Hw-5

February 17, 2022

[46]:
```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
from matplotlib import colors
```

[1]:
```
from platform import python_version

print(python_version())
```

```
3.8.8
```

[2]:
```
import pymongo
```

[3]:
```
from pymongo import MongoClient
```

[39]:
```
client = pymongo.MongoClient("mongodb+srv://Sadman:Sadman28@cluster0.zxxxh.
↪mongodb.net/sample_supplies?retryWrites=true&w=majority")
db = client.sample_supplies
```

[40]:
```
collection = db.sales
```

[41]:
```
data = pd.DataFrame(list(db.sales.find()))
```

[42]:
```
data.head()
```

[42]:

	_id	saleDate	items	storeLocation
0	5bd761dcae323e45a93ccfee	2014-11-11 02:13:51.893	[{'name': 'laptop', 'tags': ['electronics', 's…	London
1	5bd761dcae323e45a93ccff0	2017-03-21 01:54:26.657	[{'name': 'envelopes', 'tags': ['stationary', …	New York
2	5bd761dcae323e45a93ccff9	2017-11-12 20:30:15.045	[{'name': 'notepad', 'tags': ['office', 'writi…	London
3	5bd761dcae323e45a93cd03a	2017-02-24 19:17:51.731	[{'name': 'backpack', 'tags': ['school', 'trav…	Denver
4	5bd761dcae323e45a93cd05f	2017-06-26 20:57:32.151	[{'name': 'laptop', 'tags': ['electronics', 's…	New York

```
                                                customer  couponUsed  \
0  {'gender': 'F', 'age': 40, 'email': 'pan@cak.z…       False
1  {'gender': 'M', 'age': 26, 'email': 'rapifoozi…        True
2  {'gender': 'F', 'age': 50, 'email': 'velo@nuka…       False
3  {'gender': 'F', 'age': 59, 'email': 'riul@issu…       False
4  {'gender': 'M', 'age': 34, 'email': 'li@efva.g…       False

  purchaseMethod
0       In store
1       In store
2       In store
3       In store
4       In store
```

[44]:
```
df = pd.concat([data.drop(['customer'], axis=1), data['customer'].apply(pd.
↪Series)], axis=1)
df.head()
```

[44]:
```
                        _id                saleDate  \
0  5bd761dcae323e45a93ccfee 2014-11-11 02:13:51.893
1  5bd761dcae323e45a93ccff0 2017-03-21 01:54:26.657
2  5bd761dcae323e45a93ccff9 2017-11-12 20:30:15.045
3  5bd761dcae323e45a93cd03a 2017-02-24 19:17:51.731
4  5bd761dcae323e45a93cd05f 2017-06-26 20:57:32.151

                                               items storeLocation  \
0  [{'name': 'laptop', 'tags': ['electronics', 's…        London
1  [{'name': 'envelopes', 'tags': ['stationary', …      New York
2  [{'name': 'notepad', 'tags': ['office', 'writi…        London
3  [{'name': 'backpack', 'tags': ['school', 'trav…        Denver
4  [{'name': 'laptop', 'tags': ['electronics', 's…      New York

   couponUsed purchaseMethod gender  age                email  satisfaction
0       False       In store      F   40           pan@cak.zm             5
1        True       In store      M   26  rapifoozi@viupoen.bb             5
2       False       In store      F   50         velo@nukav.fr             5
3       False       In store      F   59        riul@issuiw.bq             5
4       False       In store      M   34            li@efva.gm             5
```

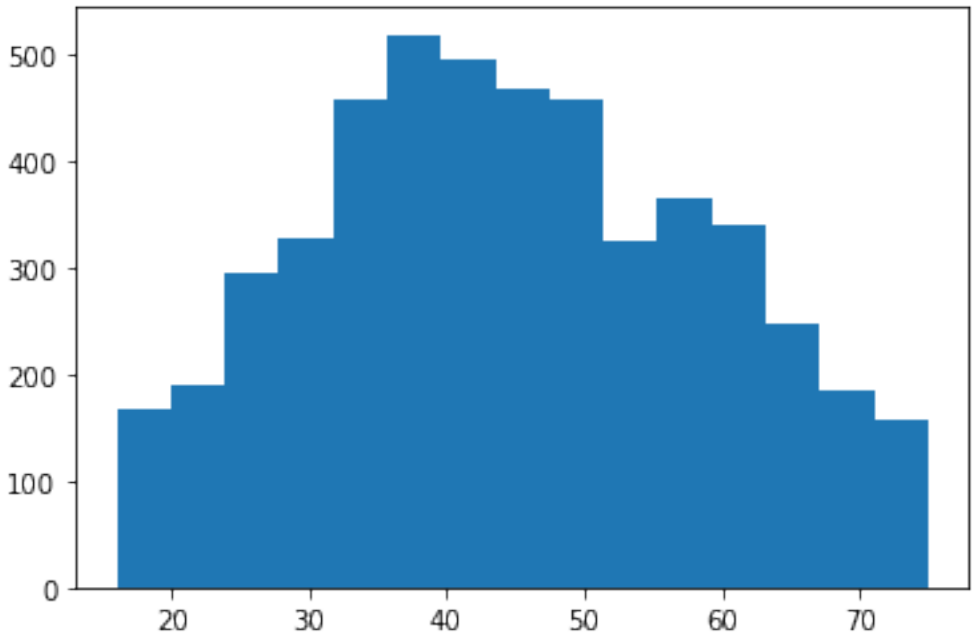
[64]:
```
df[['_id','age']]
```

[64]:
```
                         _id  age
0    5bd761dcae323e45a93ccfee   40
1    5bd761dcae323e45a93ccff0   26
2    5bd761dcae323e45a93ccff9   50
3    5bd761dcae323e45a93cd03a   59
```

```
4     5bd761dcae323e45a93cd05f   34
…                          …  …
4995  5bd761deae323e45a93ce31b   49
4996  5bd761deae323e45a93ce350   63
4997  5bd761deae323e45a93ce358   51
4998  5bd761deae323e45a93ce35b   19
4999  5bd761deae323e45a93ce35c   35

[5000 rows x 2 columns]
```

[61]:
```
plt.hist(df['age'],bins=15)
plt.show()
```



[65]:
```
plot = df[['_id','age']]
plot.head()
```

[65]:
```
                        _id  age
0  5bd761dcae323e45a93ccfee   40
1  5bd761dcae323e45a93ccff0   26
2  5bd761dcae323e45a93ccff9   50
3  5bd761dcae323e45a93cd03a   59
4  5bd761dcae323e45a93cd05f   34
```

[79]:
```
plot['bin'] = pd.cut(x=df['age'],␣
 ↪bins=[15,20,25,30,35,40,45,50,55,60,65,70,75,80])
```

```
plot
```

```
<ipython-input-79-d2e251b5fc3d>:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-
docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  plot['bin'] = pd.cut(x=df['age'],
bins=[15,20,25,30,35,40,45,50,55,60,65,70,75,80])
```

[79]:

	_id	age	bin
0	5bd761dcae323e45a93ccfee	40	(35, 40]
1	5bd761dcae323e45a93ccff0	26	(25, 30]
2	5bd761dcae323e45a93ccff9	50	(45, 50]
3	5bd761dcae323e45a93cd03a	59	(55, 60]
4	5bd761dcae323e45a93cd05f	34	(30, 35]
…	…	…	…
4995	5bd761deae323e45a93ce31b	49	(45, 50]
4996	5bd761deae323e45a93ce350	63	(60, 65]
4997	5bd761deae323e45a93ce358	51	(50, 55]
4998	5bd761deae323e45a93ce35b	19	(15, 20]
4999	5bd761deae323e45a93ce35c	35	(30, 35]

[5000 rows x 3 columns]

[80]:
```
df1 = plot.groupby("bin").count()
df1['bin'] = df1.index
df1.reset_index(drop=True, inplace=True)
df1
```

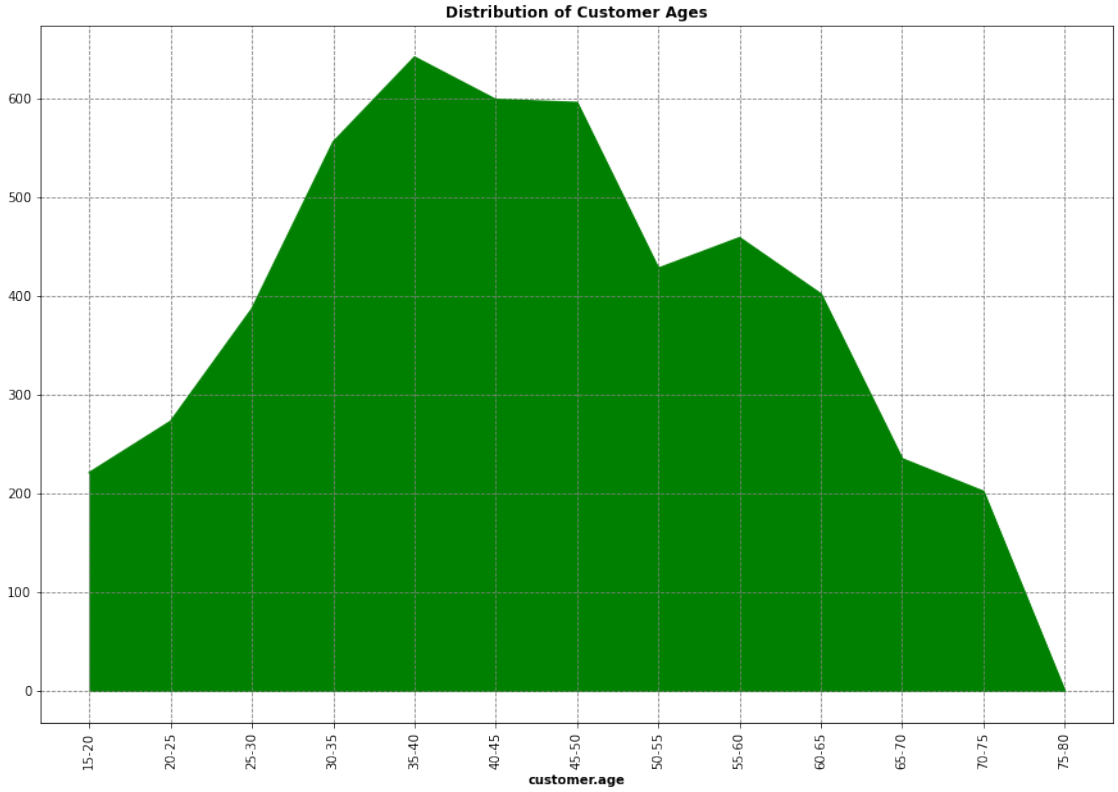
[80]:

	_id	age	bin
0	221	221	(15, 20]
1	273	273	(20, 25]
2	387	387	(25, 30]
3	556	556	(30, 35]
4	642	642	(35, 40]
5	599	599	(40, 45]
6	596	596	(45, 50]
7	428	428	(50, 55]
8	459	459	(55, 60]
9	402	402	(60, 65]
10	235	235	(65, 70]
11	202	202	(70, 75]
12	0	0	(75, 80]

[84]:
```
x = ['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']
```

```
y = df1['age']
plt.figure(figsize=(15,10))
plt.fill_between(x, y, color ='green')
plt.xticks(rotation=90)
# ax.set_axisbelow(True)
plt.grid(color ='gray', linestyle='--')
plt.plot(x,y, color ='green')
plt.title("Distribution of Customer Ages", fontweight="bold")
plt. xlabel("customer.age", fontweight="bold")
```

[84]: Text(0.5, 0, 'customer.age')



[ ]:

[ ]:

[ ]:

[ ]: