Assignment 5

1. Upload Registration.csvPreview the document and Course_info.csvPreview the document into Pandas. (5 points)

In [1]:
```
import pandas as pd
import os
import numpy as np
```

In [2]:
```
os.chdir(r'E:\GoogleDriveNew\PSU\DAAN862\Course contents\Lesson 5')
```

In [3]:
```
registration = pd.read_csv('Registration.csv')
```

In [4]:
```
registration.head()
```

Out[4]:

	Student name	semester new	coursename
0	Bill Mumy	Fall 2004	BEHAVIORAL PHARMACOLOGY
1	Bill Mumy	Fall 2000	AMERICAN FOREIGN POLICY
2	Bill Mumy	Fall 2003	DRUGS, BRAIN AND MIND
3	Bill Mumy	Fall 2005	Environmental Case Studies
4	Bill Mumy	Fall 2000	COMPUTER LINEAR ALGEBRA

In [5]:
```
courses = pd.read_excel('Course_info.xlsx')
```

In [6]:
```
courses.head()
```

Out[6]:

	Course number	Course Name	Course Type
0	ARTS400	EXPERIMENTAL WRITING SEM: The Ecology of Poetry	C
1	ARTS401	ART: ancient to 1945	C
2	ARTS465	ENVIRONMENTAL SYSTEMS II	F
3	ARTS486	COMPUTER LINEAR ALGEBRA	F
4	ARTS512	ANALYTICAL MECHANICS	F

In [7]:
```
courses.columns = ['Course_number', 'Course_name', 'Course_type']
```

2. Explore and clean Registration data. (30 points)

In [8]:
```
registration.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4900 entries, 0 to 4899
Data columns (total 3 columns):
Student name    4900 non-null object
semester new    4900 non-null object
coursename      4899 non-null object
dtypes: object(3)
memory usage: 114.9+ KB
```

In [9]:
```
registration.shape
```

Out[9]: (4900, 3)

In [10]:
```
registration.describe(include = 'all')
```

Out[10]:

	Student name	semester new	coursename
count	4900	4900	4899
unique	448	16	168
top	Harvey Golub	Spring 2002	COMPUT LINEAR ALGEBRA
freq	52	486	411

Remove missing values

In [11]:
```
registration.isnull().sum()
```

Out[11]:
```
Student name    0
semester new    0
coursename      1
dtype: int64
```

In [12]:
```
registration = registration.dropna()
```

In [13]:
```
registration.isnull().sum()
```

Out[13]:
```
Student name    0
semester new    0
coursename      0
dtype: int64
```

Remove Duplicates

In [14]:
```
registration.duplicated().sum()
```

Out[14]: 1249

In [15]:
```
registration = registration.drop_duplicates()
```

In [16]:
```
registration.shape
```

Out[16]: (3650, 3)

Optional: Clean typos

3. Explore and clean Course_info data. (10 points)

In [17]:
```
courses.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 42 entries, 0 to 41
Data columns (total 3 columns):
Course_number    42 non-null object
Course_name      41 non-null object
Course_type      42 non-null object
dtypes: object(3)
memory usage: 1.1+ KB
```

In [18]:
```
courses.shape
```

Out[18]: (42, 3)

In [19]:
```
courses.describe(include = 'all')
```

Out[19]:

	Course_number	Course_name	Course_type
count	42	41	42
unique	42	40	3
top	ARTS488	FRANCE & THE EUROP.UNION	E
freq	1	2	33

For course name column, we see it has 41 courses, while it has 40 unique values. This is an indication of duplicates. You need to understand your data to remove duplicates. Except for duplicated rows, sometimes you need to check if a column has duplicates. For example, ID number shouldn't contains duplicates.

In [20]:
```
courses.columns = ['Course_No', 'Course_Name', 'Course_type']
```

In [21]:
```
courses.head()
```

Out[21]:

	Course_No	Course_Name	Course_type
0	ARTS400	EXPERIMENTAL WRITING SEM: The Ecology of Poetry	C
1	ARTS401	ART: ancient to 1945	C
2	ARTS465	ENVIRONMENTAL SYSTEMS II	F
3	ARTS486	COMPUTER LINEAR ALGEBRA	F
4	ARTS512	ANALYTICAL MECHANICS	F

Remove missing values

In [22]:
```
courses.isnull().sum()
```

Out[22]:
```
Course_No      0
Course_Name    1
Course_type    0
dtype: int64
```

In [23]:
```
courses = courses.dropna()
```

Remove duplicates

In [24]:
```
courses.Course_Name.duplicated().sum()
```

Out[24]:
```
1
```

In [25]:
```
courses = courses.drop_duplicates(subset = 'Course_Name', keep = 'last')
```

In [26]:
```
courses.describe(include = 'all')
```

Out[26]:

	Course_No	Course_Name	Course_type
count	40	40	40
unique	40	40	3
top	ARTS488	COMPARATIVE POLITICS	E
freq	1	1	31

Explore the data

In [27]:
```
courses.Course_type.value_counts()
```

Out[27]:
```
E    31
F     7
C     2
Name: Course_type, dtype: int64
```

Calculate the string distance between raw course names with correct course names

In [28]:
```
from difflib import SequenceMatcher
courses_raw = registration.coursename.unique()
courses_Cross_ref = pd.DataFrame(index = courses_raw, columns = courses.Course_Name)
```

In [29]:
```
for i in courses_raw:
    for j in courses_Cross_ref.columns.values:
        courses_Cross_ref.loc[i, j] = SequenceMatcher(None, i, j).ratio()
```

In [30]:
```
courses_Cross_ref.columns.values
```

Out[30]:
```
array(['EXPERIMENTAL WRITING SEM: The Ecology of Poetry',
       'ART: ancient to 1945', 'ENVIRONMENTAL SYSTEMS II',
       'COMPUTER LINEAR ALGEBRA', 'ANALYTICAL MECHANICS',
       'A WORLD AT WAR', 'BEHAVIORAL PHARMACOLOGY',
       'CONTEMPORARY AFRICAN ART', 'FOOD/FEAST ARCH OF TABLE',
       "DEVIL'S PACT LIT/FILM", 'AMERICAN SOCIAL POLICY',
       'ART AND RELIGION', 'CONTEMPORARY POL.THOUGHT',
       'AFRICAN-AMERICAN LIT: AFRICAN-AMER LIT:CHANGE',
       'AMERICAN HEALTH POLICY', 'Business German: A Micro Perspective',
       'COMM and  THE PRESIDENCY', 'French Thought Till 1945',
       'CONTEMP ART ~ 1945 to PRESENT',
       '20th Century Russian Literature: Fiction and Reality',
       'COMMUNICATIONS INTERNSHP', 'FRESHWATER ECOLOGY', 'AESTHETICS',
       'French Thought Since 1945', 'BECOMING HUMAN',
       'EVIDENCED BASED CRIME AND JUSTICE POLICY',
       'EUROPE IN A WIDER WORLD', '19TH-CENTURY BRITISH LITERATURE',
       'AMERICAN SOUTH 1861-PRES', 'AUGUSTAN CULTRAL REVOLUTION',
       'Environmental Studies Research Seminar Junior Level',
       'CELL. BIOL. & BIOCHEM.', 'ANALYZING THE POL WORLD',
       'EARLY MESOPOTAM HISTORY/SOCIETY', 'FRANCE & THE EUROP.UNION',
       'EARLY BALCAN HIST/SOC', 'COMPARATIVE POLITICS',
       'BRITISH POETRY 1660-1914', 'CONTEMPORARY SOCIO THEORY',
       'ELEMENTARY ARABIC II'], dtype=object)
```

In [31]:
```
courses_Cross_ref_test = courses_Cross_ref.astype('float')
```

set a threshold for distance. First, I tried 0.8

In [32]:
```
courses_Cross_ref_test[courses_Cross_ref < 0.8] = np.nan
```

In [33]:
```
courses_Cross_ref_test
```

	COMM & THE PRESIDENCY	The Ecology of Poetry	ANCIENT to 1945	SYSTEMS II	LINEAR ALGEBRA	MECHANICS	AT WAR	PHARMACOLOGY	AFRICAN ART	ADAPT OF TABLE	LIT/FILM	...	Research Seminar Junior Level	BIOCHEM	THE POE WORLD	HISTORY/SOCIETY	EUROP.UNION	HIST/SOC	POLITICS	1
ART: ancient to 1945	NaN	1.000000	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN		
AMERICAN SOCIETY	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN		
AFRICAN LANG. & CULTURE	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN		
AMER POST-1800: BF SEM: MODERN AMERICAN CITIES	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN		
CONTEMPORARY AFRICAN-ART	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.958333	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN		
Business German - Micro Perspective	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN		
EVIDENCED																				

In [34]:
```
courses_Cross_ref_test.idxmax(axis = 1).sort_index()  # idxmax is used to find out the column names for the max value for each row.
```

Out[34]:
```
1000 YRS MUSICAL LISTENG: 1000 YRS MUSICAL LISTENG                                  NaN
19TH-CENT BRITISH LIT                                  19TH-CENTURY BRITISH LITERATURE
19TH-CENT NOVEL                                                                     NaN
1ST YR CLASSICAL CHIN II                                                            NaN
20TH-CENT POETRY                                                                    NaN
20th Century Russian Literature: Fiction and Reality    20th Century Russian Literature: Fiction and R...
A WORLD AT WAR                                                          A WORLD AT WAR
ABNORMAL PSYCHOLOGY                                                                 NaN
ACCEL INTERMD PORTUGUESE                                                            NaN
ACCEL INTERMEDIATE SPAN                                                             NaN
ACCELERATED HINDI                                                                   NaN
ACCELERATED INTERMD GRMN                                                            NaN
AESTHETICS                                                               AESTHETICS
AFGHANISTAN & ISLAMISM: AFGHANISTAN & ISLAMISM                                      NaN
AFRICAN LANG. & CULTURE                                                             NaN
AFRICAN-AMERICAN LIT                                                                NaN
AFRICAN-AMERICAN LIT: AFRICAN-AMER LIT:CHANGE       AFRICAN-AMERICAN LIT: AFRICAN-AMER LIT:CHANGE
AFRO AMER HIST 1876-PRES                                                            NaN
AMER POST-1800: BF SEM: MODERN AMERICAN CITIES                                      NaN
AMER REVOLUTION                                                                     NaN
AMERICA IN THE 1960S                                                                NaN
AMERICAN FOREIGN POLICY                                        AMERICAN SOCIAL POLICY
AMERICAN HEALT POLICY                                          AMERICAN HEALTH POLICY
AMERICAN HEALTH POLICY                                         AMERICAN HEALTH POLICY
AMERICAN MUSICAL THEATRE                                                            NaN
AMERICAN POETRY                                                                     NaN
AMERICAN SOCIETY                                                                    NaN
AMERICAN SOUTH 1861-PRES                                       AMERICAN SOUTH 1861-PRES
ANAL METH ECON, LAW MED                                                             NaN
ANALYTICAL MECHANICS                                                 ANALYTICAL MECHANICS
                                                                                    ...
EUR PRE-1800: BF SEM: UTOPIA                                                        NaN
EURO ART & CIV > 1400: RENAISSANCE TO CONTEMP                                       NaN
EURO INT'L REL SINCE WW One                                                         NaN
EURO INT'L REL SINCE WWI                                                            NaN
EURO INT'L REL SINCE WWI                                                            NaN
EURO INTELL HIST 18 C.                                                              NaN
EUROPE IN A WIDER WORLD                                          EUROPE IN A WIDER WORLD
EVIDENCED BASED CRIME & JUSTICE POLICY                 EVIDENCED BASED CRIME AND JUSTICE POLICY
EVIDENCED BASED CRIME AND JUSTICE POLICY               EVIDENCED BASED CRIME AND JUSTICE POLICY
EXPERIMENTAL WRITING SEM                                                            NaN
EXPERIMENTAL WRITING SEM: The Ecology of Poetry      EXPERIMENTAL WRITING SEM: The Ecology of Poetry
EYE, MIND AND IMAGE                                                                 NaN
Environmental Case Studies                                                          NaN
Environmental Studies Research Seminar Junior Level    Environmental Studies Research Seminar Junior ...
Environmental Studies Research Seminar for Juniors      Environmental Studies Research Seminar Junior ...
FICTION WRITING WORKSHOP                                                            NaN
FOOD/FEAST ARCH OF TABLE                                        FOOD/FEAST ARCH OF TABLE
FORENSIC ANTHROPOLOGY                                                               NaN
FORMAL LOGIC I                                                                      NaN
FORMAL SEM AND COG SCI                                                              NaN
FR FOR PROFESSIONS I                                                                NaN
FR FOR PROFESSIONS II                                                               NaN
FR LIT OF THE 19TH C: STUDIES IN THE 19TH C.                                        NaN
FRANCE & THE EUROP.UNION                                        FRANCE & THE EUROP.UNION
FRANCE AND ITS OTHERS: Anthropology and French Modernism                            NaN
FREEDOM OF EXPRESSION                                                               NaN
FRENCH PHONETICS                                                                    NaN
FRESHWATER ECOLOGY                                                   FRESHWATER ECOLOGY
Feminist Theory: Feminism, Activism, and the Body                                   NaN
French Thought Since 1945                                       French Thought Since 1945
Length: 168, dtype: object
```

From the result, we can see that it cannot match perfectly. More work needs to be done here. You can try to improve the results by yourself. There are other string distance methods. For this method, below are something you can try:

1. Maybe only use the first few charactors for each word
2. Remove words after ":".
3. Replace special characters.

4. Which course has the highest registration? (15 points)

In [35]:
```
registration.coursename.value_counts().head()
```

Out[35]:
```
COMPUT LINEAR ALGEBRA         303
Environmental Case Studies    286
A WORLD AT WAR                269
BEHAVIORAL PHARMACOLOGY       260
ANALYTICAL MECHANICS          256
Name: coursename, dtype: int64
```

5. Inner join two datasets. (20 points)

In [36]:
```
registration_all = pd.merge(registration, courses, left_on = 'coursename', right_on = 'Course_Name')
```

In [37]:
```
registration_all.head()
```

Out[37]:

	Student name	semester new	coursename	Course_No	Course_Name	Course_type
0	Bill Mumy	Fall 2004	BEHAVIORAL PHARMACOLOGY	ARTS516	BEHAVIORAL PHARMACOLOGY	F
1	Geraldine Ferraro	Summer 2004	BEHAVIORAL PHARMACOLOGY	ARTS516	BEHAVIORAL PHARMACOLOGY	F
2	Laura Lippman	Fall 2004	BEHAVIORAL PHARMACOLOGY	ARTS516	BEHAVIORAL PHARMACOLOGY	F
3	Dom DeLuise	Fall 2000	BEHAVIORAL PHARMACOLOGY	ARTS516	BEHAVIORAL PHARMACOLOGY	F
4	Sally Field	Summer 2001	BEHAVIORAL PHARMACOLOGY	ARTS516	BEHAVIORAL PHARMACOLOGY	F

In [38]:
```
registration_all = registration_all.drop('Course_Name', axis = 1)
```

In [39]:
```
registration_all.describe(include = 'all')
```

Out[39]:

	Student name	semester new	coursename	Course_No	Course_type
count	1734	1734	1734	1734	1734
unique	408	16	33	33	3
top	Ellen Burstyn	Spring 2002	A WORLD AT WAR	ARTS514	F
freq	12	158	269	269	1157

6. Create a data frame with student name as the index, course numbers as columns, and if the student registered a course as values(0, 1). (20 points)

In [40]:
```
registration_pivot = pd.pivot_table(registration_all, index = 'Student name',
                                    columns = 'Course_No',
                                    values = 'coursename',
                                    aggfunc = 'count',
                                    fill_value = 0)
```

In [41]:
```
registration_pivot.head()
```

Out[41]:

Course_No	ARTS400	ARTS401	ARTS465	ARTS484	ARTS485	ARTS486	ARTS488	ARTS491	ARTS492	ARTS493	...	ARTS553	ARTS555	ARTS559	ARTS565	ARTS569	ARTS573	ARTS577	ARTS581	ARTS583	ARTS587
Student name																					
ABella Abzug	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	1	0	0
Al Gore	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
Al Hirt	0	0	1	0	0	1	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
Al Roker	1	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
Alan Bates	0	0	0	0	0	0	0	0	0	0	...	0	1	0	0	0	0	0	0	0	0

5 rows × 33 columns

In [42]:
```
registration_pivot.shape
```

Out[42]: (408, 33)

Some students also used get_dummies to get the results, however this is not exactly what I am looking for, since the student name is not unique.Please note that you need to know the difference.'

In [43]:
```
regis_pivot2 = pd.get_dummies(registration_all[['Student name', 'Course_No']], columns = ['Course_No'])
```

In [44]:
```
regis_pivot2.loc[:,'Student name'].duplicated().sum() # duplicated student names
```

Out[44]: 1326

In [45]:
```
regis_pivot2.set_index('Student name', inplace= True)
```

In [46]:
```
regis_pivot2.head()
```

Out[46]:

Student name	Course_No_ARTS400	Course_No_ARTS401	Course_No_ARTS465	Course_No_ARTS484	Course_No_ARTS485	Course_No_ARTS486	Course_No_ARTS488	Course_No_ARTS491	Course_No_ARTS492	Course_No_ARTS493	...	Course_No_ARTS553	Course_No_ARTS555
Bill Mumy	0	0	0	0	0	0	0	0	0	0	...	0	0
Geraldine Ferraro	0	0	0	0	0	0	0	0	0	0	...	0	0
Laura Lippman	0	0	0	0	0	0	0	0	0	0	...	0	0
Dom DeLuise	0	0	0	0	0	0	0	0	0	0	...	0	0
Sally Field	0	0	0	0	0	0	0	0	0	0	...	0	0

5 rows × 33 columns

Dimension is different too.

In [45]:
```
regis_pivot2.shape
```

Out[45]: (1734, 33)