Lab4. Pandas Grouping and Aggregation

Suriya (225229140)

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

In [2]:
```
data=pd.read_csv('thanksgiving-2015-poll-data.csv',encoding='Latin-1')
data
```

Out[2]:

	RespondentID	Do you celebrate Thanksgiving?	What is typically the main dish at your Thanksgiving dinner?	What is typically the main dish at your Thanksgiving dinner? - Other (please specify)	How is the main dish typically cooked?	How is the main dish typically cooked? - Other (please specify)	What kind of stuffing/dressing do you typically have?	What kind of stuffing/dressing do you typically have? - Other (please specify)	What type of cranberry saucedo you typically have?	What type of cranberry saucedo you typically have? - Other (please specify)	...	Have you ever tried to meet up with hometown friends on Thanksgiving night?
0	4337954960	Yes	Turkey	NaN	Baked	NaN	Bread-based	NaN	None	NaN	...	Yes
1	4337951949	Yes	Turkey	NaN	Baked	NaN	Bread-based	NaN	Other (please specify)	Homemade cranberry gelatin ring	...	No
2	4337935621	Yes	Turkey	NaN	Roasted	NaN	Rice-based	NaN	Homemade	NaN	...	Yes
3	4337933040	Yes	Turkey	NaN	Baked	NaN	Bread-based	NaN	Homemade	NaN	...	Yes
4	4337931983	Yes	Tofurkey	NaN	Baked	NaN	Bread-based	NaN	Canned	NaN	...	Yes
...	...	...	...	...	...	...	...	...	...	...	...	...
1053	4335944082	Yes	Turkey	NaN	Roasted	NaN	Bread-based	NaN	Homemade	NaN	...	Yes
1054	4335943173	Yes	Turkey	NaN	Baked	NaN	Bread-based	NaN	Canned	NaN	...	No
1055	4335943060	Yes	Other (please specify)	Duck	Baked	NaN	Rice-based	NaN	None	NaN	...	Yes
1056	4335934708	Yes	Turkey	NaN	Baked	NaN	None	NaN	Homemade	NaN	...	Yes
1057	4335894916	Yes	Turkey	NaN	Baked	NaN	Bread-based	NaN	Canned	NaN	...	Yes

058 rows × 65 columns

In [6]:
```
data.head(5)
```

Out[6]:

	RespondentID	Do you celebrate Thanksgiving?	What is typically the main dish at your Thanksgiving dinner?	What is typically the main dish at your Thanksgiving dinner? - Other (please specify)	How is the main dish typically cooked?	How is the main dish typically cooked? - Other (please specify)	What kind of stuffing/dressing do you typically have?	What kind of stuffing/dressing do you typically have? - Other (please specify)	What type of cranberry saucedo you typically have?	What type of cranberry saucedo you typically have? - Other (please specify)	...	Have you ever tried to meet up with hometown friends on Thanksgiving night?
0	4337954960	Yes	Turkey	NaN	Baked	NaN	Bread-based	NaN	None	NaN	...	Yes
1	4337951949	Yes	Turkey	NaN	Baked	NaN	Bread-based	NaN	Other (please specify)	Homemade cranberry gelatin ring	...	No
2	4337935621	Yes	Turkey	NaN	Roasted	NaN	Rice-based	NaN	Homemade	NaN	...	Yes
3	4337933040	Yes	Turkey	NaN	Baked	NaN	Bread-based	NaN	Homemade	NaN	...	Yes
4	4337931983	Yes	Tofurkey	NaN	Baked	NaN	Bread-based	NaN	Canned	NaN	...	Yes

