Alex Ptacek

Data 602 - Final Project Proposal

1. Research Question

Is there a relationship between obesity and demographics? If so, how has this changed over time? What other factors affect obesity rates? The dataset I will be using for this project contains predicted U.S. yearly obesity rates by State, as well as diet and physical activity behaviors. The data also contains demographic info, such as race, gender, age, income, and more. I will be checking all variables for potential relationships with obesity.

2. Justification - why is this relevant to you or industry?

Overall, I believe this study may yield useful information for two groups: certain companies and all individuals. At the individual level, it is not only interesting, but also important for everyone to better understand risk factors for obesity, which itself is known as a significant risk factor for many health conditions. This study may also inform individuals about potential changes they can make to their lifestyle to reduce risk of obesity.

This study may also be valuable to certain companies, particularly in health-related industries. For example, fitness companies (e.g. gyms, workout gear, health blogs, etc.) may want to target people who are struggling with obesity, and the results of this study may help them piece together and understand those audiences.

3. Data Sources - did you find this data online or collect yourself? Provide links.

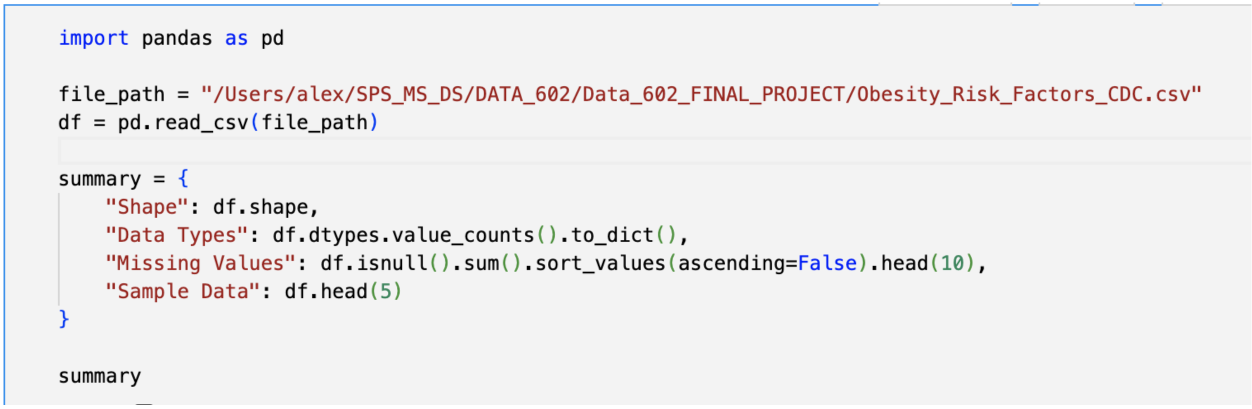
This data was published by the U.S. Centers for Disease Control and Prevention (CDC) and can be found on the DATA.gov website:

https://catalog.data.gov/dataset/nutrition-physical-activity-and-obesity-behavioral-risk-factor-surveillance-system

4. Libraries potentially being used.

- Numpy & Pandas: These packages are used together for data wrangling in a tabular structure, which will be integral for this project.
- Matplotlib: This will be important for data visualizations.
- Scikit-learn: This project may include some machine learning models, if appropriate.

5. EDA and summary statistics.



```
import pandas as pd

file_path = "/Users/alex/SPS_MS_DS/DATA_602/Data_602_FINAL_PROJECT/Obesity_Risk_Factors_CDC.csv"
df = pd.read_csv(file_path)

summary = {
    "Shape": df.shape,
    "Data Types": df.dtypes.value_counts().to_dict(),
    "Missing Values": df.isnull().sum().sort_values(ascending=False).head(10),
    "Sample Data": df.head(5)
}

summary
```

Output:

```
{'Shape': (104272, 33),
 'Data Types': {dtype('O'): 24, dtype('float64'): 6, dtype('int64'): 3},


'Missing Values': Total                  100548
 Sex                    96824
 Data_Value_Footnote         93505
 Data_Value_Footnote_Symbol    93505
 Education                89376
 Age(years)               81928
 Income                 78204
 Race/Ethnicity             74480
 Data_Value_Unit            15400
 Low_Confidence_Limit        10767
 dtype: int64,



 'Sample Data':    YearStart  YearEnd LocationAbbr LocationDesc Datasource  \
 0                        2011    2011       AK                    Alaska     BRFSS
 1                        2011    2011       AK                    Alaska     BRFSS
 2                        2011    2011       AK                    Alaska     BRFSS
 3                        2011    2011       AK                    Alaska     BRFSS
 4                        2011    2011       AK                     Alaska     BRFSS
```

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

                               Question                         Data_Value_Unit  \
0  Percent of adults aged 18 years and older who ...        2011.0
1  Percent of adults aged 18 years and older who ...        2011.0
2  Percent of adults who achieve at least 150 min...        2011.0
3  Percent of adults aged 18 years and older who ...        2011.0
4  Percent of adults aged 18 years and older who ...        2011.0

  Data_Value_Type  ...                  GeoLocation               ClassID TopicID  \
0        Value  ...  (64.845079957001, -147.722059036)    OWS    OWS1
1        Value  ...  (64.845079957001, -147.722059036)    OWS    OWS1
2        Value  ...  (64.845079957001, -147.722059036)     PA    PA1
3        Value  ...  (64.845079957001, -147.722059036)    OWS    OWS1
4        Value  ...  (64.845079957001, -147.722059036)    OWS    OWS1

  QuestionID  DataValueTypeID  LocationID  StratificationCategory1  \
0     Q036         VALUE        2        Race/Ethnicity
1     Q036         VALUE        2        Race/Ethnicity
2     Q044         VALUE        2                Sex
3     Q036         VALUE        2           Age (years)
4     Q037         VALUE        2               Income

    Stratification1 StratificationCategoryId1 StratificationID1
0   2 or more races                  RACE       RACE2PLUS
1          Other                 RACE        RACEOTH
2         Female                  SEX         FEMALE
3         35 - 44                AGEYR       AGEYR3544
4  $15,000 - $24,999                  INC         INC1525

[5 rows x 33 columns]}
```