Australian National University

COMP3430 / COMP8430 Data wrangling

Lecture 5: Data quality assessment and data profiling

(Lecturer: Peter Christen)

Lecture outline

- Data quality assessment
- Data quality dimensions

- Data profiling
- Data visualisation
- Data profiling tools
- Summary

Australian National University

Data quality

- **Real-world data are dirty**, especially personal data are prone to errors
- Various sources of errors:
  - Missing data
  - Out-of-date data
  - Data variations in different sources
  - Errors/noise/outliers during data entry
  - Misinterpretation of data

- For quality analysis and mining results, data quality is important
  - **Garbage-in garbage-out principle**

Data quality assessment

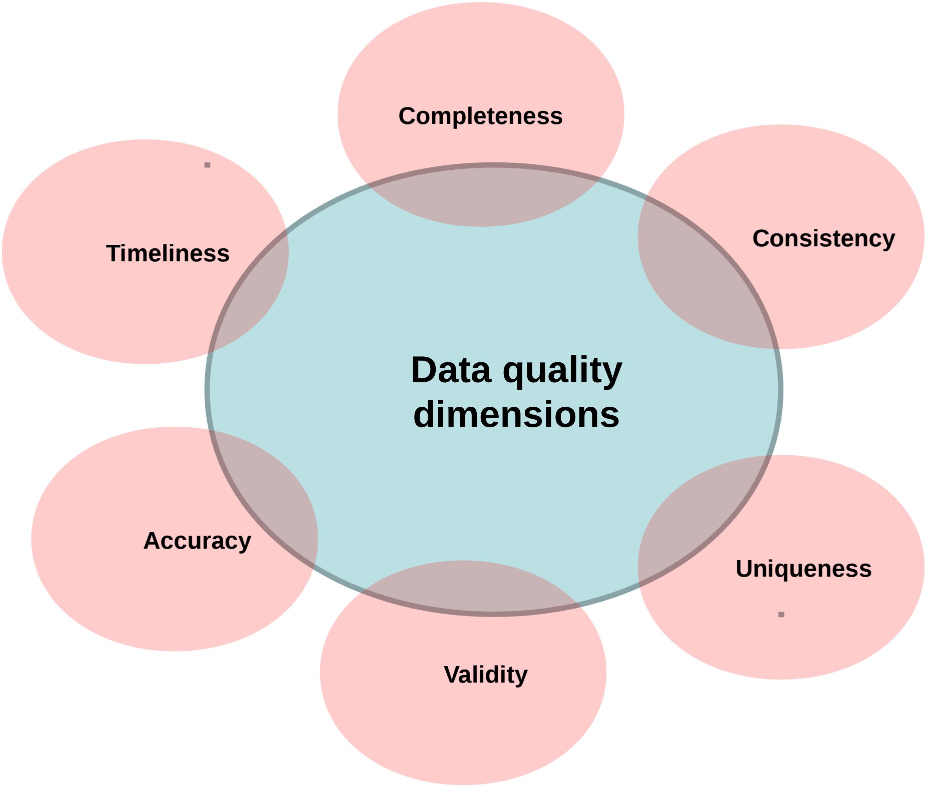
- **Data quality is specific to context**
  - High quality data for some applications may not be sufficiently good for other applications
  - Often not possible to get data of 100% quality

- Scientific and statistical evaluation of data to determine if they are adequate for intended use
- Data quality is a multi-dimensional concept
  - Both subjective and objective

Australian National University

Data quality dimensions

- Six core dimensions
  - Completeness – no missing data
  - Consistency – across different sources
  - Uniqueness – single view of dataset
  - Validity – meet constraints and rules
  - Accuracy – correct and reflect real data
  - Timeliness – no out-of-date values



The six primary dimensions for data quality assessment, DAMA UK Working Group, 2013

Data quality dimensions (2)

Dimension	Completeness
Definition	Proportion of available data against 100% complete
Measure	Percentage of missing values
Example	Emergency contact details of children in school admin database: 290 out of 300 records have the contact value yielding to 290/300 * 100 = 96.7% completeness

Data quality dimensions (3)

Dimension	Consistency
Definition	No difference between two or more representations of a record
Measure	Percentage of record representations with the same format
Example	The school admin database (300 records) and the school register database (200 records) have 400 out of 500 students' records with the same telephone numbers, resulting in 400/500 * 100 = 80% consistency

Data quality dimensions (4)

Dimension	**Uniqueness**
Definition	No duplicate records in a dataset
Measure	Ratio of number of records in a dataset and number of real entities
Example	The school admin database has 300 student records, but the number of actual students is 280, indicating a uniqueness of 280/300* 100 = 93.3%

Data quality dimensions (5)

Dimension	Validity
Definition	Data confirming to the syntax (format, type, range)
Measure	Comparison between the data and the metadata
Example	The age values in a student database must be numeric and in the range between 5 and 18; postcodes must be 4 digits containing numerical characters

Data quality dimensions (6)

