P.1. Data Preparation for Heterogeneous Datasets

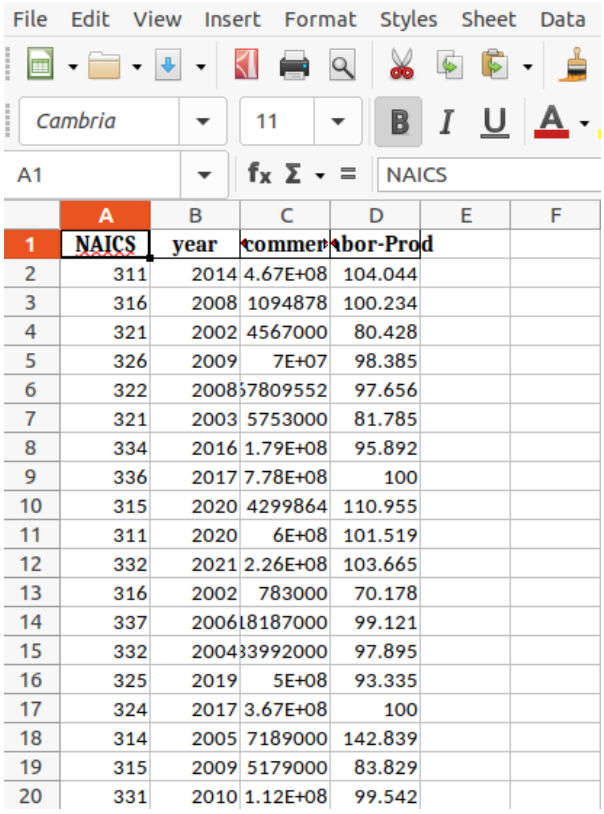
(From Remark 6.2). **Data preparation** is **difficult** because the process is *not objective*, and it is **important** because ML algorithms *learn from data*. Consider the following.

- Data preparation is one of the most **important** steps in any data-mining project – and traditionally, one of the most **time consuming**.
  - ☹ Datasets may involve missing values.
  - ☹ In many cases, datasets are saved in various formats.
- Often, it takes **up to 80% of the time**.
- Data preparation is **not a once-off process**; that is, it is iterative as you understand the problem deeper on each successive pass.

Objectives. In this project, you will combine and sort data values saved in multiple Excel files.

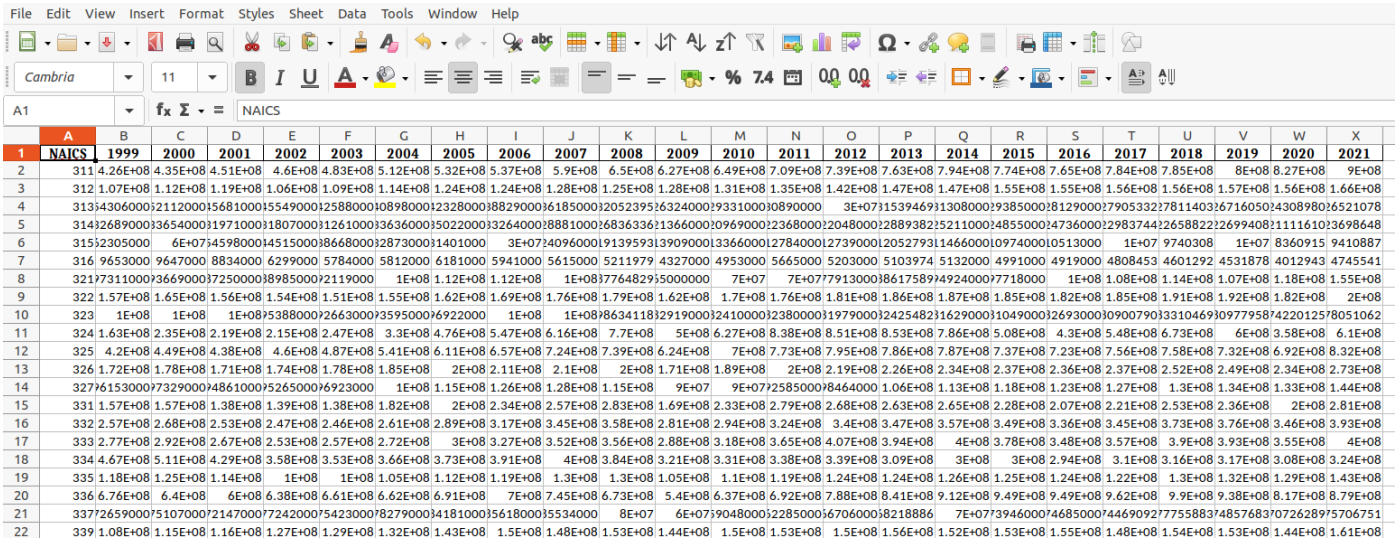
- **Excel Data**
  - Excel is easy to use and analyze data.
  - However, the whole data is often saved in multiple files.
  - If you would like to employ powerful Python libraries effectively, the data must be combined and sorted meaningfully.
  - The trimmed data can also be saved in an Excel file. In this case, you may enjoy both advantages of Python and benefits of Excel.
- **Python for Data Preparation**:
  - Use Python for combing and sorting data values.
  - Save the trimmed data into an Excel file.

A **modelcode** is implemented for your convenience. Download the code to untar. Then you will see two files and a directory including two Excel data files.



	A	B	C	D	E	F
1	NAICS	year	commer	abor-Prod		
2	311	2014	4.67E+08	104.044		
3	316	2008	1094878	100.234		
4	321	2002	4567000	80.428		
5	326	2009	7E+07	98.385		
6	322	2008	57809552	97.656		
7	321	2003	5753000	81.785		
8	334	2016	1.79E+08	95.892		
9	336	2017	7.78E+08	100		
10	315	2020	4299864	110.955		
11	311	2020	6E+08	101.519		
12	332	2021	2.26E+08	103.665		
13	316	2002	783000	70.178		
14	337	2006	18187000	99.121		
15	332	2004	33992000	97.895		
16	325	2019	5E+08	93.335		
17	324	2017	3.67E+08	100		
18	314	2005	7189000	142.839		
19	315	2009	5179000	83.829		
20	331	2010	1.12E+08	99.542		

Figure P.1: data-ECommerce-Labor_Prod.xlsx



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X
1	NAICS	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021
2	311	4.26E+08	4.35E+08	4.51E+08	4.6E+08	4.83E+08	5.12E+08	5.32E+08	5.37E+08	5.9E+08	6.5E+08	6.27E+08	6.49E+08	7.09E+08	7.39E+08	7.63E+08	7.94E+08	7.74E+08	7.65E+08	7.84E+08	7.85E+08	8E+08	8.27E+08	9E+08
3	312	1.07E+08	1.12E+08	1.19E+08	1.06E+08	1.09E+08	1.14E+08	1.24E+08	1.24E+08	1.28E+08	1.25E+08	1.28E+08	1.31E+08	1.35E+08	1.42E+08	1.47E+08	1.47E+08	1.55E+08	1.55E+08	1.56E+08	1.56E+08	1.57E+08	1.56E+08	1.66E+08
4	313	43060000	21120000	15681000	15549000	12588000	10898000	12328000	8829000	6185000	20523950	6324000	9331000	10890000	3E+07	31539469	31308000	29385000	28129000	27905332	27811403	26716050	24308980	26521078
5	314	26890000	36540000	19710000	18070000	12610000	13636000	15022000	13264000	8881000	6836336	13660000	9769000	12368000	12048000	12889382	15211000	14855000	14736000	12983744	12658822	12699408	11116610	13698648
6	315	12305000	6E+07	45980001	45150003	36680003	28730003	31401000	3E+07	24096000	19139593	13909000	13366000	12784000	12739000	12052793	11466000	10974000	10513000	1E+07	9740308	1E+07	8360915	9410887
7	316	9653000	9647000	8834000	6299000	5784000	5812000	6181000	5941000	5615000	5211979	4327000	4953000	5665000	5203000	5103974	5132000	4991000	4919000	4808453	4601292	4531878	4012943	4745541
8	321	77311000	36690003	37250003	38985000	32119000	1E+08	1.12E+08	1.12E+08	1E+08	37764829	35000000	7E+07	7E+07	77913000	86175897	94924000	77718000	1E+08	1.08E+08	1.14E+08	1.07E+08	1.18E+08	1.55E+08
9	322	1.57E+08	1.65E+08	1.56E+08	1.54E+08	1.51E+08	1.55E+08	1.62E+08	1.69E+08	1.76E+08	1.79E+08	1.62E+08	1.7E+08	1.76E+08	1.81E+08	1.86E+08	1.87E+08	1.85E+08	1.82E+08	1.85E+08	1.91E+08	1.92E+08	1.82E+08	2E+08
10	323	1E+08	1E+08	1E+08	53880003	26630003	35950003	46922000	1E+08	1E+08	86341183	29190003	24100003	23800003	19790003	24254823	16290003	10490003	26930003	09007903	33104693	09779587	42201253	78051062
11	324	1.63E+08	2.35E+08	2.19E+08	2.15E+08	2.47E+08	3.3E+08	4.76E+08	5.47E+08	6.16E+08	7.7E+08	5E+08	6.27E+08	8.38E+08	8.51E+08	8.53E+08	7.86E+08	5.08E+08	4.3E+08	5.48E+08	6.73E+08	6E+08	3.58E+08	6.1E+08
12	325	4.2E+08	4.49E+08	4.38E+08	4.6E+08	4.87E+08	5.41E+08	6.11E+08	6.57E+08	7.24E+08	7.39E+08	6.24E+08	7E+08	7.73E+08	7.95E+08	7.86E+08	7.87E+08	7.37E+08	7.23E+08	7.56E+08	7.58E+08	7.32E+08	6.92E+08	8.32E+08
13	326	1.72E+08	1.78E+08	1.71E+08	1.74E+08	1.78E+08	1.85E+08	2E+08	2.11E+08	2.1E+08	2E+08	1.71E+08	1.89E+08	2E+08	2.19E+08	2.26E+08	2.34E+08	2.37E+08	2.36E+08	2.37E+08	2.52E+08	2.49E+08	2.34E+08	2.73E+08
14	327	76153000	73290003	48610003	52650003	69230000	1E+08	1.15E+08	1.26E+08	1.28E+08	1.15E+08	9E+07	9E+07	25850003	84640003	1.06E+08	1.13E+08	1.18E+08	1.23E+08	1.27E+08	1.3E+08	1.34E+08	1.33E+08	1.44E+08
15	331	1.57E+08	1.57E+08	1.38E+08	1.39E+08	1.38E+08	1.82E+08	2E+08	2.34E+08	2.57E+08	2.83E+08	1.69E+08	2.33E+08	2.79E+08	2.68E+08	2.63E+08	2.65E+08	2.28E+08	2.07E+08	2.21E+08	2.53E+08	2.36E+08	2E+08	2.81E+08
16	332	2.57E+08	2.68E+08	2.53E+08	2.47E+08	2.46E+08	2.61E+08	2.89E+08	3.17E+08	3.45E+08	3.58E+08	2.81E+08	2.94E+08	3.24E+08	3.4E+08	3.47E+08	3.57E+08	3.49E+08	3.36E+08	3.45E+08	3.73E+08	3.76E+08	3.46E+08	3.93E+08
17	333	2.77E+08	2.92E+08	2.67E+08	2.53E+08	2.57E+08	2.72E+08	3E+08	3.27E+08	3.52E+08	3.56E+08	2.88E+08	3.18E+08	3.65E+08	4.07E+08	3.94E+08	4E+08	3.78E+08	3.48E+08	3.57E+08	3.9E+08	3.93E+08	3.55E+08	4E+08
18	334	4.67E+08	5.11E+08	4.29E+08	3.58E+08	3.53E+08	3.66E+08	3.73E+08	3.91E+08	4E+08	3.84E+08	3.21E+08	3.31E+08	3.38E+08	3.39E+08	3.09E+08	3E+08	3E+08	2.94E+08	3.1E+08	3.16E+08	3.17E+08	3.08E+08	3.24E+08
19	335	1.18E+08	1.25E+08	1.14E+08	1E+08	1E+08	1E+08	1.12E+08	1.19E+08	1.3E+08	1.3E+08	1.05E+08	1.1E+08	1.19E+08	1.24E+08	1.24E+08	1.26E+08	1.25E+08	1.24E+08	1.22E+08	1.3E+08	1.32E+08	1.29E+08	1.43E+08
20	336	6.76E+08	6.4E+08	6E+08	6.38E+08	6.61E+08	6.62E+08	6.91E+08	7E+08	7.45E+08	6.73E+08	5.4E+08	6.37E+08	6.92E+08	7.88E+08	8.41E+08	9.12E+08	9.49E+08	9.49E+08	9.62E+08	9.9E+08	9.38E+08	8.17E+08	8.79E+08
21	337	26590003	51070003	21470003	72420003	54230003	82790003	41810003	56180003	55340003	8E+07	6E+07	90480003	22850003	67060003	82188863	7E+07	39460003	46850003	44690923	77558833	48576833	70726289	75706751
22	339	1.08E+08	1.15E+08	1.16E+08	1.27E+08	1.29E+08	1.32E+08	1.43E+08	1.5E+08	1.48E+08	1.53E+08	1.44E+08	1.5E+08	1.53E+08	1.5E+08	1.56E+08	1.52E+08	1.53E+08	1.55E+08	1.48E+08	1.54E+08	1.53E+08	1.44E+08	1.61E+08

Figure P.2: data-Total-Sale.xlsx

DATA_Preparation.py

```
import numpy as np
import os,sys
from util_DATA_Prep import *

file_ELP = './dataFiles/data-ECommerce-Labor_Prod.xlsx'
file_TS  = './dataFiles/data-Toal-Sale.xlsx'

#-------------------------------------------------
# Read Excel files
#-------------------------------------------------
DATA_ELP, header_ELP = load_data(file_ELP)
DATA_TS,  header_TS  = load_data(file_TS)

#-------------------------------------------------
# Combine and Sort
#  Combine the above for <DATA> and <header>
#  in the order ['NAICS', 'year', 'Total', 'E-commerce', 'Labor-Prod']
#  Sort: First, with 'NAICS code' and then with 'year'
#-------------------------------------------------

# Implement a function or two into "util_DATA_Prep.py" to complete

#-------------------------------------------------
# You can save the trimmed "DATA" to an Excel file:
# First, you should get combined <DATA> and <header>
#-------------------------------------------------
```

util_DATA_Prep.py

```
import numpy as np
import pandas as pd

def load_data(excelfile):
    df = pd.read_excel(excelfile)
    df.fillna(0,inplace=True) #replace nan(=empty spot) by 0
    DATA = df.values; header = df.columns.tolist()
    print('@@',excelfile)
    print('   DATA.dtype,DATA.shape =',DATA.dtype,DATA.shape)
    print('   header =',header)

    return DATA,header
```

What to do

- Download a modelcode: project-Data-Preparation.tar. The data is a part of the *North American Industry Classification System* (**NAICS**) database, for which the USA Federal has been collecting data.
- Implement a function or two to complete the project.
  - Combine two excel files and sort data values.
    Combining order: ['NAICS', 'year', 'Total', 'E-commerce', 'Labor-Prod']
    Sorting: First, with 'NAICS code' and then with 'year'
  - Save the data into an Excel file, say "Trimmed-DATA.xlsx", which looks like

	A	B	C	D	E	F
1	**NAICS**	**year**	**Total**	**E-commerce**	**Labor-Prod**	
2	311	1999	4.26E+8	45757000	92.461	
3	311	2000	4.35E+8	54837000	93.886	
	⋮		⋮			
24	311	2021	9E+8	652192662	99.956	
25	312	1999	1.07E+8	35138000	118.993	
	⋮		⋮			
484	339	2021	1.61E+8	96050779	102.872	

- Open <Trimmed-DATA.xlsx> in Excel.
  For (NAICS=311), draw a figure of three curves for
  (Total, E-commerce, Labor-Prod) vs. (year).

Report: Start with a one-page summary note.

- Add your whole code.
- Export <Trimmed-DATA.xlsx> as PDF to attach.
- Attach the figure, drawn in Excel.