5 rows × 65 columns

In [8]:
```
data.shape
```

Out[8]:
```
(1058, 65)
```

In [11]:
```
data['Do you celebrate Thanksgiving?'].unique()
```

Out[11]:
```
array(['Yes', 'No'], dtype=object)
```

In [12]:
```
data.columns[1:5]
```

Out[12]:
```
Index(['Do you celebrate Thanksgiving?',
       'What is typically the main dish at your Thanksgiving dinner?',
       'What is typically the main dish at your Thanksgiving dinner? - Other (please specify)',
       'How is the main dish typically cooked?'],
      dtype='object')
```

Apply function to Series

In [13]:
```
data["What is your gender?"].value_counts(dropna=False)
```

Out[13]:
```
Female    544
Male      481
NaN        33
Name: What is your gender?, dtype: int64
```

In [14]:
```
import math
def gender_code(gender_string):
    if isinstance(gender_string,float)and math.isnan(gender_string):
        return gender_string
    return int(gender_string=="Female")
```

In [15]:
```
data["gender"] = data["What is your gender?"].apply(gender_code)
data["gender"].value_counts(dropna=False)
```

Out[15]:
```
1.0    544
0.0    481
NaN     33
Name: gender, dtype: int64
```

Applying functions to DataFrames

In [16]:
```
data.apply(lambda x: x.dtype)[0:5]
```

Out[16]:
```
RespondentID                                                                             int64
Do you celebrate Thanksgiving?                                                          object
What is typically the main dish at your Thanksgiving dinner?                            object
What is typically the main dish at your Thanksgiving dinner? - Other (please specify)   object
How is the main dish typically cooked?                                                  object
dtype: object
```

In [35]:
```
data["How much total combined money did all members of your HOUSEHOLD earn last year?"].value_counts(dropna=False)
```

Out[35]:
```
$25,000 to $49,999      180
Prefer not to answer    136
$50,000 to $74,999      135
$75,000 to $99,999      133
$100,000 to $124,999    111
$200,000 and up          80
$10,000 to $24,999       68
$0 to $9,999             66
$125,000 to $149,999     49
$150,000 to $174,999     40
NaN                      33
$175,000 to $199,999     27
Name: How much total combined money did all members of your HOUSEHOLD earn last year?, dtype: int64
```

In [36]:

```
import numpy as np
def clean_income(value):
    if value == "$200,000 and up":
        return 200000
    elif value == "Prefer not to answer":
        return np.nan
    elif isinstance(value, float) and math.isnan(value):
        return np.nan
    value = value.replace("$", "").replace(",","")

    income_high, income_low = value.split(" to ")
    return (int(income_high) + int(income_low)) / 2
```

In [37]:

```
data["income"] = data["How much total combined money did all members of your HOUSEHOLD earn last year?"].apply(clean_income)
data["income"].head()
```

Out[37]:

```
0     87499.5
1     62499.5
2      4999.5
3    200000.0
4    112499.5
Name: income, dtype: float64
```

Grouping Data with Pandas

In [38]:

```
data["What type of cranberry saucedo you typically have?"].value_counts()
```

Out[38]:

```
Canned                    502
Homemade                  301
None                      146
Other (please specify)     25
Name: What type of cranberry saucedo you typically have?, dtype: int64
```

In [39]:

```
homemade = data[data["What type of cranberry saucedo you typically have?"] == "Homemade"]
canned = data[data["What type of cranberry saucedo you typically have?"] == "Canned"]
```

In [40]:

```
print(homemade["income"].mean())
print(canned["income"].mean())
```

```
94878.1072874494
83823.40340909091
```

In [41]:

```
grouped = data.groupby("What type of cranberry saucedo you typically have?")
grouped
```

Out[41]:

```
<pandas.core.groupby.generic.DataFrameGroupBy object at 0x0000019E7F0661C0>
```

In [60]:

```
dict(grouped.groups)
```

Out[60]:

```
{'Canned': Int64Index([   4,    6,    8,   11,   12,   15,   18,   19,   26,   27,
             ...
             1040, 1041, 1042, 1044, 1045, 1046, 1047, 1051, 1054, 1057],
            dtype='int64', length=502),
 'Homemade': Int64Index([   2,    3,    5,    7,   13,   14,   16,   20,   21,   23,
             ...
             1016, 1017, 1025, 1027, 1030, 1034, 1048, 1049, 1053, 1056],
            dtype='int64', length=301),
 'None': Int64Index([   0,   17,   24,   29,   34,   36,   40,   47,   49,   51,
             ...
              980,  981,  997, 1015, 1018, 1031, 1037, 1043, 1050, 1055],
            dtype='int64', length=146),
 'Other (please specify)': Int64Index([   1,    9,  154,  216,  221,  233,  249,  265,  301,  336,  380,
              435,  444,  447,  513,  550,  749,  750,  784,  807,  860,  872,
              905, 1000, 1007],
            dtype='int64')}
```

In [54]:

```
grouped.size()
```

Out[54]:

```
What type of cranberry saucedo you typically have?
Canned                    502
Homemade                  301
None                      146
Other (please specify)     25
dtype: int64
```

In [55]:

```
for name,group in grouped:
    print(name)
    print(group.shape)
    print(type(group))
```

```
Canned
(502, 67)
<class 'pandas.core.frame.DataFrame'>
Homemade
(301, 67)
<class 'pandas.core.frame.DataFrame'>
None
(146, 67)
<class 'pandas.core.frame.DataFrame'>
Other (please specify)
(25, 67)
<class 'pandas.core.frame.DataFrame'>
```

In [56]:

```
grouped["income"]
```

Out[56]:

```
<pandas.core.groupby.generic.SeriesGroupBy object at 0x0000019E7FE81EE0>
```

In [57]:

```
grouped["income"].size()
```

Out[57]:

```
What type of cranberry saucedo you typically have?
Canned                    502
Homemade                  301
None                      146
Other (please specify)     25
Name: income, dtype: int64
```

Aggregating values in groups

In [58]:

```
grouped["income"].agg(np.mean)
```

Out[58]:

```
What type of cranberry saucedo you typically have?
Canned                    83823.403409
Homemade                  94878.107287
None                      78886.084034
Other (please specify)    86629.978261
Name: income, dtype: float64
```

In [61]:

```
grouped.agg(np.mean)
```

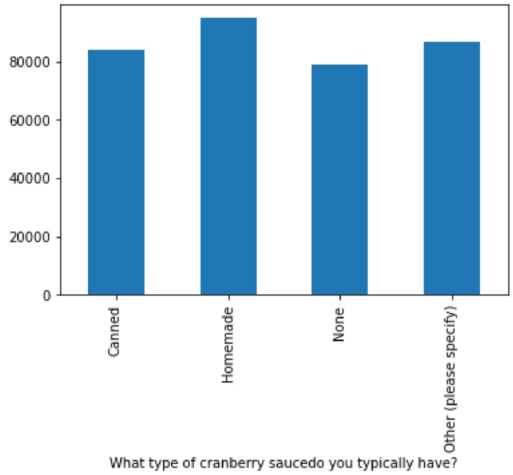
Out[61]:

	RespondentID	gender	income
What type of cranberry saucedo you typically have?			
Canned	4.336699e+09	0.552846	83823.403409
Homemade	4.336792e+09	0.533101	94878.107287
None	4.336765e+09	0.517483	78886.084034
Other (please specify)	4.336763e+09	0.640000	86629.978261

Plotting the results of aggregation

In [62]:
```
sauce = grouped.agg(np.mean)
sauce["income"].plot(kind="bar")
```

Out[62]:
```
<AxesSubplot:xlabel='What type of cranberry saucedo you typically have?'>
```



Aggregating with multiple columns

In [64]:
```
grouped = data.groupby(["What type of cranberry saucedo you typically have?", "What is typically the main dish at your Thanksgivi
grouped.agg(np.mean)
```

Out[64]:

		RespondentID	gender	income
What type of cranberry saucedo you typically have?	**What is typically the main dish at your Thanksgiving dinner?**			
Canned	**Chicken**	4.336354e+09	0.333333	80999.600000
	Ham/Pork	4.336757e+09	0.642857	77499.535714
	I don't know	4.335987e+09	0.000000	4999.500000
	Other (please specify)	4.336682e+09	1.000000	53213.785714
	Roast beef	4.336254e+09	0.571429	25499.500000
	Tofurkey	4.337157e+09	0.714286	100713.857143
	Turkey	4.336705e+09	0.544444	85242.682045
Homemade	**Chicken**	4.336540e+09	0.750000	19999.500000
	Ham/Pork	4.337253e+09	0.250000	96874.625000
	I don't know	4.336084e+09	1.000000	NaN
	Other (please specify)	4.336863e+09	0.600000	55356.642857
	Roast beef	4.336174e+09	0.000000	33749.500000
	Tofurkey	4.336790e+09	0.666667	57916.166667
	Turducken	4.337475e+09	0.500000	200000.000000
	Turkey	4.336791e+09	0.531008	97690.147982
None	**Chicken**	4.336151e+09	0.500000	11249.500000
	Ham/Pork	4.336680e+09	0.444444	61249.500000
	I don't know	4.336412e+09	0.500000	33749.500000
	Other (please specify)	4.336688e+09	0.600000	119106.678571
	Roast beef	4.337424e+09	0.000000	162499.500000
	Tofurkey	4.336950e+09	0.500000	112499.500000
	Turducken	4.336739e+09	0.000000	NaN
	Turkey	4.336784e+09	0.523364	74606.275281
Other (please specify)	**Ham/Pork**	4.336465e+09	1.000000	87499.500000
	Other (please specify)	4.337335e+09	0.000000	124999.666667
	Tofurkey	4.336122e+09	1.000000	37499.500000
	Turkey	4.336724e+09	0.700000	82916.194444

Aggregating with multiple functions

In [65]:

```
grouped["income"].agg([np.mean, np.sum, np.std]).head(10)
```

Out[65]:

		mean	sum	std
What type of cranberry saucedo you typically have?	**What is typically the main dish at your Thanksgiving dinner?**			
Canned	**Chicken**	80999.600000	404998.0	75779.481062
	Ham/Pork	77499.535714	1084993.5	56645.063944
	I don't know	4999.500000	4999.5	NaN
	Other (please specify)	53213.785714	372496.5	29780.946290
	Roast beef	25499.500000	127497.5	24584.039538
	Tofurkey	100713.857143	704997.0	61351.484439
	Turkey	85242.682045	34182315.5	55687.436102
Homemade	**Chicken**	19999.500000	59998.5	16393.596311
	Ham/Pork	96874.625000	387498.5	77308.452805
	I don't know	NaN	0.0	NaN

In [66]:

```
grouped = data.groupby("How would you describe where you live?")["What is typically the main dish at your Thanksgiving dinner?"]
grouped.apply(lambda x:x.value_counts())
```

Out[66]:

```
How would you describe where you live?
Rural                                     Turkey                    189
                                          Other (please specify)      9
                                          Ham/Pork                    7
                                          Tofurkey                    3
                                          I don't know                3
                                          Turducken                   2
                                          Chicken                     2
                                          Roast beef                  1
Suburban                                  Turkey                    449
                                          Ham/Pork                   17
                                          Other (please specify)     13
                                          Tofurkey                    9
                                          Chicken                     3
                                          Roast beef                  3
                                          Turducken                   1
                                          I don't know                1
Urban                                     Turkey                    198
                                          Other (please specify)     13
                                          Tofurkey                    8
                                          Chicken                     7
                                          Roast beef                  6
                                          Ham/Pork                    4
Name: What is typically the main dish at your Thanksgiving dinner?, dtype: int64
```

In [ ]: