String Manupulation and Aggregation

September 3, 2019

String Manupulation and Aggregation

[2]:
```
#read from tsv file
import pandas as pd
employee=pd.read_table("/home/sakil/Desktop/DataScience/Udemy/Module2/
 ↪tab_seperated_values.tsv")
employee.head()
```

```
/home/sakil/anaconda/lib/python3.7/site-packages/ipykernel_launcher.py:3:
FutureWarning: read_table is deprecated, use read_csv instead, passing sep='\t'.
  This is separate from the ipykernel package so we can avoid doing imports
until
```

[2]:
```
              Name                     Position         Office  Age  \
0      Airi Satou                    Accountant          Tokyo   33
1  Angelica Ramos  Chief Executive Officer (CEO)         London   47
2      Ashton Cox        Junior Technical Author  San Francisco   66
3   Bradley Greer              Software Engineer         London   41
4  Brenden Wagner              Software Engineer  San Francisco   28

   Start date      Salary
0  2008/11/28    $162,700
1  2009/10/09  $1,200,000
2  2009/01/12     $86,000
3  2012/10/13    $132,000
4  2011/06/07    $206,850
```

[3]:
```
#Calculating mean
employee.mean()
```

[3]:
```
Age    42.736842
dtype: float64
```

[7]:
```
#String Manipulation
employee.head()
employee.Name.str.upper().head()
```

[7]:
```
0        AIRI SATOU
1    ANGELICA RAMOS
```

```
2        ASHTON COX
3     BRADLEY GREER
4    BRENDEN WAGNER
Name: Name, dtype: object
```

[8]:

```
#String Manipulation To LowerCase
employee.Name.str.lower().head()
```

[8]:

```
0        airi satou
1    angelica ramos
2        ashton cox
3     bradley greer
4    brenden wagner
Name: Name, dtype: object
```

[10]:

```
#contains keyword
employee.Position.str.contains("Software").head()
```

[10]:

```
0    False
1    False
2    False
3     True
4     True
Name: Position, dtype: bool
```

[11]:

```
#returns row having word Software
employee[employee.Position.str.contains("Software")]
```

[11]:

```
              Name           Position         Office  Age  Start date  \
3    Bradley Greer  Software Engineer         London   41  2012/10/13
4   Brenden Wagner  Software Engineer  San Francisco   28  2011/06/07
6       Bruno Nash  Software Engineer         London   38  2011/05/03
46     Sonya Frost  Software Engineer      Edinburgh   23  2008/12/13
55   Zenaida Frank  Software Engineer       New York   63  2010/01/04
56  Zorita Serrano  Software Engineer  San Francisco   56  2012/06/01

      Salary
3   $132,000
4   $206,850
6   $163,500
46  $103,600
55  $125,250
56  $115,000
```

[12]:

```
#replace keyword
employee.Position.str.replace("Engineer","Developer").head()
```

[12]:

```
0                        Accountant
1    Chief Executive Officer (CEO)
2          Junior Technical Author
3               Software Developer
```

```
4               Software Developer
Name: Position, dtype: object
```

[13]:

```
#using aggregations
employee.Age.min()
```

[13]:

```
19
```

[14]:

```
#using max
employee.Age.max()
```

[14]:

```
66
```

[16]:

```
#using groupby
employee.groupby("Position").Age.min().head()
```

[16]:

```
Position
Accountant                         33
Chief Executive Officer (CEO)      47
Chief Financial Officer (CFO)      64
Chief Marketing Officer (CMO)      40
Chief Operating Officer (COO)      48
Name: Age, dtype: int64
```

[17]:

```
employee.groupby("Position").Age.agg(['count','min','max']).head()
```

[17]:

```
                               count  min  max
Position
Accountant                         2   33   63
Chief Executive Officer (CEO)      1   47   47
Chief Financial Officer (CFO)      1   64   64
Chief Marketing Officer (CMO)      1   40   40
Chief Operating Officer (COO)      1   48   48
```

Learn LOC and DROPNA

[18]:

```
#LOC(choose a number of rows or colums)
#rows 0,all columns
employee.loc[0,:]
```

[18]:

```
Name           Airi Satou
Position       Accountant
Office              Tokyo
Age                    33
Start date     2008/11/28
Salary           $162,700
Name: 0, dtype: object
```

[19]:

```
#rows 0-2,all columns
employee.loc[0:2,:]
```

[19]:

```
             Name                     Position         Office  Age  \
0      Airi Satou                   Accountant          Tokyo   33
1  Angelica Ramos  Chief Executive Officer (CEO)        London   47
```

```
2      Ashton Cox        Junior Technical Author  San Francisco   66

   Start date      Salary
0  2008/11/28    $162,700
1  2009/10/09  $1,200,000
2  2009/01/12     $86,000
```

[22]:
```
#row 0-2,columns 0-2
employee.loc[0:2,["Name","Position"]]
```

[22]:
```
             Name                     Position
0      Airi Satou                   Accountant
1  Angelica Ramos  Chief Executive Officer (CEO)
2      Ashton Cox        Junior Technical Author
```

[25]:
```
#rows 0-2,columns 0-3
employee.loc[0:2,'Name':'Office']
```

[25]:
```
             Name                     Position         Office
0      Airi Satou                   Accountant          Tokyo
1  Angelica Ramos  Chief Executive Officer (CEO)         London
2      Ashton Cox        Junior Technical Author  San Francisco
```

[26]:
```
#rows with certain conditions
employee.loc[employee.Position=="Accountant",:]
```

[26]:
```
               Name    Position Office  Age  Start date    Salary
0        Airi Satou  Accountant  Tokyo   33  2008/11/28  $162,700
17  Garrett Winters  Accountant  Tokyo   63  2011/07/25  $170,750
```

[28]:
```
#use of dropna
employee=pd.read_table("/home/sakil/Desktop/DataScience/Udemy/Module2/
 ↪tab_seperated_values_missing.tsv")
employee.shape
```

```
/home/sakil/anaconda/lib/python3.7/site-packages/ipykernel_launcher.py:2:
FutureWarning: read_table is deprecated, use read_csv instead, passing sep='\t'.
```

[28]:
```
(57, 6)
```

[29]:
```
employee.dropna(how="any").shape
```

[29]:
```
(53, 6)
```

[34]:
```
#drop all colums with missing values
employee.dropna(subset=["Name","Salary"],how="any").shape
```

[34]:
```
(53, 6)
```