Final Milestone 3

July 13, 2024

1 Final Milestone 3 - Kyle Kinston - DSC540

pull in our libraries

[1]:
```
from bs4 import BeautifulSoup
import requests
import pandas as pd
```

store our url in a variable and use pd.read_html to store the table

[2]:
```
url = 'https://en.wikipedia.org/wiki/Obesity_in_the_United_States'
```

[ ]:
```

```

[3]:
```
df = pd.read_html(url,header=None)[1]
```

[4]:
```
df.head()
```

[4]:
```
  States, district,  & territories Obesity rank Obese adults                  \
  States, district,  & territories Obesity rank  (mid-2000s) (2020)[90][96]
0                          Alabama            5        30.1%           36.3%
1                           Alaska            9        27.3%           34.2%
2                   American Samoa            -            -         75%[94]
3                          Arizona           30        23.3%           29.5%
4                         Arkansas            7        28.1%           35.0%

  Overweight (incl. obese) adults (mid-2000s)  \
  Overweight (incl. obese) adults (mid-2000s)
0                                       65.4%
1                                       64.5%
2                                     95%[97]
3                                       59.5%
4                                       64.7%

  Obese children and adolescents (mid-2000s)[95]
  Obese children and adolescents (mid-2000s)[95]
0                                          16.7%
1                                          11.1%
2                                    35%[94][98]
```

```
3                                            12.2%
4                                            16.4%
```

[ ]:

```

```

transformation 1: fix the headers - we see the header index is multi dimensional and I dont like how that looks so we first reassign the headers to fix this

[5]:

```
df.columns = ['State','Obesity␣
 ↪rank','adult-mid_2000s','adult_2020','Overweight(incl. obese) adult mid␣
 ↪2000s','obese children mid 2000s']
```

[6]:

```
df.head()
```

[6]:

```
             State Obesity rank adult-mid_2000s adult_2020  \
0          Alabama            5           30.1%      36.3%
1           Alaska            9           27.3%      34.2%
2   American Samoa            -               -    75%[94]
3          Arizona           30           23.3%      29.5%
4         Arkansas            7           28.1%      35.0%

  Overweight(incl. obese) adult mid 2000s obese children mid 2000s
0                                   65.4%                    16.7%
1                                   64.5%                    11.1%
2                                 95%[97]              35%[94][98]
3                                   59.5%                    12.2%
4                                   64.7%                    16.4%
```

[ ]:

```

```

[7]:

```
#### transformation 2: set state as the index value of this dataset - just out␣
 ↪of personal preference I set the index to state
```

[8]:

```
df.set_index("State")
```

[8]:

```
                         Obesity rank adult-mid_2000s adult_2020  \
State
Alabama                             5           30.1%      36.3%
Alaska                              9           27.3%      34.2%
American Samoa                      -               -    75%[94]
Arizona                            30           23.3%      29.5%
Arkansas                            7           28.1%      35.0%
California                         48           23.1%      25.1%
Colorado                           51           21.0%      22.6%
Connecticut                        42           20.8%      26.9%
Delaware                           23           25.9%      31.8%
District of Columbia               50           22.1%      23.0%
Florida                            35           23.3%      28.4%
```

Georgia	24	27.5%	31.6%
Guam	-	-	28.3%
Hawaii	49	20.7%	23.8%
Idaho	32	24.6%	29.3%
Illinois	27	25.3%	31.1%
Indiana	12	27.5%	33.6%
Iowa	4	26.3%	36.4%
Kansas	18	25.8%	32.4%
Kentucky	8	28.4%	34.3%
Louisiana	6	29.5%	36.2%
Maine	33	23.7%	29.1%
Maryland	26	25.2%	31.3%
Massachusetts	44	20.9%	25.9%
Michigan	19	27.7%	32.3%
Minnesota	35	24.8%	28.4%
Mississippi	2	34.4%	37.3%
Missouri	17	27.4%	32.5%
Montana	46	21.7%	25.3%
Nebraska	15	26.5%	32.8%
Nevada	43	23.6%	26.7%
New Hampshire	38	23.6%	28.1%
New Jersey	41	22.9%	27.3%
New Mexico	35	23.3%	28.4%
New York	45	23.5%	25.7%
North Carolina	20	27.1%	32.1%
North Dakota	13	25.9%	33.2%
Northern Mariana Islands	-	-	-
Ohio	11	26.9%	33.8%
Oklahoma	3	28.1%	36.5%
Oregon	31	25.0%	29.4%
Pennsylvania	24	25.7%	31.6%
Puerto Rico	-	-	30.7%
Rhode Island	29	21.4%	30.0%
South Carolina	10	29.2%	34.1%
South Dakota	22	26.1%	31.9%
Tennessee	15	29.0%	32.8%
Texas	14	27.2%	33.0%
Utah	46	21.8%	25.3%
Vermont	40	21.1%	27.6%
Virgin Islands (U.S.)	-	-	32.5%
Virginia	28	25.2%	30.1%
Washington	39	24.5%	27.7%
West Virginia	1	30.6%	38.1%
Wisconsin	21	25.5%	32.0%
Wyoming	34	24.0%	28.8%

Overweight(incl. obese) adult mid 2000s \

State	
Alabama	65.4%
Alaska	64.5%
American Samoa	95%[97]
Arizona	59.5%
Arkansas	64.7%
California	59.4%
Colorado	55.0%
Connecticut	58.7%
Delaware	63.9%
District of Columbia	55.0%
Florida	60.8%
Georgia	63.3%
Guam	-
Hawaii	55.3%
Idaho	61.4%
Illinois	61.8%
Indiana	62.8%
Iowa	63.4%
Kansas	62.3%
Kentucky	66.8%
Louisiana	64.2%
Maine	60.8%
Maryland	61.5%
Massachusetts	56.8%
Michigan	63.9%
Minnesota	61.9%
Mississippi	67.4%
Missouri	63.3%
Montana	59.6%
Nebraska	63.9%
Nevada	61.8%
New Hampshire	60.8%
New Jersey	60.5%
New Mexico	60.3%
New York	60.0%
North Carolina	63.4%
North Dakota	64.5%
Northern Mariana Islands	-
Ohio	63.3%
Oklahoma	64.2%
Oregon	60.8%
Pennsylvania	61.9%
Puerto Rico	-
Rhode Island	60.4%
South Carolina	65.1%
South Dakota	64.2%

Tennessee	65.0%
Texas	64.1%
Utah	56.4%
Vermont	56.9%
Virgin Islands (U.S.)	-
Virginia	61.6%
Washington	60.7%
West Virginia	66.8%
Wisconsin	62.4%
Wyoming	61.7%

	obese children mid 2000s
State	
Alabama	16.7%
Alaska	11.1%
American Samoa	35%[94][98]
Arizona	12.2%
Arkansas	16.4%
California	13.2%
Colorado	9.9%
Connecticut	12.3%
Delaware	22.8%
District of Columbia	14.8%
Florida	14.4%
Georgia	16.4%
Guam	22%[99]
Hawaii	13.3%
Idaho	10.1%
Illinois	15.8%
Indiana	15.6%
Iowa	12.5%
Kansas	14.0%
Kentucky	20.6%
Louisiana	17.2%
Maine	12.7%
Maryland	13.3%
Massachusetts	13.6%
Michigan	14.5%
Minnesota	10.1%
Mississippi	17.8%
Missouri	15.6%
Montana	11.1%
Nebraska	11.9%
Nevada	12.4%
New Hampshire	12.9%
New Jersey	13.7%
New Mexico	16.8%

```
New York                                 15.3%
North Carolina                           19.3%
North Dakota                             12.1%
Northern Mariana Islands               16%[100]
Ohio                                     14.2%
Oklahoma                                 15.4%
Oregon                                   14.1%
Pennsylvania                             13.3%
Puerto Rico                        26%[101][102]
Rhode Island                             11.9%
South Carolina                           18.9%
South Dakota                             12.1%
Tennessee                                20.0%
Texas                                    19.1%
Utah                                      8.5%
Vermont                                  11.3%
Virgin Islands (U.S.)                        -
Virginia                                 13.8%
Washington                               10.8%
West Virginia                            20.9%
Wisconsin                                13.5%
Wyoming                                   8.7%
```

[ ]:

[ ]:

[ ]:

transformation 3: drop the states outside the scope of our data - as discussed with my previous final post we are going to limit the scope of this to just the states

[9]:

```
toberemoved = ['American Samoa','Virgin Islands (U.S.)','Puerto Rico','Northern Mariana Islands','Guam','District of Columbia','American Samoa']
df = df[~df["State"].isin(toberemoved)]
```

[10]:

```
df.head()
```

[10]:

```
        State Obesity rank adult-mid_2000s adult_2020  \
0     Alabama            5           30.1%      36.3%
1      Alaska            9           27.3%      34.2%
3     Arizona           30           23.3%      29.5%
4    Arkansas            7           28.1%      35.0%
5  California           48           23.1%      25.1%

  Overweight(incl. obese) adult mid 2000s obese children mid 2000s
0                                   65.4%                    16.7%
1                                   64.5%                    11.1%
```

```
 3                                    59.5%                    12.2%
 4                                    64.7%                    16.4%
 5                                    59.4%                    13.2%
```

transformation 4: search for duplicates - I wanted to verify that the state values did not contain any duplicated values

[11]:

```python
duplicate_values = df['State'].duplicated()
print(duplicate_values)
```

```
0     False
1     False
3     False
4     False
5     False
6     False
7     False
8     False
10    False
11    False
13    False
14    False
15    False
16    False
17    False
18    False
19    False
20    False
21    False
22    False
23    False
24    False
25    False
26    False
27    False
28    False
29    False
30    False
31    False
32    False
33    False
34    False
35    False
36    False
38    False
39    False
40    False
41    False
```

```
43    False
44    False
45    False
46    False
47    False
48    False
49    False
51    False
52    False
53    False
54    False
55    False
Name: State, dtype: bool
```

```
[ ]:
```

transformation 5: sort the df by obesity ranking - Again this is just a preference because this will likely be how we will utilize the data, also I am struggling to find more data transformations because this data was rather clean to begin with found this field was stored as a string or varchar and need it to be numeric for the sort to work as intended

```
[12]: df["Obesity rank"] = df["Obesity rank"].apply(pd.to_numeric)
```

```
[ ]:
```

```
[ ]:
```

```
[13]: df.sort_values(by=['Obesity rank'])
```

```
[13]:              State  Obesity rank adult-mid_2000s adult_2020  \
 53   West Virginia             1           30.6%      38.1%
 26     Mississippi             2           34.4%      37.3%
 39        Oklahoma             3           28.1%      36.5%
 17            Iowa             4           26.3%      36.4%
 0          Alabama             5           30.1%      36.3%
 20       Louisiana             6           29.5%      36.2%
 4         Arkansas             7           28.1%      35.0%
 19        Kentucky             8           28.4%      34.3%
 1           Alaska             9           27.3%      34.2%
 44  South Carolina            10           29.2%      34.1%
 38            Ohio            11           26.9%      33.8%
 16         Indiana            12           27.5%      33.6%
 36    North Dakota            13           25.9%      33.2%
 47           Texas            14           27.2%      33.0%
 46       Tennessee            15           29.0%      32.8%
 29        Nebraska            15           26.5%      32.8%
 27        Missouri            17           27.4%      32.5%
 18          Kansas            18           25.8%      32.4%
```

```
24        Michigan            19           27.7%      32.3%
35  North Carolina            20           27.1%      32.1%
54       Wisconsin            21           25.5%      32.0%
45    South Dakota            22           26.1%      31.9%
8         Delaware            23           25.9%      31.8%
11         Georgia            24           27.5%      31.6%
41    Pennsylvania            24           25.7%      31.6%
22        Maryland            26           25.2%      31.3%
15        Illinois            27           25.3%      31.1%
51        Virginia            28           25.2%      30.1%
43    Rhode Island            29           21.4%      30.0%
3          Arizona            30           23.3%      29.5%
40          Oregon            31           25.0%      29.4%
14           Idaho            32           24.6%      29.3%
21           Maine            33           23.7%      29.1%
55         Wyoming            34           24.0%      28.8%
10         Florida            35           23.3%      28.4%
25       Minnesota            35           24.8%      28.4%
33      New Mexico            35           23.3%      28.4%
31   New Hampshire            38           23.6%      28.1%
52      Washington            39           24.5%      27.7%
49         Vermont            40           21.1%      27.6%
32      New Jersey            41           22.9%      27.3%
7      Connecticut            42           20.8%      26.9%
30          Nevada            43           23.6%      26.7%
23   Massachusetts            44           20.9%      25.9%
34        New York            45           23.5%      25.7%
48            Utah            46           21.8%      25.3%
28         Montana            46           21.7%      25.3%
5       California            48           23.1%      25.1%
13          Hawaii            49           20.7%      23.8%
6         Colorado            51           21.0%      22.6%

   Overweight(incl. obese) adult mid 2000s obese children mid 2000s
53                                   66.8%                    20.9%
26                                   67.4%                    17.8%
39                                   64.2%                    15.4%
17                                   63.4%                    12.5%
0                                    65.4%                    16.7%
20                                   64.2%                    17.2%
4                                    64.7%                    16.4%
19                                   66.8%                    20.6%
1                                    64.5%                    11.1%
44                                   65.1%                    18.9%
38                                   63.3%                    14.2%
16                                   62.8%                    15.6%
36                                   64.5%                    12.1%
```

47	64.1%	19.1%
46	65.0%	20.0%
29	63.9%	11.9%
27	63.3%	15.6%
18	62.3%	14.0%
24	63.9%	14.5%
35	63.4%	19.3%
54	62.4%	13.5%
45	64.2%	12.1%
8	63.9%	22.8%
11	63.3%	16.4%
41	61.9%	13.3%
22	61.5%	13.3%
15	61.8%	15.8%
51	61.6%	13.8%
43	60.4%	11.9%
3	59.5%	12.2%
40	60.8%	14.1%
14	61.4%	10.1%
21	60.8%	12.7%
55	61.7%	8.7%
10	60.8%	14.4%
25	61.9%	10.1%
33	60.3%	16.8%
31	60.8%	12.9%
52	60.7%	10.8%
49	56.9%	11.3%
32	60.5%	13.7%
7	58.7%	12.3%
30	61.8%	12.4%
23	56.8%	13.6%
34	60.0%	15.3%
48	56.4%	8.5%
28	59.6%	11.1%
5	59.4%	13.2%
13	55.3%	13.3%
6	55.0%	9.9%

1.0.1 What changes were made to the data?

For the most part, the data itself did not change, but we did drop a few rows that were associated with nonstate values.

1.0.2 Are there any legal or regulatory guidelines for your data or project topic?

I don't feel there are any regulatory or legal guidelines per se but it's worth noting that this data was based on CDC surveys so their definition of overweight is $25 <= \text{BMI} < 30$ and obese is $\text{BMI} >= 30$

1.0.3 What risks could be created based on the transformations done?

we lose a little bit of data for those nonstate territories but other than that I don't feel anything risky was performed

1.0.4 Did you make any assumptions in cleaning/transforming the data?

the main assumption is that I would not need that additional data that was trimmed out.

1.0.5 How was your data sourced / verified for credibility?

This data came from Wikipedia but this particular data chart was fed data from the Regards survey conducted by the CDC ### Was your data acquired in an ethical way? With it being a government agency I certainly hope so but there is not enough information to be 100% certain

1.0.6 How would you mitigate any of the ethical implications you have identified?

I think the key is being forthcoming about the source of the data and the definitions of the subgroups such as obese and overweight.

[ ]:

```

```