Lab-Exercise

November 19, 2019

1 Reza Dibaj

[1]: ```python
import pandas as pd
```

[2]: ```python
myDF=pd.read_csv('test.csv')
```

[3]: ```python
myDF.head(3)
```

[3]:
|   | ID | Range | Value | Type | Zip Code | Model | Cost | Continent |
|---|----|-------|-------|------|----------|-------|------|-----------|
| 0 | 10 | 192 | 50 | F | NXEYEM | EX85 | [56$] | Asia |
| 1 | 20 | 299 | 21 | M | VNFKZD | CT41 | [18$] | Asia |
| 2 | 30 | 378 | 28 | X | IMGMFY | VN81 | [90$] | North America |

[4]: ```python
myDF.drop('Model', axis=1)
```

[4]:
	ID	Range	Value	Type	Zip Code	Cost	Continent
0	10	192	50	F	NXEYEM	[56$]	Asia
1	20	299	21	M	VNFKZD	[18$]	Asia
2	30	378	28	X	IMGMFY	[90$]	North America
3	40	156	44	Q	RHRHJY	[89$]	Asia
4	50	468	33	J	HPQCMC	[95$]	Africa
..	...	...	...	...	...	...	...
945	9460	136	12	C	HGVBPV	[51$]	South America
946	9470	316	46	L	SHBAUD	[33$]	North America
947	9480	461	46	K	PUZUVO	[12$]	Europe
948	9490	185	13	N	UMESZO	[39$]	North America
949	9500	137	10	W	JOHHVL	[10$]	South America

[950 rows x 7 columns]

[5]: ```python
myDF = myDF.drop('Model', axis=1)
```

[6]: ```python
myDF.head()
```

[6]:
	ID	Range	Value	Type	Zip Code	Cost	Continent
0	10	192	50	F	NXEYEM	[56$]	Asia
1	20	299	21	M	VNFKZD	[18$]	Asia
2	30	378	28	X	IMGMFY	[90$]	North America
3	40	156	44	Q	RHRHJY	[89$]	Asia
4	50	468	33	J	HPQCMC	[95$]	Africa

```
[7]: myDF.drop([1, 3, 5], axis=0, inplace=True)
```

```
[8]: myDF.head(4)
```

```
[8]:    ID  Range  Value Type Zip Code   Cost       Continent
    0  10    192     50    F   NXEYEM  [56$]            Asia
    2  30    378     28    X   IMGMFY  [90$]   North America
    4  50    468     33    J   HPQCMC  [95$]          Africa
    6  70    327     24    F   NLVQPY  [25$]          Europe
```

```
[9]: myDF['Range'].mean()
```

```
[9]: 297.53326293558604
```

```
[10]: myDF['Type'].str.contains('F').sum()
```

```
[10]: 50
```

```
[11]: myDF['Type']=myDF['Type'].str.replace('F','AAA')
```

```
[12]: myDF.head()
```

```
[12]:    ID  Range  Value Type Zip Code   Cost       Continent
    0  10    192     50  AAA   NXEYEM  [56$]            Asia
    2  30    378     28    X   IMGMFY  [90$]   North America
    4  50    468     33    J   HPQCMC  [95$]          Africa
    6  70    327     24  AAA   NLVQPY  [25$]          Europe
    7  80    495     50    Z   DLUMAX  [70$]       Antarctica
```

```
[13]: myDF['Cost']=myDF['Cost'].str.replace('[\[\]\$]','')
```

```
[14]: myDF.head()
```

```
[14]:    ID  Range  Value Type Zip Code Cost       Continent
    0  10    192     50  AAA   NXEYEM   56            Asia
    2  30    378     28    X   IMGMFY   90   North America
    4  50    468     33    J   HPQCMC   95          Africa
    6  70    327     24  AAA   NLVQPY   25          Europe
    7  80    495     50    Z   DLUMAX   70       Antarctica
```

```
[15]: myDF.dtypes
```

```
[15]: ID            int64
    Range         int64
    Value         int64
    Type         object
    Zip Code     object
    Cost         object
    Continent    object
    dtype: object
```

```
[16]: myDF['Cost']=myDF['Cost'].astype(int)
```

```
[17]: myDF.dtypes
```

```
[17]: ID           int64
      Range        int64
      Value        int64
      Type         object
      Zip Code     object
      Cost         int32
      Continent    object
      dtype: object
```

```
[18]: myDF['Cost'].mean()
```

```
[18]: 54.586061246040124
```

```
[19]: myDF.groupby('Continent').Cost.mean()
```

```
[19]: Continent
      Africa           54.741259
      Antarctica       54.706897
      Asia             54.279070
      Australia        56.838028
      Europe           53.258993
      North America    55.006993
      South America    53.162963
      Name: Cost, dtype: float64
```

```
[20]: myDF.groupby('Continent').Cost.agg(['min','max', 'std'])
```

```
[20]:                min  max        std
      Continent
      Africa          11   97  25.961432
      Antarctica      10   97  24.880590
      Asia            10   99  26.109737
      Australia       10   97  26.057121
      Europe          10   99  26.185054
      North America   10   99  25.664821
      South America   10   98  25.724780
```

```
[21]: myDF.groupby('Continent').agg(['min','max', 'std'])
```

```
[21]:                 ID                       Range                    Value      \
                min   max          std    min  max          std    min max
      Continent
      Africa         50  9450  2614.387519   100  494  118.110406    10  50
      Antarctica     80  9370  2936.313786   106  495  114.401370    10  50
      Asia           10  9390  2683.431665   103  499  112.047425    10  50
      Australia     120  9410  2793.945560   104  496  116.374530    10  50
      Europe         70  9480  2623.991892   101  500  116.808685    10  50
      North America  30  9490  2841.216413   104  500  113.611333    11  50
      South America  90  9500  2656.186804   104  499  117.375623    10  50


                     Cost
```
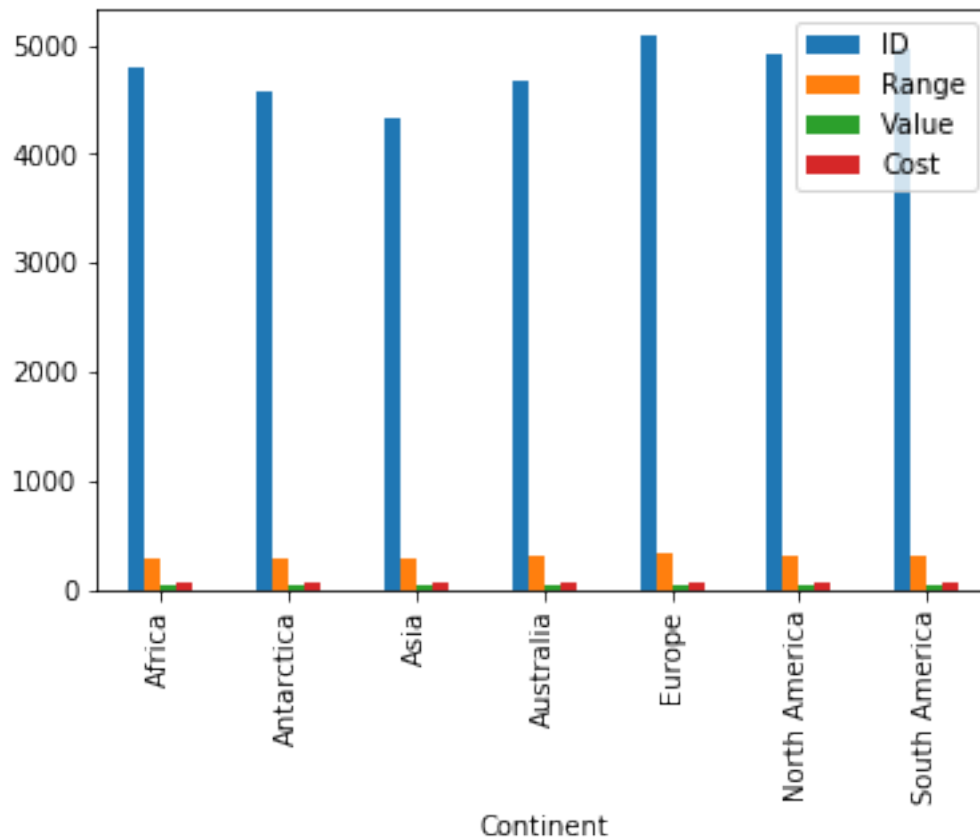
	std	min	max	std
Continent				
Africa	12.498931	11	97	25.961432
Antarctica	11.978316	10	97	24.880590
Asia	11.752819	10	99	26.109737
Australia	13.138226	10	97	26.057121
Europe	11.999544	10	99	26.185054
North America	12.319601	10	99	25.664821
South America	11.762450	10	98	25.724780

[22]:
```python
%matplotlib inline
```

[23]:
```python
myDF.groupby('Continent').mean().plot(kind='bar')
```

[23]: <matplotlib.axes._subplots.AxesSubplot at 0x202508a3d30>

[24]:
```python
myDF.Continent.value_counts()
```

[24]:
```
North America    143
Africa           143
Australia        142
Europe           139
```

```
South America      135
Asia               129
Antarctica         116
Name: Continent, dtype: int64
```

[25]: `myDF.Continent.value_counts(normalize=True)`

```
[25]: North America    0.151003
      Africa           0.151003
      Australia        0.149947
      Europe           0.146779
      South America    0.142555
      Asia             0.136220
      Antarctica       0.122492
      Name: Continent, dtype: float64
```

[26]: `myDF.Continent.unique()`

```
[26]: array(['Asia', 'North America', 'Africa', 'Europe', 'Antarctica',
             'South America', 'Australia'], dtype=object)
```

[27]: `myDF.Continent.nunique()`

[27]: 7

[28]: `myDF.Cost.value_counts()`

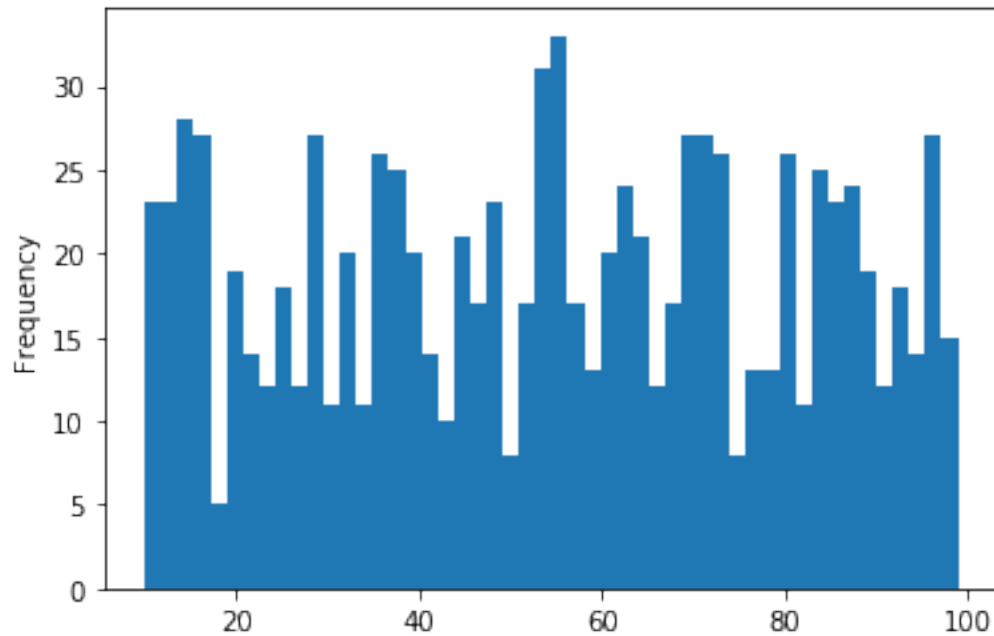```
[28]: 35   20
      56   20
      73   18
      54   17
      97   17
            ..
      77   5
      92   5
      94   5
      42   3
      24   2
      Name: Cost, Length: 90, dtype: int64
```

[29]: `%matplotlib inline`

[30]: `myDF.Cost.plot(kind='hist', bins=50)`

[30]: `<matplotlib.axes._subplots.AxesSubplot at 0x20250ca3a20>`

[31]: `myDF.Type.isnull().sum()`

[31]: 2

[32]: `myDF['Zip Code'].isnull().sum()`

[32]: 9

[33]: `myDF.isnull().sum()`

[33]:
```
ID   0
Range   0
Value   0
Type   2
Zip Code   9
Cost   0
Continent   0
dtype: int64
```

[34]: `myDF.dropna(how='any').shape`

[34]: (936, 7)

[35]: `myDF.shape`

[35]: (947, 7)

[36]: `myDF['Type'].value_counts(dropna=False)`

[36]:
```
AAA   50
N   47
Z   44
```

```
Q     44
C     43
V     41
K     41
A     40
H     39
Y     39
G     38
X     37
M     37
O     36
U     36
L     35
J     35
S     34
I     33
W     32
D     30
B     29
P     28
T     28
R     25
E     24
NaN    2
Name: Type, dtype: int64
```

[37]: `myDF['Type'].fillna(value='XXX', inplace=True)`

[38]: `myDF['Type'].value_counts(dropna=False)`

[38]:
```
AAA   50
N     47
Z     44
Q     44
C     43
V     41
K     41
A     40
H     39
Y     39
G     38
M     37
X     37
O     36
U     36
L     35
J     35
S     34
```

```
I       33
W       32
D       30
B       29
T       28
P       28
R       25
E       24
XXX      2
Name: Type, dtype: int64
```

[39]: `myDF.dropna(how='any', inplace=True)`

[40]: `myDF.shape`

[40]: (938, 7)

[41]: `myDF.set_index('ID', inplace=True)`

[42]: `myDF.index.name=None`

[43]: `myDF.head()`

[43]:
```
    Range  Value Type Zip Code  Cost       Continent
10    192     50  AAA   NXEYEM    56            Asia
30    378     28    X   IMGMFY    90   North America
50    468     33    J   HPQCMC    95          Africa
70    327     24  AAA   NLVQPY    25          Europe
80    495     50    Z   DLUMAX    70       Antarctica
```

[44]: `myDF.loc[70,'Value']`

[44]: 24

[45]: `myDF.index`

[45]:
```
Int64Index([  10,    30,    50,    70,    80,    90,   100,   110,   120,   130,
            ...
            9410, 9420, 9430, 9440, 9450, 9460, 9470, 9480, 9490, 9500],
           dtype='int64', length=938)
```

[46]: `myDF.columns`

[46]:
```
Index(['Range', 'Value', 'Type', 'Zip Code', 'Cost', 'Continent'],
      dtype='object')
```

[47]: `myDF.shape`

[47]: (938, 6)

[48]: `myDF.index.name='ID'`

[49]: `myDF.reset_index(inplace=True)`

[50]: `myDF.head()`

```
[50]:     ID  Range  Value Type Zip Code  Cost       Continent
      0   10    192     50  AAA   NXEYEM    56            Asia
      1   30    378     28    X   IMGMFY    90   North America
      2   50    468     33    J   HPQCMC    95          Africa
      3   70    327     24  AAA   NLVQPY    25          Europe
      4   80    495     50    Z   DLUMAX    70      Antarctica
```

```
[51]: myDF.index
```

```
[51]: RangeIndex(start=0, stop=938, step=1)
```

```
[52]: myDF.columns
```

```
[52]: Index(['ID', 'Range', 'Value', 'Type', 'Zip Code', 'Cost', 'Continent'],
      dtype='object')
```

```
[53]: myDF.shape
```

```
[53]: (938, 7)
```

```
[54]: myDF.describe()
```

```
[54]:                 ID        Range        Value         Cost
      count   938.000000   938.000000   938.000000   938.000000
      mean   4749.104478   297.531983    29.917910    54.611940
      std    2739.076928   115.913345    12.292447    25.742149
      min      10.000000   100.000000    10.000000    10.000000
      25%    2382.500000   194.000000    19.000000    33.000000
      50%    4725.000000   295.000000    30.000000    55.000000
      75%    7097.500000   400.000000    41.000000    76.000000
      max    9500.000000   500.000000    50.000000    99.000000
```

```
[55]: myDF.describe().loc['count','Cost']
```

```
[55]: 938.0
```

```
[56]: myDF.Continent.value_counts().sort_values()
```

```
[56]: Antarctica       115
      Asia             129
      South America    135
      Europe           138
      Africa           139
      North America    141
      Australia        141
      Name: Continent, dtype: int64
```

```
[57]: myDF.Continent.value_counts().sort_index()
```

```
[57]: Africa           139
      Antarctica       115
      Asia             129
      Australia        141
      Europe           138
```

```
North America     141
South America     135
Name: Continent, dtype: int64
```

[58]: `#quantity = pd.Series([3000000, 85000], index=['AAA', 'XXX'], name='amount')`

[59]: `myDF.loc[[1, 3, 5, 7], :]`

[59]:
	ID	Range	Value	Type	Zip Code	Cost	Continent
1	30	378	28	X	IMGMFY	90	North America
3	70	327	24	AAA	NLVQPY	25	Europe
5	90	115	30	S	KYJUHA	17	South America
7	110	301	29	A	OVINZU	69	South America

[60]: `myDF.loc[0:6, ['Type','Continent']]`

[60]:
	Type	Continent
0	AAA	Asia
1	X	North America
2	J	Africa
3	AAA	Europe
4	Z	Antarctica
5	S	South America
6	K	South America

[61]: `myDF[myDF.Type=='XXX'].Value`

[61]:
```
687    42
743    37
Name: Value, dtype: int64
```

[ ]: