In [2]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
df = pd.read_csv('https://raw.githubusercontent.com/jackiekazil/data-wrangling/master/data/chp3/data-text.csv')
```

In [3]:
```
df.shape
```

Out[3]:
```
(4656, 12)
```

In [4]:
```
df.index
```

Out[4]:
```
RangeIndex(start=0, stop=4656, step=1)
```

In [5]:
```
df.columns
```

Out[5]:
```
Index(['Indicator', 'PUBLISH STATES', 'Year', 'WHO region',
       'World Bank income group', 'Country', 'Sex', 'Display Value', 'Numeric',
       'Low', 'High', 'Comments'],
      dtype='object')
```

In [6]:
```
df.dtypes
```

Out[6]:
```
Indicator                  object
PUBLISH STATES             object
Year                        int64
WHO region                 object
World Bank income group    object
Country                    object
Sex                        object
Display Value               int64
Numeric                   float64
Low                       float64
High                      float64
Comments                  float64
dtype: object
```

In [7]:
```
df.describe(include='all')
```

Out[7]:

	Indicator	PUBLISH STATES	Year	WHO region	World Bank income group	Country	Sex	Displa Valu
count	4656	4656	4656.000000	4656	4656	4656	4656	4656.00000
unique	3	1	NaN	6	4	194	3	NaN
top	Life expectancy at birth (years)	Published	NaN	Europe	Lower-middle-income	Turkey	Both sexes	NaN
freq	1746	4656	NaN	1272	1272	24	1552	NaN
mean	NaN	NaN	2002.000000	NaN	NaN	NaN	NaN	47.194588
std	NaN	NaN	8.661184	NaN	NaN	NaN	NaN	23.843194
min	NaN	NaN	1990.000000	NaN	NaN	NaN	NaN	11.000000
25%	NaN	NaN	1997.500000	NaN	NaN	NaN	NaN	20.000000
50%	NaN	NaN	2000.000000	NaN	NaN	NaN	NaN	55.000000
75%	NaN	NaN	2012.000000	NaN	NaN	NaN	NaN	68.000000
max	NaN	NaN	2012.000000	NaN	NaN	NaN	NaN	87.000000

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

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4656 entries, 0 to 4655
Data columns (total 12 columns):
Indicator                   4656 non-null object
PUBLISH STATES              4656 non-null object
Year                        4656 non-null int64
WHO region                  4656 non-null object
World Bank income group     4656 non-null object
Country                     4656 non-null object
Sex                         4656 non-null object
Display Value               4656 non-null int64
Numeric                     4656 non-null float64
Low                         0 non-null float64
High                        0 non-null float64
Comments                    0 non-null float64
dtypes: float64(4), int64(2), object(6)
memory usage: 436.6+ KB
```

In [9]:
```
df.count()
```

Out[9]:
```
Indicator                   4656
PUBLISH STATES              4656
Year                        4656
WHO region                  4656
World Bank income group     4656
Country                     4656
Sex                         4656
Display Value               4656
Numeric                     4656
Low                            0
High                           0
Comments                       0
dtype: int64
```

In [11]:
```
df1 = pd.read_csv('https://raw.githubusercontent.com/kjam/data-wrangling-pycon/master/data/berlin_weather_oldest.csv')
df1 .head(2)
```

Out[11]:

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

2 rows × 21 columns

In [12]:
```
df1.shape
```

Out[12]:
```
(117208, 21)
```

In [13]:
```
df1.index
```

Out[13]:
```
RangeIndex(start=0, stop=117208, step=1)
```

In [14]:
```
df1.columns
```

Out[14]:
```
Index(['STATION', 'STATION_NAME', 'DATE', 'PRCP', 'SNWD', 'SNOW', 'TMAX',
       'TMIN', 'WDFG', 'PGTM', 'WSFG', 'WT09', 'WT07', 'WT01', 'WT06', 'WT05',
       'WT04', 'WT16', 'WT08', 'WT18', 'WT03'],
      dtype='object')
```

In [15]:
```
df1.dtypes
```

Out[15]:
```
STATION         object
STATION_NAME    object
DATE             int64
PRCP             int64
SNWD             int64
SNOW             int64
TMAX             int64
TMIN             int64
WDFG             int64
PGTM             int64
WSFG             int64
WT09             int64
WT07             int64
WT01             int64
WT06             int64
WT05             int64
WT04             int64
WT16             int64
WT08             int64
WT18             int64
WT03             int64
dtype: object
```

In [16]:
```
df1.columns
```

Out[16]:
```
Index(['STATION', 'STATION_NAME', 'DATE', 'PRCP', 'SNWD', 'SNOW', 'TMA
X',
       'TMIN', 'WDFG', 'PGTM', 'WSFG', 'WT09', 'WT07', 'WT01', 'WT06',
'WT05',
       'WT04', 'WT16', 'WT08', 'WT18', 'WT03'],
      dtype='object')
```

In [17]:
```
df1.describe()
```

Out[17]:

	DATE	PRCP	SNWD	SNOW	TMAX	
count	1.172080e+05	117208.000000	117208.000000	117208.000000	117208.000000	117
mean	1.956978e+07	-489.543785	-4610.883327	-9781.858363	3.150766	-68
std	2.723273e+05	2194.002669	4988.596484	1457.514451	1135.851895	109
min	1.900010e+07	-9999.000000	-9999.000000	-9999.000000	-9999.000000	-99
25%	1.937012e+07	0.000000	-9999.000000	-9999.000000	57.000000	1.0
50%	1.960040e+07	0.000000	0.000000	-9999.000000	130.000000	52.
75%	1.980032e+07	14.000000	0.000000	-9999.000000	202.000000	108
max	2.000010e+07	1247.000000	2700.000000	140.000000	748.000000	858

In [18]:
```
df1.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
```

In [19]:
```
df1.count()
```

Out[19]:
```
STATION         117208
STATION_NAME    117208
DATE            117208
PRCP            117208
SNWD            117208
SNOW            117208
TMAX            117208
TMIN            117208
WDFG            117208
PGTM            117208
WSFG            117208
WT09            117208
WT07            117208
WT01            117208
WT06            117208
WT05            117208
WT04            117208
WT16            117208
WT08            117208
WT18            117208
WT03            117208
dtype: int64
```

In [20]:
```
df.head(2)
```

Out[20]:

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

In [21]:
```
df.rename(columns = {'Indicator':'Indicator_id'}, inplace=True )
```

In [22]:
```
df.head(2)
```

Out[22]:

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

In [23]:
```
df.rename(columns = {'PUBLISH STATES':'Publication Status'}, inplace=True )
```

In [24]:
```
df.head(2)
```

Out[24]:

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

In [25]:
```
df.sort_values('Year').head(2)
```

Out[25]:

	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
1270	Life expectancy at birth (years)	Published	1990	Europe	High-income	Germany	Male	72	72.0

In [27]:
```
#df.sort_values(['Indicator_id', 'Country', 'Year', 'WHO region', 'Publication Status'], ascending=[True, True, True, True, True], inplace=False).head(2)
df.sort_values(['Indicator_id', 'Country', 'Year', 'WHO region', 'Publication Status'], ascending=
[True, True, True, True, True]).head(2)
```

Out[27]:

	Indicator_id	Publication Status	Year	WHO region	World Bank income group	Country	Sex	Display Value
2798	Healthy life expectancy (HALE) at birth (years)	Published	2000	Eastern Mediterranean	Low-income	Afghanistan	Male	45
3363	Healthy life expectancy (HALE) at birth (years)	Published	2000	Eastern Mediterranean	Low-income	Afghanistan	Both sexes	45

In [28]:
```
df.columns
```

Out[28]:
```
Index(['Indicator_id', 'Publication Status', 'Year', 'WHO region',
       'World Bank income group', 'Country', 'Sex', 'Display Value', '
Numeric',
       'Low', 'High', 'Comments'],
      dtype='object')
```

In [29]:
```
df.columns
```

Out[29]:
```
Index(['Indicator_id', 'Publication Status', 'Year', 'WHO region',
       'World Bank income group', 'Country', 'Sex', 'Display Value', '
Numeric',
       'Low', 'High', 'Comments'],
      dtype='object')
```

In [30]:
```
df = df[['Country', 'Indicator_id', 'Publication Status', 'Year', 'WHO
region',
'World Bank income group', 'Sex', 'Display Value', 'Numeric',
'Low', 'High', 'Comments']]
```

In [31]:
```
df.columns
```

Out[31]:
```
Index(['Country', 'Indicator_id', 'Publication Status', 'Year', 'WHO r
egion',
       'World Bank income group', 'Sex', 'Display Value', 'Numeric', '
Low',
       'High', 'Comments'],
      dtype='object')
```

In [32]:
```
df.as_matrix(columns=df.columns[:1])
```

Out[32]:
```
array([['Andorra'],
       ['Andorra'],
       ['Andorra'],
       ...,
       ['South Africa'],
       ['Zambia'],
       ['Zimbabwe']], dtype=object)
```

In [33]:
```
df.iloc[[11,24,37]]
```

Out[33]:

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

In [34]:
```
bad_rows = df.index.isin([5,23,34,56])
df[~bad_rows]
```

Out[34]:

	Country	Indicator_id	Publication Status	Year	WHO region	World Bank income group	Sex	Displa Valu
0	Andorra	Life expectancy at birth (years)	Published	1990	Europe	High-income	Both sexes	77
1	Andorra	Life expectancy at birth (years)	Published	2000	Europe	High-income	Both sexes	80
2	Andorra	Life expectancy at age 60 (years)	Published	2012	Europe	High-income	Female	28

3	Andorra	Life expectancy at age 60 (years)	Published	2000	Europe	High-income	Both sexes	23
4	United Arab Emirates	Life expectancy at birth (years)	Published	2012	Eastern Mediterranean	High-income	Female	78
6	Antigua and Barbuda	Life expectancy at age 60 (years)	Published	1990	Americas	High-income	Male	17
7	Antigua and Barbuda	Life expectancy at age 60 (years)	Published	2012	Americas	High-income	Both sexes	22
8	Australia	Life expectancy at birth (years)	Published	2012	Western Pacific	High-income	Male	81
9	Australia	Life expectancy at birth (years)	Published	2000	Western Pacific	High-income	Both sexes	80
10	Australia	Life expectancy at birth (years)	Published	2012	Western Pacific	High-income	Both sexes	83
11	Austria	Life expectancy at birth (years)	Published	2012	Europe	High-income	Female	83
12	Austria	Life expectancy at age 60 (years)	Published	2012	Europe	High-income	Female	25
13	Belgium	Life expectancy at birth (years)	Published	2012	Europe	High-income	Female	83

14	Bahrain	Life expectancy at birth (years)	Published	2000	Eastern Mediterranean	High-income	Male	73
15	Bahrain	Life expectancy at birth (years)	Published	1990	Eastern Mediterranean	High-income	Female	74
16	Bahrain	Life expectancy at age 60 (years)	Published	1990	Eastern Mediterranean	High-income	Male	17
17	Bahamas	Life expectancy at birth (years)	Published	2012	Americas	High-income	Male	72
18	Bahamas	Life expectancy at age 60 (years)	Published	2000	Americas	High-income	Both sexes	21
19	Barbados	Life expectancy at birth (years)	Published	1990	Americas	High-income	Male	71
20	Barbados	Life expectancy at age 60 (years)	Published	2012	Americas	High-income	Female	25
21	Barbados	Life expectancy at age 60 (years)	Published	2012	Americas	High-income	Both sexes	23
22	Brunei Darussalam	Life expectancy at age 60 (years)	Published	1990	Western Pacific	High-income	Female	20
24	Brunei Darussalam	Life expectancy at age 60 (years)	Published	2012	Western Pacific	High-income	Female	21

25	Canada	Life expectancy at birth (years)	Published	2000	Americas	High-income	Female	82
26	Canada	Life expectancy at age 60 (years)	Published	2000	Americas	High-income	Male	21
27	Canada	Life expectancy at age 60 (years)	Published	1990	Americas	High-income	Female	24
28	Switzerland	Life expectancy at birth (years)	Published	1990	Europe	High-income	Male	74
29	Switzerland	Life expectancy at birth (years)	Published	2012	Europe	High-income	Both sexes	83
30	Switzerland	Life expectancy at age 60 (years)	Published	2000	Europe	High-income	Both sexes	23
31	Cook Islands	Life expectancy at birth (years)	Published	2012	Western Pacific	High-income	Both sexes	76
...	...	...	...	...	...	...	...	...
4626	Serbia	Healthy life expectancy (HALE) at birth (years)	Published	2012	Europe	Upper-middle-income	Female	67
4627	Suriname	Healthy life expectancy (HALE) at birth (years)	Published	2012	Americas	Upper-middle-income	Both sexes	66
4628	Sweden	Healthy life expectancy (HALE) at	Published	2012	Europe	High-income	Both sexes	72

		birth (years)						
4629	Swaziland	Healthy life expectancy (HALE) at birth (years)	Published	2012	Africa	Lower-middle-income	Female	47
4630	Seychelles	Healthy life expectancy (HALE) at birth (years)	Published	2000	Africa	Upper-middle-income	Male	61
4631	Syrian Arab Republic	Healthy life expectancy (HALE) at birth (years)	Published	2000	Eastern Mediterranean	Lower-middle-income	Female	64
4632	Chad	Healthy life expectancy (HALE) at birth (years)	Published	2012	Africa	Low-income	Female	44
4633	Thailand	Healthy life expectancy (HALE) at birth (years)	Published	2000	South-East Asia	Lower-middle-income	Male	59
4634	Thailand	Healthy life expectancy (HALE) at birth (years)	Published	2000	South-East Asia	Lower-middle-income	Female	65
4635	Tajikistan	Healthy life expectancy (HALE) at birth (years)	Published	2000	Europe	Low-income	Both sexes	56
4636	Tajikistan	Healthy life expectancy (HALE) at birth (years)	Published	2012	Europe	Low-income	Female	60
4637	Tonga	Healthy life expectancy (HALE) at birth (years)	Published	2012	Western Pacific	Lower-middle-income	Female	61
4638	Trinidad and Tobago	Healthy life expectancy (HALE) at	Published	2012	Americas	High-income	Female	64

		birth (years)						
4639	Trinidad and Tobago	Healthy life expectancy (HALE) at birth (years)	Published	2012	Americas	High-income	Both sexes	61
4640	Tunisia	Healthy life expectancy (HALE) at birth (years)	Published	2000	Eastern Mediterranean	Lower-middle-income	Male	63
4641	Tuvalu	Healthy life expectancy (HALE) at birth (years)	Published	2012	Western Pacific	Upper-middle-income	Male	57
4642	Uganda	Healthy life expectancy (HALE) at birth (years)	Published	2000	Africa	Low-income	Female	40
4643	Ukraine	Healthy life expectancy (HALE) at birth (years)	Published	2000	Europe	Lower-middle-income	Both sexes	60
4644	Uruguay	Healthy life expectancy (HALE) at birth (years)	Published	2012	Americas	Upper-middle-income	Male	65
4645	Uruguay	Healthy life expectancy (HALE) at birth (years)	Published	2012	Americas	Upper-middle-income	Female	70
4646	Uruguay	Healthy life expectancy (HALE) at birth (years)	Published	2012	Americas	Upper-middle-income	Both sexes	68
4647	Saint Vincent and the Grenadines	Healthy life expectancy (HALE) at birth (years)	Published	2000	Americas	Upper-middle-income	Both sexes	61
4648	Venezuela (Bolivarian Republic	Healthy life expectancy (HALE) at	Published	2012	Americas	Upper-middle-income	Both sexes	66

	of)	birth (years)						
4649	Vanuatu	Healthy life expectancy (HALE) at birth (years)	Published	2000	Western Pacific	Lower-middle-income	Male	59
4650	Samoa	Healthy life expectancy (HALE) at birth (years)	Published	2012	Western Pacific	Lower-middle-income	Male	62
4651	Samoa	Healthy life expectancy (HALE) at birth (years)	Published	2012	Western Pacific	Lower-middle-income	Female	66
4652	Yemen	Healthy life expectancy (HALE) at birth (years)	Published	2012	Eastern Mediterranean	Low-income	Both sexes	54
4653	South Africa	Healthy life expectancy (HALE) at birth (years)	Published	2000	Africa	Upper-middle-income	Male	49
4654	Zambia	Healthy life expectancy (HALE) at birth (years)	Published	2000	Africa	Low-income	Both sexes	36
4655	Zimbabwe	Healthy life expectancy (HALE) at birth (years)	Published	2012	Africa	Low-income	Female	51

4652 rows × 12 columns

In [35]:
```
users=pd.read_csv('https://raw.githubusercontent.com/ben519/DataWrangling/master/Data/users.csv')
sessions=pd.read_csv('https://raw.githubusercontent.com/ben519/DataWrangling/master/Data/sessions.csv')
products=pd.read_csv('https://raw.githubusercontent.com/ben519/DataWrangling/master/Data/products.csv')
transactions=pd.read_csv('https://raw.githubusercontent.com/ben519/DataWrangling/master/Data/transactions.csv')
```

In [36]:
```
users.head()
```

Out[36]:

	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 [37]:
```
sessions.head()
```

Out[37]:

	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 [38]:
```
products.head()
```

Out[38]:

	ProductID	Product	Price
0	1	A	14.16
1	2	B	33.04
2	3	C	10.65
3	4	D	10.02
4	5	E	29.66

In [39]:
```
transactions.head()
```

Out[39]:

	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

In [42]:
```
pd.merge
```

Out[42]:
```
<function pandas.core.reshape.merge.merge>
```

In [43]:
```
display(pd.merge(transactions,users, on="UserID", how='left'))
```

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

In [44]:
```
display(transactions['UserID'].isin(users['UserID']))
```

```
0    False
1     True
2     True
3     True
4     True
5     True
6     True
7    False
8    False
9     True
Name: UserID, dtype: bool
```

In [45]:
```
transactions.iloc[[0,7,8]]
```

Out[45]:

	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

In [46]:
```
display(pd.merge(transactions,users, on='UserID', how='inner'))
```

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

In [47]:
```
display(pd.merge(transactions,users, on='UserID', how='outer'))
```

	TransactionID	TransactionDate	UserID	ProductID	Quantity	User	Gender	Reg
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	201 23
3	3.0	2011-06-16	3.0	3.0	1.0	Caroline	female	201 23
4	7.0	2013-12-30	3.0	4.0	1.0	Caroline	female	201 23
5	10.0	2016-05-08	3.0	4.0	4.0	Caroline	female	201 23
6	4.0	2012-08-26	1.0	2.0	3.0	Charles	male	201 21
7	5.0	2013-06-06	2.0	4.0	1.0	Pedro	male	201 01
8	6.0	2013-12-23	2.0	5.0	6.0	Pedro	male	201 01
9	8.0	2014-04-24	NaN	2.0	3.0	NaN	NaN	NaN
10	NaN	NaN	4.0	NaN	NaN	Brielle	female	201 17
11	NaN	NaN	5.0	NaN	NaN	Benjamin	male	201 25

In [48]:
```
df16=pd.merge(sessions,users, on='UserID', how='inner')
```

In [49]:
```
df16.loc[df16['SessionDate'] == df16['Registered']]
```

Out[49]:

	SessionID	SessionDate	UserID	User	Gender	Registered	Cancelled
1	2	2010-08-01	2	Pedro	male	2010-08-01	2010-08-08
7	9	2013-07-17	4	Brielle	female	2013-07-17	NaN

In [50]:
```
df17 = users.assign(foo=1).merge(products.assign(foo=1)).drop('foo', 1)
display(df17)
```

	UserID	User	Gender	Registered	Cancelled	ProductID	Product	Price
0	1	Charles	male	2012-12-21	NaN	1	A	14.16
1	1	Charles	male	2012-12-21	NaN	2	B	33.04
2	1	Charles	male	2012-12-21	NaN	3	C	10.65
3	1	Charles	male	2012-12-21	NaN	4	D	10.02
4	1	Charles	male	2012-12-21	NaN	5	E	29.66
5	2	Pedro	male	2010-08-01	2010-08-08	1	A	14.16
6	2	Pedro	male	2010-08-01	2010-08-08	2	B	33.04
7	2	Pedro	male	2010-08-01	2010-08-08	3	C	10.65
8	2	Pedro	male	2010-08-01	2010-08-08	4	D	10.02
9	2	Pedro	male	2010-08-01	2010-08-08	5	E	29.66
10	3	Caroline	female	2012-10-23	2016-06-07	1	A	14.16
11	3	Caroline	female	2012-10-23	2016-06-07	2	B	33.04
12	3	Caroline	female	2012-10-23	2016-06-07	3	C	10.65
13	3	Caroline	female	2012-10-23	2016-06-07	4	D	10.02
14	3	Caroline	female	2012-10-23	2016-06-07	5	E	29.66
15	4	Brielle	female	2013-07-17	NaN	1	A	14.16
16	4	Brielle	female	2013-07-17	NaN	2	B	33.04
17	4	Brielle	female	2013-07-17	NaN	3	C	10.65
18	4	Brielle	female	2013-07-17	NaN	4	D	10.02
19	4	Brielle	female	2013-07-17	NaN	5	E	29.66
20	5	Benjamin	male	2010-11-25	NaN	1	A	14.16
21	5	Benjamin	male	2010-11-25	NaN	2	B	33.04
22	5	Benjamin	male	2010-11-25	NaN	3	C	10.65
23	5	Benjamin	male	2010-11-25	NaN	4	D	10.02
24	5	Benjamin	male	2010-11-25	NaN	5	E	29.66

In [51]:
```
display(transactions.sort_values('UserID'))
```

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

In [52]:
```
pd.merge(transactions, transactions, on='UserID')
```

Out[52]:

	TransactionID_x	TransactionDate_x	UserID	ProductID_x	Quantity_x	TransactionID
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

In [53]:
```
data=pd.merge(users, transactions.groupby('UserID').first().reset_index
(), how='left', on='UserID'
)
data
```

Out[53]:

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

In [54]:
```
data=pd.merge(users, transactions.groupby('UserID').first().reset_index
(), how='left', on='UserID')
data
```

Out[54]:

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

In [55]:
```
list(data.dropna(thresh=int(data.shape[0] * .9), axis=1).columns)
```

Out[55]:
```
['UserID', 'User', 'Gender', 'Registered']
```

In [56]:
```
missing_info = list(data.columns[data.isnull().any()])
missing_info
```

Out[56]:
```
['Cancelled', 'TransactionID', 'TransactionDate', 'ProductID', 'Quanti
ty']
```

In [61]:
```
for col in missing_info: num_missing = data[data[col].isnull() == True]
.shape[0]
print('number missing for column {}: {}'.format(col, num_missing))
```

```
number missing for column Quantity: 2
```