**Data Preparation Documentation**

**1.** **Description of Data Source**

The dataset we use originates from the Centers for Disease Control and Prevention of the US Department of Health and Human Services. The title of our dataset is Nutrition, Physical Activity, and Obesity - Behavioral Risk Factor Surveillance System.  This dataset includes national and state specific annual data between the years of 2011 to 2015 on adults aged 18 years or older with subcategories including gender, race, education and income. In particular, the dataset reports the rates of obesity, overweight status, fruit and vegetable consumption beyond a certain threshold, and exercise being a certain intensity level, on all of these subcategories over 5 years for each state.

Nutrition, Physical Activity, and Obesity - Behavioral Risk Factor Surveillance System. (2017,

October 13). Retrieved October 30, 2017, from https://chronicdata.cdc.gov/api/views/hn4x-zwk7/rows.csv?accessType=DOWNLOAD.

**2. Intellectual Property and Licensing**

The Data User Agreement of the Centers for Disease Control and Prevention focuses on protecting the privacy of people surveyed for its data, stating that any effort to determine the identity of any reported case is prohibited by law. By using this data, we agree to use it only for statistical reporting and analysis purposes, make no use of inadvertently discovered identities of persons and inform the Director of the National Center for Health Statistics of any such discovery, and not link the dataset with personally identifying information from other datasets. The details of the license and terms of used can be accessed from the link below:Retrieved October 30, 2017, from https://www.cdc.gov/nchs/data\_access/restrictions.htm.

**3. Metadata**

The metadata available on the data download page includes a quick snapshot of the size of the dataset - 53,400 rows and 33 columns, and lists the columns with their data types. There is also contact information available for the owner of the dataset. The dataset includes percentages of adults who are overweight and percentages of adults who are obese. Initially we were unclear on whether “obese” was a subcategory of “overweight” or its own separate category. An article on the CDC’s website states that overweight is an Adult Body Mass Index (BMI) of 25.0 to < 30, and obese is a BMI of 30.0 or higher (CDC 2016), clarifying the definitions of these terms in the dataset as well as the fact that they are separate categories. This data comes from a much larger survey, the Behavioral Risk Favor Surveillance System (BRFSS). The survey overview (CDC 2017) describes the entire survey collection process, from how the questions are formulated each year, to how sample households are chosen, to how interviews are conducted, to how the data is processed. The data pertains to non-institutionalized adults in the United States who are 18 or older, and responses are obtained from landline and cellular telephone interviews that last an average of 18 minutes. BRFSS uses iterative proportional fitting, or raking, to adjust for demographic differences between respondents and the population they represent. The CDC’s Division of Nutrition, physical Activity, and Obesity takes information from the BRFSS to create this more specific dataset, to provide national and state specific data on obesity, nutrition, physical activity, and breastfeeding.

Centers for Disease Control and Prevention (CDC). (Updated 2016, June 16). Defining Adult

Overweight and Obesity. Retrieved October 30, 2017, from https://www.cdc.gov/obesity/adult/defining.html.

Centers for Disease Control and Prevention (CDC). (2017, June 29). Overview: BRFSS 2016.

Retrieved October 30, 2017, from https://www.cdc.gov/brfss/annual\_data/2016/pdf/overview\_2016.pdf.

**4 & 5. Issues with Data and Rationale for Steps to Resolve**

There were some issues with the dataset that we encountered during the cleaning process. For instance, the Age(years), Education, Gender, Income, and Race/Ethnicity columns are unstandardized. Some of the rows are empty, some are not. We thought that it would be better if we combine the columns together. Then, we realized that the unit of observation (row) in this dataset is “state,” which means that we do not have individual level data. This style of data reporting prevents us from observing the joint distribution of subcategories and the rates of, for example, obesity or high-level fruit consumption. In other words, we can look at the obesity rates among females (for all years and states) but we cannot look at the obesity rates for females who are high school graduates. For this reason, our analyses will be based on the relationship of *one* survey question to *one* of the subcategories at a time. For example, we can look at the relationship between income levels and obesity rates across all the states for the years of 2011 to 2016, but we cannot see the same relationship (between income levels and obesity rates) for non-Hispanic white females. Consequently, we decided not to merge all of the columns into one as if the data are reported coming from one individuals’ survey responses or in other words as if the individual is the unit of observation.

The data set also has some missing values. We observe the statement of “Data not available because sample size is insufficient” with respect to race/ethnicity survey question (4603 of 48772 records). When we looked for “Non-Hispanic White” with “Data not available because sample size is insufficient”, all the 14 insufficient data records were from Puerto Rico. When we run the same analysis for “Non-Hispanic Black” with insufficient data, the 368 records were from differing states where non-Hispanic Blacks are typically not high in population. These missing values do not concern our analyses because the metadata clearly states a careful methodology for choosing a representative sample population, so we trust that the values marked insufficient sample size will not impact the accuracy of our overall conclusions. The column “Data\_Value\_Footnote\_Symbol” has “~ “ values on all the rows which have insufficient sample size. Also, the column “DataValueTypeID” can be deleted since the word “value” is entered for all the records.

**6. Step-by-Step Data Cleaning Process**

**Cleaning / Deleting / Renaming the Columns**

1-# import the dataset to RStudio by running this script:

obesity <- read\_csv("~/Zeynep/~UMD\_Grad/INFM600/obesity.csv")

2- View(obesity)

3- # check if the columns YearStart and YearEnd are the same (find the difference)

obesity$year\_start\_end\_diff <- obesity$YearEnd - obesity$YearStart

unique(obesity$year\_start\_end\_diff)

4- # delete column YearEnd

obesity$YearEnd <- NULL

5- # rename column YearStart as Year

names(obesity)[names(obesity) == 'YearStart'] <- 'Year'

6- # delete column LocationAbbr is the abbreviations of the state names which is unnecessary information. There is another column called “LocationDesc” which is the state names that we want to use and keep.

obesity$LocationAbbr <- NULL

7- # delete column Datasource. We know that all the entries are the same, and we know that our data source is survey.

obesity$Datasource <- NULL

8- # delete column Class. Class had three sub-categories (Fruits and vegetables, Obesity/Weight Status, Physical Activity) since we are not interested in Class, we are interested in specific questions.

obesity$Class <- NULL

9- # delete column Topic. Topic had three sub-categories (Fruits and vegetables, Obesity/Weight Status, Physical Activity) since we are not interested in Topic, we are interested in specific questions.

obesity$Topic <- NULL

10-  # check if the values are same for Data\_Value and Data\_Value\_Alt

obesity$data\_value\_diff <- obesity$Data\_Value - obesity$Data\_Value\_Alt

unique(obesity$data\_value\_diff)

11- # the values are same for Data\_Value and Data\_Value\_Alt; so, delete Data\_Value\_Alt

obesity$Data\_Value\_Alt <- NULL

12- View(obesity)

13- # check if there is any data other than "Value" in Data\_Value\_Type

unique(obesity$Data\_Value\_Type)

14- # there is no data other than  "Value" in Data\_Value\_Type; so, delete column Data\_Value\_Type

obesity$Data\_Value\_Type <- NULL

15- # check if there is any useful data in Data\_Value\_Unit, Data\_Value\_Footnote, and Data\_Value\_Footnote\_Symbol by running the following scripts:

unique(obesity$Data\_Value\_Unit)

unique(obesity$Data\_Value\_Footnote)

unique(obesity$Data\_Value\_Footnote\_Symbol)

Or this:

unique(c(obesity$Data\_Value\_Unit, obesity$Data\_Value\_Footnote, obesity$Data\_Value\_Footnote\_Symbol))

16- # there is no useful data in Data\_Value\_Unit, Data\_Value\_Footnote, and Data\_Value\_Footnote\_Symbol; so, delete them all by using the following scripts:

obesity$Data\_Value\_Unit <- NULL

obesity$Data\_Value\_Footnote <- NULL

obesity$Data\_Value\_Footnote\_Symbol <- NULL

17- # delete all of the following columns since we are not going to use them for our analyses, and they are repetition of other entries: GeoLocation, ClassID, TopicID, QuestionID, DataValueTypeID, LocationID, StratificationCategory1, Stratification1, StratificationCategoryId1, and StratificationID1.

obesity$GeoLocation <- NULL

obesity$ClassID <- NULL

obesity$TopicID <- NULL

obesity$QuestionID <- NULL

obesity$DataValueTypeID <- NULL

obesity$LocationID <- NULL

obesity$StratificationCategory1 <- NULL

obesity$Stratification1 <- NULL

obesity$StratificationCategoryId1 <- NULL

obesity$StratificationID1 <- NULL

18- Also # delete the following columns since we will not need them for our analyses: Low\_confidence\_Limit, High\_confidence\_Limit, Sample\_Size, Total, `Age(years)`, Education, Gender, `Race/Ethnicity`, and year\_start\_end\_diff

obesity$Low\_Confidence\_Limit <- NULL

obesity$High\_Confidence\_Limit <- NULL

obesity$Sample\_Size <- NULL

obesity$Total <- NULL

obesity$`Age(years)` <- NULL

obesity$Education <- NULL

obesity$Gender <- NULL

obesity$`Race/Ethnicity` <- NULL

obesity$year\_start\_end\_diff <- NULL

**Cleaning and Deleting the Rows**

19- Find all the rows # not including “Data not reported”, and delete those rows.

obesity <- obesity[!(obesity$Income %in% c("Data not reported")),]

20- View(obesity)

21- Find all the rows # not including NA, and delete those rows.

obesity <- obesity[!(obesity$Income %in% c(NA)),]

22- Find rows # including only “Percent of adults aged 18 years and older who have obesity”, and keep those rows.

obesity <- obesity[(obesity$Question %in% c("Percent of adults aged 18 years and older who have obesity")),]

Word Count (not including citations or word count): 1,423