HW 5

February 18, 2022

HW 5

Jinglei (Lesly) Liu

Q1

[ ]:
```
# 1. Find list of distinct (that is do not repeat same name) directors in DB -movies/Neo4j Sandbox
match (p:Person) - [:DIRECTED] -> (m:Movie) return distinct p
```

[ ]:
```
# 2. Find list of distinct (that is do not repeat same name) actors in DB -movies/Neo4j Sandbox
match (p:Person) - [:ACTED_IN] -> (m:Movie) return distinct p
```

Q2

[1]:
```
import pymongo
from pymongo import MongoClient
import dns
import pprint
```

[2]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

[3]:
```
client = pymongo.MongoClient("mongodb+srv://lesly3:Ljl720115@cluster0.bfjz0.mongodb.net/sample_supplies?retryWrites=true&w=majority")
```

[4]:
```
db = client['sample_supplies']
```

[9]:
```
collection = db['sales']
```

[10]:
```
pipeline = [
    {
        "$bucket": {
            "groupBy": "$customer.age",
            "boundaries": [15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75,80],
            "output": {
```

```
            "count": { "$sum": 1 }
            }
        }
    }
]
```

[11]:
```
results = collection.aggregate(pipeline)
```

[12]:
```
data = pd.DataFrame(results)
data
```

[12]:

	_id	count
0	15	168
1	20	233
2	25	407
3	30	498
4	35	650
5	40	612
6	45	586
7	50	474
8	55	440
9	60	412
10	65	257
11	70	226
12	75	37

[13]:
```
l = [str(data['_id'][i])+' - '+str(data['_id'][i]+5) for i in list(data.index)]
```

[14]:
```
data.insert(1,"interval",l)
data
```

[14]:

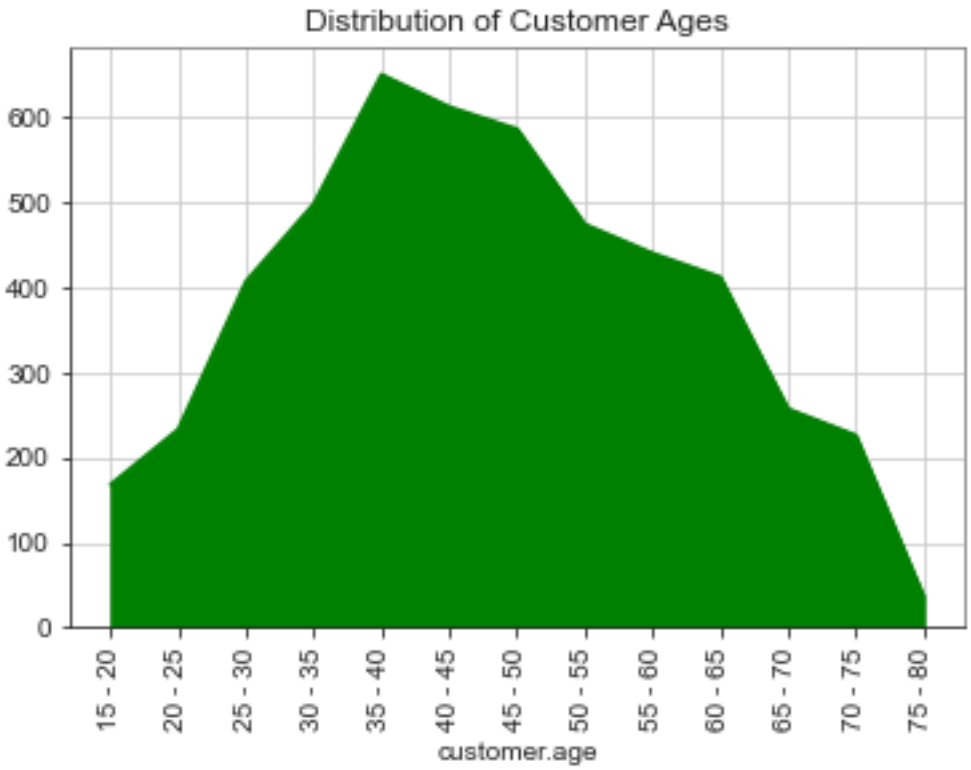
	_id	interval	count
0	15	15 - 20	168
1	20	20 - 25	233
2	25	25 - 30	407
3	30	30 - 35	498
4	35	35 - 40	650
5	40	40 - 45	612
6	45	45 - 50	586
7	50	50 - 55	474
8	55	55 - 60	440
9	60	60 - 65	412
10	65	65 - 70	257
11	70	70 - 75	226
12	75	75 - 80	37

[15]:
```
l=[]
for i in list(data['interval']):
    l.append(str(i))
l
```

[15]:
```
['15 - 20',
 '20 - 25',
 '25 - 30',
 '30 - 35',
 '35 - 40',
 '40 - 45',
 '45 - 50',
 '50 - 55',
 '55 - 60',
 '60 - 65',
 '65 - 70',
 '70 - 75',
 '75 - 80']
```

[16]:
```
sns.set_style('ticks')
```

[17]:
```
data.plot.area(x="interval", y = 'count', title='Distribution of Customer␣
 ↪Ages', legend=None, color = 'green')
plt.xticks(range(0,13), l, rotation = 90)
plt.xlabel('customer.age')
plt.grid()
plt.gca( ).spines['top'].set_color('grey')
plt.gca( ).spines['right'].set_color('grey')
plt.show()
```


Distribution of Customer Ages
600
500
400
300
200
100
0
15 - 20
20 - 25
25 - 30
30 - 35
35 - 40
40 - 45
45 - 50
50 - 55
55 - 60
60 - 65
65 - 70
70 - 75
75 - 80
customer.age