In [1]:

```
mport pandas as pd
mport numpy as np
mport matplotlib.pyplot as plt
matplotlib inline
```

In [2]:

```
f=pd.read_csv('https://raw.githubusercontent.com/jackiekazil/data-wrangling/m
ster/data/chp3/data-text.csv')
f.head(2)
```

Out[2]:

	Indicator	PUBLISH STATES	Year	WHO region	World Bank income group	Country	Sex	Display Value	Numeric	Low	High
0	Life expectancy at birth (years)	Published	1990	Europe	High-income	Andorra	Both sexes	77	77.0	NaN	NaN
1	Life expectancy at birth (years)	Published	2000	Europe	High-income	Andorra	Both sexes	80	80.0	NaN	NaN

In [3]:

```
f1=pd.read_csv('https://raw.githubusercontent.com/kjam/data-wrangling-pycon/m
ster/data/berlin_weather_oldest.csv')
f1.head(2)
```

Out[3]:

	STATION	STATION_NAME	DATE	PRCP	SNWD	SNOW	TMAX	TMIN	WDFG
0	GHCND:GME00111445	BERLIN TEMPELHOF GM	19310101	46	-9999	-9999	-9999	-11	-9999
1	GHCND:GME00111445	BERLIN TEMPELHOF GM	19310102	107	-9999	-9999	50	11	-9999

2 rows × 21 columns

1. Get the Metadata from the above files

In [4]:
```
f.info()
```

```
class 'pandas.core.frame.DataFrame'>
angeIndex: 4656 entries, 0 to 4655
ata columns (total 12 columns):
ndicator                   4656 non-null object
UBLISH STATES              4656 non-null object
ear                        4656 non-null int64
HO region                  4656 non-null object
orld Bank income group     4656 non-null object
ountry                     4656 non-null object
ex                         4656 non-null object
isplay Value               4656 non-null int64
umeric                     4656 non-null float64
ow                         0 non-null float64
igh                        0 non-null float64
omments                    0 non-null float64
types: float64(4), int64(2), object(6)
emory usage: 436.6+ KB
```

In [6]:
```
f1.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 117208 entries, 0 to 117207
Data columns (total 21 columns):
STATION         117208 non-null object
STATION_NAME    117208 non-null object
DATE            117208 non-null int64
PRCP            117208 non-null int64
SNWD            117208 non-null int64
SNOW            117208 non-null int64
TMAX            117208 non-null int64
TMIN            117208 non-null int64
WDFG            117208 non-null int64
PGTM            117208 non-null int64
WSFG            117208 non-null int64
WT09            117208 non-null int64
WT07            117208 non-null int64
WT01            117208 non-null int64
WT06            117208 non-null int64
WT05            117208 non-null int64
WT04            117208 non-null int64
WT16            117208 non-null int64
WT08            117208 non-null int64
WT18            117208 non-null int64
WT03            117208 non-null int64
dtypes: int64(19), object(2)
memory usage: 18.8+ MB
```

2. Get the row names from the above files

In [21]:
```
f.index.values
```

Out[21]:
```
array([   0,    1,    2, ..., 4653, 4654, 4655], dtype=int64)
```

In [11]:
```
f1.index.values
```

Out[11]:
```
array([     0,      1,      2, ..., 117205, 117206, 117207], dtype=int64)
```

3. Change the column name from any of the above file

Considering the first file stored in the df dataframe

In [14]:
```
Get ndArray of all column names
olumnsNamesArr_df = df.columns.values

 Modify first Column Name
olumnsNamesArr_df[0] = 'Indicator_id'
```

In [15]:
```
f.head(2)
```

Out[15]:

	Indicator_id	PUBLISH STATES	Year	WHO region	World Bank income group	Country	Sex	Display Value	Numeric	Low	High
0	Life expectancy at birth (years)	Published	1990	Europe	High-income	Andorra	Both sexes	77	77.0	NaN	NaN
1	Life expectancy at birth (years)	Published	2000	Europe	High-income	Andorra	Both sexes	80	80.0	NaN	NaN

4.Change the column name from any of the above file and store the changes made permanently

Considering the first file stored in the df dataframe

In [17]:

```
Permanently changing the column name of the first column
f.rename(columns={'Indicator':'Indicator_id'}, inplace=True)
f.head(2)
```

Out[17]:

	Indicator_id	PUBLISH STATES	Year	WHO region	World Bank income group	Country	Sex	Display Value	Numeric	Low	High
0	Life expectancy at birth (years)	Published	1990	Europe	High-income	Andorra	Both sexes	77	77.0	NaN	NaN
1	Life expectancy at birth (years)	Published	2000	Europe	High-income	Andorra	Both sexes	80	80.0	NaN	NaN

5. Change the names of multiple columns

In [18]:

```
f.rename(columns={'PUBLISH STATES':'Publication Status','WHO region':'WHO Reg
on'}, inplace=True)
f.head(2)
```

Out[18]:

	Indicator_id	Publication Status	Year	WHO Region	World Bank income group	Country	Sex	Display Value	Numeric	Low	Hig
0	Life expectancy at birth (years)	Published	1990	Europe	High-income	Andorra	Both sexes	77	77.0	NaN	Nal
1	Life expectancy at birth (years)	Published	2000	Europe	High-income	Andorra	Both sexes	80	80.0	NaN	Nal

6. Arrange values of a particular column in ascending order

In [24]:

```
f.sort_values(by=['Year'])
```

Out[24]:

	Indicator_id	Publication Status	Year	WHO Region	World Bank income group	Country	Sex	Display Value	Nume
0	Life expectancy at birth (years)	Published	1990	Europe	High-income	Andorra	Both sexes	77	7
1270	Life expectancy at birth (years)	Published	1990	Europe	High-income	Germany	Male	72	7
3193	Life expectancy at birth (years)	Published	1990	Europe	Lower-middle-income	Republic of Moldova	Male	65	6
3194	Life expectancy at birth (years)	Published	1990	Europe	Lower-middle-income	Republic of Moldova	Both sexes	68	6
3197	Life expectancy at age 60 (years)	Published	1990	Europe	Lower-middle-income	Republic of Moldova	Male	15	1
1264	Life expectancy at birth (years)	Published	1990	Europe	High-income	Cyprus	Both sexes	76	7
3199	Life expectancy at age 60 (years)	Published	1990	Europe	Lower-middle-income	Republic of Moldova	Both sexes	17	1
1262	Life expectancy at age 60 (years)	Published	1990	Western Pacific	High-income	Cook Islands	Male	17	1
1259	Life expectancy at birth (years)	Published	1990	Western Pacific	High-income	Cook Islands	Male	67	6
3203	Life expectancy at age 60 (years)	Published	1990	South-East Asia	Lower-middle-income	Maldives	Female	12	1
1273	Life expectancy at age 60 (years)	Published	1990	Europe	High-income	Denmark	Both sexes	20	2
3204	Life expectancy at birth (years)	Published	1990	Western Pacific	Lower-middle-income	Marshall Islands	Female	65	6
1253	Life expectancy at birth (years)	Published	1990	Western Pacific	High-income	Brunei Darussalam	Both sexes	73	7

	Indicator_id	Publication Status	Year	WHO Region	World Bank income group	Country	Sex	Display Value	Nume
1247	Life expectancy at age 60 (years)	Published	1990	Americas	High-income	Bahamas	Male	17	1
3219	Life expectancy at age 60 (years)	Published	1990	Western Pacific	Lower-middle-income	Vanuatu	Both sexes	16	1
3226	Life expectancy at birth (years)	Published	1990	Europe	Upper-middle-income	Bulgaria	Both sexes	71	7
1240	Life expectancy at age 60 (years)	Published	1990	Europe	High-income	Belgium	Female	23	2
1239	Life expectancy at birth (years)	Published	1990	Europe	High-income	Belgium	Both sexes	76	7
1238	Life expectancy at birth (years)	Published	1990	Europe	High-income	Belgium	Female	79	7
1237	Life expectancy at birth (years)	Published	1990	Europe	High-income	Austria	Both sexes	76	7
1236	Life expectancy at birth (years)	Published	1990	Europe	High-income	Austria	Male	72	7
3207	Life expectancy at birth (years)	Published	1990	Western Pacific	Lower-middle-income	Mongolia	Female	64	6
3231	Life expectancy at birth (years)	Published	1990	Europe	Upper-middle-income	Belarus	Female	76	7
3188	Life expectancy at age 60 (years)	Published	1990	Africa	Lower-middle-income	Lesotho	Female	17	1
1277	Life expectancy at birth (years)	Published	1990	Europe	High-income	Estonia	Both sexes	70	7
1302	Life expectancy at birth (years)	Published	1990	Europe	High-income	Hungary	Male	65	6

	Indicator_id	Publication Status	Year	WHO Region	World Bank income group	Country	Sex	Display Value	Nume
3158	Life expectancy at birth (years)	Published	1990	Europe	Lower-middle-income	Georgia	Male	67	6
1300	Life expectancy at birth (years)	Published	1990	Europe	High-income	Croatia	Male	69	6
3159	Life expectancy at age 60 (years)	Published	1990	Europe	Lower-middle-income	Georgia	Both sexes	19	1
3160	Life expectancy at age 60 (years)	Published	1990	Americas	Lower-middle-income	Guatemala	Female	19	1
...	...	...	...	...	...	...	...	...	
3175	Life expectancy at age 60 (years)	Published	2012	Eastern Mediterranean	Lower-middle-income	Iran (Islamic Republic of)	Male	19	1
3174	Life expectancy at birth (years)	Published	2012	Eastern Mediterranean	Lower-middle-income	Iran (Islamic Republic of)	Female	76	7
1285	Life expectancy at birth (years)	Published	2012	Europe	High-income	France	Both sexes	82	8
1286	Life expectancy at age 60 (years)	Published	2012	Europe	High-income	France	Both sexes	25	2
3171	Life expectancy at birth (years)	Published	2012	Eastern Mediterranean	Lower-middle-income	Iran (Islamic Republic of)	Male	72	7
1288	Life expectancy at age 60 (years)	Published	2012	Europe	High-income	United Kingdom of Great Britain and Northern I...	Female	25	2
1290	Life expectancy at age 60 (years)	Published	2012	Europe	High-income	United Kingdom of Great Britain and Northern I...	Both sexes	24	2
1292	Life expectancy at birth (years)	Published	2012	Africa	High-income	Equatorial Guinea	Female	57	5

	Indicator_id	Publication Status	Year	WHO Region	World Bank income group	Country	Sex	Display Value	Nume
3166	Life expectancy at age 60 (years)	Published	2012	Americas	Lower-middle-income	Honduras	Male	21	2
3165	Life expectancy at birth (years)	Published	2012	Americas	Lower-middle-income	Honduras	Both sexes	74	7
3163	Life expectancy at age 60 (years)	Published	2012	Americas	Lower-middle-income	Guyana	Male	13	1
3162	Life expectancy at birth (years)	Published	2012	Americas	Lower-middle-income	Guyana	Female	67	6
1301	Life expectancy at age 60 (years)	Published	2012	Europe	High-income	Croatia	Both sexes	21	2
3137	Life expectancy at birth (years)	Published	2012	Africa	Lower-middle-income	Cameroon	Male	55	5
1303	Life expectancy at birth (years)	Published	2012	Europe	High-income	Hungary	Both sexes	75	7
3155	Life expectancy at birth (years)	Published	2012	Western Pacific	Lower-middle-income	Micronesia (Federated States of)	Male	68	6
3154	Life expectancy at age 60 (years)	Published	2012	Eastern Mediterranean	Lower-middle-income	Egypt	Male	16	1
1304	Life expectancy at age 60 (years)	Published	2012	Europe	High-income	Hungary	Both sexes	20	2
1306	Life expectancy at birth (years)	Published	2012	Europe	High-income	Ireland	Female	83	8
3150	Life expectancy at age 60 (years)	Published	2012	Americas	Lower-middle-income	Ecuador	Male	21	2
3148	Life expectancy at birth (years)	Published	2012	Americas	Lower-middle-income	Ecuador	Female	78	7

	Indicator_id	Publication Status	Year	WHO Region	World Bank income group	Country	Sex	Display Value	Nume
3147	Life expectancy at age 60 (years)	Published	2012	Eastern Mediterranean	Lower-middle-income	Djibouti	Female	17	1
3146	Life expectancy at birth (years)	Published	2012	Eastern Mediterranean	Lower-middle-income	Djibouti	Both sexes	61	6
3145	Life expectancy at birth (years)	Published	2012	Eastern Mediterranean	Lower-middle-income	Djibouti	Female	63	6
1309	Life expectancy at age 60 (years)	Published	2012	Europe	High-income	Ireland	Female	25	2
1316	Life expectancy at birth (years)	Published	2012	Europe	High-income	Italy	Both sexes	83	8
3141	Life expectancy at birth (years)	Published	2012	Africa	Lower-middle-income	Cabo Verde	Both sexes	74	7
3139	Life expectancy at age 60 (years)	Published	2012	Africa	Lower-middle-income	Cameroon	Female	17	1
3156	Life expectancy at age 60 (years)	Published	2012	Western Pacific	Lower-middle-income	Micronesia (Federated States of)	Male	16	1
4655	Healthy life expectancy (HALE) at birth (years)	Published	2012	Africa	Low-income	Zimbabwe	Female	51	5

4656 rows × 12 columns

7. Arrange multiple column values in ascending order

In [26]:

```
Creating a temporary dataframe from the main by dropping few columns
f_temp=df.loc[:,['Indicator_id','Country','Year','WHO Region','Publication St
tus']]
```

In [27]:

```
sorting the temporary dataframe by country and year and showing the first 4 r
sults
f_temp.sort_values(by=['Country','Year'])
f_temp.head(4)
```

Out[27]:

	Indicator_id	Country	Year	WHO Region	Publication Status
0	Life expectancy at birth (years)	Andorra	1990	Europe	Published
1	Life expectancy at birth (years)	Andorra	2000	Europe	Published
2	Life expectancy at age 60 (years)	Andorra	2012	Europe	Published
3	Life expectancy at age 60 (years)	Andorra	2000	Europe	Published

8. Make country as the first column of the dataframe

In [34]:

```
ountry_1 = list(df)
ountry_1.insert(0, country_1.pop(country_1.index('Country')))
f_update = df.loc[:, cols]
f_update.head()
```

Out[34]:

	Country	Indicator_id	Publication Status	Year	WHO Region	World Bank income group	Sex	Display Value	Numeric	L
0	Andorra	Life expectancy at birth (years)	Published	1990	Europe	High-income	Both sexes	77	77.0	N
1	Andorra	Life expectancy at birth (years)	Published	2000	Europe	High-income	Both sexes	80	80.0	N
2	Andorra	Life expectancy at age 60 (years)	Published	2012	Europe	High-income	Female	28	28.0	N
3	Andorra	Life expectancy at age 60 (years)	Published	2000	Europe	High-income	Both sexes	23	23.0	N
4	United Arab Emirates	Life expectancy at birth (years)	Published	2012	Eastern Mediterranean	High-income	Female	78	78.0	N

9. Get the column array using a variable

In [30]:

```
ol2=df['WHO Region'].values
ol2
```

Out[30]:

```
array(['Europe', 'Europe', 'Europe', ..., 'Africa', 'Africa', 'Africa'],
      dtype=object)
```

10.Get the subset rows 11, 24, 37

In [41]:

```
Assigning the rows 11,24,37 to a new dataframe called df_rows
f_rows=df.iloc[[11,24,37],:]
f_rows
```

Out[41]:

	Indicator_id	Publication Status	Year	WHO Region	World Bank income group	Country	Sex	Display Value	Numeric	Lo
11	Life expectancy at birth (years)	Published	2012	Europe	High-income	Austria	Female	83	83.0	Na
24	Life expectancy at age 60 (years)	Published	2012	Western Pacific	High-income	Brunei Darussalam	Female	21	21.0	Na
37	Life expectancy at age 60 (years)	Published	2012	Europe	High-income	Cyprus	Female	26	26.0	Na

11. Get the subset rows excluding 5, 12, 23, and 56

In [47]:

```
dropping the rows 5,12,23,56

f_rows1=df.drop([5,12,23,56])
f_rows1.head(25)
```

Out[47]:

	Indicator_id	Publication Status	Year	WHO Region	World Bank income group	Country	Sex	Display Value	Numeric
0	Life expectancy at birth (years)	Published	1990	Europe	High-income	Andorra	Both sexes	77	77.0
1	Life expectancy at birth (years)	Published	2000	Europe	High-income	Andorra	Both sexes	80	80.0
2	Life expectancy at age 60 (years)	Published	2012	Europe	High-income	Andorra	Female	28	28.0
3	Life expectancy at age 60 (years)	Published	2000	Europe	High-income	Andorra	Both sexes	23	23.0
4	Life expectancy at birth (years)	Published	2012	Eastern Mediterranean	High-income	United Arab Emirates	Female	78	78.0
6	Life expectancy at age 60 (years)	Published	1990	Americas	High-income	Antigua and Barbuda	Male	17	17.0
7	Life expectancy at age 60 (years)	Published	2012	Americas	High-income	Antigua and Barbuda	Both sexes	22	22.0
8	Life expectancy at birth (years)	Published	2012	Western Pacific	High-income	Australia	Male	81	81.0
9	Life expectancy at birth (years)	Published	2000	Western Pacific	High-income	Australia	Both sexes	80	80.0
10	Life expectancy at birth (years)	Published	2012	Western Pacific	High-income	Australia	Both sexes	83	83.0
11	Life expectancy at birth (years)	Published	2012	Europe	High-income	Austria	Female	83	83.0
13	Life expectancy at birth (years)	Published	2012	Europe	High-income	Belgium	Female	83	83.0
14	Life expectancy at birth (years)	Published	2000	Eastern Mediterranean	High-income	Bahrain	Male	73	73.0

	Indicator_id	Publication Status	Year	WHO Region	World Bank income group	Country	Sex	Display Value	Numeri
15	Life expectancy at birth (years)	Published	1990	Eastern Mediterranean	High-income	Bahrain	Female	74	74.(
16	Life expectancy at age 60 (years)	Published	1990	Eastern Mediterranean	High-income	Bahrain	Male	17	17.(
17	Life expectancy at birth (years)	Published	2012	Americas	High-income	Bahamas	Male	72	72.(
18	Life expectancy at age 60 (years)	Published	2000	Americas	High-income	Bahamas	Both sexes	21	21.(
19	Life expectancy at birth (years)	Published	1990	Americas	High-income	Barbados	Male	71	71.(
20	Life expectancy at age 60 (years)	Published	2012	Americas	High-income	Barbados	Female	25	25.(
21	Life expectancy at age 60 (years)	Published	2012	Americas	High-income	Barbados	Both sexes	23	23.(
22	Life expectancy at age 60 (years)	Published	1990	Western Pacific	High-income	Brunei Darussalam	Female	20	20.(
24	Life expectancy at age 60 (years)	Published	2012	Western Pacific	High-income	Brunei Darussalam	Female	21	21.(
25	Life expectancy at birth (years)	Published	2000	Americas	High-income	Canada	Female	82	82.(
26	Life expectancy at age 60 (years)	Published	2000	Americas	High-income	Canada	Male	21	21.(
27	Life expectancy at age 60 (years)	Published	1990	Americas	High-income	Canada	Female	24	24.(

Load datasets from CSV

In [48]:

```
sers=pd.read_csv('https://raw.githubusercontent.com/ben519/DataWrangling/mast
r/Data/users.csv')
essions =pd.read_csv('https://raw.githubusercontent.com/ben519/DataWrangling/
aster/Data/sessions.csv')
roducts =pd.read_csv('https://raw.githubusercontent.com/ben519/DataWrangling/
aster/Data/products.csv')
ransactions =pd.read_csv('https://raw.githubusercontent.com/ben519/DataWrangl
ng/master/Data/transactions.csv')
```

In [49]:

```
sers.head()
```

Out[49]:

	UserID	User	Gender	Registered	Cancelled
0	1	Charles	male	2012-12-21	NaN
1	2	Pedro	male	2010-08-01	2010-08-08
2	3	Caroline	female	2012-10-23	2016-06-07
3	4	Brielle	female	2013-07-17	NaN
4	5	Benjamin	male	2010-11-25	NaN

In [50]:

```
essions.head()
```

Out[50]:

	SessionID	SessionDate	UserID
0	1	2010-01-05	2
1	2	2010-08-01	2
2	3	2010-11-25	2
3	4	2011-09-21	5
4	5	2011-10-19	4

In [51]:

```
ransactions.head()
```

Out[51]:

	TransactionID	TransactionDate	UserID	ProductID	Quantity
0	1	2010-08-21	7.0	2	1
1	2	2011-05-26	3.0	4	1
2	3	2011-06-16	3.0	3	1
3	4	2012-08-26	1.0	2	3
4	5	2013-06-06	2.0	4	1

12. Join users to transactions, keeping all rows from transactions and only matching rows from users (left join)

In [53]:

```
sers_trans = pd.merge(transactions, users, on='UserID', how='left')
sers_trans
```

Out[53]:

	TransactionID	TransactionDate	UserID	ProductID	Quantity	User	Gender	Registered	Ca
0	1	2010-08-21	7.0	2	1	NaN	NaN	NaN	
1	2	2011-05-26	3.0	4	1	Caroline	female	2012-10-23	2
2	3	2011-06-16	3.0	3	1	Caroline	female	2012-10-23	2
3	4	2012-08-26	1.0	2	3	Charles	male	2012-12-21	
4	5	2013-06-06	2.0	4	1	Pedro	male	2010-08-01	2
5	6	2013-12-23	2.0	5	6	Pedro	male	2010-08-01	2
6	7	2013-12-30	3.0	4	1	Caroline	female	2012-10-23	2
7	8	2014-04-24	NaN	2	3	NaN	NaN	NaN	
8	9	2015-04-24	7.0	4	3	NaN	NaN	NaN	
9	10	2016-05-08	3.0	4	4	Caroline	female	2012-10-23	2

13.Which transactions have a UserID not in users?

In [54]:

```
ransactions[~transactions['UserID'].isin(users['UserID'])]
```

Out[54]:

	TransactionID	TransactionDate	UserID	ProductID	Quantity
0	1	2010-08-21	7.0	2	1
7	8	2014-04-24	NaN	2	3
8	9	2015-04-24	7.0	4	3

14.Join users to transactions, keeping only rows from transactions and users that match via UserID (inner join)

In [57]:

```
sers_trans2 = pd.merge(transactions, users, on='UserID', how='inner', sort=Fa
se)
sers_trans2
```

Out[57]:

	TransactionID	TransactionDate	UserID	ProductID	Quantity	User	Gender	Registered	Ca
0	2	2011-05-26	3.0	4	1	Caroline	female	2012-10-23	2
1	3	2011-06-16	3.0	3	1	Caroline	female	2012-10-23	2
2	7	2013-12-30	3.0	4	1	Caroline	female	2012-10-23	2
3	10	2016-05-08	3.0	4	4	Caroline	female	2012-10-23	2
4	4	2012-08-26	1.0	2	3	Charles	male	2012-12-21	
5	5	2013-06-06	2.0	4	1	Pedro	male	2010-08-01	2
6	6	2013-12-23	2.0	5	6	Pedro	male	2010-08-01	2

15. Join users to transactions, displaying all matching rows AND all non-matching rows (full outer join)

In [58]:

```
sers_trans3= pd.merge(transactions, users, on='UserID', how='outer', sort=False)
sers_trans3
```

Out[58]:

	TransactionID	TransactionDate	UserID	ProductID	Quantity	User	Gender	Registered
0	1.0	2010-08-21	7.0	2.0	1.0	NaN	NaN	NaN
1	9.0	2015-04-24	7.0	4.0	3.0	NaN	NaN	NaN
2	2.0	2011-05-26	3.0	4.0	1.0	Caroline	female	2012-10-23
3	3.0	2011-06-16	3.0	3.0	1.0	Caroline	female	2012-10-23
4	7.0	2013-12-30	3.0	4.0	1.0	Caroline	female	2012-10-23
5	10.0	2016-05-08	3.0	4.0	4.0	Caroline	female	2012-10-23
6	4.0	2012-08-26	1.0	2.0	3.0	Charles	male	2012-12-21
7	5.0	2013-06-06	2.0	4.0	1.0	Pedro	male	2010-08-01
8	6.0	2013-12-23	2.0	5.0	6.0	Pedro	male	2010-08-01
9	8.0	2014-04-24	NaN	2.0	3.0	NaN	NaN	NaN
10	NaN	NaN	4.0	NaN	NaN	Brielle	female	2013-07-17
11	NaN	NaN	5.0	NaN	NaN	Benjamin	male	2010-11-25

16. Determine which sessions occurred on the same day each user registered

In [59]:

```
sers.merge(sessions, left_on=['UserID', 'Registered'], right_on=['UserID', 'SssionDate'])
```

Out[59]:

	UserID	User	Gender	Registered	Cancelled	SessionID	SessionDate
0	2	Pedro	male	2010-08-01	2010-08-08	2	2010-08-01
1	4	Brielle	female	2013-07-17	NaN	9	2013-07-17

17. Build a dataset with every possible (UserID, ProductID) pair (cross join)

In [60]:

```
sers_1 = users
sers_1['key'] = 0

roducts_1 = products
roducts_1['key'] = 0

d.merge(users_1, products_1, on='key', how="outer")[['UserID', 'ProductID']]
```

Out[60]:

	UserID	ProductID
0	1	1
1	1	2
2	1	3
3	1	4
4	1	5
5	2	1
6	2	2
7	2	3
8	2	4
9	2	5
10	3	1
11	3	2
12	3	3
13	3	4
14	3	5
15	4	1
16	4	2
17	4	3
18	4	4
19	4	5
20	5	1
21	5	2
22	5	3
23	5	4
24	5	5

18. Determine how much quantity of each product was purchased by each user

In [63]:

```
sers.merge(products, how='outer').merge(transactions, on=['UserID','ProductI
'], how="outer").loc[:, ["UserID", "ProductID", "Quantity"]].fillna(0)
```

Out[63]:

	UserID	ProductID	Quantity
0	1.0	1	0.0
1	1.0	2	3.0
2	1.0	3	0.0
3	1.0	4	0.0
4	1.0	5	0.0
5	2.0	1	0.0
6	2.0	2	0.0
7	2.0	3	0.0
8	2.0	4	1.0
9	2.0	5	6.0
10	3.0	1	0.0
11	3.0	2	0.0
12	3.0	3	1.0
13	3.0	4	1.0
14	3.0	4	1.0
15	3.0	4	4.0
16	3.0	5	0.0
17	4.0	1	0.0
18	4.0	2	0.0
19	4.0	3	0.0
20	4.0	4	0.0
21	4.0	5	0.0
22	5.0	1	0.0
23	5.0	2	0.0
24	5.0	3	0.0
25	5.0	4	0.0
26	5.0	5	0.0
27	7.0	2	1.0
28	0.0	2	3.0
29	7.0	4	3.0

19. For each user, get each possible pair of pair transactions (TransactionID1,TransacationID2)

In [64]:

```
d.merge(transactions, transactions, on='UserID')
```

Out[64]:

	TransactionID_x	TransactionDate_x	UserID	ProductID_x	Quantity_x	TransactionID_y	Transa
0	1	2010-08-21	7.0	2	1	1	
1	1	2010-08-21	7.0	2	1	9	
2	9	2015-04-24	7.0	4	3	1	
3	9	2015-04-24	7.0	4	3	9	
4	2	2011-05-26	3.0	4	1	2	
5	2	2011-05-26	3.0	4	1	3	
6	2	2011-05-26	3.0	4	1	7	
7	2	2011-05-26	3.0	4	1	10	
8	3	2011-06-16	3.0	3	1	2	
9	3	2011-06-16	3.0	3	1	3	
10	3	2011-06-16	3.0	3	1	7	
11	3	2011-06-16	3.0	3	1	10	
12	7	2013-12-30	3.0	4	1	2	
13	7	2013-12-30	3.0	4	1	3	
14	7	2013-12-30	3.0	4	1	7	
15	7	2013-12-30	3.0	4	1	10	
16	10	2016-05-08	3.0	4	4	2	
17	10	2016-05-08	3.0	4	4	3	
18	10	2016-05-08	3.0	4	4	7	
19	10	2016-05-08	3.0	4	4	10	
20	4	2012-08-26	1.0	2	3	4	
21	5	2013-06-06	2.0	4	1	5	
22	5	2013-06-06	2.0	4	1	6	
23	6	2013-12-23	2.0	5	6	5	
24	6	2013-12-23	2.0	5	6	6	
25	8	2014-04-24	NaN	2	3	8	

20. Join each user to his/her first occuring transaction in the transactions table

In [65]:

```
irst_transactions = transactions[transactions['UserID'].isin(users['UserID'
)].groupby('UserID').first().reset_index()

ata = users.merge(first_transactions, on='UserID', how="outer")

ata
```

Out[65]:

	UserID	User	Gender	Registered	Cancelled	key	TransactionID	TransactionDate	Produ
0	1	Charles	male	2012-12-21	NaN	0	4.0	2012-08-26	
1	2	Pedro	male	2010-08-01	2010-08-08	0	5.0	2013-06-06	
2	3	Caroline	female	2012-10-23	2016-06-07	0	2.0	2011-05-26	
3	4	Brielle	female	2013-07-17	NaN	0	NaN	NaN	
4	5	Benjamin	male	2010-11-25	NaN	0	NaN	NaN	

21. Test to see if we can drop columns

In [72]:

```
olumns=list(data.columns)
olumns
```

Out[72]:

```
['UserID',
 'User',
 'Gender',
 'Registered',
 'Cancelled',
 'key',
 'TransactionID',
 'TransactionDate',
 'ProductID',
 'Quantity']
```

In [83]:

```
ist(data.dropna(axis=1))
```

Out[83]:

```
['UserID', 'User', 'Gender', 'Registered', 'key']
```

In [84]:

```
issing_cols = list(data.columns[data.isnull().any()])
issing_cols
```

Out[84]:

```
['Cancelled', 'TransactionID', 'TransactionDate', 'ProductID', 'Quantity']
```

In [ ]: