Final Milestone 4

July 28, 2024

Perform at least 5 data transformation and/or cleansing steps to your API data. The below examples are not required - they are just potential transformations you could do. If your data doesn't work for these scenarios, complete different transformations. You can do the same transformation multiple times if needed to clean your data. The goal is a clean dataset at the end of the milestone. As a reminder - you cannot export your API data to CSV to work with it, you must do all the work directly against the API/JSON source.

[ ]:

It almost feels too easy but the cdc provided this code snippit with the connection string to the api, though it required some minor tweeking

[20]:
```
#!/usr/bin/env python

# make sure to install these packages before running:
# pip install pandas
# pip install sodapy

import pandas as pd
from sodapy import Socrata

# Unauthenticated client only works with public data sets. Note 'None'
# in place of application token, and no username or password:
client = Socrata("data.cdc.gov", None)

# Example authenticated client (needed for non-public datasets):
# client = Socrata(data.cdc.gov,
#                  MyAppToken,
#                  username="user@example.com",
#                  password="AFakePassword")

# First 2000 results, returned as JSON from API / converted to Python list of
# dictionaries by sodapy.
results = client.get("fqb7-mgjf", limit=80000)

# Convert to pandas DataFrame
results_df = pd.DataFrame.from_records(results)
```

```
WARNING:root:Requests made without an app_token will be subject to strict
```

throttling limits.

[22]:
```
results_df.head()
```

[22]:
```
                                          geolocation break_out data_value  \
0  {'latitude': '32.84057112200048', 'human_addre…   Overall        2.0
1  {'latitude': '32.84057112200048', 'human_addre…   Overall       26.1
2  {'latitude': '32.84057112200048', 'human_addre…   Overall       33.7
3  {'latitude': '32.84057112200048', 'human_addre…   Overall       38.3
4  {'latitude': '32.84057112200048', 'human_addre…      Male        2.0

  breakoutid confidence_limit_high responseid breakoutcategoryid  \
0        BO1                   2.6    RESP042               CAT1
1        BO1                  28.0    RESP041               CAT1
2        BO1                  35.7    RESP040               CAT1
3        BO1                  40.3    RESP039               CAT1
4       SEX1                   3.1    RESP042               CAT2

                                        question datasource  \
0  Weight classification by Body Mass Index (BMI)…      BRFSS
1  Weight classification by Body Mass Index (BMI)…      BRFSS
2  Weight classification by Body Mass Index (BMI)…      BRFSS
3  Weight classification by Body Mass Index (BMI)…      BRFSS
4  Weight classification by Body Mass Index (BMI)…      BRFSS

  data_value_unit  …  topicid break_out_category           topic  \
0               %  …  TOPIC09            Overall  BMI Categories
1               %  …  TOPIC09            Overall  BMI Categories
2               %  …  TOPIC09            Overall  BMI Categories
3               %  …  TOPIC09            Overall  BMI Categories
4               %  …  TOPIC09             Gender  BMI Categories

                         class locationdesc                      response  \
0  Overweight and Obesity (BMI)      Alabama    Underweight (BMI 12.0-18.4)
1  Overweight and Obesity (BMI)      Alabama  Normal Weight (BMI 18.5-24.9)
2  Overweight and Obesity (BMI)      Alabama     Overweight (BMI 25.0-29.9)
3  Overweight and Obesity (BMI)      Alabama        Obese (BMI 30.0 - 99.8)
4  Overweight and Obesity (BMI)      Alabama    Underweight (BMI 12.0-18.4)

  confidence_limit_low sample_size data_value_footnote  \
0                  1.3          69                 NaN
1                 24.1        1016                 NaN
2                 31.7        1460                 NaN
3                 36.2        1663                 NaN
4                  1.0          24                 NaN

  data_value_footnote_symbol
```

```
0                        NaN
1                        NaN
2                        NaN
3                        NaN
4                        NaN

[5 rows x 27 columns]
```

[ ]:

[ ]:

[ ]:

[ ]:

[ ]:

0.0.1 transformation 1 - remove non relevent location values

Here I pulled a series of the values for location description to find the values that need to be elimiated. Then I removed those values from the dataset

[30]:
```
unique_values = results_df['locationdesc'].unique()
```

[32]:
```
unique_values
```

[32]:
```
array(['Alabama', 'Alaska', 'Arizona', 'Arkansas', 'California',
       'Colorado', 'Connecticut', 'Delaware', 'District of Columbia',
       'Florida', 'Georgia', 'Hawaii', 'Idaho', 'Illinois', 'Indiana',
       'Iowa', 'Kansas', 'Kentucky', 'Louisiana', 'Maine', 'Maryland',
       'Massachusetts', 'Michigan', 'Minnesota', 'Mississippi',
       'Missouri', 'Montana', 'Nebraska', 'Nevada', 'New Hampshire',
       'New Jersey', 'New Mexico', 'New York', 'North Carolina',
       'North Dakota', 'Ohio', 'Oklahoma', 'Oregon', 'Pennsylvania',
       'Rhode Island', 'South Carolina', 'South Dakota', 'Tennessee',
       'Texas', 'Utah', 'Vermont', 'Virginia', 'Washington',
       'West Virginia', 'Wisconsin', 'Wyoming', 'Guam', 'Puerto Rico',
       'Virgin Islands', 'All States and DC (median) **',
       'All States, DC and Territories (median) **'], dtype=object)
```

[34]:
```
toBeRemoved = ['District of Columbia','Guam','Puerto Rico','All States and DC␣
↪(median) **','All States, DC and Territories (median) **','Virgin Islands']
```

[36]:
```
results_df = results_df[~results_df['locationdesc'].isin(toBeRemoved)]
```

[38]:
```
results_df.head()
```

[38]:

```
                                          geolocation break_out data_value  \
0  {'latitude': '32.84057112200048', 'human_addre…   Overall        2.0
1  {'latitude': '32.84057112200048', 'human_addre…   Overall       26.1
2  {'latitude': '32.84057112200048', 'human_addre…   Overall       33.7
3  {'latitude': '32.84057112200048', 'human_addre…   Overall       38.3
4  {'latitude': '32.84057112200048', 'human_addre…      Male        2.0

  breakoutid confidence_limit_high responseid breakoutcategoryid  \
0        BO1                   2.6    RESP042               CAT1
1        BO1                  28.0    RESP041               CAT1
2        BO1                  35.7    RESP040               CAT1
3        BO1                  40.3    RESP039               CAT1
4       SEX1                   3.1    RESP042               CAT2

                                        question datasource  \
0  Weight classification by Body Mass Index (BMI)…      BRFSS
1  Weight classification by Body Mass Index (BMI)…      BRFSS
2  Weight classification by Body Mass Index (BMI)…      BRFSS
3  Weight classification by Body Mass Index (BMI)…      BRFSS
4  Weight classification by Body Mass Index (BMI)…      BRFSS

  data_value_unit  …  topicid break_out_category          topic  \
0               %  …  TOPIC09            Overall  BMI Categories
1               %  …  TOPIC09            Overall  BMI Categories
2               %  …  TOPIC09            Overall  BMI Categories
3               %  …  TOPIC09            Overall  BMI Categories
4               %  …  TOPIC09             Gender  BMI Categories

                         class locationdesc                      response  \
0  Overweight and Obesity (BMI)      Alabama    Underweight (BMI 12.0-18.4)
1  Overweight and Obesity (BMI)      Alabama  Normal Weight (BMI 18.5-24.9)
2  Overweight and Obesity (BMI)      Alabama     Overweight (BMI 25.0-29.9)
3  Overweight and Obesity (BMI)      Alabama        Obese (BMI 30.0 - 99.8)
4  Overweight and Obesity (BMI)      Alabama    Underweight (BMI 12.0-18.4)

  confidence_limit_low sample_size data_value_footnote  \
0                  1.3          69                 NaN
1                 24.1        1016                 NaN
2                 31.7        1460                 NaN
3                 36.2        1663                 NaN
4                  1.0          24                 NaN

  data_value_footnote_symbol
0                        NaN
1                        NaN
2                        NaN
3                        NaN
```

```
4                          NaN

[5 rows x 27 columns]
```

0.0.2 Transformation 2 - remove covid effected data 2020,2021,2022

Using the same method as the previous setp we are going to remove the data for 20,21,22

[41]:
```
# results_df.to_csv('temp.csv')
```

[43]:
```
toBeRemoved = ['2020','2021','2022']
```

[45]:
```
results_df = results_df[~results_df['year'].isin(toBeRemoved)]
```

[47]:
```
results_df.head()
```

[47]:
```
                                            geolocation            break_out  \
17520  {'latitude': '32.84057112200048', 'human_addre…  Black, non-Hispanic
17521  {'latitude': '32.84057112200048', 'human_addre…                25-34
17522  {'latitude': '32.84057112200048', 'human_addre…     College graduate
17523  {'latitude': '32.84057112200048', 'human_addre…                55-64
17524  {'latitude': '32.84057112200048', 'human_addre…                  65+

      data_value breakoutid confidence_limit_high responseid  \
17520       46.7     RACE02                  49.9    RESP039
17521       31.6      AGE02                  35.9    RESP040
17522       32.0     EDUCA4                  34.5    RESP039
17523       41.2      AGE05                  44.4    RESP039
17524       39.0      AGE09                  41.4    RESP040

      breakoutcategoryid                                          question  \
17520               CAT4  Weight classification by Body Mass Index (BMI)…
17521               CAT3  Weight classification by Body Mass Index (BMI)…
17522               CAT5  Weight classification by Body Mass Index (BMI)…
17523               CAT3  Weight classification by Body Mass Index (BMI)…
17524               CAT3  Weight classification by Body Mass Index (BMI)…

      datasource data_value_unit  …  topicid  break_out_category  \
17520      BRFSS               %  …  TOPIC09      Race/Ethnicity
17521      BRFSS               %  …  TOPIC09           Age Group
17522      BRFSS               %  …  TOPIC09  Education Attained
17523      BRFSS               %  …  TOPIC09           Age Group
17524      BRFSS               %  …  TOPIC09           Age Group

               topic                          class locationdesc  \
17520  BMI Categories  Overweight and Obesity (BMI)      Alabama
17521  BMI Categories  Overweight and Obesity (BMI)      Alabama
17522  BMI Categories  Overweight and Obesity (BMI)      Alabama
```

```
17523  BMI Categories  Overweight and Obesity (BMI)      Alabama
17524  BMI Categories  Overweight and Obesity (BMI)      Alabama

                        response confidence_limit_low sample_size  \
17520     Obese (BMI 30.0 - 99.8)                 43.4         780
17521  Overweight (BMI 25.0-29.9)                 27.2         192
17522     Obese (BMI 30.0 - 99.8)                 29.5         685
17523     Obese (BMI 30.0 - 99.8)                 38.0         586
17524  Overweight (BMI 25.0-29.9)                 36.5         950

      data_value_footnote data_value_footnote_symbol
17520                 NaN                       NaN
17521                 NaN                       NaN
17522                 NaN                       NaN
17523                 NaN                       NaN
17524                 NaN                       NaN

[5 rows x 27 columns]
```

[49]:
```
results_df.head()
```

[49]:
```
                                          geolocation            break_out  \
17520  {'latitude': '32.84057112200048', 'human_addre…  Black, non-Hispanic
17521  {'latitude': '32.84057112200048', 'human_addre…                25-34
17522  {'latitude': '32.84057112200048', 'human_addre…     College graduate
17523  {'latitude': '32.84057112200048', 'human_addre…                55-64
17524  {'latitude': '32.84057112200048', 'human_addre…                  65+

      data_value breakoutid confidence_limit_high responseid  \
17520       46.7     RACE02                  49.9    RESP039
17521       31.6      AGE02                  35.9    RESP040
17522       32.0     EDUCA4                  34.5    RESP039
17523       41.2      AGE05                  44.4    RESP039
17524       39.0      AGE09                  41.4    RESP040

      breakoutcategoryid                                          question  \
17520               CAT4  Weight classification by Body Mass Index (BMI)…
17521               CAT3  Weight classification by Body Mass Index (BMI)…
17522               CAT5  Weight classification by Body Mass Index (BMI)…
17523               CAT3  Weight classification by Body Mass Index (BMI)…
17524               CAT3  Weight classification by Body Mass Index (BMI)…

      datasource data_value_unit  …  topicid  break_out_category  \
17520      BRFSS               %  …  TOPIC09      Race/Ethnicity
17521      BRFSS               %  …  TOPIC09           Age Group
17522      BRFSS               %  …  TOPIC09  Education Attained
17523      BRFSS               %  …  TOPIC09           Age Group
```

```
17524      BRFSS                %  …  TOPIC09           Age Group

                topic                          class locationdesc  \
17520  BMI Categories  Overweight and Obesity (BMI)      Alabama
17521  BMI Categories  Overweight and Obesity (BMI)      Alabama
17522  BMI Categories  Overweight and Obesity (BMI)      Alabama
17523  BMI Categories  Overweight and Obesity (BMI)      Alabama
17524  BMI Categories  Overweight and Obesity (BMI)      Alabama

                       response confidence_limit_low sample_size  \
17520     Obese (BMI 30.0 - 99.8)                 43.4         780
17521  Overweight (BMI 25.0-29.9)                 27.2         192
17522     Obese (BMI 30.0 - 99.8)                 29.5         685
17523     Obese (BMI 30.0 - 99.8)                 38.0         586
17524  Overweight (BMI 25.0-29.9)                 36.5         950

      data_value_footnote data_value_footnote_symbol
17520                 NaN                        NaN
17521                 NaN                        NaN
17522                 NaN                        NaN
17523                 NaN                        NaN
17524                 NaN                        NaN

[5 rows x 27 columns]
```

[ ]:

0.0.3 Transformation 3 - remove irellivent collumns

[53]:
```
#results_df
```

[55]:
```
results_df = results_df.drop(columns␣
↪=['geolocation','confidence_limit_high','question', 'datasource',␣
↪'questionid', 'data_value_type', 'topicid', 'data_value_footnote',␣
↪'data_value_footnote_symbol'])
```

0.0.4 transformation 4 - remove results with sample size <=100

filter the data to anything more then 100 sample size and save the data to a new df

[58]:
```
pd.to_numeric(results_df['sample_size'])
```

[58]:
```
17520     780
17521     192
17522     685
17523     586
17524     950
          …
```

```
64581      13
64582     640
64583    1100
64584    1276
64585      42
Name: sample_size, Length: 44248, dtype: int64
```

[60]:
```
results_df2 = results_df
```

[62]:
```
# rslt_df = dataframe[dataframe['Percentage'] > 70]
```

[64]:
```
results_df2['sample_size'] = pd.to_numeric(results_df['sample_size'])
```

[66]:
```
results_df2 = results_df2[results_df2['sample_size'] > 10]
```

[68]:
```
results_df2.head()
```

[68]:

	break_out	data_value	breakoutid	responseid	\
17520	Black, non-Hispanic	46.7	RACE02	RESP039	
17521	25-34	31.6	AGE02	RESP040	
17522	College graduate	32.0	EDUCA4	RESP039	
17523	55-64	41.2	AGE05	RESP039	
17524	65+	39.0	AGE09	RESP040	

	breakoutcategoryid	data_value_unit	locationid	display_order	year	\
17520	CAT4	%	01	1	2019	
17521	CAT3	%	01	1	2019	
17522	CAT5	%	01	1	2019	
17523	CAT3	%	01	1	2019	
17524	CAT3	%	01	1	2019	

	locationabbr	classid	break_out_category	topic	\
17520	AL	CLASS14	Race/Ethnicity	BMI Categories	
17521	AL	CLASS14	Age Group	BMI Categories	
17522	AL	CLASS14	Education Attained	BMI Categories	
17523	AL	CLASS14	Age Group	BMI Categories	
17524	AL	CLASS14	Age Group	BMI Categories	

	class	locationdesc	response	\
17520	Overweight and Obesity (BMI)	Alabama	Obese (BMI 30.0 - 99.8)	
17521	Overweight and Obesity (BMI)	Alabama	Overweight (BMI 25.0-29.9)	
17522	Overweight and Obesity (BMI)	Alabama	Obese (BMI 30.0 - 99.8)	
17523	Overweight and Obesity (BMI)	Alabama	Obese (BMI 30.0 - 99.8)	
17524	Overweight and Obesity (BMI)	Alabama	Overweight (BMI 25.0-29.9)	

	confidence_limit_low	sample_size
17520	43.4	780

```
17521                27.2          192
17522                29.5          685
17523                38.0          586
17524                36.5          950
```

0.0.5 transformation 5 - only keep the survey for each state/year/subgroup/ that has the highest sample size

I made a new df because there was lots of trial and error with this step and reruning the pull for the inital data is quite a process.

[71]:
```
results_df3 = results_df2
```

create a field for uniqueid that will essentally be our groupid

[73]:
```
results_df3["uniqueid"] = results_df3["locationdesc"] + results_df3["year"] +␣
↪results_df3["break_out_category"]
```

```
C:\Users\kings\AppData\Local\Temp\ipykernel_29064\4042686915.py: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
  results_df3["uniqueid"] = results_df3["locationdesc"] + results_df3["year"] +
results_df3["break_out_category"]
```

[75]:
```
results_df3.head()
```

[75]:
```
                break_out data_value breakoutid responseid  \
17520  Black, non-Hispanic       46.7     RACE02    RESP039
17521                25-34       31.6      AGE02    RESP040
17522     College graduate       32.0     EDUCA4    RESP039
17523                55-64       41.2      AGE05    RESP039
17524                  65+       39.0      AGE09    RESP040

      breakoutcategoryid data_value_unit locationid display_order  year  \
17520               CAT4                %         01             1  2019
17521               CAT3                %         01             1  2019
17522               CAT5                %         01             1  2019
17523               CAT3                %         01             1  2019
17524               CAT3                %         01             1  2019

      locationabbr  classid  break_out_category           topic  \
17520           AL  CLASS14      Race/Ethnicity  BMI Categories
17521           AL  CLASS14           Age Group  BMI Categories
17522           AL  CLASS14  Education Attained  BMI Categories
17523           AL  CLASS14           Age Group  BMI Categories
```

```
17524           AL  CLASS14           Age Group  BMI Categories

                              class locationdesc                    response  \
17520  Overweight and Obesity (BMI)      Alabama     Obese (BMI 30.0 - 99.8)
17521  Overweight and Obesity (BMI)      Alabama  Overweight (BMI 25.0-29.9)
17522  Overweight and Obesity (BMI)      Alabama     Obese (BMI 30.0 - 99.8)
17523  Overweight and Obesity (BMI)      Alabama     Obese (BMI 30.0 - 99.8)
17524  Overweight and Obesity (BMI)      Alabama  Overweight (BMI 25.0-29.9)

      confidence_limit_low  sample_size                      uniqueid
17520                 43.4          780      Alabama2019Race/Ethnicity
17521                 27.2          192           Alabama2019Age Group
17522                 29.5          685  Alabama2019Education Attained
17523                 38.0          586           Alabama2019Age Group
17524                 36.5          950           Alabama2019Age Group
```

create a new df with the grouping

[77]:
```
grouped_df = results_df3.groupby(['uniqueid'],group_keys=True)['sample_size'].
↪max()
```

merge the two dfs to get all the values from the original df with the filtering of the second

[81]:
```
merged_df = pd.merge(grouped_df,results_df3,on␣
↪=['uniqueid','sample_size'],how='inner')
```

[83]:
```
merged_df.head()
```

[83]:
```
                       uniqueid  sample_size       break_out data_value  \
0           Alabama2011Age Group          950             65+       39.0
1  Alabama2011Education Attained          835  H.S. or G.E.D.       34.1
2             Alabama2011Gender         1591          Female       31.8
3    Alabama2011Household Income          722        $50,000+       37.5
4             Alabama2011Overall         2529         Overall       34.7

  breakoutid responseid breakoutcategoryid data_value_unit locationid  \
0      AGE09    RESP040               CAT3               %          1
1     EDUCA2    RESP040               CAT5               %          1
2       SEX2    RESP039               CAT2               %          1
3    INCOME5    RESP040               CAT6               %          1
4        BO1    RESP040               CAT1               %          1

  display_order  year locationabbr  classid  break_out_category  \
0            34  2011           AL  CLASS14           Age Group
1            62  2011           AL  CLASS14  Education Attained
2             9  2011           AL  CLASS14              Gender
3            90  2011           AL  CLASS14    Household Income
4             2  2011           AL  CLASS14             Overall
```

```
            topic                          class locationdesc  \
0  BMI Categories  Overweight and Obesity (BMI)      Alabama
1  BMI Categories  Overweight and Obesity (BMI)      Alabama
2  BMI Categories  Overweight and Obesity (BMI)      Alabama
3  BMI Categories  Overweight and Obesity (BMI)      Alabama
4  BMI Categories  Overweight and Obesity (BMI)      Alabama

                     response confidence_limit_low
0  Overweight (BMI 25.0-29.9)                 36.6
1  Overweight (BMI 25.0-29.9)                 31.4
2     Obese (BMI 30.0 - 99.8)                 29.9
3  Overweight (BMI 25.0-29.9)                 34.6
4  Overweight (BMI 25.0-29.9)                 33.1
```

- Transformation 1 - Remove non-relevant location values . The first transformation removes the non-relevant locations. This is the same process that we did in the previous data sets, we are removing any values that are not us states - Transformation 2 - remove covidaeffected data 2020,2021,202 The second transformation was to remove the data that was possibly affected by the COVID-19 pandemic. We also did this in the previous data sets.

- Transformation 3 - remove non-relevant columns The third transformation was to remove the non-relevant columns for this particular data set
- transformation 4 - remove results with sample size <=100 Transformation 4 was to remove any surveys that were conducted with less than 100 recipients. This was an attempt to limit the data sets involved and make the data overall more manageable. This did not accomplish what I would have liked, and that led to the 5th transformation. - transformation 5 - only keep the survey for each state/year/subgroup/ that has the highest sample siz After seeing the amount of data remaining with the 4th transformation, I regrouped and came up with a new strategy to limit the data. Here I decided I wanted to keep 1 survey for each state/year/subgroup. So I created a column that combined those three values and grouped them by that column, looking at the max survey size value. Once I had that dataset I then merged the two to limit the results as desired. I don't feel this is the most efficient way to do this, but I could not get it to work trying other methods.
- Are there any legal or regulatory guidelines for your data or project tropic? I don't feel there are any guidelines to abide by aside from the assumption that this data was gathered ethically
- What risks could be created based on the transformations done? I don't think any major risks should be created. The only potential risks I see are limiting the survey results that we are looking at may have skewed the data in some way.
- Did you make any assumptions in cleaning/transforming the data? Again I made the assumption that the data would not be skewed too much based on the largest survey for those sub groups.
- How was your data Sourced/verified for credibility? This data was sourced through the CDC. Being from a government agency, I assume they are regulated to the point that the data is credible.

- Was your data acquired in an ethical way? Again, I certainly hope so with it being gathered by a government agency.
- how would you mitigate any ethical implications you identified? I would find a way to create a set of rules for limiting the data sets instead of just doing it blindly based on the largest survey for each group. I think there has to be a better way to do this but I'm not sure what that would be.

e2s

[ ]: