TITLE: PHYSICAL ACTIVITY IN ADULTS AS A SOCIAL DETERMINANT OF HEALTH

MID-PROJECT REPORT

DATA WRANGLING AND EXPLORATORY DATA ANALYSIS (CAP5320)

DEPARTMENT OF DATA SCIENCE & BUSINESS ANALYTICS

FLORIDA POLYTECHNIC UNIVERSITY

BY

JOSHUA OLABISI

SPRING 2020

Table of Contents

LIST OF FIGURES

INTRODUCTION

1.1 Definition

Physical activity is defined as any bodily movement produced by contraction of skeletal muscles that increases energy expenditure above resting levels and comprises routine daily tasks such as commuting, occupational tasks, or household activities, as well as purposeful health-enhancing movements/activities (Diaz & Shimbo, 2013). It is important to note that physical activity is a first-line prevention or treatment in almost all diagnoses, but especially chronic diseases, including Cardiovascular Disease, Diabetes, Hypertension, Cancers, and Mental illness — accounting for most of the negative health outcomes and healthcare spending.

1.2 Methodology

The motivation for this project is to help identify the physical activities as it relates to the social determinants of health. According to the World Health Organization (WHO), the social determinants of health (SDH) are the conditions in which people are born, grow, work, live, and age, and the wider set of forces and systems shaping the conditions of daily life (*WHO | Social determinants of health*, n.d.). The goal of this project is to merge, clean, and analyze the available data on social determinants of health particularly for physical activities in adults.

DATA COLLECTION AND VISUALIZATION

The data used for this project was collected from the Center for Disease Control and Prevention (CDC). The title of the dataset is "Nutrition, Physical Activity, and Obesity; the data source is the Behavioral Risk Factor Surveillance System (BRFSS)". The dataset includes data on adult's diet, physical activity, and weight status from Behavioral Risk Factor Surveillance System. This data is used for Division of Nutrition, Physical Activity, and Obesity (DNPAO) Data, Trends, and Maps database, which provides national and state specific data on obesity, nutrition, physical activity, and breastfeeding. However, for the purpose of this project, the dataset was limited to data on physical activities levels for adults in Florida from 2011 to 2017.

YearStart	YearEnd	LocationAbb	LocationDesc	Datasource	Class	Topic	Question	Data_Value_Unit
2011	2011	FL	Florida	Behavioral Ri	Physical Acti	Physical Acti	Percent of adults who enga	ge in muscle-strengthening activities o
2011	2011	FL	Florida	Behavioral Ri	Physical Acti	Physical Acti	Percent of adults who achieve at least 150 minutes a week of mo	
2011	2011	FL	Florida	Behavioral Ri	Physical Acti	Physical Acti	Percent of adults who achieve at least 150 minutes a week of mo	
2011	2011	FL	Florida	Behavioral Ri	Physical Acti	Physical Acti	Percent of adults who engage in no leisure-time physical activity	
2011	2011	FL	Florida	Behavioral Ri	Physical Acti	Physical Acti	Percent of adults who achieve at least 150 minutes a week of mo	
2011	2011	FL	Florida	Behavioral Ri	Physical Acti	Physical Acti	Percent of adults who engage in no leisure-time physical activity	
2011	2011	FL	Florida	Behavioral Ri	Physical Acti	Physical Acti	Percent of adults who achieve at least 150 minutes a week of mo	
2011	2011	FL	Florida	Behavioral Ri	Physical Acti	Physical Acti	Percent of adults who achieve at least 300 minutes a week of mo	
2011	2011	FL	Florida	Behavioral Ri	Physical Acti	Physical Acti	Percent of adults who engage in no leisure-time physical activity	
2011	2011	FL	Florida	Behavioral Ri	Physical Acti	Physical Acti	Percent of adults who achieve at least 150 minutes a week of mo	
2011	2011	FL	Florida	Behavioral Ri	Physical Acti	Physical Acti	Percent of adults who engage in muscle-strengthening activities o	
2011	2011	FL	Florida	Behavioral Ri	Physical Acti	Physical Acti	Percent of adults who achieve at least 300 minutes a week of mo	
2011	2011	FL	Florida	Behavioral Ri	Physical Acti	Physical Acti	Percent of adults who achieve at least 300 minutes a week of mo	
2011	2011	FL	Florida	Behavioral Ri	Physical Acti	Physical Acti	Percent of adults who achieve at least 150 minutes a week of mo	
2011	2011	FL	Florida	Behavioral Ri	Physical Acti	Physical Acti	Percent of adults who engage in muscle-strengthening activities o	
2011	2011	FL	Florida	Behavioral Ri	Physical Acti	Physical Acti	Percent of adults who achieve at least 300 minutes a week of mo	
2011	2011	FL	Florida	Behavioral Ri	Physical Acti	Physical Acti	Percent of adults who engage in muscle-strengthening activities o	
2011	2011	FL	Florida	Behavioral Ri	Physical Acti	Physical Acti	Percent of adults who achieve at least 150 minutes a week of mo	
2011	2011	FL	Florida	Behavioral Ri	Physical Acti	Physical Acti	Percent of adults who achieve at least 300 minutes a week of mo	

Figure 1: Visual of Dataset

2.1 Data Cleaning

Data cleaning is the process of preparing data for analysis by removing or modifying data that is incorrect, incomplete, irrelevant, duplicated, or improperly formatted (*What is Data Cleaning?*, n.d.). The initial dataset retrieved from the CDC was made up of about 63 thousand rows and 33 columns with data collected from every state in the United States from 2011 to 2018. In order to simplify the dataset for better analysis, the data was limited to measuring the rate of physical activity levels for adults in Florida. In addition, some of the rows in the dataset had missing information; these rows were deleted to avoid unnecessary lines. Data cleaning is especially important in that, the presence of unnecessary data may hinder the data analysis process or provide inaccurate results.

Value			~	Data not available because sample size is insufficient.					
Value	27.4	27.4		23.8	31.3	938		25 - 34	
Value	41.3	41.3		39.2	43.5	4,261		65 or older	
Value	20.6	20.6		17.7	23.9	1,416			
Value	28.2	28.2		26.8	29.8	8,707			
Value	33.7	33.7		32.3	35.1	10,998	Total		
Value	29.2	29.2		27.8	30.5	11,377	Total		
Value			~	Data not available because sample size is insufficient.					
Value	25.6	25.6		22.3	29.3	1,033			
Value	32.9	32.9		22.8	44.9	105			
Value			~	Data not available because sample size is insufficient.					
Value	35.3	35.3		31.6	39.2	1,434			
Value	15.7	15.7		12.1	20.1	466		18 - 24	
Value	20.4	20.4		16.9	24.4	1,101			Less than hig
Value	57.3	57.3		53.5	61.1	1,422			
Value	28.9	28.9		25.5	32.4	1,456			
Value	25	25		21.7	28.6	1,487			
Value	20	20		17.5	22.9	1,833		45 - 54	
Value			~	Data not available because sample size is insufficient.					

Figure 2: Visual of Unnecessary Data

2.2 Understanding the Data

Furthermore, just as data cleaning is important, it is also very important to understand the dataset in its entirety; that is, all the attributes/columns and rows. Understanding the dataset makes it easier to interpret what each row or column means in order to provide better analysis of the data. One of the challenges during the process of understanding what each attribute of the data means is knowing what each attribute measures. The data dictionary of the dataset did not explain in detail the meaning of each attribute or what they measure. For example, one of the columns in the dataset is the "Data_Value"; this measures the percentage of adults who engage in no leisure-time physical activity in relation to race/ethnicity, age, education, income, etc. The dataset used for this project was available and ready to use on the CDC website without any special requirement to gain access and use the data.

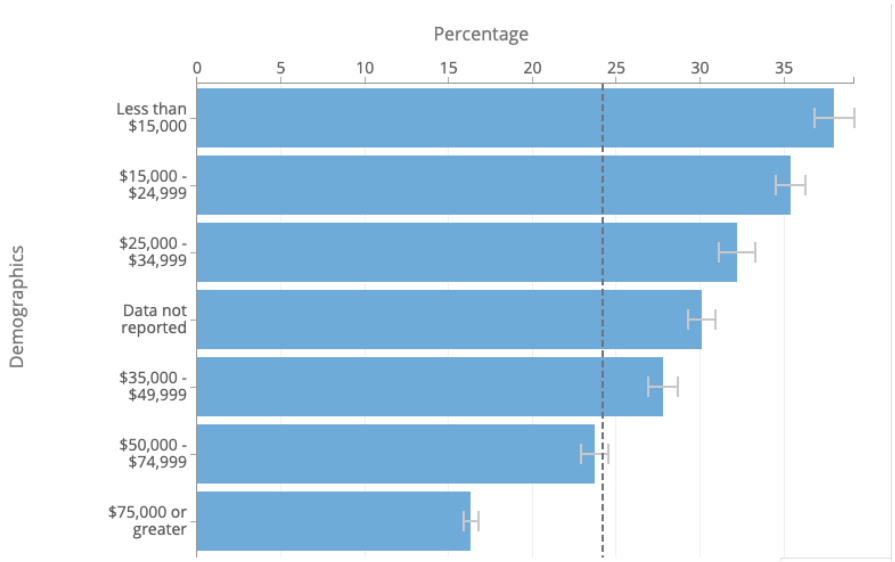
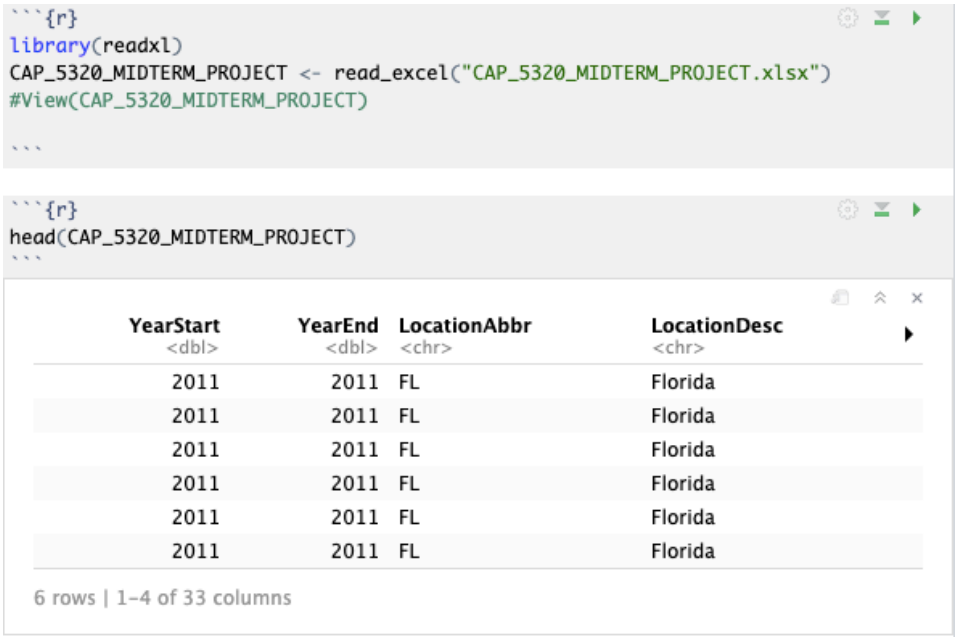


Figure 3: Percentage of adults who engage in no leisure-time physical activity

2.3 R Packages

For this project, the "tidyverse" was the major package used for data analysis, visualization and exploration. The tidyverse is also made up of some core packages such as: "dplyr, tidyr, ggplot2, readr, stringr, tibble, etc." these packages have various important functions. "dplyr provides a grammar of data manipulation, providing a consistent set of verbs that solve the most common data manipulation challenges. tidyr provides a set of functions that helps produce tidy data. Tidy data is data with a consistent form: in brief, every variable goes in a column, and every column is a variable. ggplot2 is a system for declaratively creating graphics, based on The Grammar of Graphics. readr provides a fast and friendly way to read rectangular data (like csv, tsv, fwf, etc.)" (*Tidyverse*, n.d.). For example, the dataset from the CDC is an excel file that needs to be read into R using the "read_excel" function which is included in the readr package.



```{r}
library(readxl)
CAP_5320_MIDTERM_PROJECT <- read_excel("CAP_5320_MIDTERM_PROJECT.xlsx")
#View(CAP_5320_MIDTERM_PROJECT)

```

```{r}
head(CAP_5320_MIDTERM_PROJECT)
```

YearStart <dbl>	YearEnd <dbl>	LocationAbbr <chr>	LocationDesc <chr>
2011	2011	FL	Florida
2011	2011	FL	Florida
2011	2011	FL	Florida
2011	2011	FL	Florida
2011	2011	FL	Florida
2011	2011	FL	Florida

6 rows | 1-4 of 33 columns

Figure 4: 'read_excel' function in R

2.4 Data Visualization using R

In the process of understanding what the "Data_Value" measures, it is important to note that the dataset comprises of various questions that compliment what was being measured when the data was collected. For example, some of the questions posed include: what percentage of adults engage in muscle-strengthening activities on 2 or more days a week? What percentage of adults achieve at least 150 minutes a week of moderate-intensity aerobic physical activity or 75 minutes a week of vigorous-intensity aerobic activity (or an equivalent combination)? What percentage of adults engage in no leisure-time physical activity? In order to effectively answer some of these questions, the following code was used in R to show the percentage of adults who engage in no leisure-time physical activity based on race/ethnicity.

```r
CAP_5320_MIDTERM_PROJECT %>%
  filter(Question == "Percent of adults who engage in no leisure-time physical activity") %>%
  filter(YearEnd == 2017) %>%
  group_by(`Race/Ethnicity`) %>%
  summarise(mean(Data_Value, na.rm = T),
            mean(High_Confidence_Limit, na.rm = T))
```

Race/Ethnicity <chr>	mean(Data_Value, na.rm = T) <dbl>
2 or more races	29.90
American Indian/Alaska Native	40.60
Asian	19.40
Hawaiian/Pacific Islander	NaN
Hispanic	30.60
Non-Hispanic Black	33.50
Non-Hispanic White	28.00
Other	20.40

Figure 5: Percentage of adults who engage in no leisure-time physical activity

2.4.1 Data Visualization using R

In addition to the above code that showed a list of the percentage of adults who engage in no leisure-time activities in relation to race/ethnicity. In order to show a better visual of the results of the codes tested as well as the relationship between the attributes of the dataset, creating a boxplot proved to be very useful. For example, the boxplot below, displays a clear depiction of the relationship between the percentage of adults (male or female) who engage in no leisure-time activities based on their levels of education. This code was used to produce the boxplot:

```r
ggplot(data = CAP_5320_MIDTERM_PROJECT, aes(x = Gender, y = Data_Value, color =
         Education)) +
  geom_boxplot() +
  labs(title = "Percent of adults who engage in no leisure-time physical activity")
```

Figure 6: Comparing gender and educational levels

2.5 CONCLUSION/FUTURE WORK

To this end, social determinants of health are basically the economic and social conditions that influence individual and group differences in health status. It is imperative to understand the importance of physical activity in that helps prevent chronic diseases like cardiovascular diseases and diabetes. Conversely, the goal of focusing on physical activity is to help reduce the cost of healthcare in general and promote a healthier lifestyle. The future work proposed for this project is to simply limit the age group to young adults (18-44 years). As stated earlier, this will help prevent/control the prevalence of chronic diseases in the future and reduce healthcare costs.

References

Diaz, K. M., & Shimbo, D. (2013). Physical Activity and the Prevention of Hypertension. *Current Hypertension Reports*, *15*(6), 659–668. https://doi.org/10.1007/s11906-013-0386-8

Tidyverse. (n.d.). Retrieved February 24, 2020, from https://www.tidyverse.org/packages/

What is Data Cleaning? (n.d.). Sisense. Retrieved February 23, 2020, from https://www.sisense.com/glossary/data-cleaning/

WHO | Social determinants of health. (n.d.). WHO. Retrieved February 23, 2020, from http://www.who.int/social_determinants/en/

Github: https://github.com/jolabisi/Mid-Term-Report.git