Final Milestone 2

June 26, 2024

0.1 DSC 540 Final Milestone 2 - Kyle Kingston

[1]:
```
import numpy as np
import pandas as pd
```

[2]:
```
df = pd.read_csv("Nutrition__Physical_Activity__and_Obesity.csv")
```

[3]:
```
print(df.head())
print(len(df))
```

```
   YearStart  YearEnd LocationAbbr LocationDesc  \
0       2020     2020           US     National
1       2014     2014           GU         Guam
2       2013     2013           US     National
3       2013     2013           US     National
4       2015     2015           US     National

                                  Datasource                    Class  \
0  Behavioral Risk Factor Surveillance System         Physical Activity
1  Behavioral Risk Factor Surveillance System  Obesity / Weight Status
2  Behavioral Risk Factor Surveillance System  Obesity / Weight Status
3  Behavioral Risk Factor Surveillance System  Obesity / Weight Status
4  Behavioral Risk Factor Surveillance System         Physical Activity

                         Topic  \
0  Physical Activity - Behavior
1       Obesity / Weight Status
2       Obesity / Weight Status
3       Obesity / Weight Status
4  Physical Activity - Behavior

                                          Question  Data_Value_Unit  \
0  Percent of adults who engage in no leisure-tim…              NaN
1  Percent of adults aged 18 years and older who …              NaN
2  Percent of adults aged 18 years and older who …              NaN
3  Percent of adults aged 18 years and older who …              NaN
4  Percent of adults who achieve at least 300 min…              NaN

  Data_Value_Type  …              GeoLocation  ClassID TopicID QuestionID  \
```

```
0           Value  …                       NaN       PA     PA1       Q047
1           Value  …  (13.444304, 144.793731)      OWS    OWS1       Q036
2           Value  …                       NaN      OWS    OWS1       Q036
3           Value  …                       NaN      OWS    OWS1       Q037
4           Value  …                       NaN       PA     PA1       Q045

   DataValueTypeID  LocationID  StratificationCategory1       Stratification1  \
0            VALUE          59           Race/Ethnicity              Hispanic
1            VALUE          66                Education  High school graduate
2            VALUE          59                   Income     $50,000 - $74,999
3            VALUE          59                   Income     Data not reported
4            VALUE          59                   Income     Less than $15,000

  StratificationCategoryId1 StratificationID1
0                      RACE           RACEHIS
1                       EDU         EDUHSGRAD
2                       INC           INC5075
3                       INC             INCNR
4                       INC         INCLESS15

[5 rows x 33 columns]
93249
```

- transformation 1: remove puerto rico,virgin islands,guam from data set
- transformation 2: remove columns with only null values
- transformation 3: remove irrelevant data(data value null values)
- transformation 4: eliminate data from years(2020,2021,2022) due to it being skewed due to the covid19 pandemic
- transformation 5: break data into smaller sets based on class(may join together again later to make a wider data set but we will see how our other data joins in before making this determination

- transformation 1: remove national based data as we want to isolate data sets to the individual states Here I am going to remove the values for puerto rico, guam, and the virgin islands to keep the data set narrowed down to the 50 states. This is just to keep the data a little simpler because I feel the other data sets will possibly not contain as complete of data to include these areas.

[4]:
```
toberemoved = ['PR','GU','VI']
df2 = df[~df["LocationAbbr"].isin(toberemoved)]
print(len(df2))
```

```
90029
```

- transformation 2: remove columns with only null values using the sum of the isnull valuesand comparing that to the length of the df we can determine what collumns are completly empty

[5]:
```
nan_values_count = df2.isnull().sum()
print(nan_values_count)
print(len(df2))
```

```
YearStart                        0
YearEnd                          0
LocationAbbr                     0
LocationDesc                     0
Datasource                       0
Class                            0
Topic                            0
Question                         0
Data_Value_Unit                90029
Data_Value_Type                    0
Data_Value                      8380
Data_Value_Alt                  8380
Data_Value_Footnote_Symbol     81649
Data_Value_Footnote            81649
Low_Confidence_Limit            8380
High_Confidence_Limit           8380
Sample_Size                     8380
Total                          86814
Age(years)                     70739
Education                      77169
Gender                         83599
Income                         67524
Race/Ethnicity                 64309
GeoLocation                     1736
ClassID                            0
TopicID                            0
QuestionID                         0
DataValueTypeID                    0
LocationID                         0
StratificationCategory1            9
Stratification1                    9
StratificationCategoryId1          9
StratificationID1                  9
dtype: int64
90029
```

[6]:
```
df3 = df2.drop(columns=['Data_Value_Footnote_Symbol',␣
  ↪'Data_Value_Footnote','Data_Value_Unit'])
print(df3.head())
```

```
   YearStart  YearEnd LocationAbbr LocationDesc  \
0       2020     2020           US     National
2       2013     2013           US     National
3       2013     2013           US     National
```

```
4       2015     2015           US     National
6       2012     2012           WY      Wyoming

                                  Datasource                    Class  \
0  Behavioral Risk Factor Surveillance System        Physical Activity
2  Behavioral Risk Factor Surveillance System  Obesity / Weight Status
3  Behavioral Risk Factor Surveillance System  Obesity / Weight Status
4  Behavioral Risk Factor Surveillance System        Physical Activity
6  Behavioral Risk Factor Surveillance System  Obesity / Weight Status

                         Topic  \
0  Physical Activity - Behavior
2       Obesity / Weight Status
3       Obesity / Weight Status
4  Physical Activity - Behavior
6       Obesity / Weight Status

                                          Question Data_Value_Type  \
0  Percent of adults who engage in no leisure-tim…           Value
2  Percent of adults aged 18 years and older who …           Value
3  Percent of adults aged 18 years and older who …           Value
4  Percent of adults who achieve at least 300 min…           Value
6  Percent of adults aged 18 years and older who …           Value

   Data_Value  …                    GeoLocation  ClassID  TopicID  \
0        30.6  …                            NaN       PA      PA1
2        28.8  …                            NaN      OWS     OWS1
3        32.7  …                            NaN      OWS     OWS1
4        26.6  …                            NaN       PA      PA1
6        48.5  …  (43.235541343, -108.109830353)      OWS     OWS1

   QuestionID DataValueTypeID LocationID StratificationCategory1  \
0        Q047           VALUE         59          Race/Ethnicity
2        Q036           VALUE         59                  Income
3        Q037           VALUE         59                  Income
4        Q045           VALUE         59                  Income
6        Q037           VALUE         56          Race/Ethnicity

                Stratification1 StratificationCategoryId1 StratificationID1
0                      Hispanic                      RACE          RACEHIS
2             $50,000 - $74,999                       INC          INC5075
3             Data not reported                       INC            INCNR
4             Less than $15,000                       INC         INCLESS15
6  American Indian/Alaska Native                      RACE          RACENAA

[5 rows x 30 columns]
```

transformation 3: remove irreliivent data(data value null values) The data value collumn contains the relevent data for each state. I am goign to eliminate these value because it renders that row of data seemingly useless

[7]:
```
df3 = df3[df3['Data_Value'].notna()]
print(len(df3))
```

```
81649
```

- transformation 4: eliminate data from years(2020,2021,2022) due to it being skewed due to the covid-19 pandemic I went back and forth on this a bit. I realize that its hard to get rid of years worth of data but I've seen a variety of areas that this data is skewed greatly by the pandemic and its hard to use that data for decision making becuase of how unprecidented it is im comparison to the rest of the data set.

[8]:
```
toberemoved = [2020,2021,2022]
df3 = df3[~df3["YearStart"].isin(toberemoved)]
print(len(df3))
```

```
67172
```

- transformation 5: break data into smaller sets based on class(may join together again later to make a wider data set but we will see how our other data joins in before making this determinatation

I wanted to break the data into the three categories for relevence. I am considering joining this data back into a wider dataframe but I want to wait and see what my other data shows us and determine if that would be fruitful or unneeded.

[11]:
```
PhysicalBehavior = df3[df3["Topic"] == "Physical Activity - Behavior"]
ObesityWeightStatus = df3[df3["Topic"] == "Obesity / Weight Status"]
FruitsVegBehavior = df3[df3["Topic"] == "Fruits and Vegetables - Behavior"]
print(len(PhysicalBehavior),len(ObesityWeightStatus),len(FruitsVegBehavior))
```

```
38163 23782 5227
```

[12]:
```
PhysicalBehavior.head()
```

[12]:

	YearStart	YearEnd	LocationAbbr	LocationDesc	Datasource	Class
4	2015	2015	US	National	Behavioral Risk Factor Surveillance System	Physical Activity
13	2012	2012	WY	Wyoming	Behavioral Risk Factor Surveillance System	Physical Activity
20	2017	2017	NJ	New Jersey	Behavioral Risk Factor Surveillance System	Physical Activity
23	2013	2013	US	National	Behavioral Risk Factor Surveillance System	Physical Activity
25	2011	2011	US	National	Behavioral Risk Factor Surveillance System	Physical Activity

```
25  Behavioral Risk Factor Surveillance System  Physical Activity

                          Topic  \
4   Physical Activity - Behavior
13  Physical Activity - Behavior
20  Physical Activity - Behavior
23  Physical Activity - Behavior
25  Physical Activity - Behavior

                                              Question Data_Value_Type  \
4   Percent of adults who achieve at least 300 min…           Value
13  Percent of adults who engage in no leisure-tim…           Value
20  Percent of adults who engage in muscle-strengt…           Value
23  Percent of adults who engage in no leisure-tim…           Value
25  Percent of adults who engage in no leisure-tim…           Value

    Data_Value  …                    GeoLocation  ClassID  TopicID  \
4         26.6  …                            NaN       PA      PA1
13        32.3  …  (43.235541343, -108.109830353)       PA      PA1
20        19.8  …   (40.130570048, -74.273691288)       PA      PA1
23        27.9  …                            NaN       PA      PA1
25        16.9  …                            NaN       PA      PA1

    QuestionID DataValueTypeID LocationID StratificationCategory1  \
4        Q045           VALUE         59                   Income
13       Q047           VALUE         56                   Income
20       Q046           VALUE         34          Race/Ethnicity
23       Q047           VALUE         59                   Gender
25       Q047           VALUE         59              Age (years)

      Stratification1 StratificationCategoryId1 StratificationID1
4   Less than $15,000                      INC         INCLESS15
13  Less than $15,000                      INC         INCLESS15
20              Other                     RACE           RACEOTH
23             Female                      GEN            FEMALE
25           18 - 24                     AGEYR         AGEYR1824

[5 rows x 30 columns]
```

[13]:
```
ObesityWeightStatus.head()
```

[13]:
```
   YearStart  YearEnd LocationAbbr          LocationDesc  \
2       2013     2013           US              National
3       2013     2013           US              National
6       2012     2012           WY               Wyoming
7       2012     2012           DC  District of Columbia
9       2011     2011           AL               Alabama
```

```
                                    Datasource                    Class  \
2  Behavioral Risk Factor Surveillance System  Obesity / Weight Status
3  Behavioral Risk Factor Surveillance System  Obesity / Weight Status
6  Behavioral Risk Factor Surveillance System  Obesity / Weight Status
7  Behavioral Risk Factor Surveillance System  Obesity / Weight Status
9  Behavioral Risk Factor Surveillance System  Obesity / Weight Status

                    Topic                                          Question  \
2  Obesity / Weight Status  Percent of adults aged 18 years and older who …
3  Obesity / Weight Status  Percent of adults aged 18 years and older who …
6  Obesity / Weight Status  Percent of adults aged 18 years and older who …
7  Obesity / Weight Status  Percent of adults aged 18 years and older who …
9  Obesity / Weight Status  Percent of adults aged 18 years and older who …

  Data_Value_Type  Data_Value  …                    GeoLocation  ClassID  \
2           Value        28.8  …                            NaN      OWS
3           Value        32.7  …                            NaN      OWS
6           Value        48.5  …  (43.235541343, -108.109830353)      OWS
7           Value        31.6  …   (38.890371385, -77.031961127)      OWS
9           Value        35.2  …   (32.840571122, -86.631860762)      OWS

   TopicID  QuestionID DataValueTypeID LocationID StratificationCategory1  \
2     OWS1        Q036           VALUE         59                   Income
3     OWS1        Q037           VALUE         59                   Income
6     OWS1        Q037           VALUE         56          Race/Ethnicity
7     OWS1        Q036           VALUE         11               Education
9     OWS1        Q036           VALUE          1             Age (years)

                Stratification1 StratificationCategoryId1 StratificationID1
2             $50,000 - $74,999                       INC           INC5075
3             Data not reported                       INC             INCNR
6  American Indian/Alaska Native                      RACE           RACENAA
7          Less than high school                       EDU             EDUHS
9                        25 - 34                     AGEYR         AGEYR2534

[5 rows x 30 columns]
```

[14]:
```
FruitsVegBehavior.head()
```

[14]:
```
     YearStart  YearEnd LocationAbbr   LocationDesc  \
22        2017     2017           WA     Washington
75        2017     2017           RI   Rhode Island
80        2017     2017           MA  Massachusetts
280       2017     2017           HI         Hawaii
342       2017     2017           NM     New Mexico
```

```
                                    Datasource                  Class  \
22   Behavioral Risk Factor Surveillance System  Fruits and Vegetables
75   Behavioral Risk Factor Surveillance System  Fruits and Vegetables
80   Behavioral Risk Factor Surveillance System  Fruits and Vegetables
280  Behavioral Risk Factor Surveillance System  Fruits and Vegetables
342  Behavioral Risk Factor Surveillance System  Fruits and Vegetables

                              Topic  \
22   Fruits and Vegetables - Behavior
75   Fruits and Vegetables - Behavior
80   Fruits and Vegetables - Behavior
280  Fruits and Vegetables - Behavior
342  Fruits and Vegetables - Behavior

                                          Question Data_Value_Type  \
22   Percent of adults who report consuming fruit l…           Value
75   Percent of adults who report consuming vegetab…           Value
80   Percent of adults who report consuming fruit l…           Value
280  Percent of adults who report consuming vegetab…           Value
342  Percent of adults who report consuming fruit l…           Value

     Data_Value  …                    GeoLocation  ClassID  TopicID  \
22         36.1  …   (47.522278629, -120.47001079)       FV      FV1
75         26.0  …   (41.708280193, -71.522470314)       FV      FV1
80         29.9  …    (42.27687047, -72.082690675)       FV      FV1
280        17.1  …  (21.304850435, -157.857749403)       FV      FV1
342        41.2  …  (34.520880952, -106.240580985)       FV      FV1

     QuestionID DataValueTypeID LocationID StratificationCategory1  \
22         Q018           VALUE         53                  Income
75         Q019           VALUE         44                  Income
80         Q018           VALUE         25          Race/Ethnicity
280        Q019           VALUE         15             Age (years)
342        Q018           VALUE         35                  Gender

       Stratification1 StratificationCategoryId1 StratificationID1
22   Less than $15,000                      INC         INCLESS15
75   $25,000 - $34,999                      INC           INC2535
80            Hispanic                     RACE           RACEHIS
280            45 - 54                    AGEYR         AGEYR4554
342               Male                      GEN              MALE

[5 rows x 30 columns]
```

1 paragraph of the ethical implications of data wrangling specific to your datasource and the steps you completed answering the following questions: What changes were made to the data? Are there any legal or regulatory guidelines for your data or project topic? What risks could be created based on the transformations done? Did you make any assumptions in cleaning/transforming the data?

How was your data sourced / verified for credibility? Was your data acquired in an ethical way? How would you mitigate any of the ethical implications you have identified?

0.1.1 what changes were made to the data

the following transfermation were made to the data: - transformation 1: remove puerto rico,virgin islands,guam from data set - transformation 2: remove columns with only null values - transformation 3: remove irrelevant data(data value null values) - transformation 4: eliminate data from years(2020,2021,2022) due to it being skewed due to the covid19 pandemic - transformation 5: break data into smaller sets based on class(may join together again later to make a wider data set but we will see how our other data joins in before making this determination

0.1.2 what risks could be created based on these transformations

With any data manipulation, we risk losing some of the "story" associated. For instance, we are losing that COVID data, which is good data, but where we are looking at trend data over time, it may have skewed the perception of the overall data trend.

0.1.3 did you make any assumptions during cleaning/transforming the data

I did make assumptoins on what fields would be relevant for the data and the assumption that some of my other data sources would likely not include Puerto Rico, the Virgin Islands, and Guam.

0.1.4 how was the data sourced

This data was sourced from the Center for Disease Control(CDC) ### was the data acquired in an ethical way I am unable to locate information on how this data was gathered ### how would these ethical implications be mitigated I am hopeful that there were ethical practices in place when gathering the data; everything in this data source is aggregated and anonymized so that helps from zeroing out at the individual level. This helps to already mitigate some of the ethical risks with this dataset. I feel that overall, this data, to my knowledge, is within a degree ethical already, though it's hard to be certain without knowing how the data was gathered.

[ ]: