INF1340 Midterm Write-up

CUNRONG LIU

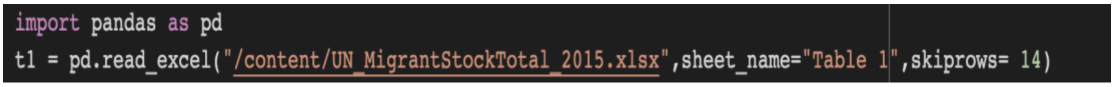
Introduction:
This excel contains nine parts, and we mainly focus on six of them. Each table counts different things, and some tables contain different gender, year, Major area, region, country or area of destination. Some tables also include year range, etc. Details as follows：

Table 1: International migrant stick at mid-year by sex and by major area, region country or area, 1990-2015
Table 2: Total population at mid-year by sex and major area, region, country or area, 1990-2015(thousands)
Table 3: International migrant stock as a percentage of the total population by sex and by major area, region, country or area, 1990-2015
Table 4: Female migrants as a percentage of the international migrant stock by major area, region, country or area, 1990-2015
Table 5: Annual rate of change of the migrant stock by sex and by major area, region, country or area, 1990-2015 (percentage)
Table 6: Estimated refugee stock at mid-year by major area, region, country or area, 1990-2015

None of the six tables are tidy data. These six forms basically have similar problems, for example, each variable does not form a column, and secondly, each observation does not form a row, and the same observation unit should not be distributed in multiple tables, and so on. I will revise these questions according to the principles of tidy data.
Tidy data principle 1: Column names need to be informative, variable names and not values
Tidy data principle 2: Each column needs to consist of one and only one variable
Tidy data principle 3: Variables need to be in cells, not rows and columns
Tidy data principle 4: Each table column needs to have a singular data type
Tidy data principle 5: A single observational units must be in one table

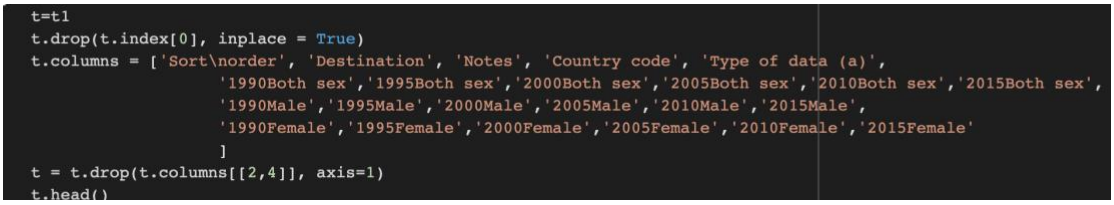
Tidy data process：

First, do some preparatory work. While reading the table through pandas, skip the first 14 rows. These 14 rows include some titles, etc., which we don't need. This step serves as the first step in tables 1 to 6.
For example, table 1:



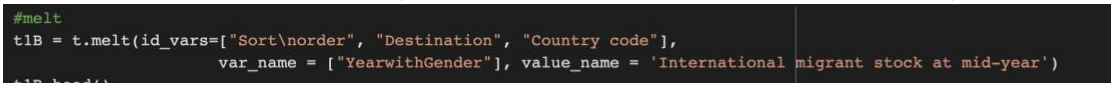
```
import pandas as pd
t1 = pd.read_excel("/content/UN_MigrantStockTotal_2015.xlsx",sheet_name="Table 1",skiprows= 14)
```

Table1:
When using pandas to read the table, some columns are unnamed, so we need to add column names, here I use year plus gender to represent, such as "1990both sex". Second I removed some unnecessary columns like "notes" and "type of data".



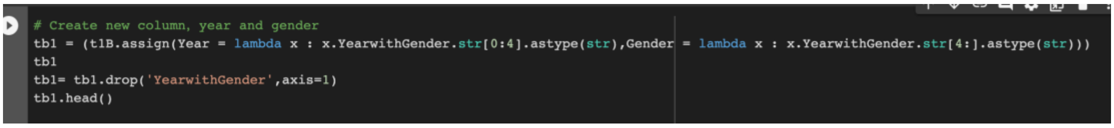
```
t=t1
t.drop(t.index[0], inplace = True)
t.columns = ['Sort\norder', 'Destination', 'Notes', 'Country code', 'Type of data (a)',
             '1990Both sex','1995Both sex','2000Both sex','2005Both sex','2010Both sex','2015Both sex',
             '1990Male','1995Male','2000Male','2005Male','2010Male','2015Male',
             '1990Female','1995Female','2000Female','2005Female','2010Female','2015Female'
             ]
t = t.drop(t.columns[[2,4]], axis=1)
t.head()
```

Column names need to be informative, variable names and not values, so use the Pandas melt() function to change the Data Frame format from wide to long.



```
#melt
t1B = t.melt(id_vars=["Sort\norder", "Destination", "Country code"],
                    var_name = ["YearwithGender"], value_name = 'International migrant stock at mid-year')
```

At the same time, each column only needs to contain one variable. I will divide "yearwithgender" into two new columns, "year" and "gender". Remove previous "yearwithgender" due to new year and gender columns.



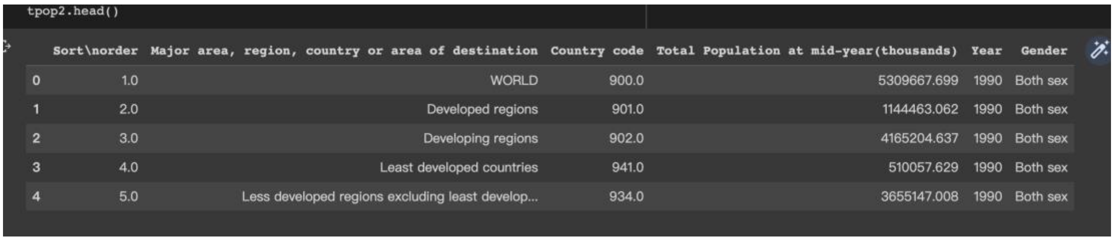
```
# Create new column, year and gender
tbl = (t1B.assign(Year = lambda x : x.YearwithGender.str[0:4].astype(str),Gender = lambda x : x.YearwithGender.str[4:].astype(str)))
tbl
tbl= tbl.drop('YearwithGender',axis=1)
tbl.head()
```

Table 2-5:

The problems in these four tables are the same as those in Table 1. I perform the same steps to perform tidy data.

Final version:

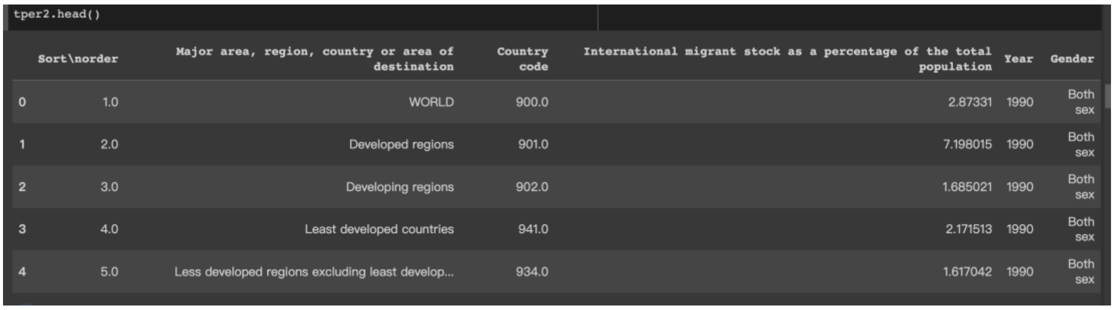
Table2:



```
tpop2.head()
```

	Sort\norder	Major area, region, country or area of destination	Country code	Total Population at mid-year(thousands)	Year	Gender
0	1.0	WORLD	900.0	5309667.699	1990	Both sex
1	2.0	Developed regions	901.0	1144463.062	1990	Both sex
2	3.0	Developing regions	902.0	4165204.637	1990	Both sex
3	4.0	Least developed countries	941.0	510057.629	1990	Both sex
4	5.0	Less developed regions excluding least develop...	934.0	3655147.008	1990	Both sex

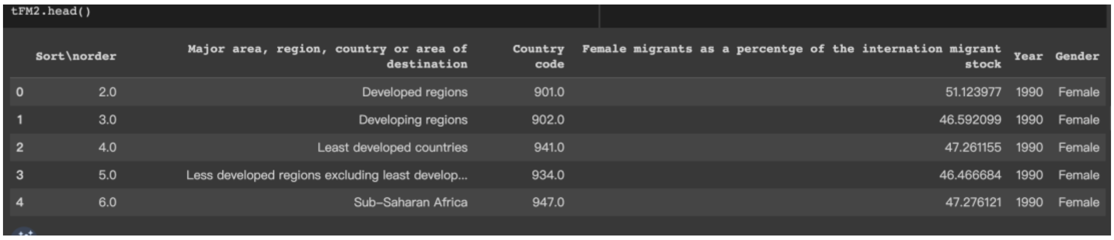
Table3:



```
tper2.head()
```

	Sort\norder	Major area, region, country or area of destination	Country code	International migrant stock as a percentage of the total population	Year	Gender
0	1.0	WORLD	900.0	2.87331	1990	Both sex
1	2.0	Developed regions	901.0	7.198015	1990	Both sex
2	3.0	Developing regions	902.0	1.685021	1990	Both sex
3	4.0	Least developed countries	941.0	2.171513	1990	Both sex
4	5.0	Less developed regions excluding least develop...	934.0	1.617042	1990	Both sex

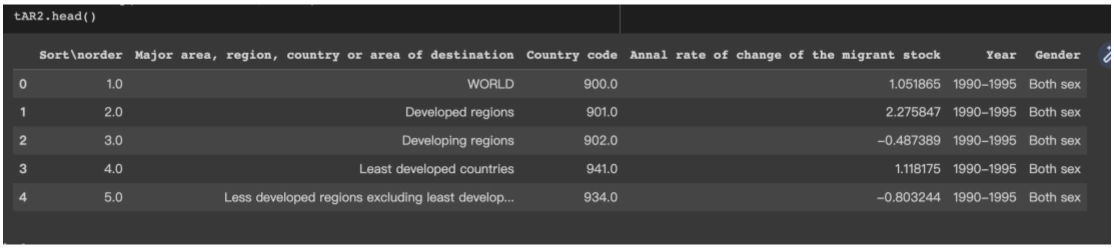
Table4:



```
tFM2.head()
```

	Sort\norder	Major area, region, country or area of destination	Country code	Female migrants as a percentge of the internation migrant stock	Year	Gender
0	2.0	Developed regions	901.0	51.123977	1990	Female
1	3.0	Developing regions	902.0	46.592099	1990	Female
2	4.0	Least developed countries	941.0	47.261155	1990	Female
3	5.0	Less developed regions excluding least develop...	934.0	46.466684	1990	Female
4	6.0	Sub-Saharan Africa	947.0	47.276121	1990	Female

Table 5:

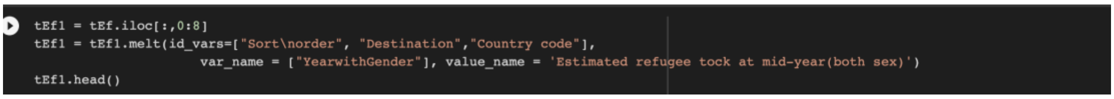


```
tAR2.head()
```

	Sort\norder	Major area, region, country or area of destination	Country code	Annal rate of change of the migrant stock	Year	Gender
0	1.0	WORLD	900.0	1.051865	1990–1995	Both sex
1	2.0	Developed regions	901.0	2.275847	1990–1995	Both sex
2	3.0	Developing regions	902.0	–0.487389	1990–1995	Both sex
3	4.0	Least developed countries	941.0	1.118175	1990–1995	Both sex
4	5.0	Less developed regions excluding least develop...	934.0	–0.803244	1990–1995	Both sex

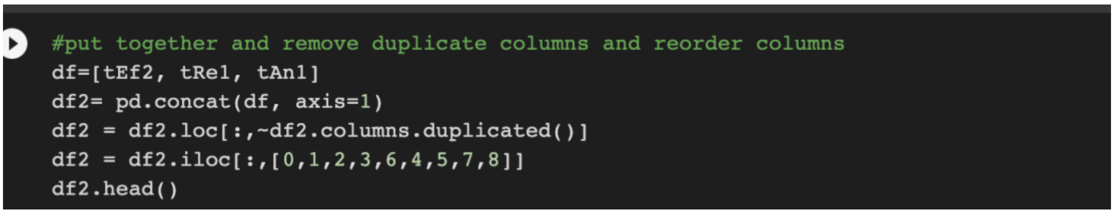
Table 6:

Table 6 includes three types of data, I divided it into three tables to tidy. Here I use “iloc” function to extract required columns. For example, extract the first eight rows as the first table.



```
tEf1 = tEf.iloc[:,0:8]
tEf1 = tEf1.melt(id_vars=["Sort\norder", "Destination","Country code"],
                 var_name = ["YearwithGender"], value_name = 'Estimated refugee tock at mid-year(both sex)')
tEf1.head()
```

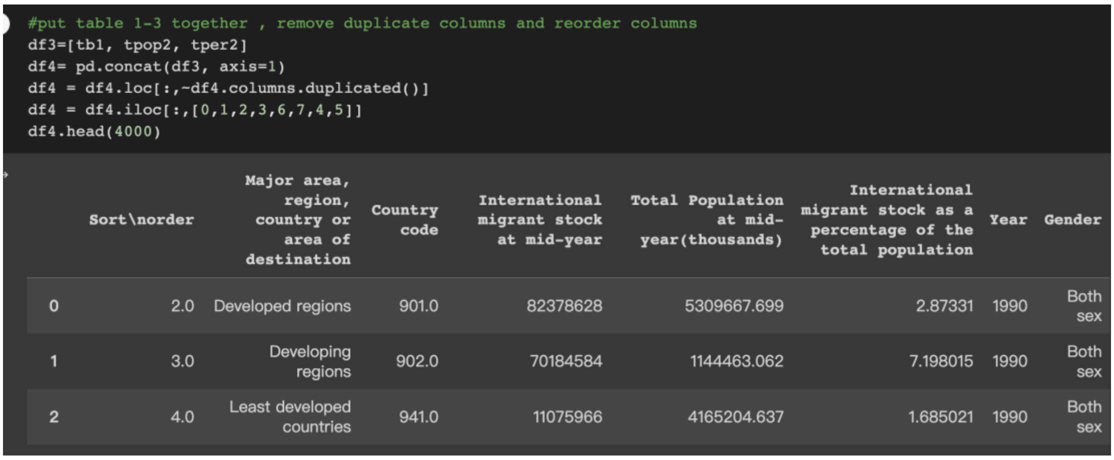
The remaining steps are also the same as the previous table processing method. Finally, I merge the three separate tables together and remove duplicate columns and reorder these columns.



```
#put together and remove duplicate columns and reorder columns
df=[tEf2, tRe1, tAn1]
df2= pd.concat(df, axis=1)
df2 = df2.loc[:,~df2.columns.duplicated()]
df2 = df2.iloc[:,[0,1,2,3,6,4,5,7,8]]
df2.head()
```

The remaining steps of table 6 are the same as the previous table

By the end I put table 1-3 together.



```
#put table 1-3 together , remove duplicate columns and reorder columns
df3=[tb1, tpop2, tper2]
df4= pd.concat(df3, axis=1)
df4 = df4.loc[:,~df4.columns.duplicated()]
df4 = df4.iloc[:,[0,1,2,3,6,7,4,5]]
df4.head(4000)
```

	Sort\norder	Major area, region, country or area of destination	Country code	International migrant stock at mid-year	Total Population at mid-year(thousands)	International migrant stock as a percentage of the total population	Year	Gender
0	2.0	Developed regions	901.0	82378628	5309667.699	2.87331	1990	Both sex
1	3.0	Developing regions	902.0	70184584	1144463.062	7.198015	1990	Both sex
2	4.0	Least developed countries	941.0	11075966	4165204.637	1.685021	1990	Both sex

What I learned:

First, I learned how to judge whether the data is clean or not. There are some basic principles for clean data, such as each variable forms a column, each observation forms a row, each type of observational unit forms a table, and so on. And learn to find the corresponding problems through these guidelines and use pandas and some functions to clean up, such as "pandas.DataFrame.melt". Also, this assignment helped me clear up a misconception. At first, I thought that all the codes needed to be memorized, but now I learned that I just need to open another webpage and find the codes you need.

Conclusion:

In summary, this write up includes all the steps I have taken to tidy the data this time, although some repeated steps have not been expressed again. According to the principles of tidy data, the data is now tidy. I combined table 1-3, and the rest of the tables were not together because they analyzed different content.