Thanksgiving analysis

October 14, 2016

1 Turkey Time

1.0.1 Initial imports

```
In [28]: import csv
         import numpy as np
         import pandas as pd
         import matplotlib.pyplot as plt
         import seaborn as sns
         from __future__ import division
```

1.0.2 Importing csv file into list

```
In [29]: with open('thanksgiving-2015-poll-data.csv', 'rU') as csvfile:
             reader=csv.reader(csvfile)
             rows=[]
             for row in reader:
                 rows.append(row)

         '''for r in rows[0]:
             print r[:50]'''

         for r in rows[0][:2]+rows[0][-4:]:
             print r

         print len(rows)
```

```
RespondentID
Do you celebrate Thanksgiving?
Age
What is your gender?
How much total combined money did all members of your HOUSEHOLD earn last year?
US Region
1059
```

1.0.3 Counting percentage of people who celebrate Thanksgiving

```
In [30]: arr_answers=np.array(rows)[1:]

         def count_in_column(col_number, answer):
             count_answer=0
             for i in range(arr_answers.shape[0]):
                 if arr_answers[i,col_number]==answer:
                     count_answer+=1
             return count_answer

         print 'Percentage Yes:', round(count_in_column(1, 'Yes')/arr_answers.shape
         print 'Percentage No:', round(count_in_column(1, 'No')/arr_answers.shape[0
```

```
Percentage Yes: 92.63
Percentage No: 7.37
```

1.0.4 Check for duplicates in respondent ID column

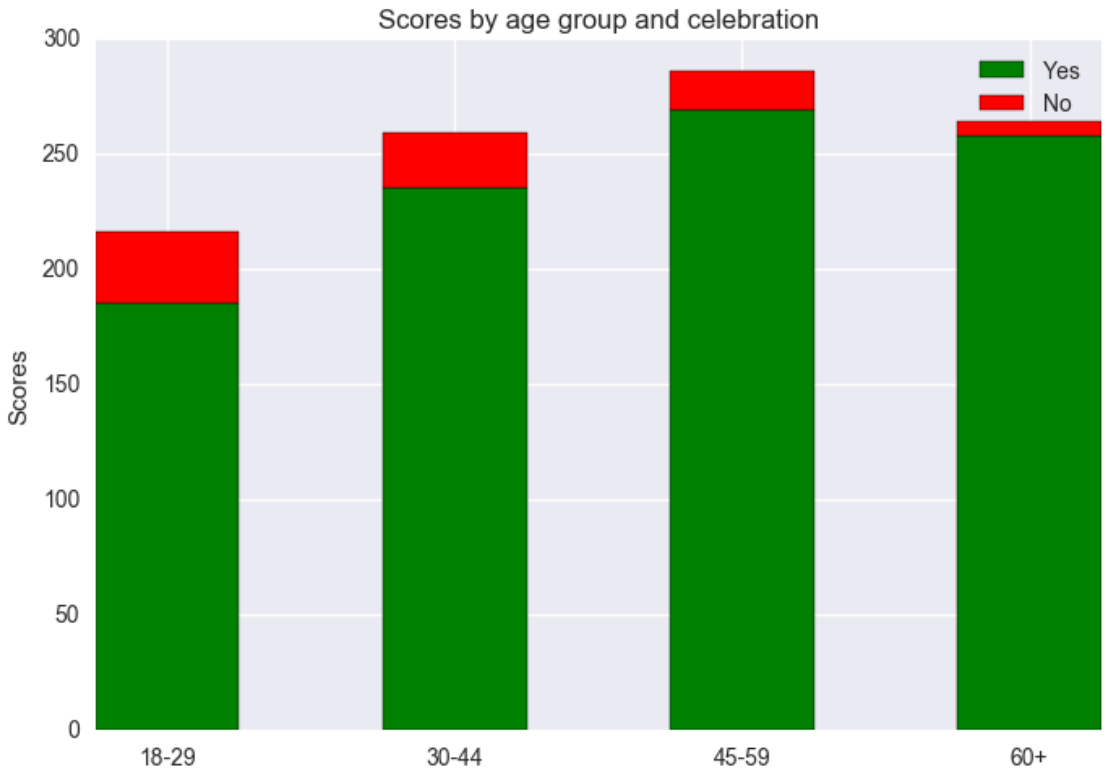
```
In [31]: all_answers_pd=pd.read_csv('thanksgiving-2015-poll-data.csv')
         print 'Number of duplicates:', all_answers_pd['RespondentID'].duplicated()

         answers_pd=all_answers_pd[['RespondentID',
                                    'Do you celebrate Thanksgiving?',
                                    'Age',
                                    'What is your gender?',
                                    'How much total combined money did all members
                                    'US Region']]

         print answers_pd[:10]
```

```
Number of duplicates: 0
   RespondentID Do you celebrate Thanksgiving?      Age What is your gender?  \
0    4337954960                             Yes  18 - 29                 Male
1    4337951949                             Yes  18 - 29               Female
2    4337935621                             Yes  18 - 29                 Male
3    4337933040                             Yes  30 - 44                 Male
4    4337931983                             Yes  30 - 44                 Male
5    4337929779                             Yes  18 - 29                 Male
6    4337924420                             Yes  18 - 29                 Male
7    4337916002                             Yes  18 - 29                 Male
8    4337914977                             Yes  30 - 44                 Male
9    4337899817                             Yes  30 - 44                 Male

  How much total combined money did all members of your HOUSEHOLD earn last year?
0                                  $75,000 to $99,999
1                                  $50,000 to $74,999
2                                        $0 to $9,999
```

```
3                                  $200,000 and up
4                               $100,000 to $124,999
5                                       $0 to $9,999
6                                 $25,000 to $49,999
7                               Prefer not to answer
8                                 $75,000 to $99,999
9                                 $25,000 to $49,999

            US Region
0     Middle Atlantic
1  East South Central
2            Mountain
3             Pacific
4             Pacific
5             Pacific
6  East North Central
7            Mountain
8     Middle Atlantic
9  East South Central
```

```
In [32]: x = answers_pd.pivot_table(values='RespondentID', index='Age', columns='Do
         print x

         celeb_no = x['No']
         celeb_yes = x['Yes']
         ind = np.arange(4)    # the x locations for the groups
         width = 0.50       # the width of the bars: can also be len(x) sequence

         p1 = plt.bar(ind, celeb_yes, width, color='g')
         p2 = plt.bar(ind, celeb_no, width, color='r', bottom=celeb_yes)

         #find dimention for y
         #print max(x['Yes'])+max(x['No'])

         plt.ylabel('Scores')
         plt.title('Scores by age group and celebration')
         plt.xticks(ind + width/2., ('18-29', '30-44', '45-59', '60+'))
         plt.yticks(np.arange(0, 301, 50))
         plt.legend((p1[0], p2[0]), ('Yes', 'No'))
         plt.show()
```

```
Do you celebrate Thanksgiving?  No  Yes
Age
18 - 29                         31  185
30 - 44                         24  235
45 - 59                         17  269
60+                              6  258
```


Scores by age group and celebration
Yes
No
Scores
300
250
200
150
100
50
0
18-29
30-44
45-59
60+

In [ ]: