```
In [ ]: # Data Manipulation
        ## Data Import
```

```
In [1]: import pandas as pd
        # Customer Detail Data
        cusdet=pd.read_csv('/Users/xiyongzhang/documents/MQ/RA_ACST890_notes/w10_example_pandas_
        cusdet
```

```
Out[1]:   Customer  gender Country   ID   age  item
        0     Gary    Male      AU  342  25.0    35
        1     Anny  Female      US  135  45.0    45
        2  Yi-lung  Female      US  346  23.0   234
        3   Duncan    Male      US  121   NaN    23
        4    Kevin    Male      AU  223  31.0    85
        5    Angel  Female      AU  432  11.0     5
```

```
In [2]: cusdet.head(2)  # first 2 lines
        # also head for lines from the back
```

```
Out[2]:   Customer  gender Country   ID   age  item
        0     Gary    Male      AU  342  25.0    35
        1     Anny  Female      US  135  45.0    45
```

```
In [3]: cusdet.describe() # numerical description
```

```
Out[3]:                ID        age        item
        count    6.000000   5.000000    6.000000
        mean   266.500000  27.000000   71.166667
        std    126.305582  12.409674   84.138972
        min    121.000000  11.000000    5.000000
        25%    157.000000  23.000000   26.000000
        50%    282.500000  25.000000   40.000000
        75%    345.000000  31.000000   75.000000
        max    432.000000  45.000000  234.000000
```

```
In [ ]:
```

```
In [ ]: ## Data filtering and subsetting
```

```
In [4]: cusdet['ID'] # to call variables
```

```
Out[4]: 0    342
        1    135
        2    346
        3    121
        4    223
        5    432
        Name: ID, dtype: int64
```

```
In [5]: cusdet[2:4] # subsetting
```

```
Out[5]:   Customer  gender Country   ID   age  item
        2  Yi-lung  Female      US  346  23.0   234
        3   Duncan    Male      US  121   NaN    23
```

```
In [6]: cusdet[cusdet['age'] >30] # filtering
```

```
Out[6]:   Customer  gender Country   ID   age  item
        1     Anny  Female      US  135  45.0    45
        4    Kevin    Male      AU  223  31.0    85
```

```
In [7]: # little quiz
        cusdet[cusdet['Country']=='AU']['age']
        cusdet[cusdet['Country']=='AU']['age'].sum()
        cusdet[cusdet['age'].isnull()]

        # common aggregation functions
        # count() Number of non-null observations sum() Sum of values
        # mean() Mean of values
        # median() Arithmetic median of values min()   Minimum
        # max()   Maximum
        # prod() Product of values
        # std() Unbiased standard deviation
        # var() Unbiased variance
```

```
Out[7]:   Customer gender Country   ID  age  item
        3   Duncan   Male      US  121  NaN    23
```

```
In [ ]:
```

```
In [ ]: ## Data Modification
```

```
In [8]: # we can apply many functions
        def sq(x):
            return(x**2)
        cusdet['age'].apply(sq)
```

```
Out[8]: 0     625.0
        1    2025.0
        2     529.0
        3       NaN
        4     961.0
        5     121.0
        Name: age, dtype: float64
```

```
In [9]: # Adding a row
        cusdet=cusdet.append({'Customer':'Eddy','ID':250,'age':12},ignore_index=True)
        cusdet
```

```
Out[9]:   Customer  gender Country   ID   age   item
        0     Gary    Male      AU  342  25.0   35.0
        1     Anny  Female      US  135  45.0   45.0
        2  Yi-lung  Female      US  346  23.0  234.0
        3   Duncan    Male      US  121   NaN   23.0
        4    Kevin    Male      AU  223  31.0   85.0
        5    Angel  Female      AU  432  11.0    5.0
        6     Eddy     NaN     NaN  250  12.0    NaN
```

```
In [10]: # Deleting a row
         cusdet.drop([1,2,3])
         cusdet.drop(cusdet['Country']=='AU')
```

```
Out[10]:   Customer  gender Country   ID   age   item
         2  Yi-lung  Female      US  346  23.0  234.0
         3   Duncan    Male      US  121   NaN   23.0
         4    Kevin    Male      AU  223  31.0   85.0
         5    Angel  Female      AU  432  11.0    5.0
         6     Eddy     NaN     NaN  250  12.0    NaN
```

```
In [11]: # Treat missing data
         # Fill missing values
         cusdet.fillna(0)
```

```
Out[11]:   Customer  gender Country   ID   age   item
         0     Gary    Male      AU  342  25.0   35.0
         1     Anny  Female      US  135  45.0   45.0
         2  Yi-lung  Female      US  346  23.0  234.0
         3   Duncan    Male      US  121   0.0   23.0
         4    Kevin    Male      AU  223  31.0   85.0
         5    Angel  Female      AU  432  11.0    5.0
         6     Eddy       0       0  250  12.0    0.0
```

```
In [12]: # Deleting values
         cusdet.dropna()
```

```
Out[12]:   Customer  gender Country   ID   age   item
         0     Gary    Male      AU  342  25.0   35.0
         1     Anny  Female      US  135  45.0   45.0
         2  Yi-lung  Female      US  346  23.0  234.0
         4    Kevin    Male      AU  223  31.0   85.0
         5    Angel  Female      AU  432  11.0    5.0
```

```
In [ ]: # Quiz: guess what these does
        cusdet.dropna(subset = ['age'])
        cusdet.fillna(value={'Country':'Missing'})
        cusdet.fillna(value={'Country':'Missing','age':0})
```

```
In [13]: # Data Sorting
         cusdet.sort_values(by='age',ascending=0)
         # inplace = True option will overwrite data
```

```
Out[13]:   Customer  gender Country   ID   age   item
         1     Anny  Female      US  135  45.0   45.0
         4    Kevin    Male      AU  223  31.0   85.0
         0     Gary    Male      AU  342  25.0   35.0
         2  Yi-lung  Female      US  346  23.0  234.0
         6     Eddy     NaN     NaN  250  12.0    NaN
         5    Angel  Female      AU  432  11.0    5.0
         3   Duncan    Male      US  121   NaN   23.0
```

```
In [14]: # Data Grouping
         cusdet[['age','Country']].groupby('Country').mean()
         # Will cusdet['age'].groupby('Country').mean() work?
```

```
Out[14]:                age
         Country
         AU       22.333333
         US       34.000000
```

```
In [ ]:
```

```
In [ ]: ## Pivot table
```

```
In [15]: # Tabulate of data
         d1=pd.pivot_table(cusdet,values='item',index='Country',columns='gender')
         d1
         # What does these number present?
```

```
Out[15]: gender   Female  Male
         Country
         AU          5.0  60.0
         US        139.5  23.0
```

```
In [16]: # Option aggfunc=sum gives the sum
         pd.pivot_table(cusdet,values='item',index='Country',columns='gender',aggfunc=sum)
```

```
Out[16]: gender   Female   Male
         Country
         AU          5.0  120.0
         US        279.0   23.0
```

```
In [18]: # ix() extracts element of the table
         d1.ix[['AU'],['Female']]
```

```
Out[18]: gender   Female
         Country
         AU          5.0
```

```
In [ ]: # Guess what these table look like
        pd.pivot_table(cusdet,values='age',index=['Country','Customer'])
        # why would not this work?
        pd.pivot_table(cusdet,values='Country',index=['Customer'])
```

```
In [ ]:

In [ ]: ## Ranking

In [20]: # Ranking
         cusdet=pd.read_csv('/Users/xiyongzhang/documents/MQ/RA_ACST890_notes/w10_example_pandas
         r1=cusdet.rank(ascending = False)
         r1
         # it produces rank for every column

Out[20]:           gender  Country   ID  age  item
         Customer
         Gary         2.0      5.0  3.0  3.0   4.0
         Anny         5.0      2.0  5.0  1.0   3.0
         Yi-lung      5.0      2.0  2.0  4.0   1.0
         Duncan       2.0      2.0  6.0  NaN   5.0
         Kevin        2.0      5.0  4.0  2.0   2.0
         Angel        5.0      5.0  1.0  5.0   6.0

In [21]: # to look at age only
         r1['age'].sort_values()

Out[21]: Customer
         Anny       1.0
         Kevin      2.0
         Gary       3.0
         Yi-lung    4.0
         Angel      5.0
         Duncan     NaN
         Name: age, dtype: float64
```