CHAPTER 1: INTRODUCTION TO DATA WRANGLING

DR. MOHD KHAIRUL BAZLI BIN MOHD AZIZ

PUSAT SAINS MATEMATIK, UNIVERSITI MALAYSIA PAHANG

SYNOPSIS

Data wrangling is the process of cleaning, structuring and enriching complex raw data into a desired format for analysis and better decision making. This course introduces the knowledge and skills to wrangle data from diverse sources and shape it to enable data-driven applications. In this course, some main topics are covered including introduction of data wrangling, dynamics of data wrangling and data transformation. Students will learn how to gather and extract data from widely used data formats. Python will be used for implementation.

CONTENT

1.1 Definition and terminology

1.2 Data Wrangling process

1.3 Tools for Data Wrangling

1.4 Data Wrangling Application

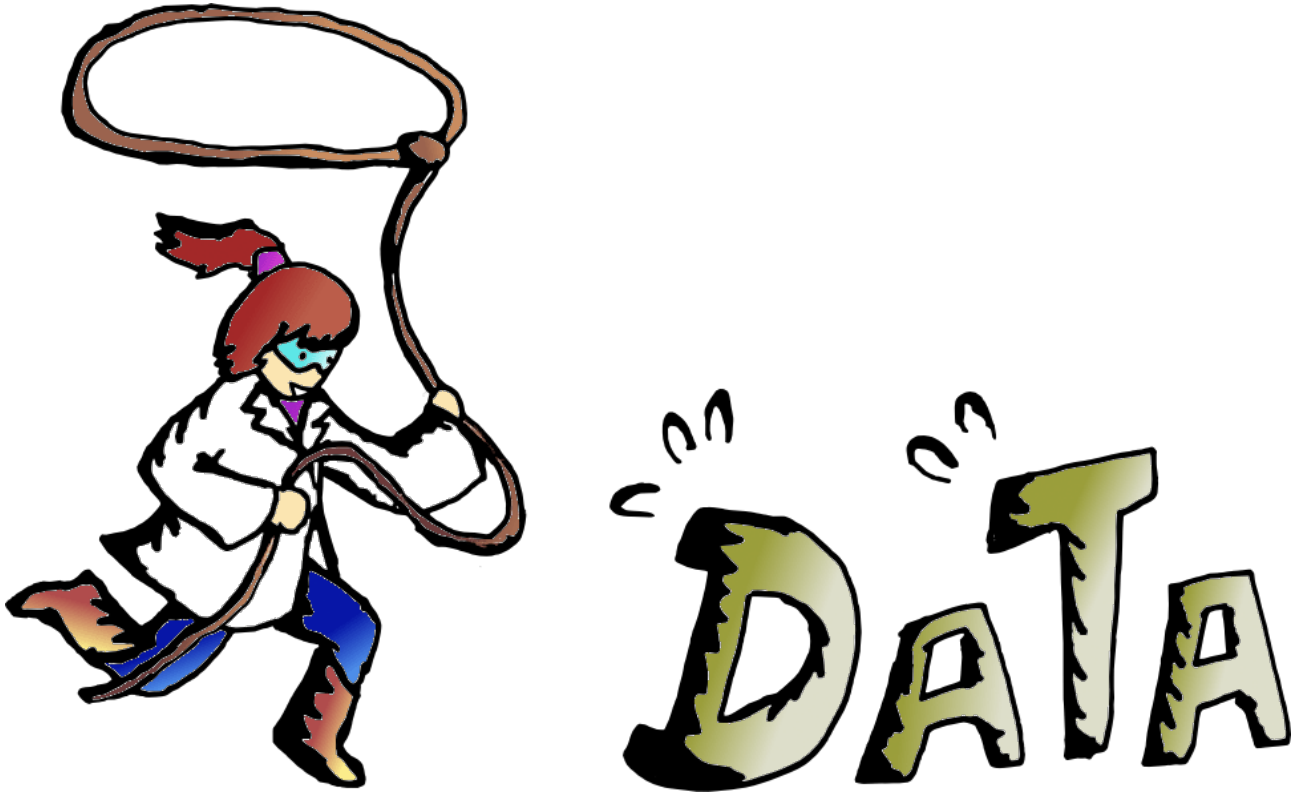
1.5 Data Wrangling using Python

1.1 DEFINITION AND TERMINOLOGY

By the end of this topic, you should be able to:

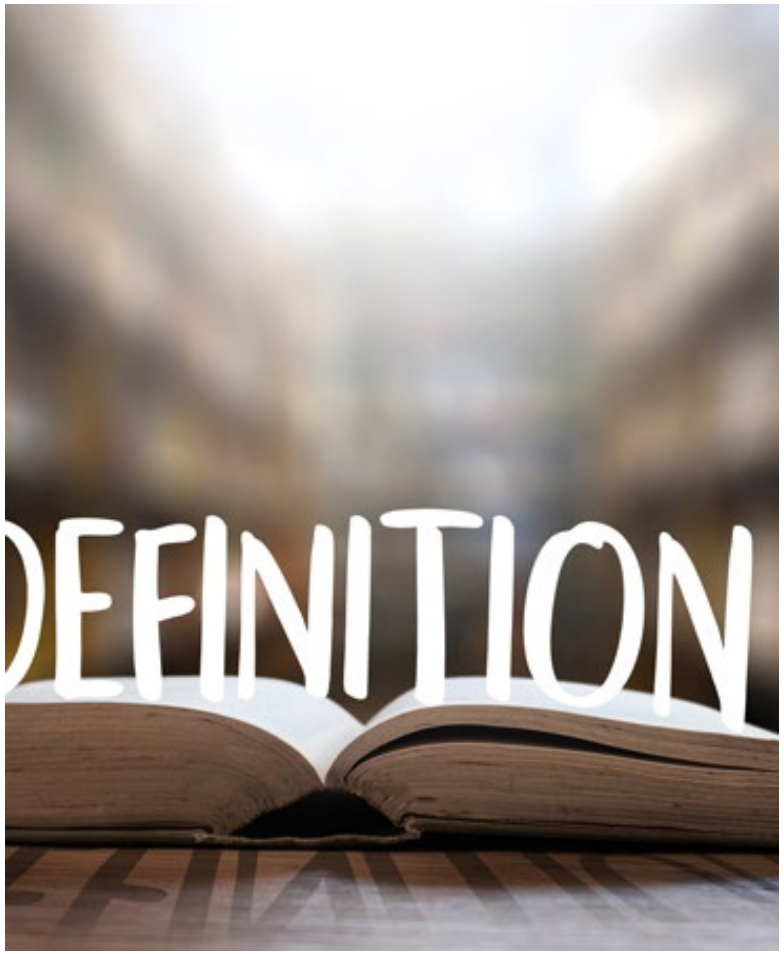
- Understand the definitions and terminologies used in data wrangling.
- Differentiate various data wrangling terminologies.

WHAT IS DATA WRANGLING?



Suppose you are working on some training data set. You decide to use your favorite classification algorithm only to realize that the training data set contains a mixture of continuous and categorical variables and you'll need to transform some of the variables into a suitable format. You realize that the raw data you have can't be used for your analysis without some manipulation — what you'll soon know as data wrangling. You'll need to clean this messy data to get anywhere with it.

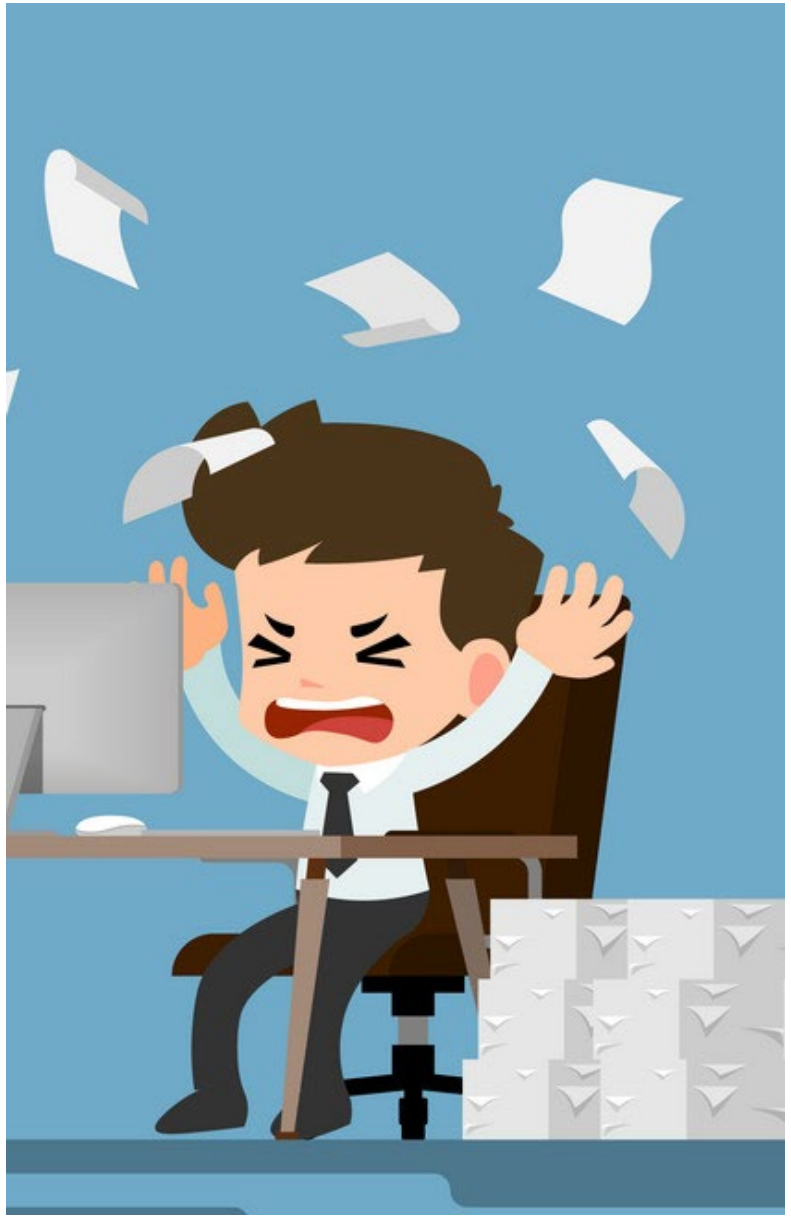
DATA WRANGLING DEFINITION



- Data wrangling, also known as data munging, is an iterative process that involves data exploration, transformation, validation, and making it available for a credible and meaningful analysis. It includes a range of tasks involved in preparing raw data for a clearly defined purpose, where raw data at this stage is data that has been collated through various data sources in a data repository. Data wrangling captures a range of tasks involved in preparing data for analysis.

DATA WRANGLING VS DATA MINING

- Some people struggle to understand the difference between data wrangling (or data munging) and data mining.
- Data mining is a process of finding patterns and relationships hidden in large data sets. Data mining helps businesses to decipher meaningful patterns in their data, whether it is open source data or not.
- Data wrangling is a superset of data mining and requires multiple other processes, such as cleaning, transforming, integrating, etc. for decision-making. The purpose of a data wrangling project is to help deliver intelligible insights.



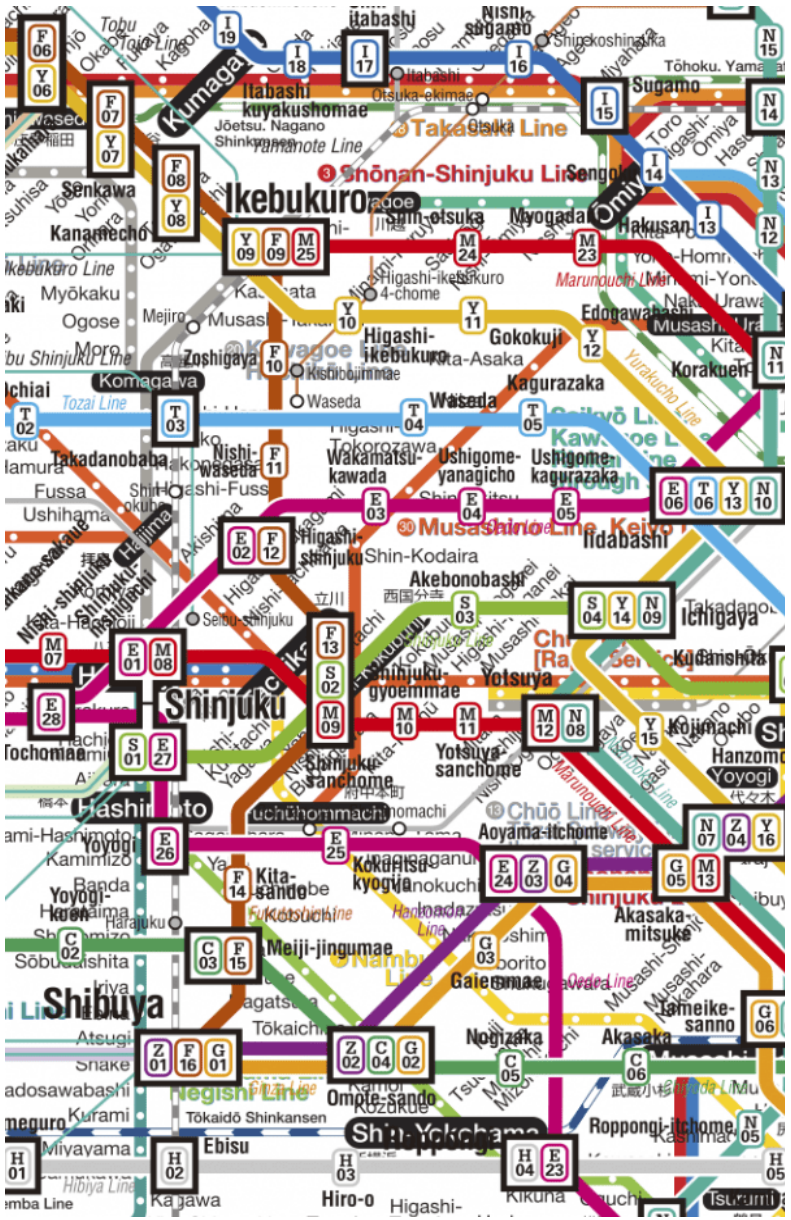
WHY DO YOU NEED DATA WRANGLING?

- Did you know, data professionals spend almost 73% of their time just wrangling the data? This means it's an indispensable aspect of data processing. It helps business users make concrete, timely decisions by cleaning and structuring raw data into the required format. As the data is becoming more and more unstructured and diverse, data wrangling is becoming a common practice among top organizations.
- Accurately wrangled data ensures that quality data is entered into analytics or downstream processes for consolidation and collaboration.
- Data wrangling is important to fasten the data-to-insight journey and support timely decision-making.
- It can be arranged into a consistent and repeatable procedure using data integration tools with automation capabilities that clean and convert source data into a format that be reused as per the end requirements. After converting data to a standard format, you can perform crucial, cross-data set analytics. Moreover, data wrangling with Python is common as Python employs different methods to wrangle the data stored in different data sets.

DATA WRANGLING GOALS

To understand why data wrangling is so essential, let's take a look at how automation tools help to achieve data-wrangling goals.

- Reduces time: As briefly mentioned, data analysts spend a bulk of their time in the data wrangling process. For some, it takes up most of their time. Imagine piecing together various data sources and manually filling in the blanks. Or, even if code is used, it takes a lot of time to string it together accurately. An automated solution like Solvexia can 10x productivity through automation.
- Data analysts can focus on analysis: Once a data analyst has freed up all their time they would have otherwise spent managing data wrangling, they can leverage the data to focus on why they were hired - to perform analysis. With the help of automation tools, data analytics and reporting can be created in an instant.
- Better decision-making in a shorter time: Business decisions rely on information promptly. By utilising automation tools for data wrangling and analytics, you can make the most informed decision quickly.
- More in-depth intelligence: Data is used in every aspect of business and will impact every department, from sales to marketing to finance. By utilising data and data wrangling, you'll be able to understand the current status of your business better and focus energy on wherever issues reside.
- Accurate, actionable data: With good data wrangling, you will have peace of mind that your data is correct, and, in turn, you can rely on it to take action.



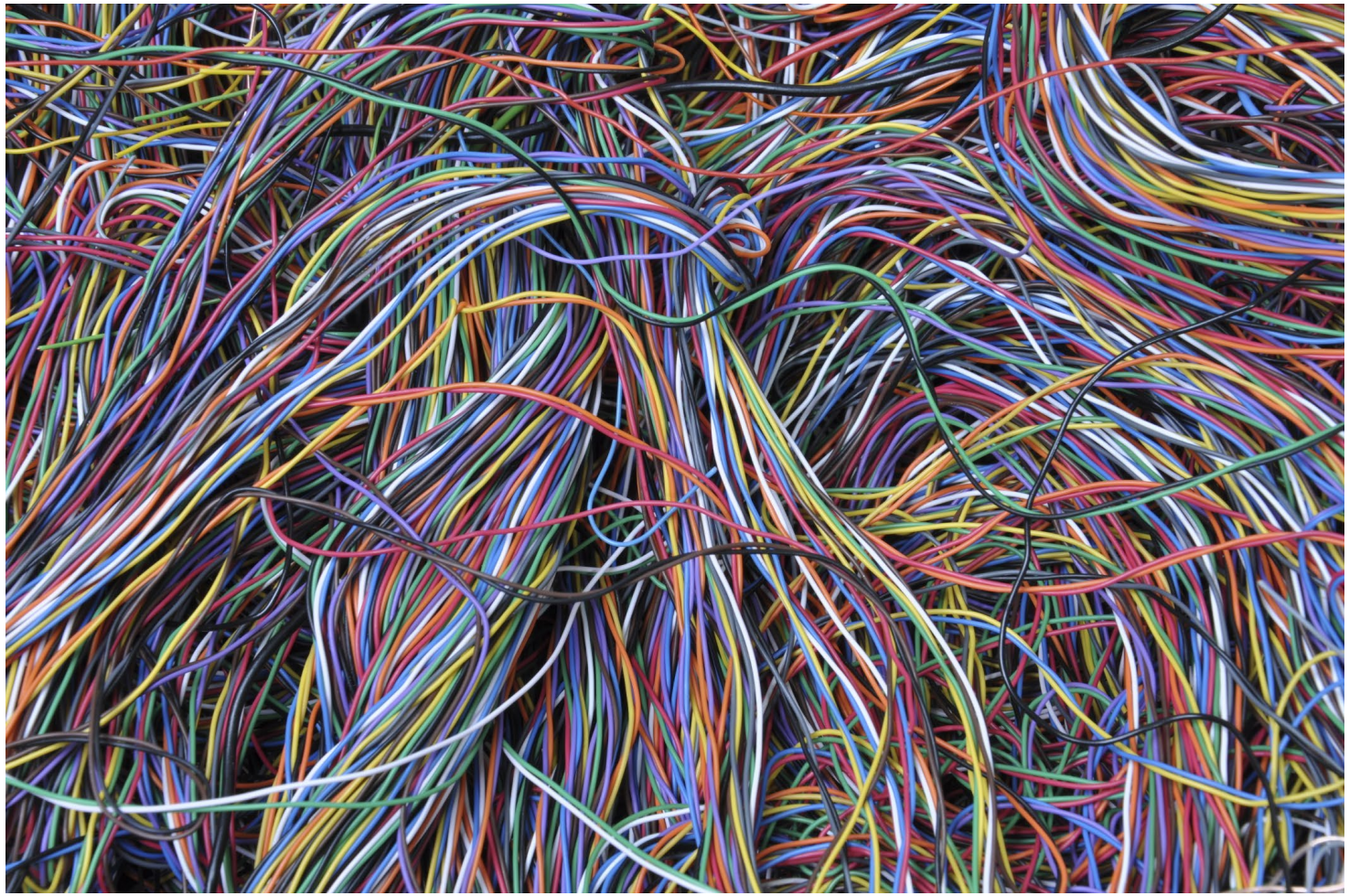
1.2 DATA WRANGLING PROCESS

By the end of this topic, you should be able to:

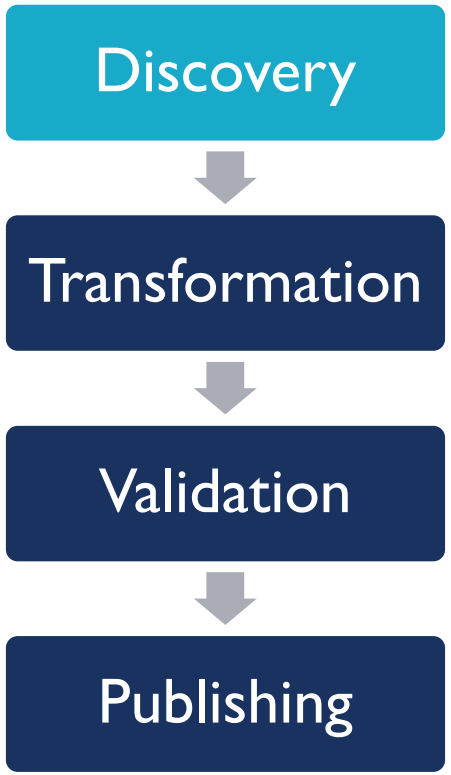
- Understand the definitions and terminologies used in data wrangling.
- Differentiate various data wrangling terminologies.

DATA WRANGLING PROCESS

- Discovery
- Transformation
- Validation
- Publishing



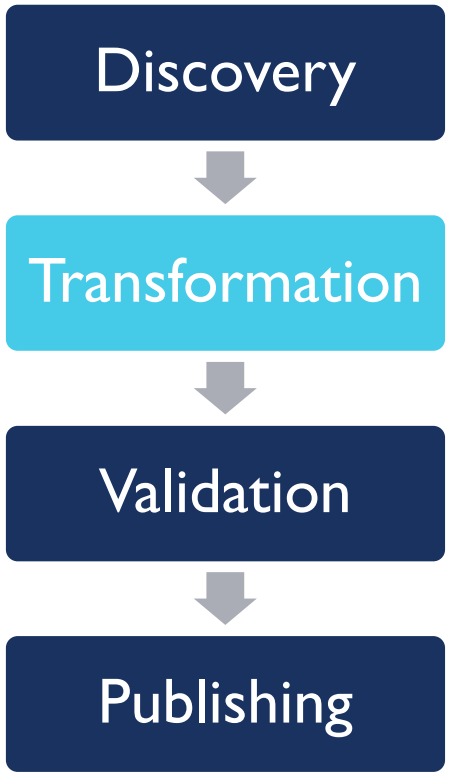
DISCOVERY PROCESS



The Discovery phase, also known as the Exploration phase, is about understanding your data better with respect to your use case. The objective is to figure out specifically how best you can clean, structure, organize, and map the data you have for your use case.

- Examining and understanding your data with respect to your use case.
- Creating a plan for cleaning, structuring, organizing and mapping your data.

TRANSFORMATION PHASE



Structuring Data

Normalizing and Denormalizing Data

Cleaning data

Enriching Data

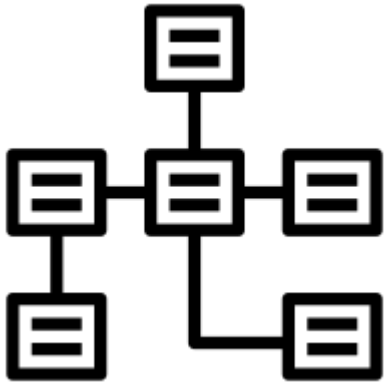
The next phase, which is the Transformation phase, forms the bulk of the data wrangling process. It involves the tasks you undertake to transform the data, such as structuring, normalizing, denormalizing, cleaning, and enriching the data.

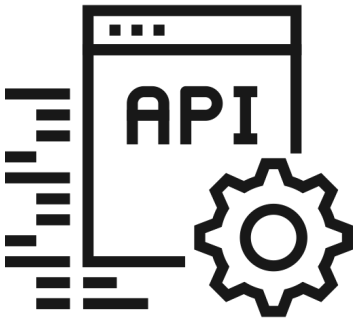
TRANSFORMATION PHASE

- Discovery
- Transformation
- Validation
- Publishing

Structuring Data | Normalizing and Denormalizing Data | Cleaning data | Enriching Data

This task includes actions that change the form and schema of your data. The incoming data can be in varied formats. You might, for example, have some data coming from a relational database and some data from Web APIs. In order to merge them, you will need to change the form or schema of your data. This change may be as simple as changing the order of fields within a record or dataset or as complex as combining fields into complex structures.



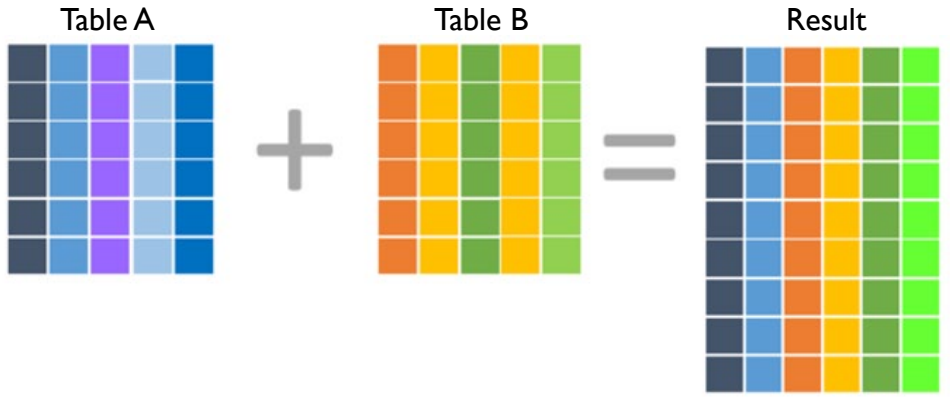


TRANSFORMATION PHASE

- Discovery
- Transformation
- Validation
- Publishing

Structuring Data | Normalizing and Denormalizing Data | Cleaning data | Enriching Data

- Joins and Unions are the most common structural transformations used to combine data from one or more tables. How they combine the data is different.
- Joins combine columns. When two tables are joined together, columns from the first source table are combined with columns from the second source table – in the same row. So, each row in the resultant table contains columns from both tables.

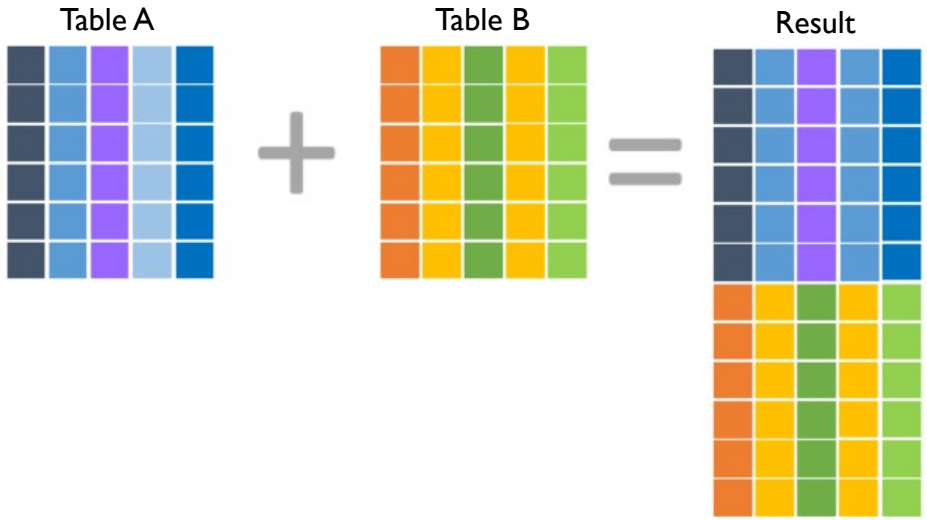


TRANSFORMATION PHASE

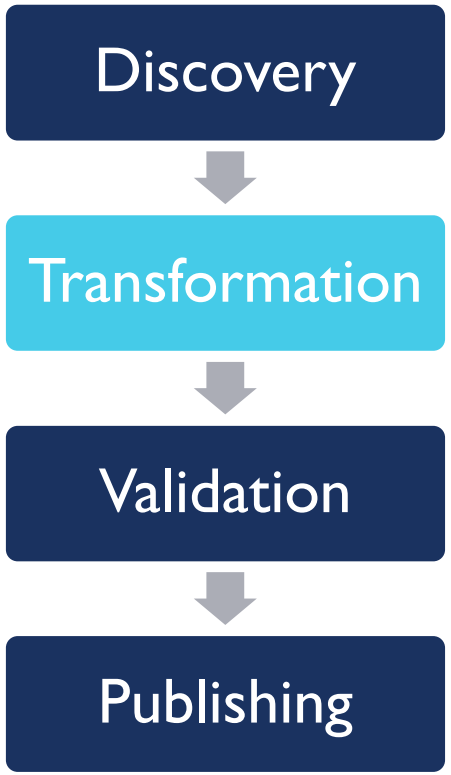
- Discovery
- Transformation
- Validation
- Publishing

Structuring Data | Normalizing and Denormalizing Data | Cleaning data | Enriching Data

Unions combine rows. Rows of data from the first source table are combined with rows of data from the second source table into a single table. Each row in the resultant table is from one source table or another.



TRANSFORMATION PHASE



Structuring Data

Normalizing and Denormalizing Data

Cleaning data

Enriching Data

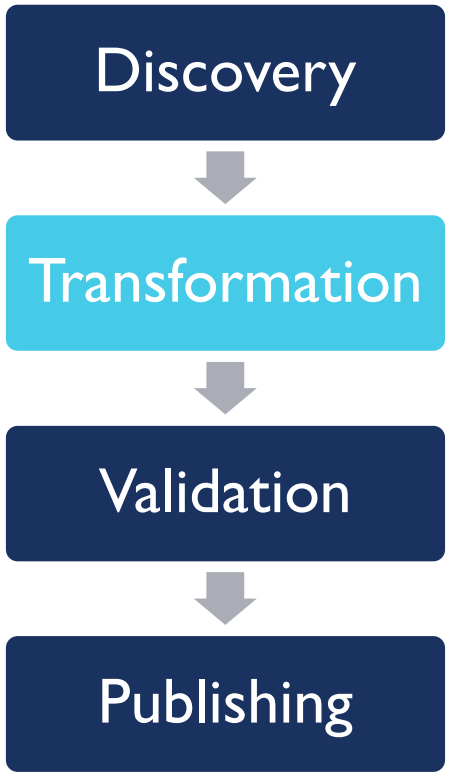
Transformation can also include normalization and denormalization of data.

Normalizing data include:

- Cleaning unused data
- Reducing redundancy
- Reducing inconsistency

Data coming from transactional systems, for example, where a number of insert, update, and delete operations are performed on an ongoing basis, are highly normalized.

TRANSFORMATION PHASE



Structuring Data

Normalizing and Denormalizing Data

Cleaning data

Enriching Data

Denormalization is used to combine data from multiple tables into a single table so that it can be queried faster. For example, normalized data coming from transactional systems is typically denormalized before running queries for reporting and analysis.

TRANSFORMATION PHASE

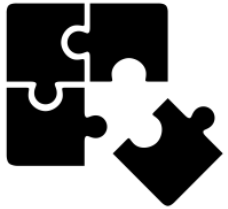
- Discovery
- Transformation
- Validation
- Publishing

Structuring Data | Normalizing and Denormalizing Data | Cleaning data | Enriching Data

Another transformation type is Cleaning. Cleaning tasks are actions that fix irregularities in data in order to produce a credible and accurate analysis. The types of data are:

Inaccurate Data

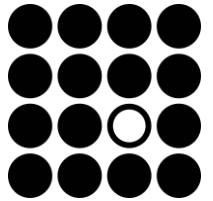
Missing Data



Incomplete Data

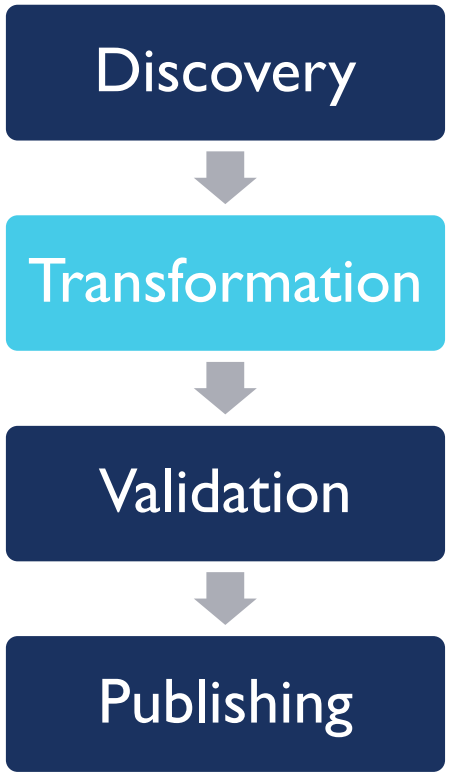
Biases Data

Null Values



Outliers

TRANSFORMATION PHASE



Structuring Data

Normalizing and Denormalizing Data

Cleaning data

Enriching Data

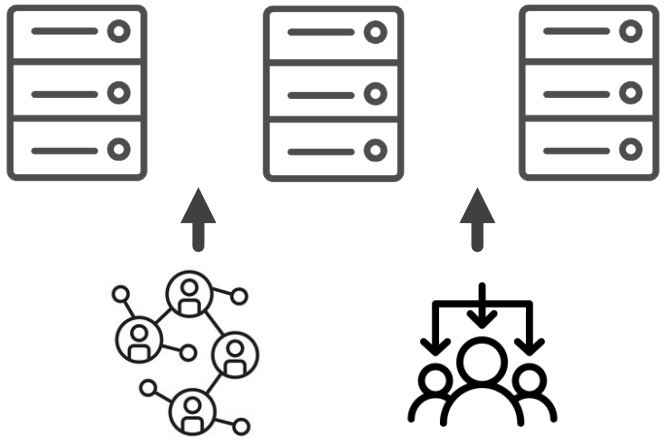
For example, you may want to find out the demographic information on the sale of a certain product, but the data you have received does not capture the gender. You either need to source this data point and merge it with your existing dataset, or you may need to remove, and not consider the records with this field missing. We will explore many more examples of data cleaning further on in the course.

TRANSFORMATION PHASE

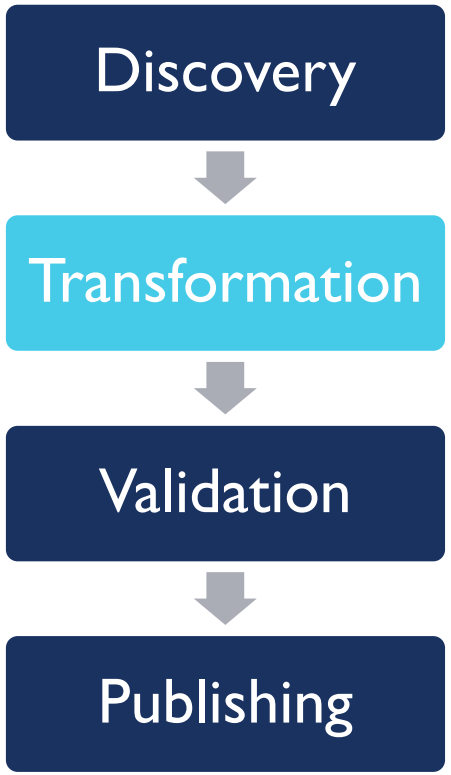
- Discovery
- Transformation
- Validation
- Publishing

Structuring Data | Normalizing and Denormalizing Data | Cleaning data | Enriching Data

Enriching the data—is the fourth type of transformation. When you consider the data you have, to look at additional data points that could make your analysis more meaningful, you are looking at enriching your data. For example, in a large organization with information fragmented across systems, you may need to enrich the dataset provided by one system with information available in other systems, or even public datasets.



TRANSFORMATION PHASE



Structuring Data

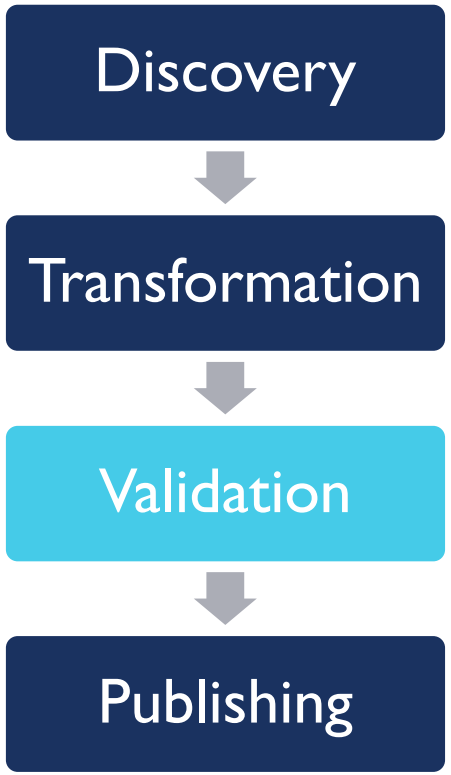
Normalizing and Denormalizing Data

Cleaning data

Enriching Data

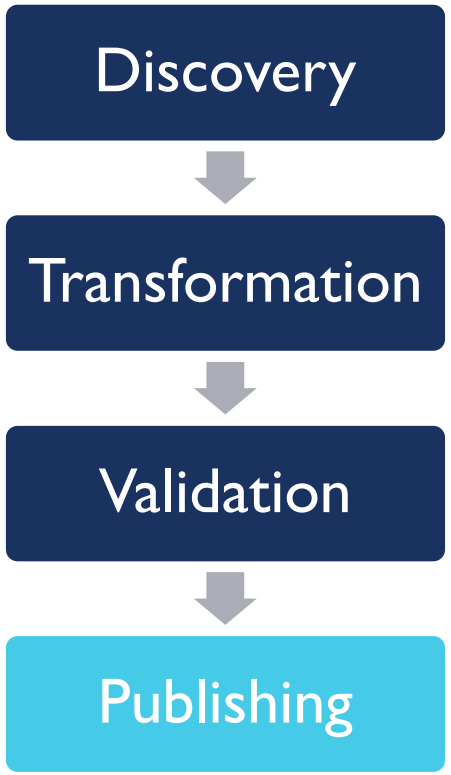
Consider a scenario where you sell IT peripherals to businesses and want to analyze the buying patterns of your customers over the last five years. You have the customer master and transaction tables from where you've captured the customer information and purchase history. Supplementing your dataset with the performance data of these businesses, possibly available as a public dataset, could be valuable for you to understand factors influencing their purchase decisions. Inserting metadata also enriches data. For example, computing a sentiment score from a customer feedback log, collecting geo-based weather data from a resorts location to analyze occupancy trends, or capturing published time and tags for a blog post.

TRANSFORMATION PHASE



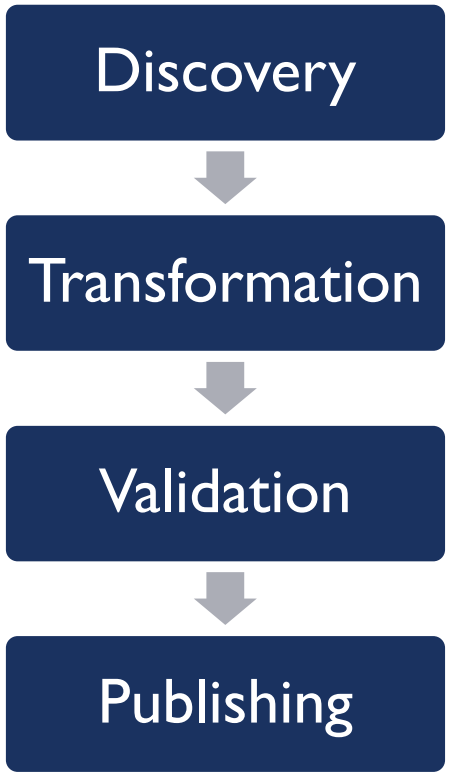
- Checking the quality of the data post structuring, normalizing, cleaning, and enriching.
- Validation rules refer to repetitive programming steps used to verify the consistency, quality, and security of the data you have.

TRANSFORMATION PHASE

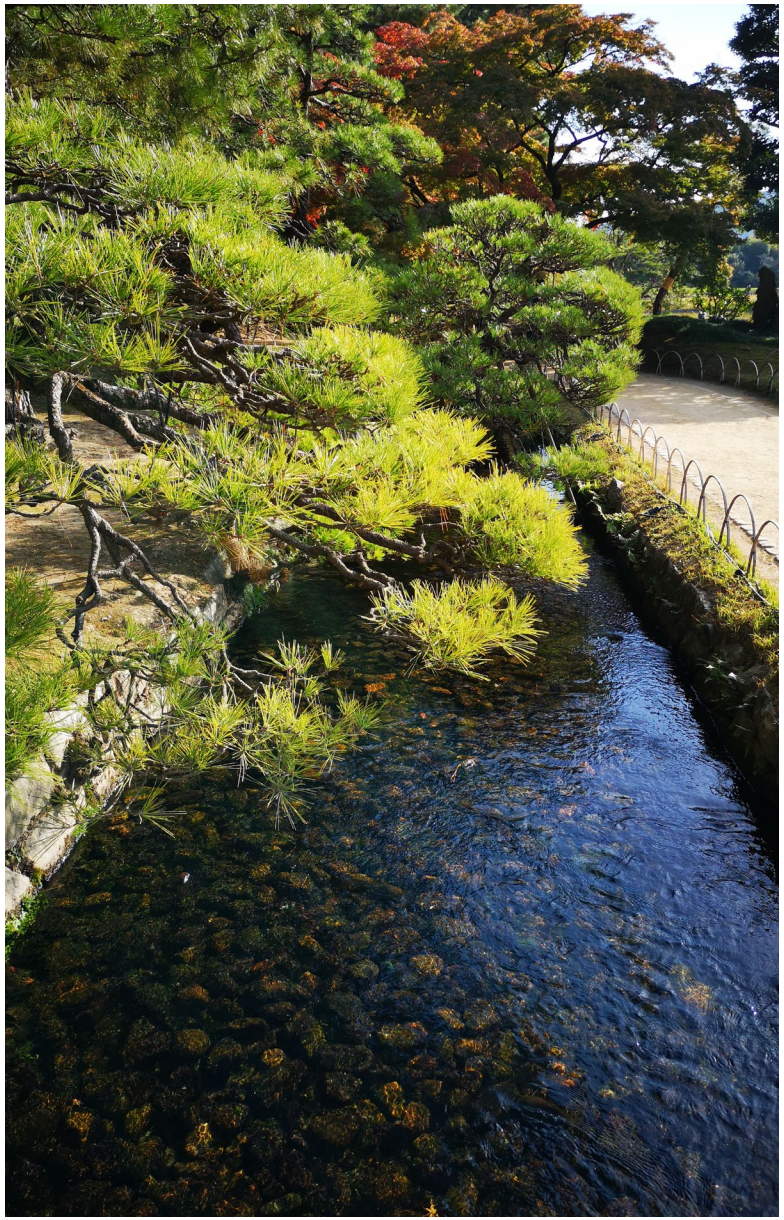


- Delivering the output of the wrangled data for downstream project needs.
- What is published is the transformed and validated version of the input dataset along with the metadata about the dataset.

DATA WRANGLING PROCESS DOCUMENTATION



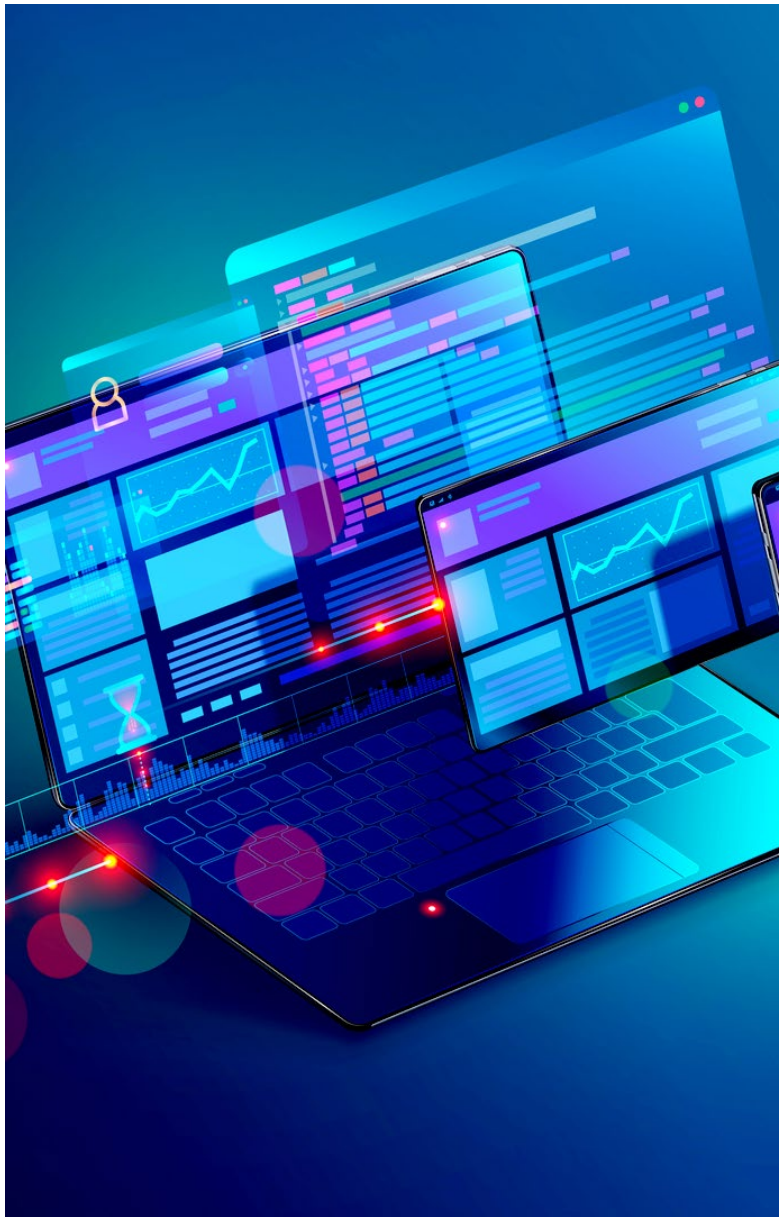
- Lastly, it is important to note the criticality of documenting the steps and considerations you have taken to convert the raw data to analysis-ready data. All phases of data wrangling are iterative in nature. In order to replicate the steps and to revisit your considerations for performing these steps, it is vital that you document all considerations and actions.



1.3 TOOLS FOR DATA WRANGLING

By the end of this topic, you should be able to:

- Understand the definitions and terminologies used in data wrangling.
- Differentiate various data wrangling terminologies.



TOOLS FOR DATA WRANGLING

Some of the popularly used data wrangling software and tools, such as:

- Excel Power Query / Spreadsheets
- OpenRefine
- Google DataPrep
- Watson Studio Refinery
- Trifacta Wrangler
- R
- Python

EXCEL POWER QUERY / SPREADSHEETS

- Spreadsheets such as Microsoft Excel and Google Sheets have a host of features and in-built formulae that can help you identify issues, clean, and transform data.
- Add-ins are available that allow you to import data from several different types of sources and clean and transform data as needed—such as Microsoft Power Query for Excel and Google Sheets Query function for Google Sheets.

OPENREFINE

- OpenRefine is an open-source tool
- Can import and export data in a wide variety of formats, such as TSV, CSV, XLS, XML, and JSON.
- Can clean data, transform it from one format to another, and extend data with web services and external data.
- Easy to learn
- Easy to use
- Offers menu-based operations, which means you don't need to memorize commands or syntax.

GOOGLE DATAPREP

- an intelligent cloud data service
- Can visually explore, clean, and prepare both structured and unstructured data for analysis.
- A fully managed service, which means you don't need to install or manage the software or the infrastructure.
- Extremely easy to use.
- With every action that you take, you get suggestions on what your ideal next step should be.
- Automatically detect schemas, data types, and anomalies.

WATSON STUDIO REFINERY

- Available via IBM Watson Studio
- Allows you to discover, cleanse, and transform data with built-in operations.
- Transforms large amounts of raw data into consumable, quality information that is ready for analytics.
- Offers the flexibility of exploring data residing in a spectrum of data sources.
- Detects data types and classifications automatically
- Enforces applicable data governance policies automatically.

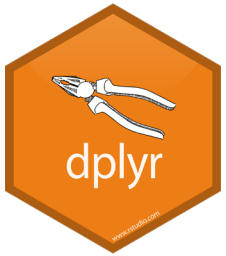
TRIFACTA WRANGLER

- An interactive cloud-based service for cleaning and transforming data.
- Takes messy, real-world data and cleans and rearranges it into data tables
- Can export to Excel, Tableau, and R.
- Known for its collaboration features, allowing multiple team members to work simultaneously.

R

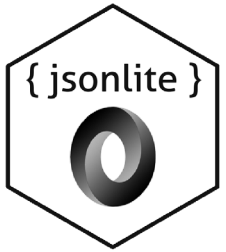
- Offers a series of libraries and packages that are explicitly created for wrangling messy data
  - Dplyr
  - Data.table
  - Jsonlite.
- Using these libraries, you can investigate, manipulate, and analyze data.

R



- Dplyr – A powerful library for data wrangling. It has a precise and straightforward syntax.

- Data.table – Helps to aggregate large data sets quickly.



- Jsonlite – A robust JSON parsing tool, great for interacting with web APIs.

PYTHON

Python has a huge library and set of packages that offer powerful data manipulation capabilities.

- Jupyter Notebook is an open-source web application widely used for data cleaning and transformation, statistical modeling, also data visualization.

PYTHON

Python has a huge library and set of packages that offer powerful data manipulation capabilities.

- Numpy, or Numerical Python, is the most basic package that Python offers.
- It is fast, versatile, interoperable, and easy to use.
- It provides support for large, multi-dimensional arrays and matrices, and high-level mathematical functions to operate on these arrays.

PYTHON

Python has a huge library and set of packages that offer powerful data manipulation capabilities.

- Pandas is designed for fast and easy data analysis operations.
- Allows complex operations such as merging, joining, and transforming huge chunks of data, performed using simple, single-line commands.
- Can prevent common errors that result from misaligned data coming in from different sources.

TOOLS FOR DATA WRANGLING
SUMMARY

Tools for data wrangling come with varying capabilities and dimensions. Your decision regarding the best tool for your needs will depend on factors that are specific to your use case, infrastructure, and teams such as:

- Supported data size
- Data structures
- Cleaning and transformation capabilities
- Infrastructure needs
- Ease of use
- Learnability

1.4 DATA WRANGLING APPLICATION

By the end of this topic, you should be able to:

- Understand the definitions and terminologies used in data wrangling.
- Differentiate various data wrangling terminologies.

DATA WRANGLING APPLICATION

Data wrangling or data munging is used for diverse use-cases. However, two of the most common use-cases of data wrangling are:

- Fraud Detection
- Customer Behavior Analysis

FRAUD DETECTION

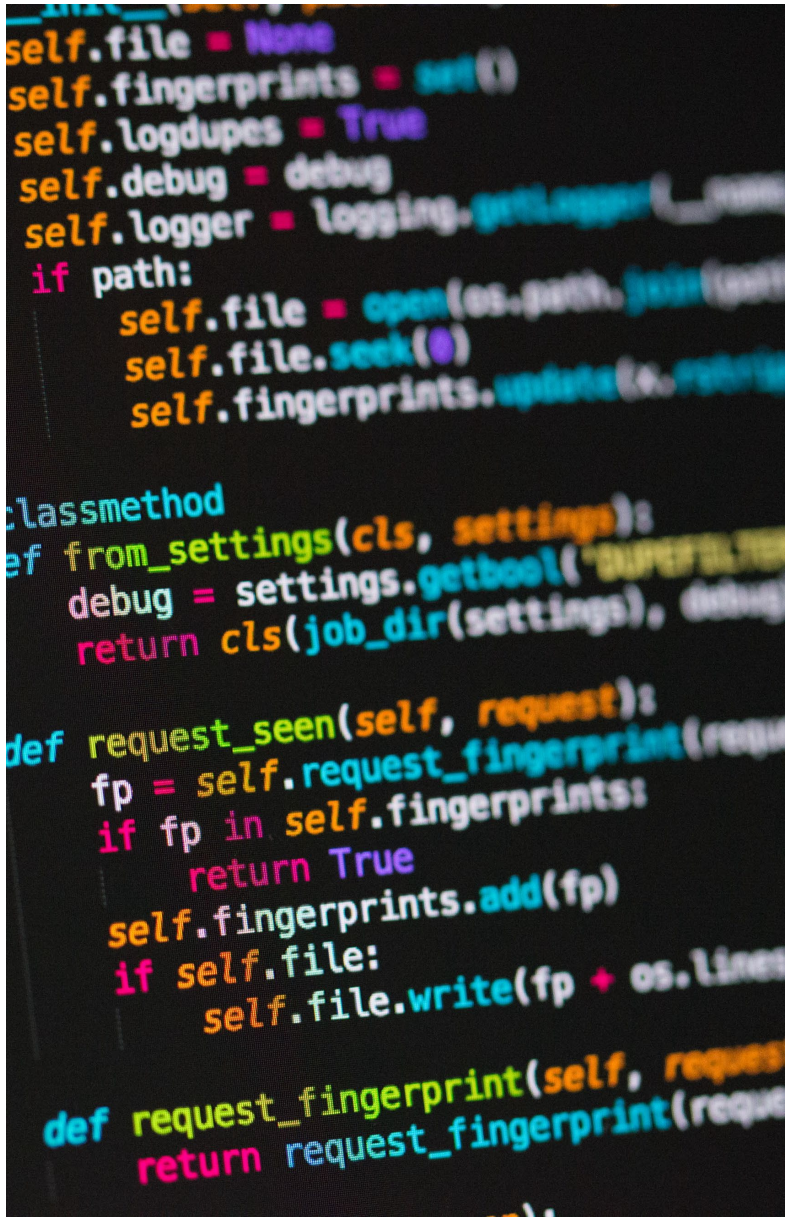
Using a data wrangling tool, a business can perform the following:

- Distinguish corporate fraud by identifying unusual behavior by examining intricate information like multi-party and multi-layered emails or web chats.
- Support data security by allowing non-technical operators to examine and wrangle data quickly to keep pace with billions of daily security tasks.
- Ensure precise and repeatable modeling outcomes by standardizing and quantifying structured and unstructured datasets.
- Enhance compliance by ensuring your business is complying with industry and government standards by following security protocols during integration.

CUSTOMER BEHAVIOR ANALYSIS

A data wrangler can help your business processes get more precise insights quickly via customer behavior analysis. It empowers the marketing team to take business decisions into their hands and make the best of it. You can use it to:

- Decrease the time spent on data preparation for analysis
- Quickly understand the business value of your data
- Allow your analytics team to utilize the customer behavior data directly
- Empower data analysts to find out data trends via data discovery and visual profiling



1.5 DATA WRANGLING USING PYTHON

By the end of this topic, you should be able to:

- Understand the definitions and terminologies used in data wrangling.
- Differentiate various data wrangling terminologies.

THE GOALS OF DATA WRANGLING WITH PYTHON:

- Gathering data from numerous sources to reveal a more profound intelligence within it
- Provide actionable and accurate data in the hands of business/data analysts in a timely matter
- Reduce the time spent collecting and organizing, in short cleaning unruly data before it can be used
- Enable data analysts and scientists to focus on the analysis of data, not the wrangling part
- Help senior leaders in an organization to take better decisions

DATA WRANGLING WITH PYTHON USING PANDAS LIBRARY

- One of the preferred tools for data visualisation in Python is Pandas Library. It used for data manipulation and analysis. It was originally built by Numpy. The data structure offered by Pandas is fast, expressive and flexible. These are specifically designed to make real-world data analysis easier.
- However, it is not that easy to use Pandas Library for the beginners as it may seem quite elaborate and hard to find a single point entry to the material. To start with, you may read books like and Pandas Cookbook by Julia Evan and understand the basics of Python. Anaconda and other video tutorials can also be used to interact with Pandas easily, even for the non-coders.

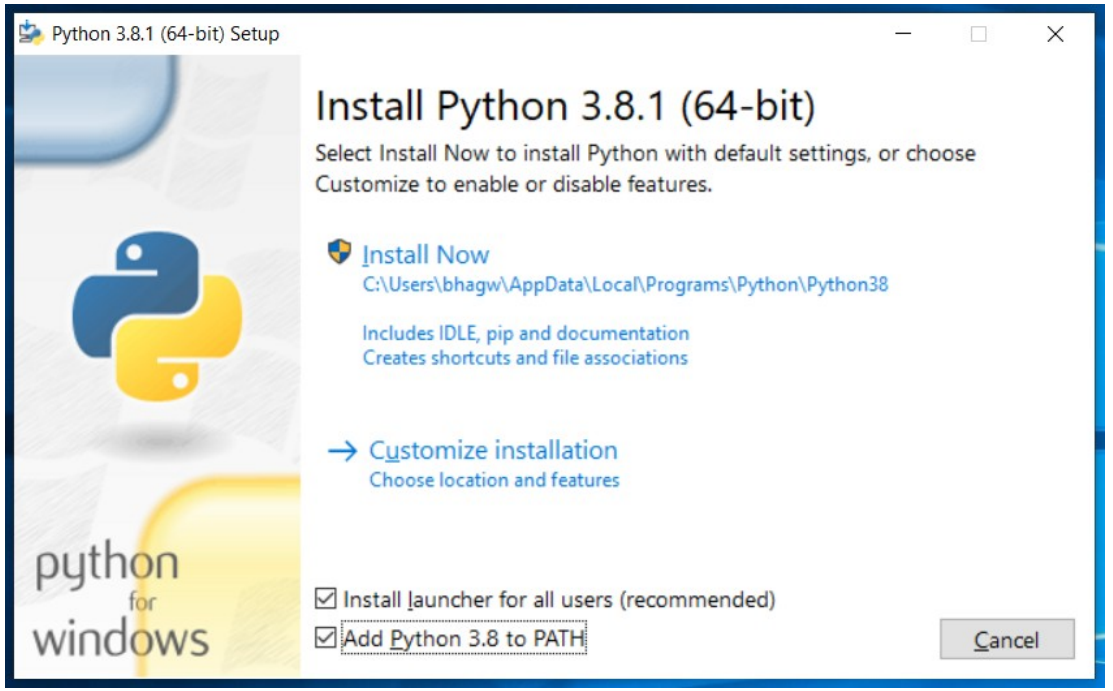
DATA WRANGLING USING PYTHON

Data wrangling in python deals with the below functionalities:

- **Data exploration:** In this process, the data is studied, analyzed and understood by visualizing representations of data.
- **Dealing with missing values:** Most of the datasets having a vast amount of data contain missing values of *NaN, they are needed to be taken* care of by replacing them with mean, mode, the most frequent value of the column or simply by dropping the row having a *NaN* value.
- **Reshaping data:** In this process, data is manipulated according to the requirements, where new data can be added, or pre-existing data can be modified.
- **Filtering data:** Sometimes datasets are comprised of unwanted rows or columns which are required to be removed or filtered
- **Other:** After dealing with the raw dataset with the above functionalities we get an efficient dataset as per our requirements and then it can be used for a required purpose like data analyzing, machine learning, data visualization, model training etc.

Python Installation

1. To download the installation package, go to https://www.python.org
2. Download the latest version of **Python 3** package The latest version is Python 3.8.2.
3. Make sure to download the compatible version with your operating system (Windows/MacOS).



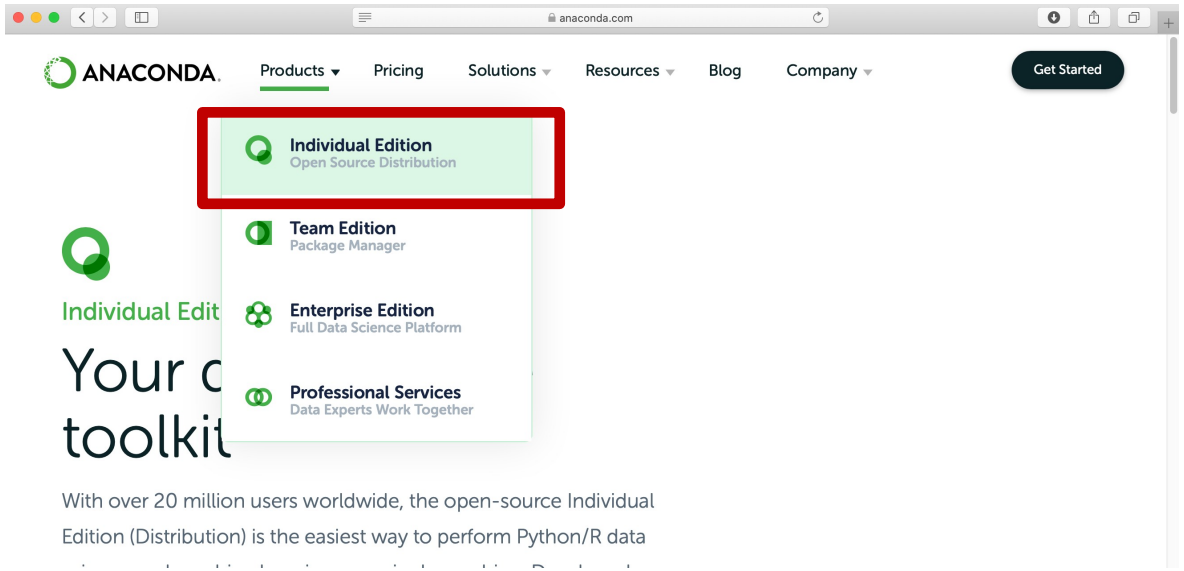
Anaconda Distribution Installation

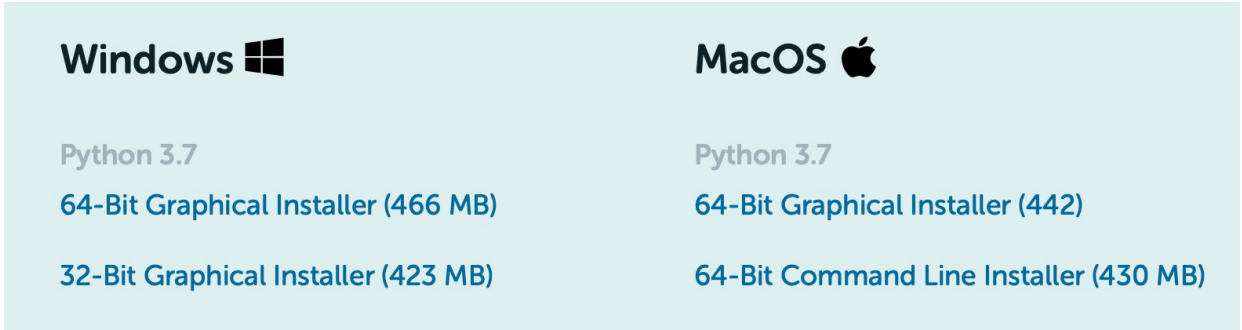
1. To download the installation package, go to https://www.anaconda.com/products/individual

2. Install the *Individual Edition*.

3. Scroll to the bottom part of the this page and select for the **ONLY INSTALL THE VERSION PYTHON 3.7**. DO NOT INSTALL VERSION PYTHON 2.7.

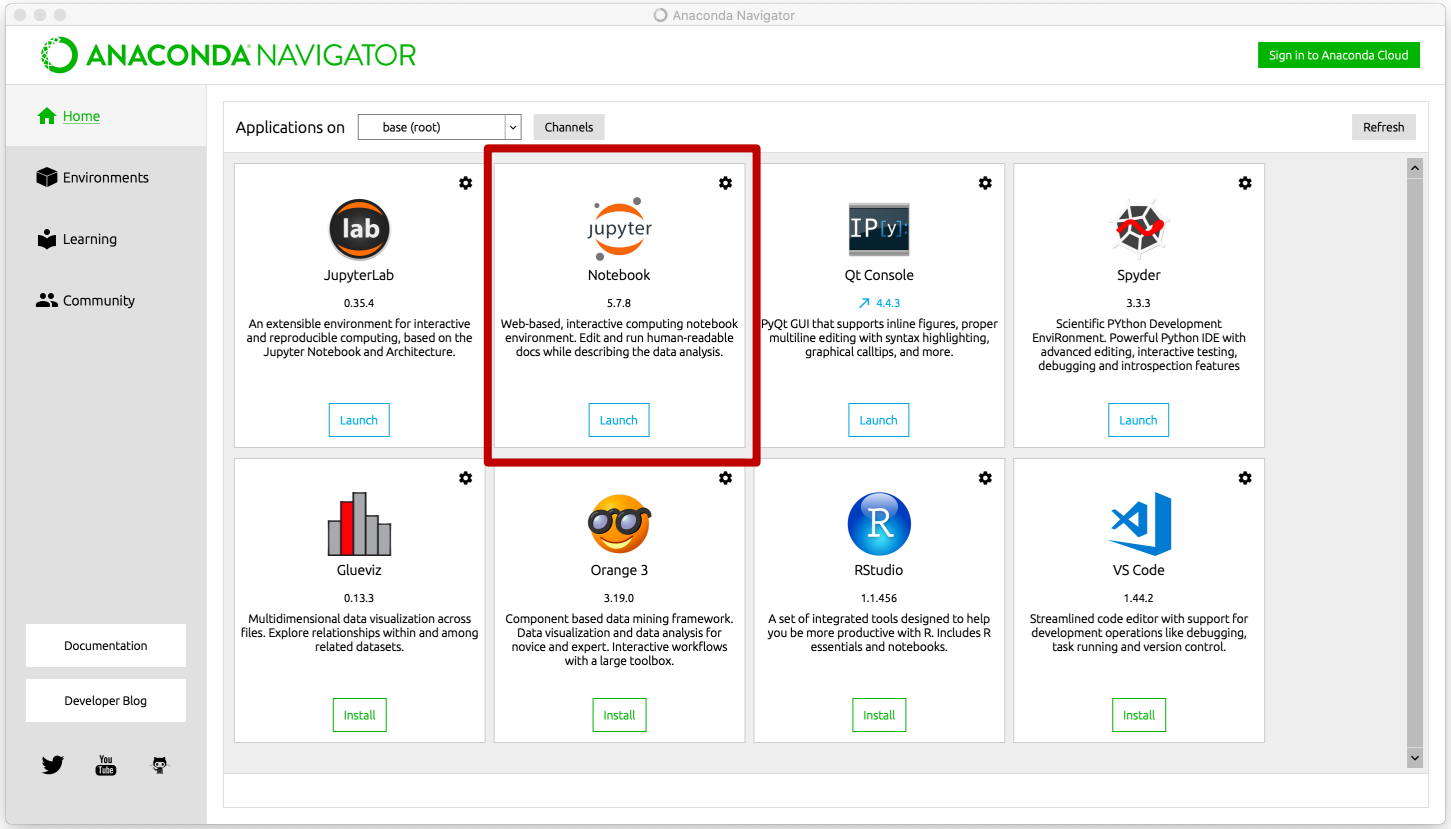
4. Make sure it compatible with your operating system (Windows/MacOS) and computer system (64 or 32 bit)



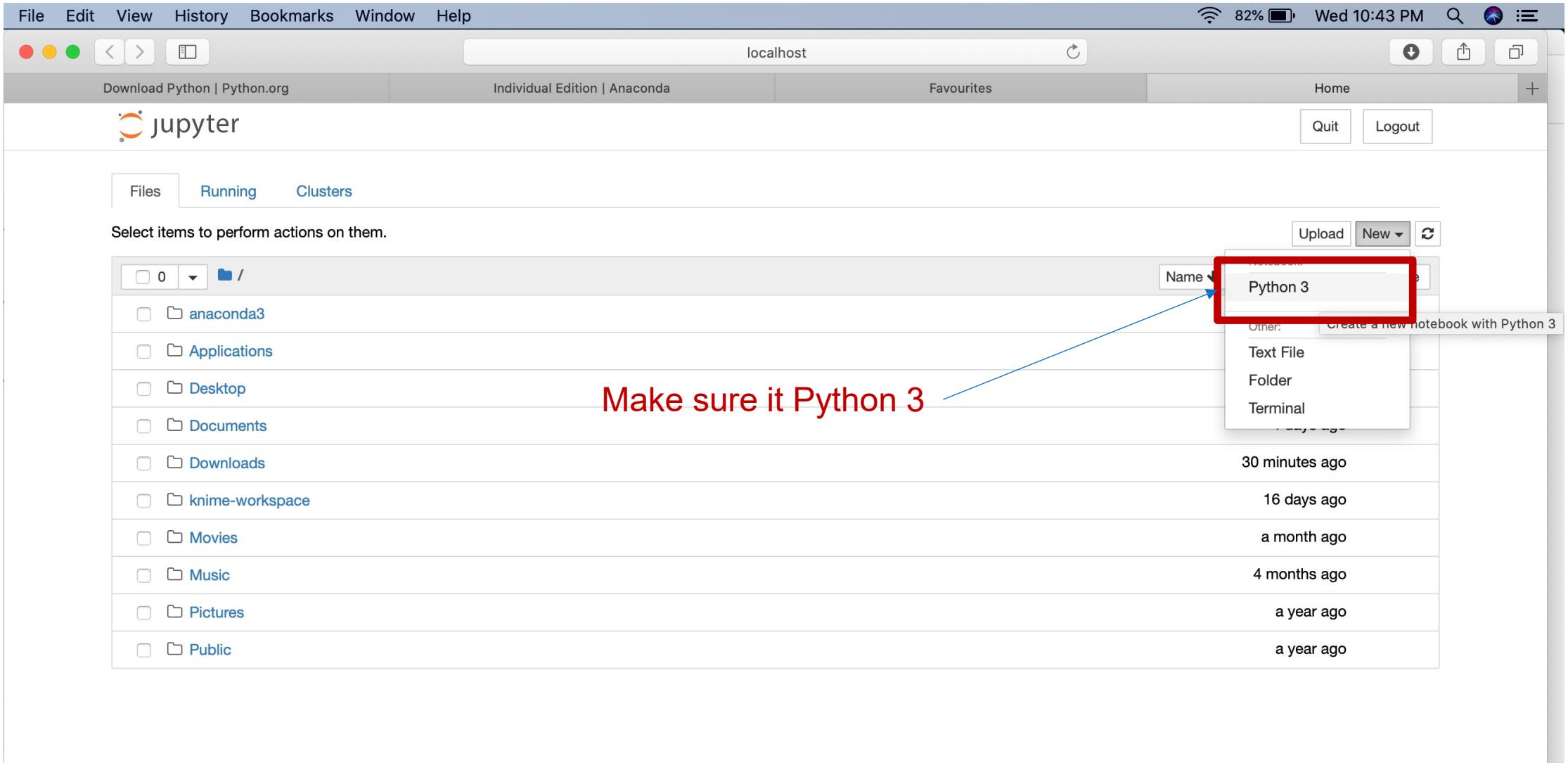


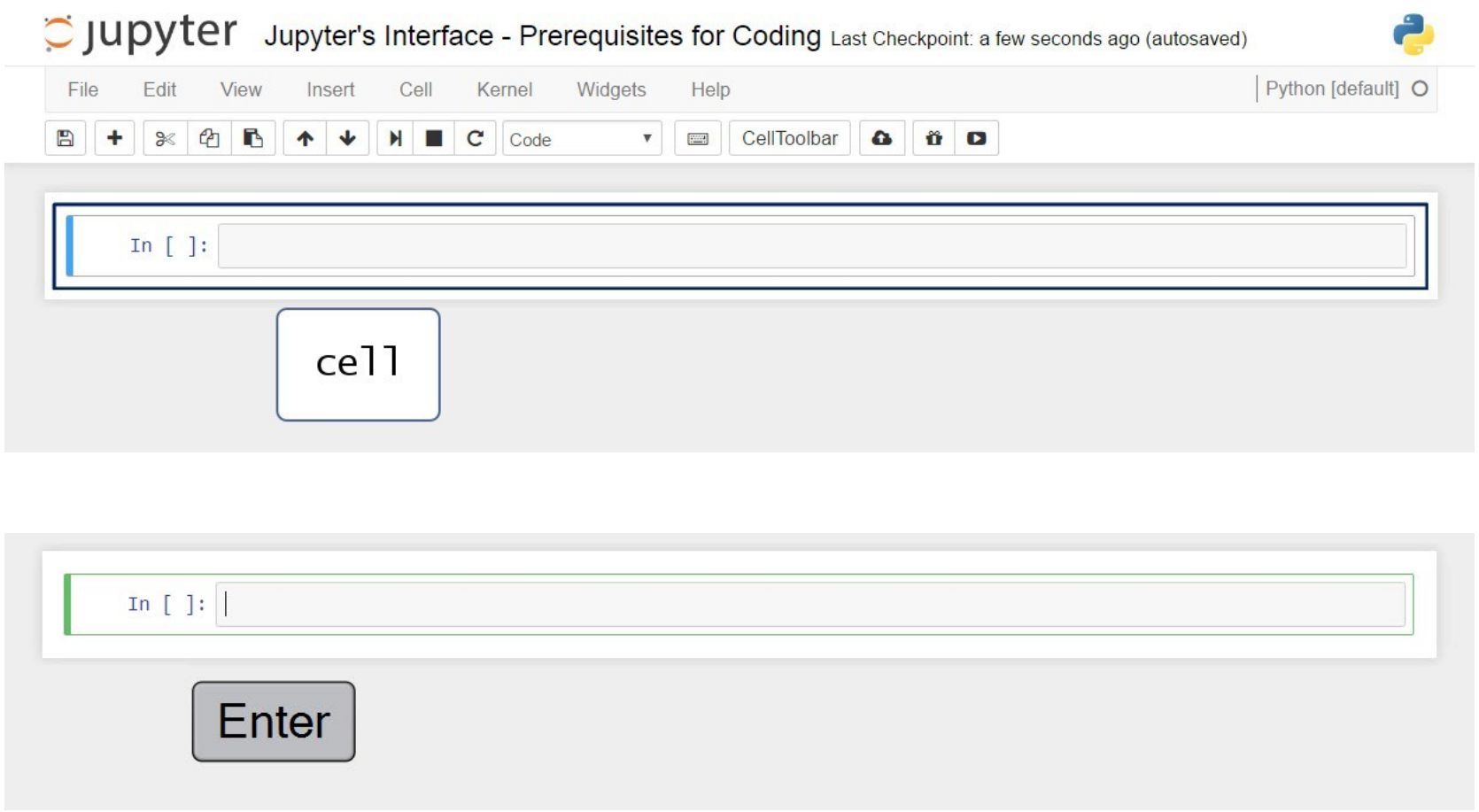
Install the - Jupyter Notebook

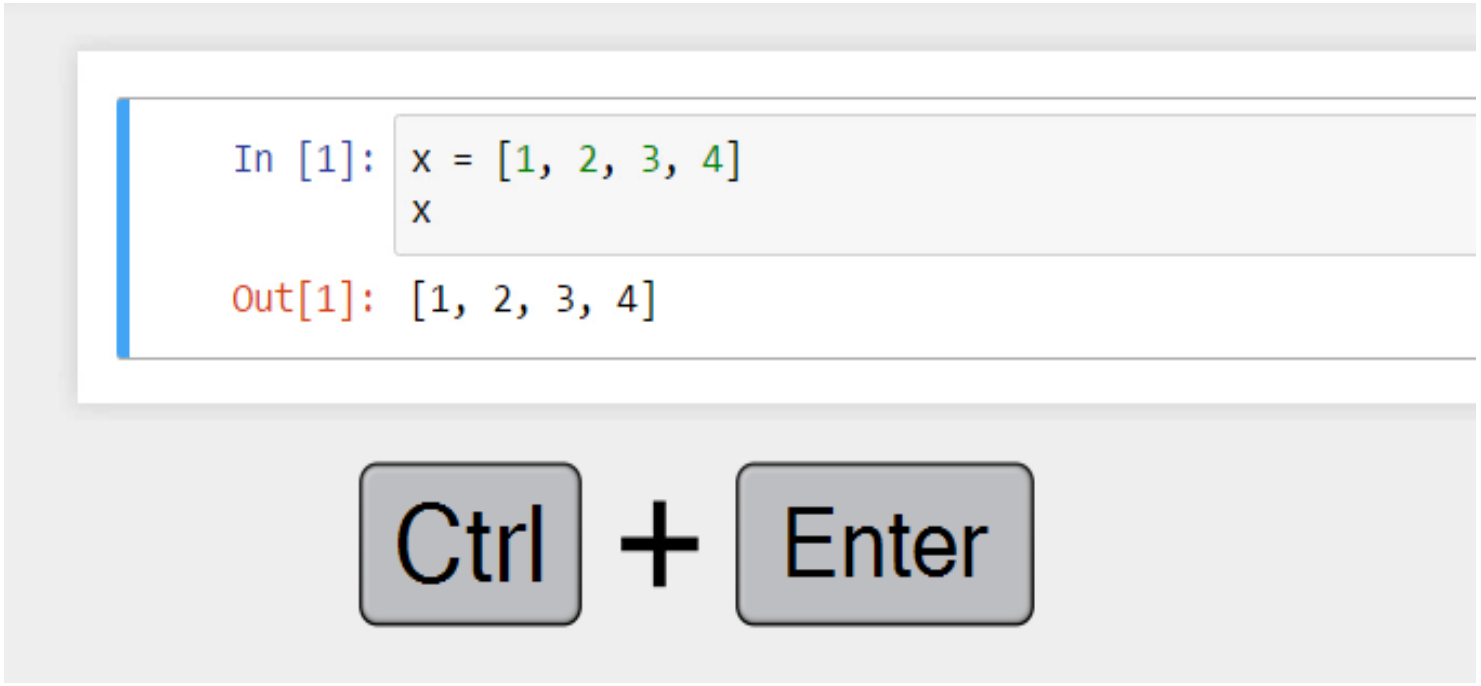
1. Open your Anaconda program.
2. Click the *Install* button at Jupyter Notebook selection.
3. Launch your Jupyter Notebook.
4. Test your Jupyter Notebook as provided in

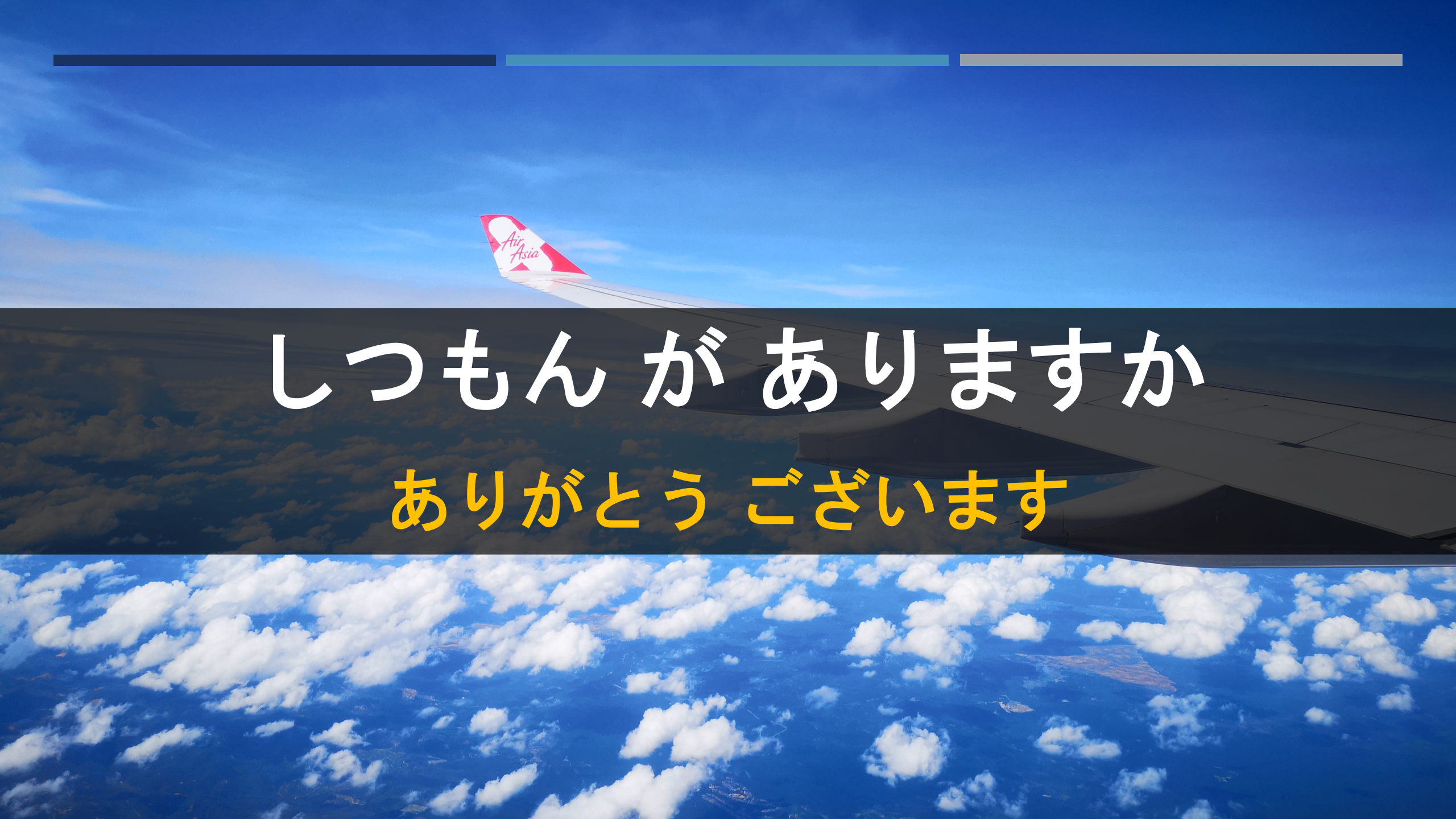


Testing Your Jupyter Notebook




jupyter
Jupyter's Interface - Prerequisites for Coding
Last Checkpoint: a few seconds ago (autosaved)
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Python [default]
Code
CellToolbar
In [ ]:
cell
In [ ]:
Enter


In [1]:
x = [1, 2, 3, 4]
x
Out[1]:
[1, 2, 3, 4]
Ctrl
+
Enter



しつもん が ありますか

ありがとう ございます

FLASH CARD GAME

Here are four items.

1. Use either pen and paper, or any document on your computer. Write a short description of these four objects. Entirely free form. About 3 or four sentences only.
2. Discuss your description in your breakout room (with partner). Circle or highlight nouns and adjectives.
3. Narrow down to two kinds of adjectives.
    a. make color one of them.
4. Draw a picture with boxes for each noun and adjective, links with lines.

Data wrangling, also known as data munging, is an iterative process that involves data exploration, transformation, validation, and making it available for a credible and meaningful analysis. It includes a range of tasks involved in preparing raw data for a clearly defined purpose, where raw data at this stage is data that has been collated through various data sources in a data repository. Data wrangling captures a range of tasks involved in preparing data for analysis. Typically, it is a 4-step process that involves—Discovery, Transformation, Validation, and Publishing. The Discovery phase, also known as the Exploration phase, is about understanding your data better with respect to your use case. The objective is to figure out specifically how best you can clean, structure, organize, and map the data you have for your use case. The next phase, which is the Transformation phase, forms the bulk of the data wrangling process. It involves the tasks you undertake to transform the data, such as structuring, normalizing, denormalizing, cleaning, and enriching the data. Let's begin with the first transformation task – Structuring. This task includes actions that change the form and schema of your data. The incoming data can be in varied formats. You might, for example, have some data coming from a relational database and some data from Web APIs. In order to merge them, you will need to change the form or schema of your data. This change may be as simple as changing the order of fields within a record or dataset or as complex as combining fields into complex structures. Joins and Unions are the most common structural transformations used to combine data from one or more tables. How they combine the data is different. Joins combine columns. When two tables are joined together, columns from the first source table are combined with columns from the second source table—in the same row. So, each row in the resultant table contains columns from both tables. Unions combine rows. Rows of data from the first source table are combined with rows of data from the second source table into a single table. Each row in the resultant table is from one source table or another. Transformation can also include normalization and denormalization of data. Normalization focuses on cleaning the database of unused data and reducing redundancy and inconsistency. Data coming from transactional systems, for example, where a number of insert, update, and delete operations are performed on an ongoing basis, are highly normalized. Denormalization is used to combine data from multiple tables into a single table so that it can be queried faster. For example, normalized data coming from transactional systems is typically denormalized before running queries for reporting and analysis. Another transformation type is Cleaning. Cleaning tasks are actions that fix irregularities in data in order to produce a credible and accurate analysis. Data that is inaccurate, missing, or incomplete can skew the results of your analysis and need to be considered. It could also be that the data is biased, or has null values in relevant fields, or have outliers. For example, you may want to find out the demographic information on the sale of a certain product, but the data you have received does not capture the gender. You either need to source this data point and merge it with your existing dataset, or you may need to remove, and not consider the records with this field missing. We will explore many more examples of data cleaning further on in the course. Enriching the data—is the fourth type of transformation. When you consider the data you have, to look at additional data points that could make your analysis more meaningful, you are looking at enriching your data. For example, in a large organization with information fragmented across systems, you may need to enrich the dataset provided by one system with information available in other systems, or even public datasets. Consider a scenario where you sell IT peripherals to businesses and want to analyze the buying patterns of your customers over the last five years. You have the customer master and transaction tables from where you've captured the customer information and purchase history. Supplementing your dataset with the performance data of these businesses, possibly available as a public dataset, could be valuable for you to understand factors influencing their purchase decisions. Inserting metadata also enriches data. For example, computing a sentiment score from a customer feedback log, collecting geo-based weather data from a resorts location to analyze occupancy trends, or capturing published time and tags for a blog post. After transformation, the next phase in Data Wrangling is Validation. This is where you check the quality of the data post structuring, normalizing, cleaning, and enriching. Validation rules refer to repetitive programming steps used to verify the consistency, quality, and security of the data you have. This brings us to Publishing—the fourth phase of the data wrangling process. Publishing involves delivering the output of the wrangled data for downstream project needs. What is published is the transformed and validated version of the input dataset along with the metadata about the dataset. Lastly, it is important to note the criticality of documenting the steps and considerations you have taken to convert the raw data to analysis-ready data. All phases of data wrangling are iterative in nature. In order to replicate the steps and to revisit your considerations for performing these steps, it is vital that you document all considerations and actions.

VIEWPOINTS: DATA PREPARATION AND RELIABILITY

In this segment, data professional share what portion of their job involves gathering, cleaning, and preparing data for analysis. I would say, a relatively big proportion of my job involves gathering, preparing, and cleaning data for analysis. I work at a company with a really great data engineering team. So I don't have to do this kind of work as much as some other data scientists do. But still, any person that is working closely with data, be they're a data scientist, a data analyst, machine learning engineer, really needs to get comfortable understanding where the data comes from. Inevitably, no dataset is perfect. There's always going to be compromises or small errors. So it's really important to spend a significant portion of your time, understanding the underlined data that was used to generate the dataset and what some potential problems might be with that data. My job as a CPA involves a lot of analysis. Financial statements, account activity, assessing processes, and controls. The gathering piece can be pretty simple as long as, the accounting information resides in a general ledger system or a central repository where the data is easy to gather. Probably, about 30 percent of the job is laying everything out. So when you get into analytics of it, you can just dive right into the meat and potatoes of it. So you need to track the data, make sure it's accurate, make sure things are adding up. Make sure you have all mumps of information. So for example, on financial statements, I need to make sure that people have given me 12 months of [inaudible] statements, I'm not missing any data and that if I am, that I have enough information to be able to project or to forecast or even look back to estimate what was done in the [inaudible] based on what I have. That is definitely helpful. In this segment, data professionals talk about the steps they take to ensure data is reliable. One of the essential steps to making sure your data is reliable, is to run summary statistics on individual columns in your data and make sure that they're consistent with reality. For example, if you have a column somewhere that records visits per month to a website and you run summary statistics on that column, you get the minimum, the mean, the median, the max, and you see something funky like, one month there's negative visits or something like this. You know, that data isn't reliable. Financial information in particular must be reliable. It must be non-bias. It must be free from error. Those are just a few of the many attributes that are necessary for data to be relied upon. So doing what I call a logic check before you get into the details of a transaction. Does it make sense at a high level? If you expected top-line revenue to increase, but you see that it has drastically decreased, then figure that part out first. Is my source correct? Am I running a query in the right period? Am I pulling the right general ledger account? So start there, make sure that basic data integrity questions have been addressed first. Once we know that the data is reliable, then we can start to deep dive into the reviews and form conclusions about the financial performance based on our analysis of the data.