Data Analysis

Kelly Im, Brian Wang, Alvin Zou

Slide 1: Datasets

1. One option here is to use data from 2011-2020 only and avoid missing years in the prior year data. Will that work? If not, why?
    a. Previously, we had come to the decision to exclude data from the years 2001 and 2011 because they did not have the response variable we needed. The datasets we would use were from the years 2003, 2005, 2007, 2009, and 2012-2020. However, after taking into consideration the issue we had with inconsistent data, we narrowed it down to the years 2012-2020.

2. How many records per year when it comes to 2011-2020? This information can further justify the rationale of whether can we drop the earlier years or not?
    a. In the remaining years we have selected (2012-2020), we have the following dimensions:
    CHIS 2012: (20355, 346)
    CHIS 2013: (20724, 340)
    CHIS 2014: (19516, 397)
    CHIS 2015: (21034, 373)
    CHIS 2016: (21055, 407)
    CHIS 2017: (21153, 405)
    CHIS 2018: (21177, 413)
    CHIS 2019: (22160, 478)
    CHIS 2020: (21949, 523)
    b. As of now, we have 189123 rows and 815 columns.

3. Would it be possible that the same person appears in different year? Have you read into the definition of each data field? What are your thoughts?
    a. While it is possible for the same person to appear it different years, we are unable to determine this fact. The only identifying marker in the dataset is a “Public Use File ID”. However, this ID is reset every year, so we cannot determine any consistency across the years.

Slide 2: Inconsistent Variables

1. Any descriptive statistics about these missing values? (i.e. how many are missing, what's the % of missing values compare to the records which have the values?)
    a. Example 1: The number of variables included in 2012, but not the subsequent years, for each subsequent year is as follows - 68 (2013), 68 (2014), 92 (2015), 95 (2016), 108 (2017), 129 (2018), 153 (2019), and 151 (2020).
    b. Based on this single year, we can see that there is already a large inconsistency in variables between 2012 and other years. (There were 68 variables present in 2012 that were not included in 2013).
    c. Example 2: The number of variables included in 2013, but not the other years, for each year is as follows - 62 (2012), 7 (2014), 46 (2015), 48 (2016), 75 (2017), 96 (2018), 122 (2019), and 121 (2020).
    d. Google Sheet Link: To view all of the inconsistent variables across the years https://docs.google.com/spreadsheets/d/1eQFsYVO4fn1O012aNNufrbGJ3fkJyezJ9PfNTJjyICI/edit?usp=sharing