Dimension	Accuracy
Definition	Degree to which data correctly describes the real entity
Measure	Percentage of accurate representations of entities in a dataset
Example	50 records of students in a database of 300 records have wrong values for postcode and/or suburb values (such as Braddon, 7612), giving 250/300 * 100 = 83.3% accuracy

Data quality dimensions (7)

Dimension	Timeliness
Definition	Degree to which data represent a real entity in a point in time
Measure	Percentage of records with up-to-date values
Example	30 out of 300 students in a student database have not updated their address change, resulting in 270/300 * 100 = 90% records with timely information

Data quality dimensions (8)

- Other dimensions
  - Usability – relevant and accessible
  - Understandability – easy to comprehend
  - Flexibility – compatible and easy to manipulate
  - Volume – appropriate amount of data for the application
  - Privacy / confidence – data protection and security
  - Value – cost/benefit of data

Data profiling

- Examining, exploring, and collecting statistics and information about data
- To determine the *metadata* about a data set
- Data profiling provides insights and allows identifying data quality requirements
  - For more thorough data quality assessment
  - A process of discovery

Data profiling versus data mining

- Data profiling
  - **Goal**: discovers information and metadata
  - **Input**: raw data
  - **Output**: information about attributes (columns)

- Data mining / analytics
  - **Goal**: discovers interesting knowledge and patterns
  - **Input**: pre-processed and cleaned data
  - **Output**: information about records (rows)

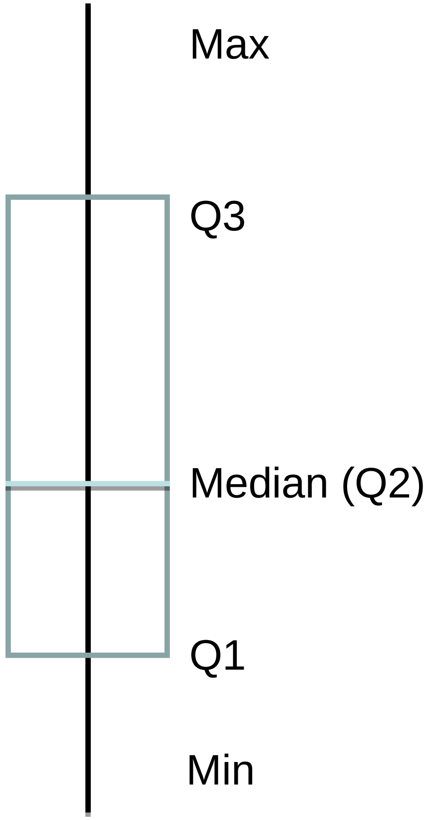
Single versus multiple column profiling

- Single column
  - Basic statistics of a single column
  - Discover common properties and statistics of a single attribute that are assumed to be of same type
  - Complexity: Number of rows
- Multiple column
  - Discover joint properties, dependencies and correlations, and statistics of multiple attributes
  - Complexity: Number of columns * Number of rows

Statistics (single column)

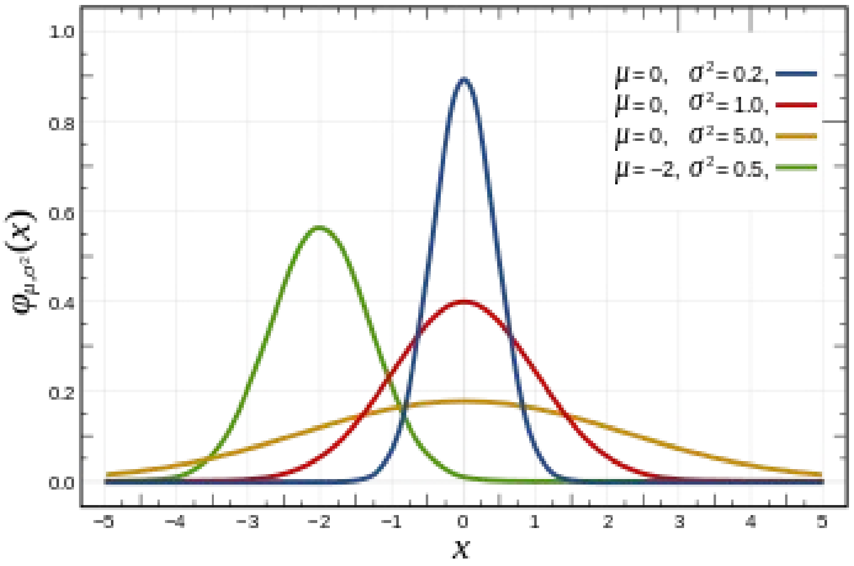
- Number of unique (distinct) values
- Number of missing values
- Minimum and maximum values
- Average (mean) and median
- Quartiles (25%, 75%)
- Variance and standard deviation

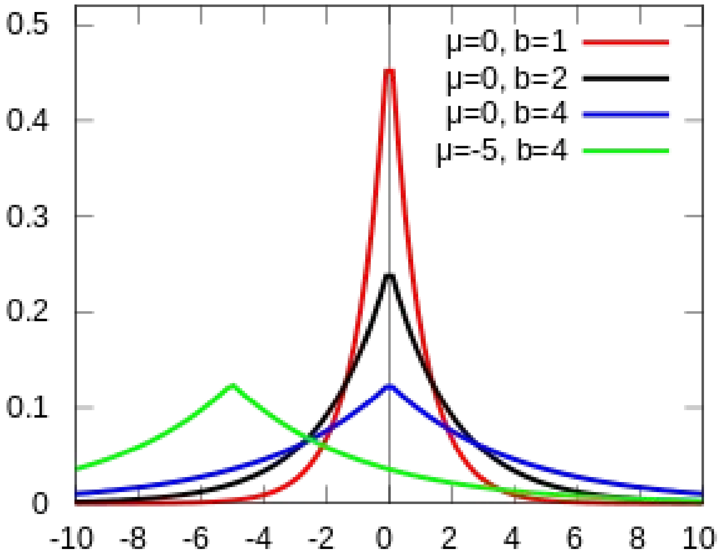
5-number summary

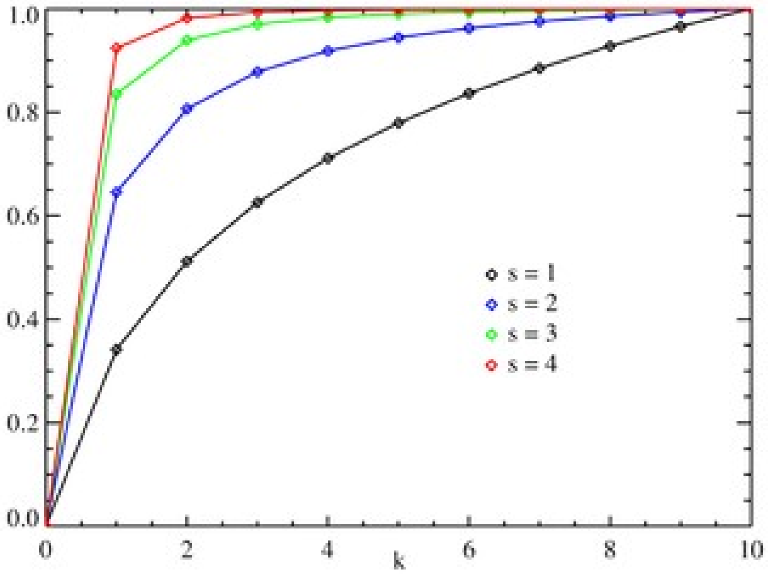


Distributions

- Examine whether data follow some well-known distributions (such as normal or Laplace, skewed, symmetric)
- Names generally follow Zipf distribution – few frequent and many rare

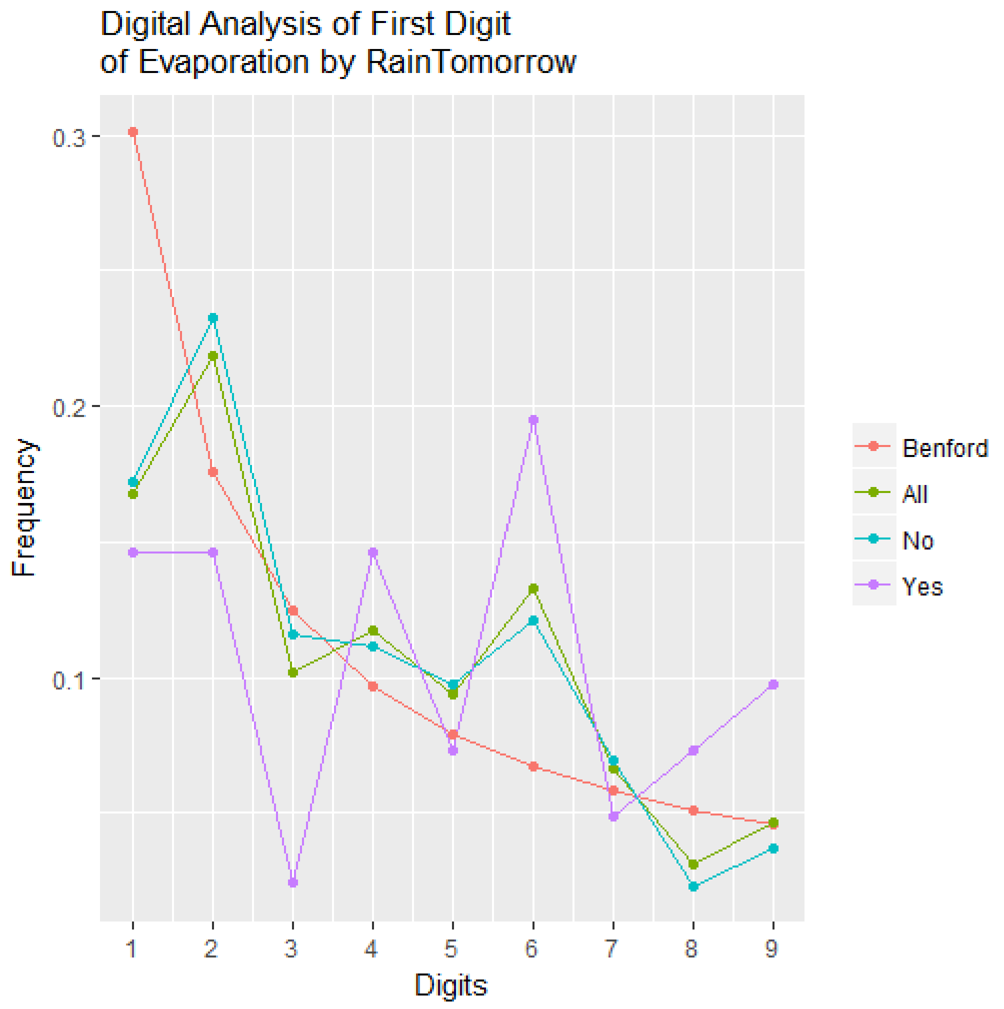






Benford's law

- First digit law
  - Distribution of first digits in natural numbers
  - Digit 1 occurs in about 30% (much greater than uniform distribution of 11.1% for each of the 9 digits)
  - Digit 9 only occurs in about 5%
  - Occurs in street numbers, stock prices, death rates, etc.
- Can be used in fraud detection (how? Think about it! To be discussed..)



Dependencies

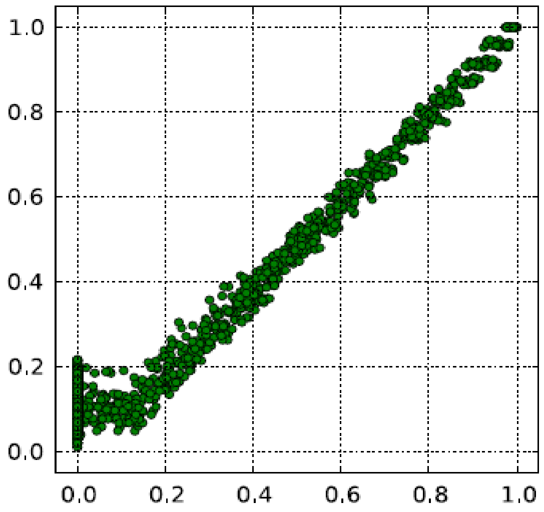
- Dependencies / correlations between attributes
  - Example: employment and income, age and weight

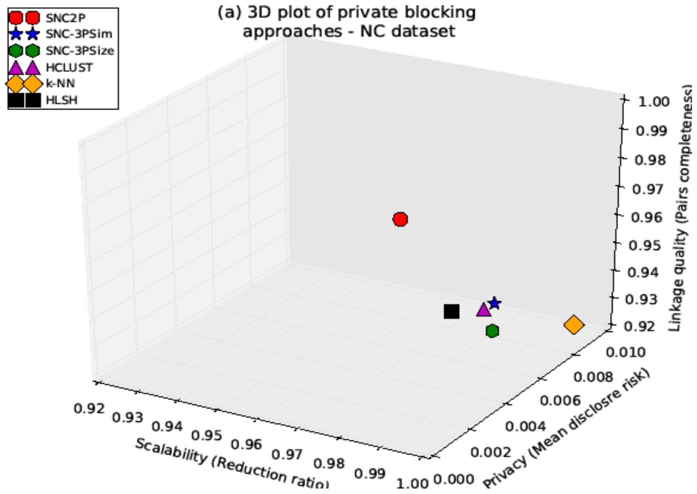
- The extent to which two variables (attributes) have a linear or non-linear relationship with each other

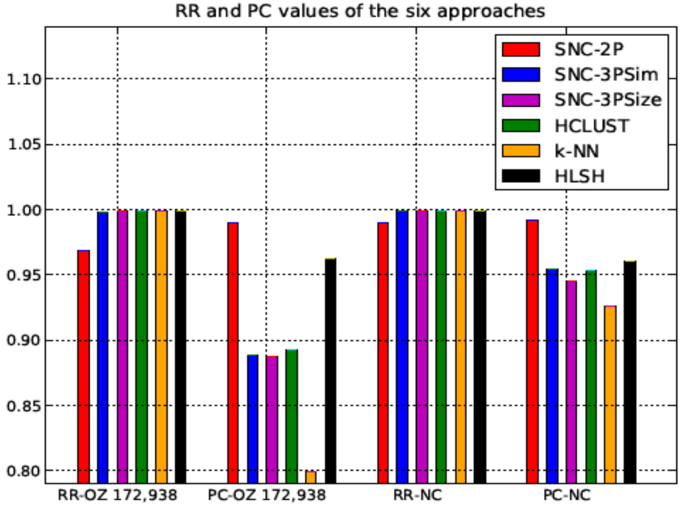
- Several correlation coefficients, including the Pearson coefficient, can be used to measure the correlation and dependency between attributes

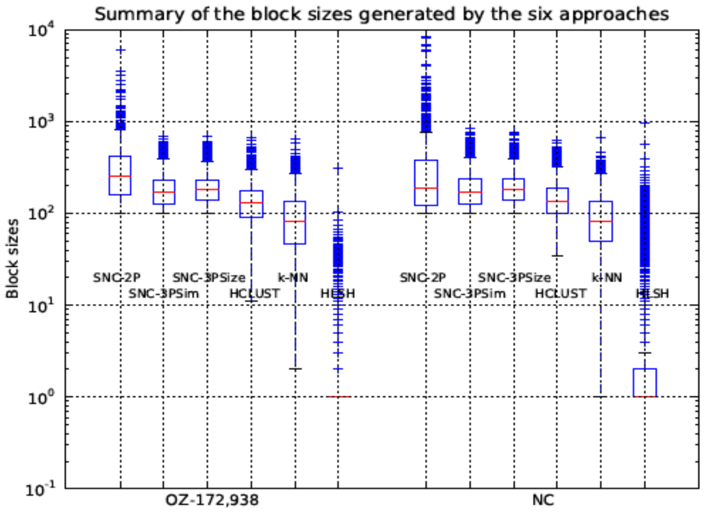
Data visualisation

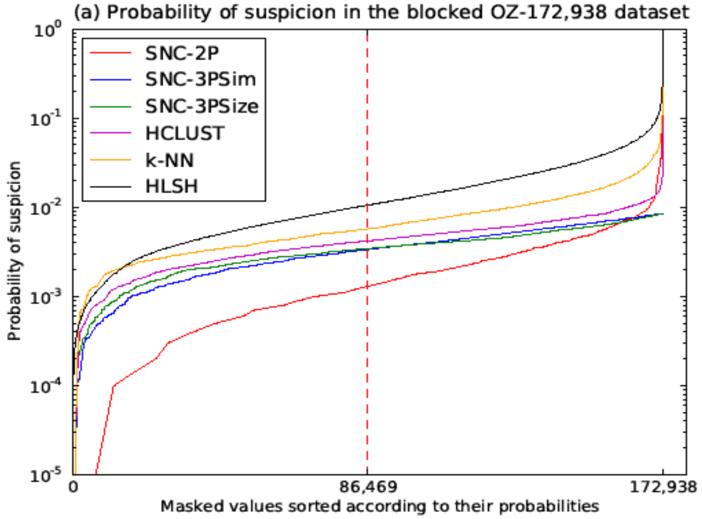
- Bar plots
- Box plots
- Scatter plots
- Line plots









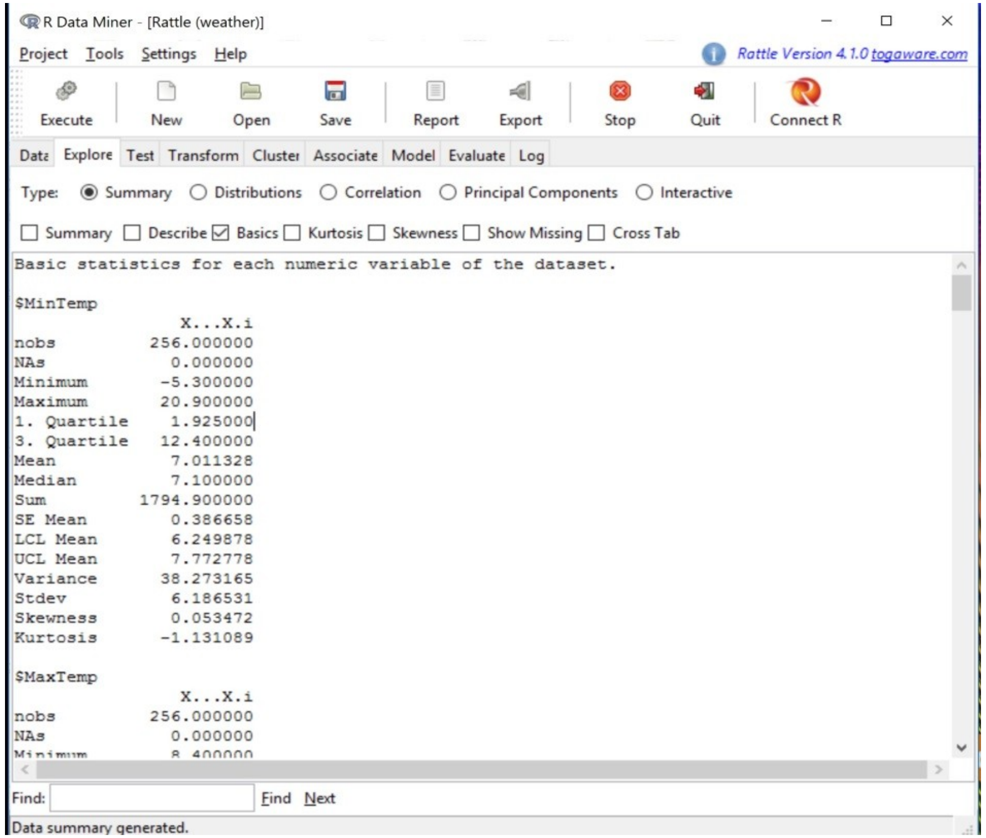


Data profiling tools

- Various commercial software:
  - IBM InfoSphere Information Analyzer, Oracle Enterprise Data Quality, SAP, Informatica Data Explorer, Trillium Software Data Profiling, Microsoft SQL Server Integration Services Data Profiling Task and Viewer
- Open source software:
  - Rattle (based on R programming language)
  - Python modules such as Pandas

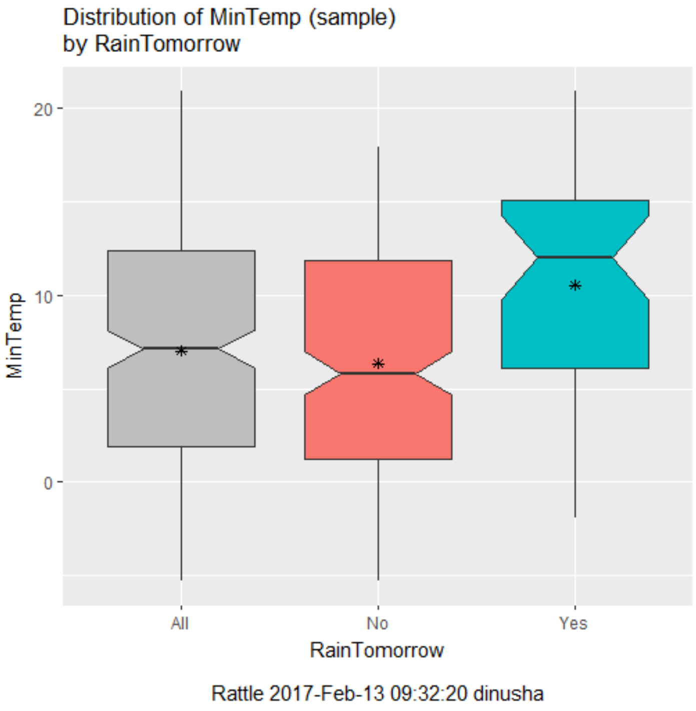
Data profiling with Rattle (1)

- Rattle weather dataset
  - Basic statistics
  - Kurtosis
  - Skewness
  - Missing values
  - Cross tab

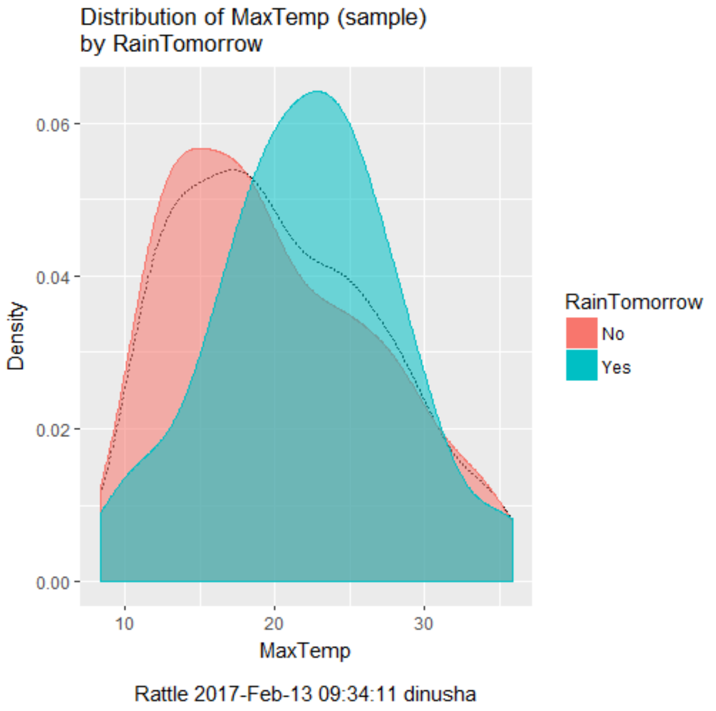


Data profiling with Rattle (2)

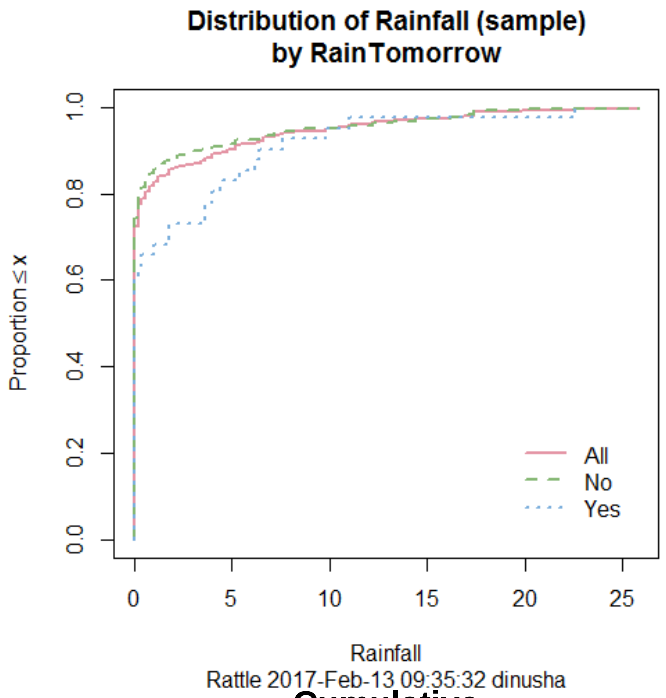
- Numerical data distributions



Box plot



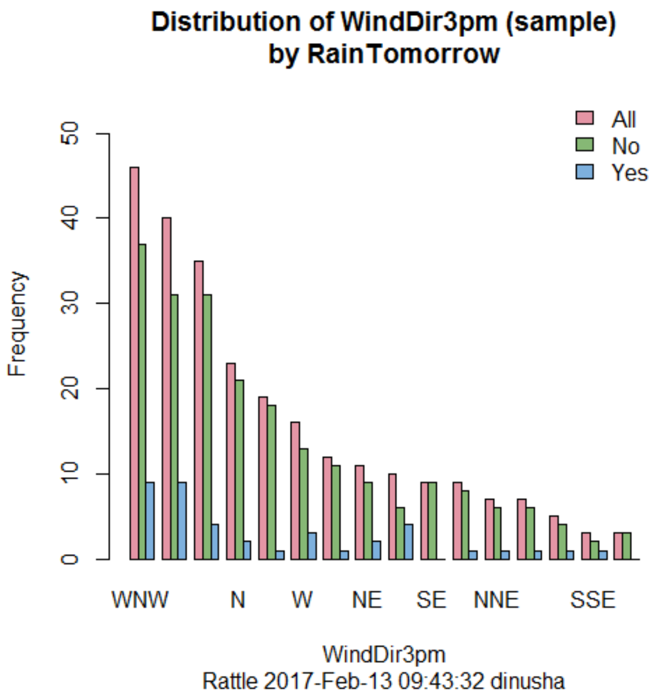
Histogram



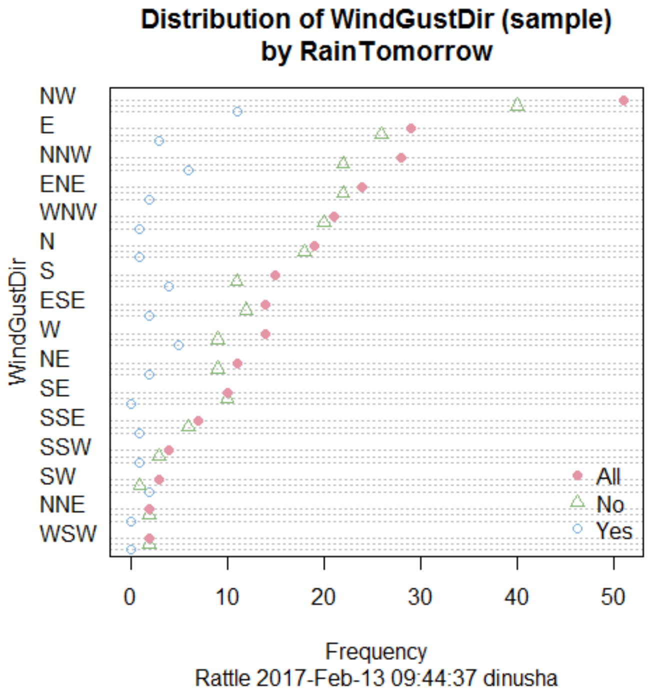
Cumulative

Data profiling with Rattle (3)

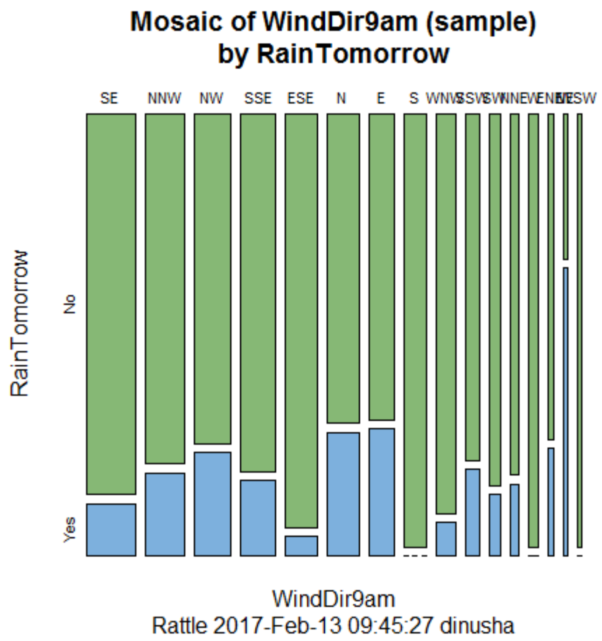
- Categorical data distributions



Bar plot



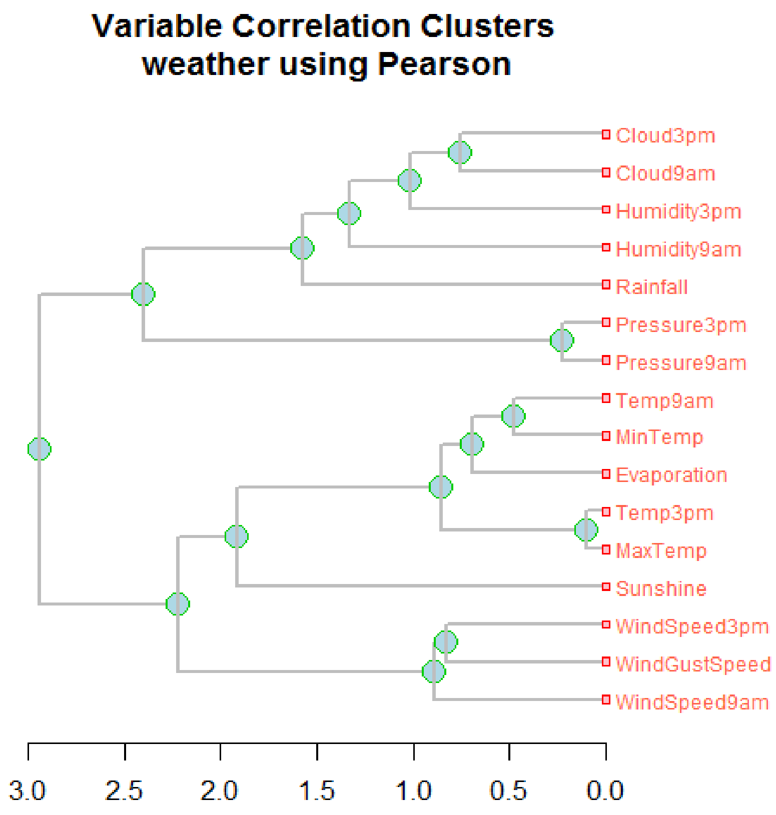
Dot plot

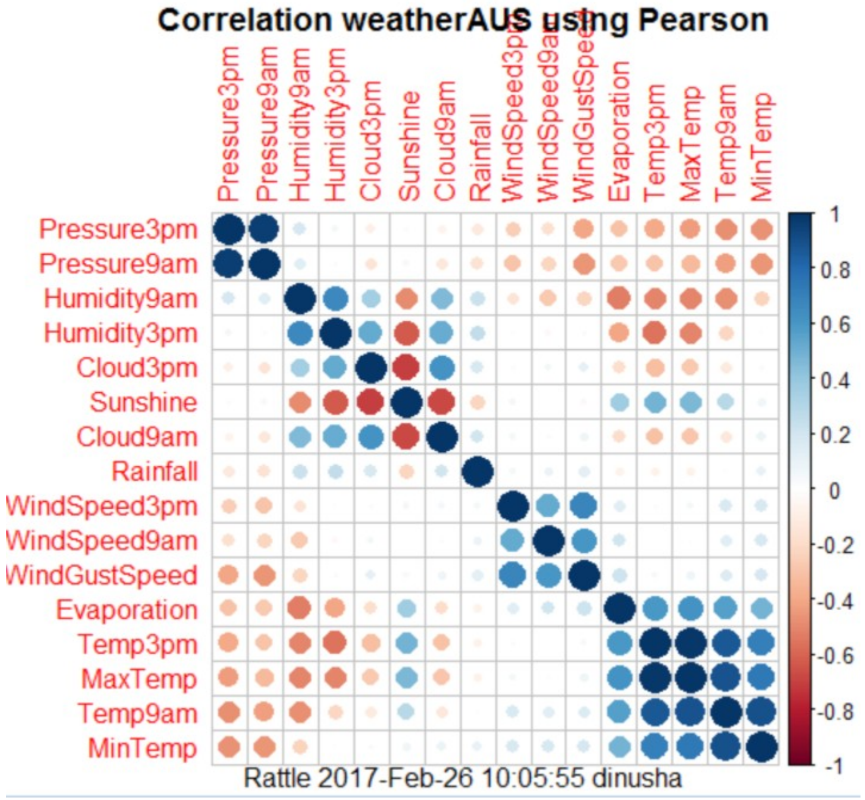


Mosaic

Data profiling with Rattle (4)

- Correlations





Data profiling with Python (1)

- Data exploration using pandas

  import pandas as pd

  df = pd.read_csv("weather.csv")

- First 10 rows

  df.head(10)

- Summary of numerical attributes

  df.describe()

- Frequency table for categorical attributes

  df['WindDir3pm'].value_counts()

Data profiling with Python (2)

- Data distributions

  df[‘MaxTemp’].hist(bins=50)

  df.boxplot(column=‘MaxTemp’)

  df.boxplot(column=‘MaxTemp’, by=‘Location’)

- Check missing values

  df.apply(lambda x: sum(x.isnull()),axis=0)

- Cross tab

  ct = pd.crosstab(df[‘WindDir9am'], df[‘RainToday'])

  ct.plot(kind='bar', stacked=True, color=['red','blue'], grid=False)

Summary

- Data profiling is a crucial step in the data wrangling pipeline
- The goal is to discover, assess, and understand meta-data of a data set
- Next generation data profiling tools and techniques:
  - Automated data profiling
  - Active learning in data profiling and cleaning
  - Advanced and interactive data visualisation