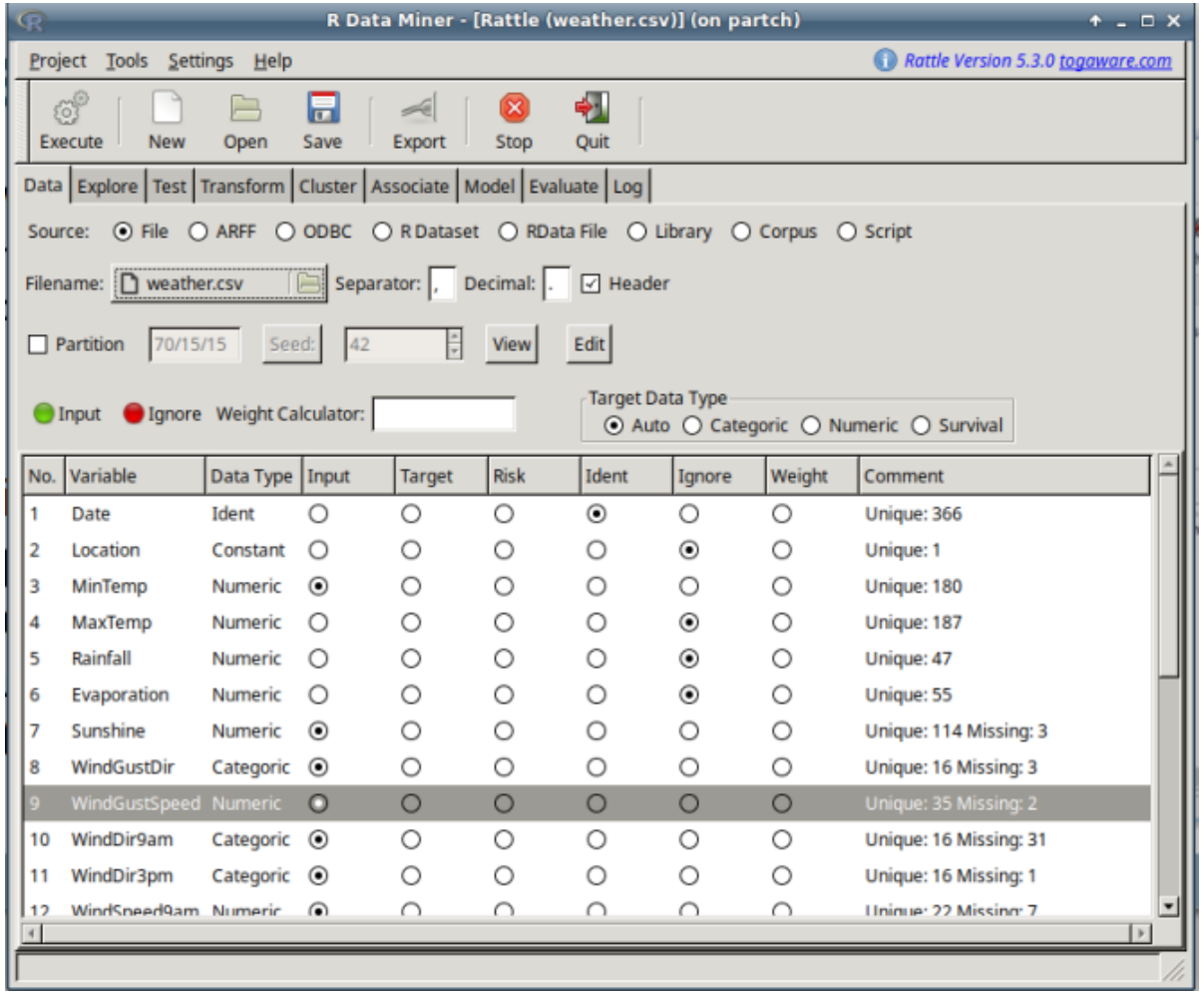
- Data cleaning and transformation using Rattle - 1 hour
- Data cleaning and transformation using Pandas - 1 hour
- Summary

Preliminaries

- **Before you begin, go back over the work from lab 1 and remind yourself about Rattle and the Python Pandas library.**
- Go through the lab tutorial pdf document provided under week 4.
- Ensure you already have Rattle and Python Pandas library running 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.

Data cleaning and transformation using Rattle

- Part one of the lab basically consists of data transformation with Rattle.
- Follow the instructions in the tutorial document to start Rattle.
- We will be using the example weather data set comes with Rattle in this lab, but feel free to use any other data sets.



Questions to discuss

- How can you identify the attributes with missing values?
- What are the different ways you can impute missing values in Rattle?

- What different transformation functions are available in Rattle?
- What transformation functions are more suitable for skewed distributions?
- Can you rank the data according to the values in an attribute?
- Can you plot these transformed distributions in Rattle and how do they differ compared to original data distributions?

Data cleaning and transformation using Pandas

- Before we start the lab questions **start the Anaconda 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.
- Follow the instruction in the tutorial document and import the necessary modules.

Questions to discuss

- How can you identify the attributes with missing values?
- What are the different ways you can impute missing values?
- What different transformation functions are available in Pandas?
- What transformation functions are more suitable for skewed distributions?
- Can you plot these transformed distributions in Pandas and how do they differ compared to original data distributions?

- Extra task - can you rank the data in a data frame according to the values in an attribute?

Summary

- In this lab we discussed how we can use Rattle and Python Pandas to clean a data set.
- We also learnt how to use different transformations in Rattle and Python Pandas on attributes.
- Also, we learnt how we can visualise distributions in a data set.

- Starting from the next lab we will be building a complete record linkage Python program.
- So it is important you familiarise yourself with basic Python programming before the next lab.