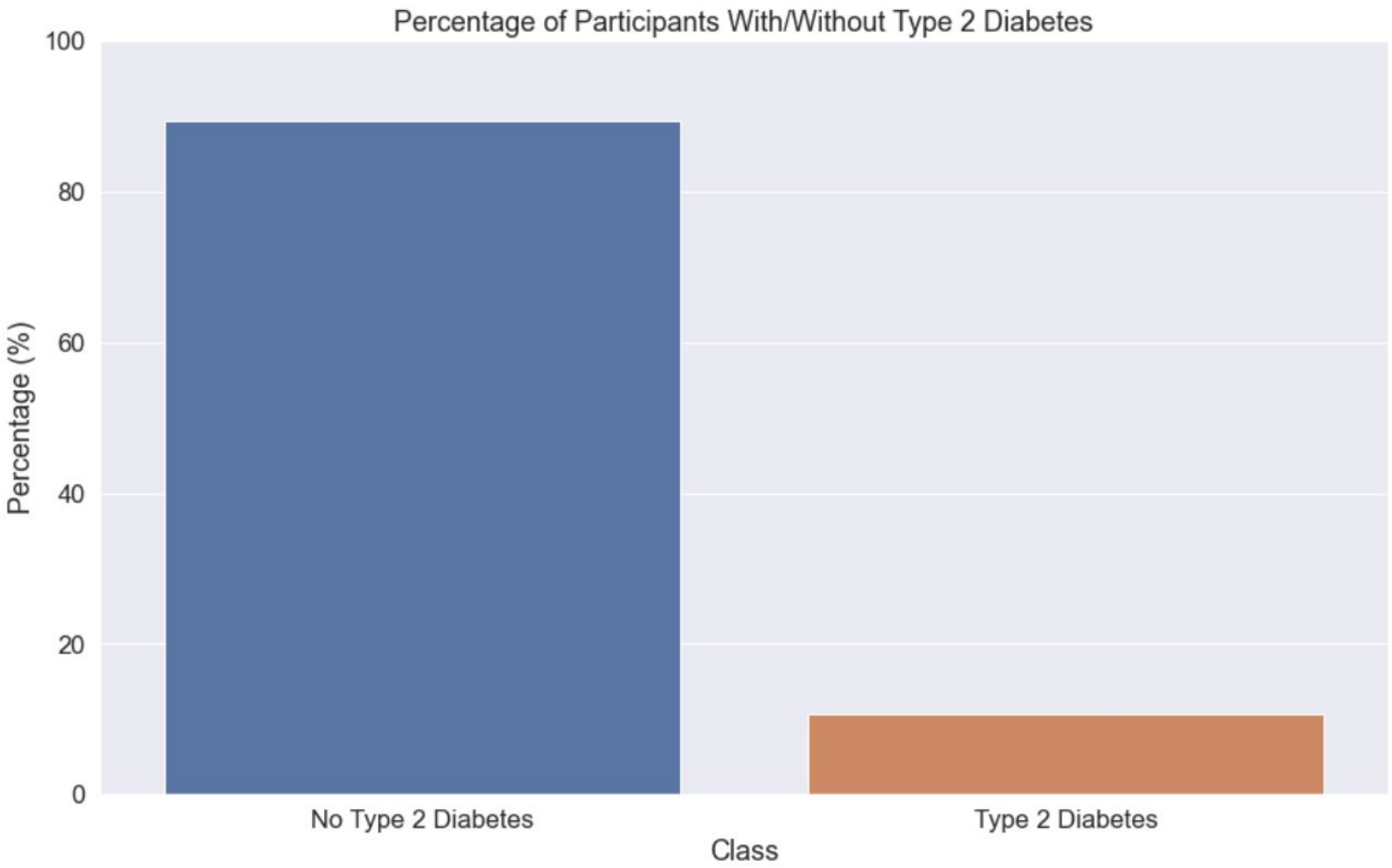
2. Is there any cutoffs to decide whether or not to include the variables?
    a. We have come to the decision that if a variable is not present in 3 or more years, the variable will the dropped from the dataset. There are a total of 9 years worth of data and so we chose to have the cutoff be ⅓ of the available data. However, at this time, we have not yet implemented this completely.

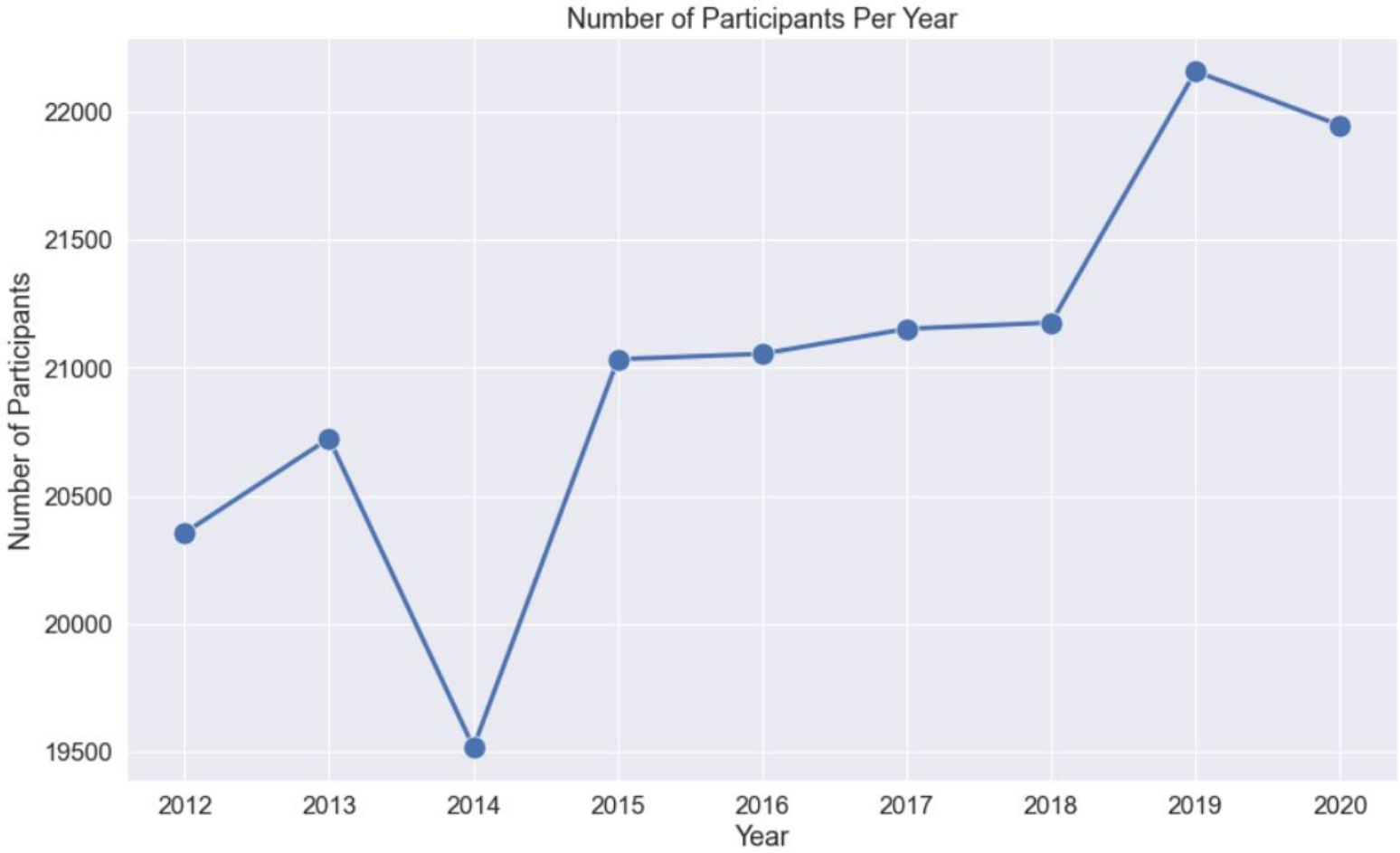
Note: We have also determined that there are variables that have different names, but the same definition and scaling (or similar enough). For these variables, we have gone through and individually changed the name to match other years. However, there are some that we are not sure if they should be combined due to being in different universes (different age limits). These variables are stated to be not trendable with other years. Other variables are scaled too differently (where some categories are lost or combined with others) and we have determined that they shouldn't be combined, despite having the same variable definition. For those that are scaled differently, but can be altered to match other years, we will be combining them and changing the values to match the scales from other years.

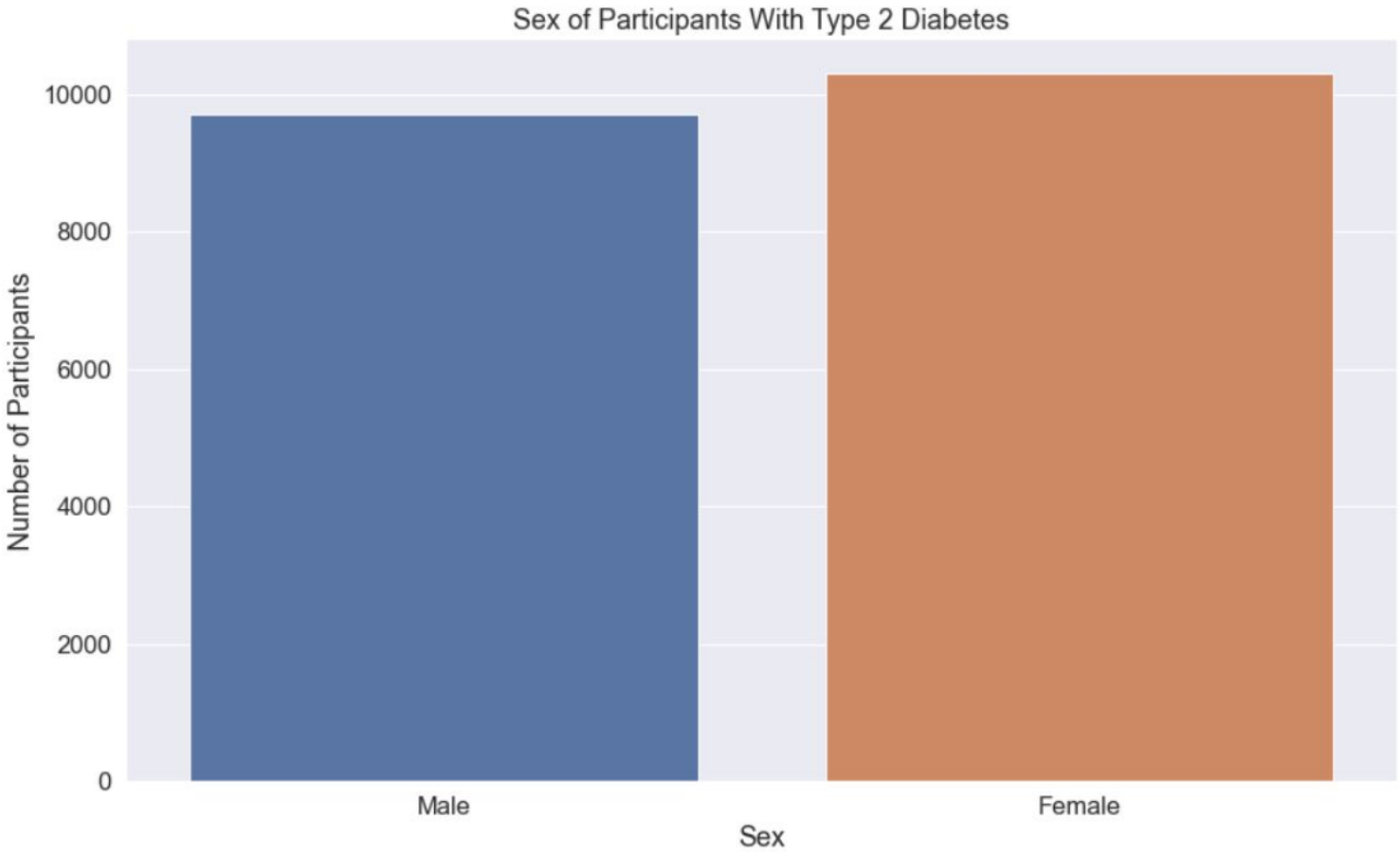
Slide 3: Class Imbalance

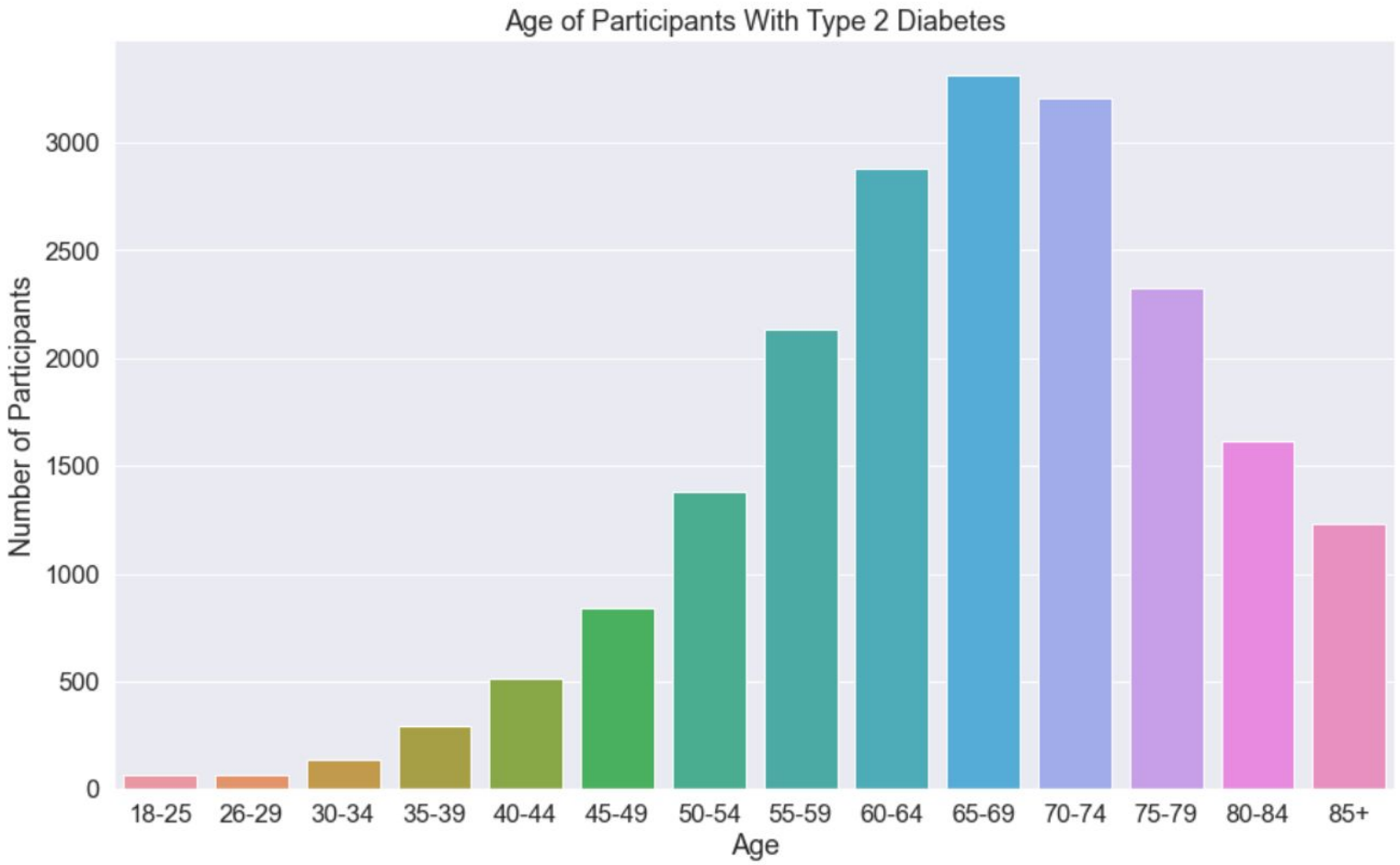
1. Are mentioning the fact that you have lot less data for the people with diabetes? If that's the case, can we use the data to predict "no diabetes" since you have plenty of data available there?

- 169,128 surveyees with no Type 2 diabetes
- 19,995 surveyees with Type 2 diabetes


Percentage of Participants With/Without Type 2 Diabetes
100
80
60
40
20
0
Percentage (%)
No Type 2 Diabetes
Type 2 Diabetes
Class


Number of Participants Per Year
Number of Participants
22000
21500
21000
20500
20000
19500
2012
2013
2014
2015
2016
2017
2018
2019
2020
Year


Sex of Participants With Type 2 Diabetes
Number of Participants
10000
8000
6000
4000
2000
0
Male
Female
Sex


Age of Participants With Type 2 Diabetes
Number of Participants
0
500
1000
1500
2000
2500
3000
18-25
26-29
30-34
35-39
40-44
45-49
50-54
55-59
60-64
65-69
70-74
75-79
80-84
85+
Age

PostgreSQL Databases

1. We have created multiple databases according to the different survey questionnaire sections.
    a. Screening Information (Ex: Household Size, Language of Interview, Proxy Interview, etc.)
    b. Demographic Information, Part I (Ex: Age, Gender, Race, Ethnicity, Marital Status, etc.)
    c. General Health Condition (Ex: Asthma, Diabetes, High Blood Pressure, Heart Disease, etc.)
    d. Health Behaviors (Ex: # Times Ate Fruit, # Times Ate Vegetables, # of Cigarettes per day, etc.)
    e. General Health, Disability & Sexual Health (Ex: Height, Weight, BMI, Type of Birth Control, etc.)
    f. Women's Health (Ex: Currently Pregnant)
    g. Mental Health (Ex: Feeling Nervous, Hopeless, Depressed, Restless, Chore Impairment, etc.)
    h. Demographic Information, Part II & Child Care (Ex: Citizen, English Proficiency, Education, etc.)
    i. Health Insurance (Ex: Covered by Medicare/Medi-cal, Covered for Prescription Drugs, etc.)
    j. Health Care Utilization/Access & Dental Health (Ex: Usual Source of Care, # Doctor Visits, etc.)
    k. Employment, Income, Poverty Status & Food Security (Ex: Work Status, # Hours Worked, etc.)
    l. Public Program Participation (Ex: Receiving TANF/CALWORKS, Food Stamps, SSI, Pension, etc.)
    m. Housing & Community Involvement (Ex: Currently Paying Off Mortgage, Active Member in Community, Own or Rent Home, Feeling Safe in Neighborhood, etc.)
    n. Demographic Information, Part III, Geographic Information (Ex: Rural or Urban)
    o. Voter Engagement (Ex: Frequency of Voting, Voter Engagement)