In [1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as pyplot
%matplotlib inline

df=pd.read_csv('https://raw.githubusercontent.com/jackiekazil/data-wrangling/master/data/chp3/data-text.csv')
df1=pd.read_csv('https://raw.githubusercontent.com/kjam/data-wrangling-pycon/master/data/berlin_weather_oldest.csv')
df.head(2)
```

Out[1]:

| | 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 [2]:
```
df1.head(2)
```

Out[2]:

| | 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

In [17]:
```
#1. Get the Metadata from the above files.
print("Metadata for df")
print("-------------------")
df.info()
```

```
Metadata for df
-------------------
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4656 entries, 0 to 4655
Data columns (total 12 columns):
Indicator_1                4656 non-null object
Publication status         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: 327.4+ KB
```

In [18]:
```
print("Metadata for df1")
print("-------------------")
df1.info()
```

```
Metadata for df1
-------------------
<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: 17.9+ MB
```

In [20]:
```
# 2. Get the row names from the above files.
print("All the rows for df are :")
df.index.values
```

```
All the rows for df are :
```

Out[20]:
```
array([   0,    1,    2, ..., 4653, 4654, 4655], dtype=int64)
```

In [21]:
```
print("All the rows for df1 are :")
df1.index.values
```

```
All the rows for df1 are :
```

Out[21]:
```
array([     0,      1,      2, ..., 117205, 117206, 117207], dtype=int64)
```

In [22]:
```
# 3. Change the column name from any of the above file.
df.rename(columns={"Indicator":"Indicator_1"}).head(2)
```

Out[22]:

| | Indicator_1 | Publication status | 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 [23]:
```
# 4. Change the column name from any of the above file and store the changes made permanently.
df.rename(columns={"Indicator":"Indicator_1"},inplace=True)
df.head(2)
```

Out[23]:

| | Indicator_1 | Publication status | 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 [24]:
```
#5. Change the names of multiple columns. (Is this going to be a perm change?
 any other method for changing column names?)
df.rename(columns={"PUBLISH STATES":"Publication status","WHO region":"WHO Reg
ion"},inplace=True)
df.head(2)
```

Out[24]:

| | Indicator_1 | Publication status | 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 [25]:
```
# 6. Arrange values of a particular column in ascending order.
df.sort_values('Year').head(5)
```

Out[25]:

| | Indicator_1 | Publication status | 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 |
| **1270** | Life expectancy at birth (years) | Published | 1990 | Europe | High-income | Germany | Male | 72 | 72.0 | NaN |
| **3193** | Life expectancy at birth (years) | Published | 1990 | Europe | Lower-middle-income | Republic of Moldova | Male | 65 | 65.0 | NaN |
| **3194** | Life expectancy at birth (years) | Published | 1990 | Europe | Lower-middle-income | Republic of Moldova | Both sexes | 68 | 68.0 | NaN |
| **3197** | Life expectancy at age 60 (years) | Published | 1990 | Europe | Lower-middle-income | Republic of Moldova | Male | 15 | 15.0 | NaN |

In [26]:
```
# 7. Arrange multiple column values in ascending order.
df[0:3].sort_values(['Indicator_1','Country','Year','WHO Region','Publication
 status'])
df[['Indicator_1','Country','Year','WHO Region','Publication status']].head(3)
```

Out[26]:

| | Indicator_1 | 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 |

In [28]:
```
#8. Make country as the first column of the dataframe.
df[['Country','Indicator_1', 'Publication status', 'Year', 'WHO Region',
       'World Bank income group', 'Sex','Display Value', 'Numeric',
       'Low', 'High', 'Comments']].head()
```

Out[28]:

| | Country | Indicator_1 | Publication status | Year | WHO Region | World Bank income group | Sex | Display Value | Numeric | Lo |
|---|---|---|---|---|---|---|---|---|---|---|
| **0** | Andorra | Life expectancy at birth (years) | Published | 1990 | Europe | High-income | Both sexes | 77 | 77.0 | Na |
| **1** | Andorra | Life expectancy at birth (years) | Published | 2000 | Europe | High-income | Both sexes | 80 | 80.0 | Na |
| **2** | Andorra | Life expectancy at age 60 (years) | Published | 2012 | Europe | High-income | Female | 28 | 28.0 | Na |
| **3** | Andorra | Life expectancy at age 60 (years) | Published | 2000 | Europe | High-income | Both sexes | 23 | 23.0 | Na |
| **4** | United Arab Emirates | Life expectancy at birth (years) | Published | 2012 | Eastern Mediterranean | High-income | Female | 78 | 78.0 | Na |

In [29]:
```
#9. Get the column array using a variable
np.array(df[['WHO Region']])
```

Out[29]:
```
array([['Europe'],
       ['Europe'],
       ['Europe'],
       ...,
       ['Africa'],
       ['Africa'],
       ['Africa']], dtype=object)
```

In [30]:
```
#10 Get the subset rows 11, 24, 37 Expected Output:
df[11:38:13]
```

Out[30]:

| | Indicator_1 | Publication status | Year | WHO Region | World Bank income group | Country | Sex | Display Value | Numeric | Low |
|---|---|---|---|---|---|---|---|---|---|---|
| **11** | Life expectancy at birth (years) | Published | 2012 | Europe | High-income | Austria | Female | 83 | 83.0 | NaN |
| **24** | Life expectancy at age 60 (years) | Published | 2012 | Western Pacific | High-income | Brunei Darussalam | Female | 21 | 21.0 | NaN |
| **37** | Life expectancy at age 60 (years) | Published | 2012 | Europe | High-income | Cyprus | Female | 26 | 26.0 | NaN |

In [31]:
```
# 11.Get the subset rows excluding 5, 12, 23, and 56
df.drop([5,12,23,56])[0:54]
```

Out[31]:

| | Indicator_1 | 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_1 | Publication status | Year | WHO Region | World Bank income group | Country | Sex | Display Value | Numeric |
|---|---|---|---|---|---|---|---|---|---|
| **15** | Life expectancy at birth (years) | Published | 1990 | Eastern Mediterranean | High-income | Bahrain | Female | 74 | 74.0 |
| **16** | Life expectancy at age 60 (years) | Published | 1990 | Eastern Mediterranean | High-income | Bahrain | Male | 17 | 17.0 |
| **17** | Life expectancy at birth (years) | Published | 2012 | Americas | High-income | Bahamas | Male | 72 | 72.0 |
| **18** | Life expectancy at age 60 (years) | Published | 2000 | Americas | High-income | Bahamas | Both sexes | 21 | 21.0 |
| **19** | Life expectancy at birth (years) | Published | 1990 | Americas | High-income | Barbados | Male | 71 | 71.0 |
| **20** | Life expectancy at age 60 (years) | Published | 2012 | Americas | High-income | Barbados | Female | 25 | 25.0 |
| **21** | Life expectancy at age 60 (years) | Published | 2012 | Americas | High-income | Barbados | Both sexes | 23 | 23.0 |
| **22** | Life expectancy at age 60 (years) | Published | 1990 | Western Pacific | High-income | Brunei Darussalam | Female | 20 | 20.0 |
| **24** | Life expectancy at age 60 (years) | Published | 2012 | Western Pacific | High-income | Brunei Darussalam | Female | 21 | 21.0 |
| **25** | Life expectancy at birth (years) | Published | 2000 | Americas | High-income | Canada | Female | 82 | 82.0 |
| **26** | Life expectancy at age 60 (years) | Published | 2000 | Americas | High-income | Canada | Male | 21 | 21.0 |
| **27** | Life expectancy at age 60 (years) | Published | 1990 | Americas | High-income | Canada | Female | 24 | 24.0 |
| **28** | Life expectancy at birth (years) | Published | 1990 | Europe | High-income | Switzerland | Male | 74 | 74.0 |

| | Indicator_1 | Publication status | Year | WHO Region | World Bank income group | Country | Sex | Display Value | Numeric |
|---|---|---|---|---|---|---|---|---|---|
| **29** | Life expectancy at birth (years) | Published | 2012 | Europe | High-income | Switzerland | Both sexes | 83 | 83.0 |
| **30** | Life expectancy at age 60 (years) | Published | 2000 | Europe | High-income | Switzerland | Both sexes | 23 | 23.0 |
| **31** | Life expectancy at birth (years) | Published | 2012 | Western Pacific | High-income | Cook Islands | Both sexes | 76 | 76.0 |
| **32** | Life expectancy at age 60 (years) | Published | 2012 | Western Pacific | High-income | Cook Islands | Female | 22 | 22.0 |
| **33** | Life expectancy at age 60 (years) | Published | 2000 | Western Pacific | High-income | Cook Islands | Both sexes | 18 | 18.0 |
| **34** | Life expectancy at birth (years) | Published | 2000 | Europe | High-income | Cyprus | Female | 79 | 79.0 |
| **35** | Life expectancy at birth (years) | Published | 2000 | Europe | High-income | Cyprus | Both sexes | 77 | 77.0 |
| **36** | Life expectancy at age 60 (years) | Published | 2000 | Europe | High-income | Cyprus | Female | 22 | 22.0 |
| **37** | Life expectancy at age 60 (years) | Published | 2012 | Europe | High-income | Cyprus | Female | 26 | 26.0 |
| **38** | Life expectancy at birth (years) | Published | 2012 | Europe | High-income | Czech Republic | Male | 75 | 75.0 |
| **39** | Life expectancy at birth (years) | Published | 1990 | Europe | High-income | Czech Republic | Female | 75 | 75.0 |
| **40** | Life expectancy at birth (years) | Published | 1990 | Europe | High-income | Germany | Female | 79 | 79.0 |
| **41** | Life expectancy at age 60 (years) | Published | 1990 | Europe | High-income | Germany | Male | 18 | 18.0 |

| | Indicator_1 | Publication status | Year | WHO Region | World Bank income group | Country | Sex | Display Value | Numeric |
|---|---|---|---|---|---|---|---|---|---|
| **42** | Life expectancy at age 60 (years) | Published | 2000 | Europe | High-income | Germany | Male | 20 | 20.0 |
| **43** | Life expectancy at birth (years) | Published | 2012 | Europe | High-income | Denmark | Both sexes | 80 | 80.0 |
| **44** | Life expectancy at age 60 (years) | Published | 1990 | Europe | High-income | Denmark | Male | 18 | 18.0 |
| **45** | Life expectancy at age 60 (years) | Published | 2000 | Europe | High-income | Denmark | Male | 19 | 19.0 |
| **46** | Life expectancy at age 60 (years) | Published | 2000 | Europe | High-income | Denmark | Both sexes | 21 | 21.0 |
| **47** | Life expectancy at age 60 (years) | Published | 2012 | Europe | High-income | Denmark | Both sexes | 23 | 23.0 |
| **48** | Life expectancy at birth (years) | Published | 2012 | Europe | High-income | Spain | Both sexes | 82 | 82.0 |
| **49** | Life expectancy at age 60 (years) | Published | 2000 | Europe | High-income | Spain | Female | 25 | 25.0 |
| **50** | Life expectancy at age 60 (years) | Published | 1990 | Europe | High-income | Spain | Both sexes | 22 | 22.0 |
| **51** | Life expectancy at birth (years) | Published | 1990 | Europe | High-income | Estonia | Female | 75 | 75.0 |
| **52** | Life expectancy at birth (years) | Published | 2012 | Europe | High-income | Estonia | Female | 81 | 81.0 |
| **53** | Life expectancy at age 60 (years) | Published | 2012 | Europe | High-income | Estonia | Male | 18 | 18.0 |
| **54** | Life expectancy at age 60 (years) | Published | 2000 | Europe | High-income | Estonia | Female | 21 | 21.0 |

| | Indicator_1 | Publication status | Year | WHO Region | World Bank income group | Country | Sex | Display Value | Numeric |
|---|---|---|---|---|---|---|---|---|---|
| **55** | Life expectancy at age 60 (years) | Published | 2012 | Europe | High-income | Estonia | Female | 24 | 24.0 |
| **57** | Life expectancy at age 60 (years) | Published | 2012 | Europe | High-income | Finland | Male | 22 | 22.0 |

In [32]:
```
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')
users.head()
```

Out[32]:

| | 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 [33]:
```
sessions.head()
```

Out[33]:

| | 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 [34]:
```
products.head()
```

Out[34]:

| | 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 [35]:
```
transactions.head()
```

Out[35]:

| | 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 [36]:
```
#12. Join users to transactions, keeping all rows from transactions and only m
atching rows from users (left join)
pd.merge(transactions,users, how="left")
```

Out[36]:

| | 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 |

In [40]:
```
# 13. Which transactions have a UserID not in users?

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

Out[40]:

| | 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 [38]:
```
#14. Join users to transactions, keeping only rows from transactions and users
 that match via UserID (inner join)
pd.merge(transactions,users,how="inner",on="UserID")
```

Out[38]:

| | 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 |

In [41]:
```
# 15. Join users to transactions, displaying all matching rows AND all non-matching rows (full outer join)
pd.merge(transactions,users,how='outer')
```

Out[41]:

| | 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 |

In [42]:
```
# 16. Determine which sessions occurred on the same day each user registered

pd.merge(left=users,right=sessions,how="inner",left_on=['UserID','Registered'],right_on=['UserID','SessionDate'])
```

Out[42]:

| | 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 |

In [43]:
```
# 17. Build a dataset with every possible (UserID, ProductID) pair (cross joi
n)
df_userid = pd.DataFrame({"UserID":users["UserID"]})
df_Tran = pd.DataFrame({"ProductID":products["ProductID"]})
#create new column Key with value as 1 for both the dataframe as this would be
come the common key to be merged
df_userid['Key'] = 1
df_Tran['Key'] = 1
dataset=pd.merge(df_userid,df_Tran,how='outer',on="Key")[['UserID','ProductID'
]]
dataset
```

Out[43]:

| | 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 |

In [44]:
```
# 18. Determine how much quantity of each product was purchased by each user

#do a left join on the output table df_out from previous step with transaction
s table on the keys ['UserID','ProductID]
df_user_prod_quant = pd.merge(dataset,transactions,how='left',on=['UserID','Pr
oductID'])

#Groupby the table on ['UserID','ProductID] and calculate the sum of Qunatity
 entity for each group
df_user_quantity = df_user_prod_quant.groupby(['UserID','ProductID'])['Quantit
y'].sum()

#reset index so that the index column will have consecutive default numbers an
d fill NAN values with 0
df_user_quantity.reset_index().fillna(0)
```

Out[44]:

| | UserID | ProductID | Quantity |
|---|---|---|---|
| **0** | 1 | 1 | 0.0 |
| **1** | 1 | 2 | 3.0 |
| **2** | 1 | 3 | 0.0 |
| **3** | 1 | 4 | 0.0 |
| **4** | 1 | 5 | 0.0 |
| **5** | 2 | 1 | 0.0 |
| **6** | 2 | 2 | 0.0 |
| **7** | 2 | 3 | 0.0 |
| **8** | 2 | 4 | 1.0 |
| **9** | 2 | 5 | 6.0 |
| **10** | 3 | 1 | 0.0 |
| **11** | 3 | 2 | 0.0 |
| **12** | 3 | 3 | 1.0 |
| **13** | 3 | 4 | 6.0 |
| **14** | 3 | 5 | 0.0 |
| **15** | 4 | 1 | 0.0 |
| **16** | 4 | 2 | 0.0 |
| **17** | 4 | 3 | 0.0 |
| **18** | 4 | 4 | 0.0 |
| **19** | 4 | 5 | 0.0 |
| **20** | 5 | 1 | 0.0 |
| **21** | 5 | 2 | 0.0 |
| **22** | 5 | 3 | 0.0 |
| **23** | 5 | 4 | 0.0 |
| **24** | 5 | 5 | 0.0 |

In [46]:
```
# 19. For each user, get each possible pair of pair transactions (TransactionID1,TransacationID2)
pd.merge(transactions,transactions,how='outer',on='UserID')
```

Out[46]:

| | 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 | |

In [47]:
```
# 20. Join each user to his/her first occuring transaction in the transactions
 table

#do an left outer join on user and transactions dataframe on the UserID column
df_usertran = pd.merge(users,transactions,how='left',on='UserID')
# craete a new dataframe df_ with all duplicates on UserID being dropped , onl
y keeping the first entry
df_ = df_usertran.drop_duplicates(subset='UserID')
#reset the index to the default integer index.
df_ = df_.reset_index(drop=True)

#display the contents of the dataframe df_
df_
```

Out[47]:

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

In [48]:
```
# 21. Test to see if we can drop columns
#Retieve the column list for the dataframe df_ created in problem statement 20
my_columns = list(df_.columns)
print(my_columns)
```

```
['UserID', 'User', 'Gender', 'Registered', 'Cancelled', 'TransactionID', 'Tra
nsactionDate', 'ProductID', 'Quantity']
```

In [54]:
```
list(df_.dropna(thresh=int(df_.shape[0] * .9), axis=1).columns) #set threshold
 to drop NAs
```

Out[54]:
```
['UserID', 'User', 'Gender', 'Registered']
```

In [55]:
```
missing_info = list(df_.columns[df_.isnull().any()])
missing_info
```

Out[55]:
```
['Cancelled', 'TransactionID', 'TransactionDate', 'ProductID', 'Quantity']
```

In [51]:
```
for col in missing_info:
    num_missing = df_[df_[col].isnull() == True].shape[0]
    print('number missing for column {}: {}'.format(col, num_missing))
```

```
number missing for column Cancelled: 3
number missing for column TransactionID: 2
number missing for column TransactionDate: 2
number missing for column ProductID: 2
number missing for column Quantity: 2
```

In [56]:
```
#count of missing df_
for col in missing_info:
    percent_missing = df_[df_[col].isnull() == True].shape[0] / df_.shape[0]
    print('percent missing for column {}: {}'.format(col, percent_missing))
```

```
percent missing for column Cancelled: 0.6
percent missing for column TransactionID: 0.4
percent missing for column TransactionDate: 0.4
percent missing for column ProductID: 0.4
percent missing for column Quantity: 0.4
```