Objective:
To introduce Data Wrangling approaches and descriptive analytics on large data sets..

Syllabus:
Data Wrangling : Data acquisition, data formats, imputation, the split-apply-combine paradigm.

Learning Outcomes:
The student will be able to

- describe Data Wrangling approaches and descriptive analytics on large data sets.

Learning Material
UNIT - V: Data Wrangling
Data acquisition, data formats, imputation, the split-apply-combine paradigm.

1. What is Data wrangling?

Data wrangling involves getting and reading data, cleaning data, merging and shaping data. Data wrangling is important because, data in its original raw format is rarely prepared for its end use to begin with. Data wrangling is the process of cleaning and unifying messy and complex data sets for easy access and analysis. Data wrangling refers to the process of cleaning, restructuring and enriching the raw data available into a more usable format.

This process typically includes manually converting/mapping data from one raw form into another format to allow for more convenient consumption and organization of the data.

If you have 9 lakh birth year values of the format mm/dd/yyyy and 1 lakh of the format yyyy-mm-dd, doing Data wrangling, we can convert the later format to look like the former format, so that we can use all-together in a common format.

The key steps to data wrangling:

- Data Acquisition: Identify and obtain access to the data within your sources
- Joining Data : Combine the edited data for further use and analysis

- Data Cleansing: Redesign the data into a usable/functional format and correct/remove any bad data

This will help the scientist quicken the process of decision making, and thus get better insights in less time. This practice is being followed by a large number of top firms in the field, partly owing to the benefits it has and partly because of large amounts of data which is supposed to be analysed. Organizing and cleaning data before analysis has been shown to be extremely useful and helps the firms quickly analyse larger amounts of data. Data cleaning can be applied to remove noise and correct inconsistencies in the data. Data integration merges data from multiple sources into a coherent data store, such as a data warehouse.

Data transformations, such as normalization, may be applied. For example, normalization may improve the accuracy and efficiency of mining algorithms involving distance measurements.
Data reduction can reduce the data size by aggregating, eliminating redundant features, or clustering, for instance. These techniques are not mutually exclusive; they may work together.
For example, data cleaning can involve transformations to correct wrong data, such as by transforming all entries for a date field to a common format. Data processing techniques, when applied before mining, can substantially improve the overall quality of the patterns mined and/or the time required for the actual mining

https://github.com/okfn/handbook/tree/master

1.2 An introduction to the data pipeline

1. Acquisition describes gaining access to data, either through any of the methods or by generating fresh data, e.g through a survey or observations.
2. • In the extraction stage, data is converted from whatever input format has been acquired (e.g. XLS files, PDFs or even plain text documents) into a form that can be used for further processing and analysis. This often involves loading data into a database system, such as MySQL or PostgreSQL.

3. • Cleaning and transforming the data often involves removing invalid records and translating all the columns to use a sane set of values. You may also combine two different datasets into a single table, remove duplicate entries or apply any number of other normalizations. As you acquire data, you will notice that such data often has many inconsistencies: names are used inconsistently, amounts will be stated in badly formatted numbers, while some data may not be usable at all due to file corruptions. In short: data always needs to be cleaned and processed. In fact, processing, augmenting and cleaning the data is very likely to be the most time- and labour-intensive aspect of your project.
4. • Analysis of data to answer particular questions. The aspects of analysis are automated and large-scale analysis, showing tips and tricks for getting and using data, and having a machine do a lot of the work, for example: network analysis or natural language processing.
5. • Presentation of data only has impact when it is packaged in an appropriate way for the audiences it needs to aim at.

Data wrangling, like most data analytics processes, is an iterative one – the practitioner will need to carry out these steps repeatedly in order to produce the results he desires

What is Tidy Data?

A dataset is said to be tidy if it satisfies the following conditions

1. observations are in rows
2. variables are in columns
3. contained in a single dataset.

Tidy data makes it easy to carry out data analysis.

What is dplyr: An R package full of data verbs. Some examples of things it can do

Select (rows or columns)○

Sort (rearrange data)○

Filter (remove rows)○

Summarize (e.g., mean)○

Transform (e.g., add columns)○

The group_by() operation

Qualitative data is data telling you something about qualities: e.g. description, colors etc. Interviews count as qualitative data

Quantitative data tells you something about a measure or quantification. Such as the quantity of things you have, the size (if measured) etc.

JSON JavaScript Object Notation. A common format to exchange data. Although it is derived from Javascript, libraries to parse JSON data exist for many programming languages. Its compact style and ease of use has made it widespread. To make viewing JSON in a browser easier you can install a plugin such as JSONView in Chrome and JSONView in Firefox.

Machine-readable Formats that are machine readable are ones which are able to have their data extracted by computer programs easily. PDF documents are not machine readable. Computers can display the text nicely, but have great difficulty understanding the context that surrounds the text. Common machine-readable file formats are CSV and Excel Files. Tab-separated values Tab-separated values (TSV) are a very common form of text file format for sharing tabular data. The format is extremely simple and highly machine-readable.

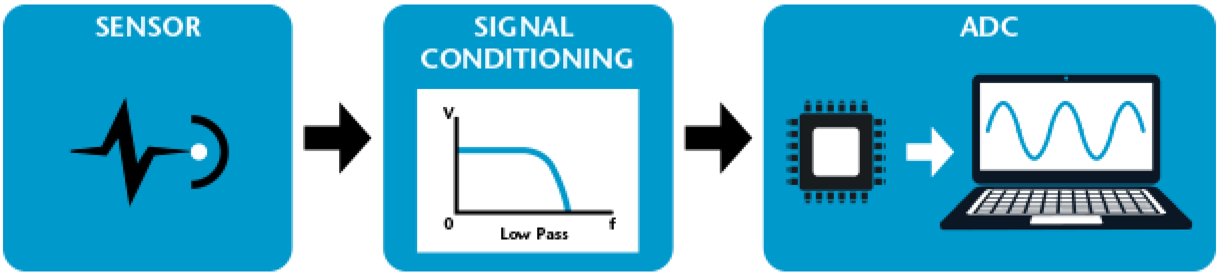
Free Data Wrangling Tools

1. *• Tabula: convert PDF table into a spreadsheet.*
2. *• OpenRefine: Friendly GUI for describing and manipulating data*
3. *• R package*
4. *• DataWrangler*
5. *• CSVKit*
6. *• Python and Pandas*
7. *• Mr. Data Converter*

Data acquisition is the process of digitizing data from the world around us so it can be displayed, analyzed, and stored in a computer.

Components of a Data Acquisition System

All data acquisition systems consist of three essential elements – Sensor, Signal Conditioning, and Analog-to-Digital Converter (ADC).



Analog-to-Digital Converter

At the core of all data acquisition systems is an Analog to Digital Converter (ADC). As the name implies, this chip takes data from the environment and converts it to discrete levels that can be interpreted by a processor. These discrete levels correspond to the smallest detectable change in the signal being measured. The higher the number of "bits" of an ADC (12-bit, 16-bit, 18-bit etc.), the greater the number of discrete levels that can represent an analog signal and the greater the resolution of the ADC. The resolution of an ADC is essentially analogous to the ticks on a measuring stick. A measuring stick with mm tick marks has more resolution than a measuring stick with only cm tick marks. Whether you need mm or cm tick marks depends on what you are measuring – the same is true for ADC resolution.

Sensors (Transducers)

Sensors, often called Transducers, convert real-world phenomenon like temperature, force, and movement to voltage or current signals that can be used as inputs to the ADC. Common sensors include thermocouples, thermistors, and RTDs to measure temperature, accelerometers to measure movement, and strain gauges to measure force. When choosing the right sensor for your measurement system, it's important to consider factors like the accuracy of the sensor and the signal conditioning required to record a readable signal.

Signal Conditioning

To make quality measurements on transducers, additional circuitry is often needed between the transducer and the ADC. This circuitry is generally referred to as signal conditioning and can include amplification/attenuation, filtering, Wheatstone bridge completion, excitation, linearization, calibration, and cold-junction-compensation (CJC). Different sensors have different signal conditioning needs. For instance, signal conditioning for a strain gauge requires excitation, bridge completion and calibration. Thermocouples, which output signals in the mV range, need to be amplified as well as filtered before going through the ADC. Many times, signal conditioning circuitry is contained within a data acquisition device, but signal conditioning may also be part of the transducer. Load cells, for example, contain the bridge completion, calibration circuitry, and amplification. Many MEM (micro-electro-mechanical) sensors also contain signal conditioning.

Scraping: The process of extracting data in machine-readable formats of non-pure data sources e.g.: webpages or PDF documents. Often prefixed with the source (web-scraping PDF-scraping).

We use imputation when we don't have very much data, or where removing our missing values would compromise the representativeness of our sample. Throwing away a bunch of entries just because they're missing values could severely impact the statistical power of whatever analysis we were trying to perform. In this case, it likely makes sense to make an intelligent guess at the missing values in our data like approximation. We could for example replace all missing values by the mean of all others or using linear regression to estimate the missing values. However, imputation introduces biases and inaccuracies into the data set. It is a really hard problem and new techniques are constantly being developed. Dealing with data files statisticians often have to consider the problem of missing data due to unit nonresponse (complete nonresponse) and item nonresponse (partial nonresponse).

imputating missing values is an iterative process. naniar aims to make it easier to manage imputed values by providing the nabular data structure to simplify managing missingness. This vignette provides some useful recipes for imputing and exploring imputed data.

naniar implements a few imputation methods to facilitate exploration and visualisations, which were not otherwise available: impute_below, and impute_mean. For single imputation, the R package simputation works very well with naniar, and provides the main example given.

Imputing and tracking missing values

Using impute_below

impute_below imputes values below the minimum of the data, with some noise to reduce overplotting. The amount data is imputed below, and the amount of jitter, can be changed by changing the arguments prop_below and jitter.

```
library(dplyr)
#>
#> Attaching package: 'dplyr'
#> The following objects are masked from 'package:stats':
#>
#>     filter, lag
#> The following objects are masked from 'package:base':
#>
#>     intersect, setdiff, setequal, union
library(naniar)

airquality %>%
  impute_below_at(vars(Ozone)) %>%
  select(Ozone, Solar.R) %>%
  head()
#>       Ozone Solar.R
#> 1  41.00000     190
#> 2  36.00000     118
#> 3  12.00000     149
#> 4  18.00000     313
#> 5 -19.72321      NA
#> 6  28.00000      NA
```

Using impute_mean

The mean can be imputed using impute_mean, and is useful to explore structure in missingness, but are not recommended for use in analysis. Similar to simputation, each impute_ function returns the data with values imputed.

Imputation functions in naniar implement "scoped variants" for imputation: _all, _at and _if.

This means:

- _all operates on all columns
- _at operates on specific columns, and
- _if operates on columns that meet some condition (such as is.numeric or is.character).

If the impute_ functions are used as-is - e.g., impute_mean, this will work on a single vector, but not a data.frame.

1.1 The problem of missing data

1.1.1 Current practice

The mean of the numbers 1, 2 and 4 can be calculated in R as

```
y <- c(1, 2, 4)
mean(y)
[1] 2.33
```

where y is a vector containing three numbers, and where mean(y) is the R expression that returns their mean. Now suppose that the last number is missing. R indicates this by the symbol NA, which stands for “not available”:

```
y <- c(1, 2, NA)
mean(y)
[1] NA
```

We can track the imputed values using the nabular format of the data.

Track imputed values using nabular data

We can track the missing values by combining the verbs bind_shadow, impute_, add_label_shadow. We can then refer to missing values by their shadow variable, _NA.
The add_label_shadow function adds an additional column called any_missing, which tells us if any observation has a missing value.

Imputing values using simputation

We can impute the data using the easy-to-use simputation package, and then track the missingness using bind_shadow and add_label_shadow:

What is split-apply-combine?

Split-apply-combine is a commonly used strategy for dealing with repeated calculations over large data sets.
Many data analysis tasks can be approached using the *split-apply-combine* paradigm:

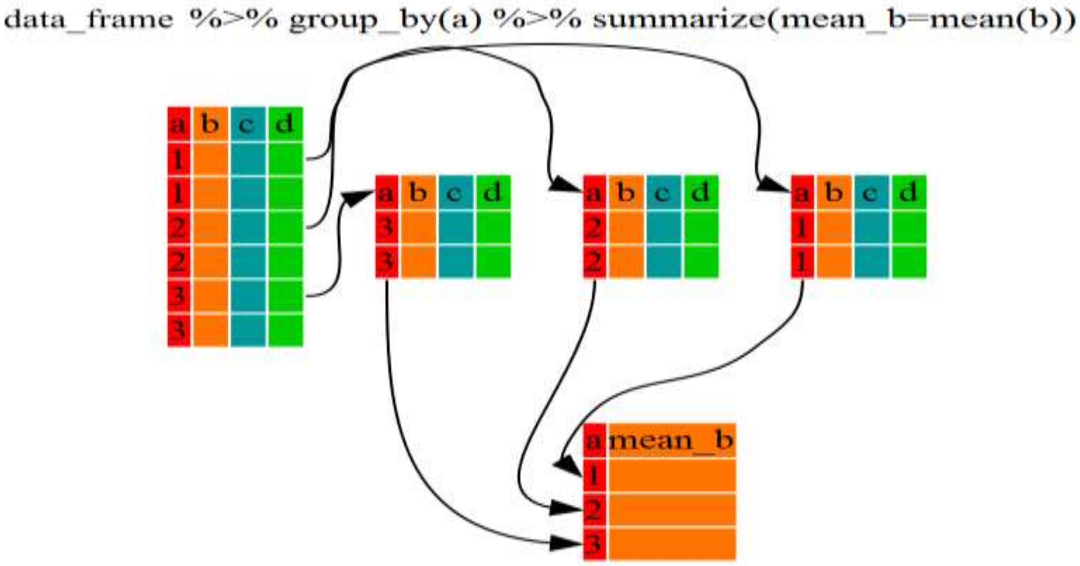
1. split the data into groups,
2. apply some analysis to each group, and then
3. combine the results.

- It is helpful conceptually :

i. what are the main chunks of data,

ii. what are the main functions to apply, iii. how to put things back together in a sensible way)

- It is also helpful computationally for large data sets
- Turns out there are a whole family of apply() like functions that will make your life even easier, and these are also extra helpful for large data sets (next time)



```
data_frame %>% group_by(a) %>% summarize(mean_b=mean(b))
```

Assignment-Cum-Tutorial Questions

I) Objective Questions (10 to 15)

1) In____________, the data are modeled to fit a straight line. [ ]

A. simple linear regression B. Logistic regression C. Thymine D. Quadratic regression

2) A random variable,y (called a response variable), can be modeled as a linear function of another random variable,x (called a predictor variable), with the equation y=wx+b, where the variance of y is assumed to be constant. In the context of data mining,x and y are numeric database attributes. The coefficients,w and b are called ____

A. regression coefficients B. response variable C. dependent variables

3) Terabyte is ----- number of petabytes

4. ____can be applied to remove noise and correct inconsistencies in the data. [ ]

A. Data cleaning B. Data integration C. Cytosine D. Data acquisition

2______ merges data from multiple sources into a coherent data store, such as a data warehouse.

A. Data cleaning B. Data integration C. Data acquisition D. Signal conditioning

4.Which of the following are TRUE?

1. Data transformations, such as normalization, may be applied and may improve the accuracy and efficiency of mining algorithms involving distance measurements.

2. Data reduction can reduce the data size by aggregating, eliminating redundant features, or clustering, for instance.

3. The techniques in data pre-processing are not mutually exclusive; they may work together.

A. 1 only B. 2 only C. 2& 3 only D. 1, 2 and 3

6. Data cleaning can involve transformations to correct wrong data, such as by transforming all entries for a date field to a common format.

7. Data processing techniques, when applied before mining, can substantially improve the overall quality of the patterns mined and/or the time required for the actual mining
8. Raw data in the real-world is often untidy and poorly formatted. [True/False]

II) Descriptive Questions(6 to 8)

1. Discuss about Data Wrangling
2. Explain about Data acquisition,
3. Explain different data formats,
4. Discuss about imputation,
5. Describe the split-apply-combine paradigm.