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

FINAL 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

[INTRODUCTION 4](#_Toc38840554)

[1.1 DEFINITION 4](#_Toc38840555)

[1.2 METHODOLOGY 4](#_Toc38840556)

[DATA COLLECTION AND VISUALIZATION 5](#_Toc38840557)

[2.1 DATA CLEANING 5](#_Toc38840558)

[2.2 UNDERSTANDING THE DATA 6](#_Toc38840559)

[2.3 R PACKAGES 7](#_Toc38840560)

[2.4 DATA VISUALIZATION USING R 7](#_Toc38840561)

[**2.4.1 DATA VISUALIZATION USING R** 8](#_Toc38840562)

[2.5 CONCLUSION/FUTURE WORK 9](#_Toc38840563)

[EXPLORATORY DATA ANALYSIS 10](#_Toc38840564)

[3.1 INTRODUCTION 10](#_Toc38840565)

[3.2 DATA DESCRIPTION 10](#_Toc38840566)

[CONCLUSION 13](#_Toc38840567)

[REFERENCES 14](#_Toc38840568)

LIST OF FIGURES

Figure Title Page

Figure 1 Visual of Dataset 5

Figure 2 Visual of Unnecessary Data 6

Figure 3 Percentage of adults who engage in no leisure-time physical activity 6

Figure 4 ‘read\_excel’ function in R 7

Figure 5 Percentage of adults who engage in no leisure-time physical activity 8

Figure 6 Comparing gender and educational levels 9

Figure 7 Aggregated Data for Age(years) 10

Figure 8 Filter for Age(years) 11

Figure 9 R code and Bar plot 12

Figure 10 R code and Jitter plot 13

# **INTRODUCTION**

## **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.

A close up of a piece of paper

Description automatically generated

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.

A screenshot of a cell phone

Description automatically generated

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.

A screenshot of a cell phone

Description automatically generated

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.

A screenshot of a cell phone

Description automatically generated

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.

A screenshot of a cell phone

Description automatically generated

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:

**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")**

**![A screenshot of a cell phone

Description automatically generated]()**

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.

# **EXPLORATORY DATA ANALYSIS**

## **3.1 INTRODUCTION**

For this part of the project, it is important to note that there are no changes to the dataset that was used for the wrangling portion of this project. In other words, the same dataset from the Center for Disease Control and Prevention (CDC) which focuses on “Nutrition, Physical Activity, and Obesity levels among adults. As stated during the early stages of this project, the primary focus is understanding the physical activity levels of young adults particularly in Florida. In order to have a more accurate understanding of physical activity levels in young adults; the following are some of the questions asked based on a given sample size. 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? For the purpose of this project, one important research question to ask is “To what extent should physical activities be increased in order to reduce chronic diseases or hospitalization?”

## **3.2 DATA DESCRIPTION**

Although the same dataset used for the wrangling portion of this project, some novel discoveries were made about the data. For example, in the process of creating an exploratory analysis of the dataset, it was determined that the entire dataset is a collection of observations. In other words, the dataset has no combinations and to avoid having a huge amount of data, the dataset was aggregated. Aggregated data refers to numerical or non-numerical information (1) obtained from multiple sources and/or from multiple steps, variables or individuals and (2) compiled in summaries of data or summary reports, usually for public reporting or statistical analysis purposes (Partnership, 2015).

A screenshot of a cell phone

Description automatically generated

Figure 7: Aggregated Data for Age(years)

**3.3 DATA ANALYSIS**

One of the first steps taken in order to provide a quality/efficient analysis of the data is to create data frames for each individual aggregated attribute. The reason for this is to simplify the data and present a more accurate data by eliminating all missing values. It is important to note that in the presence of aggregated rows or columns, it is better to analyze each attribute individually in order to avoid a large number of “NA” values. For the purpose of this project, the filter function was simply used to focus on each individual attribute for a more effective analysis. For example, the filter function was used in R to focus only on the “Age(years)” column. This same function was used to focus on other attributes like “Education, Gender, Income, and Race/Ethnicity”. Although this might seem like the filter function is being used repeatedly, this is the only way to get meaningful data that can be analyzed effectively.

A screenshot of a cell phone

Description automatically generated

Figure 8: Filter for Age(years)

In addition, one of the first analysis that was tested for this project was to determine the average percentage of individuals in a given sample size and also determine how these individuals answer the questions asked in the dataset based on their level of education. One of the questions from the dataset is to determine what percentage of adults engage in no leisure-time physical activity? According to the World Health Organization (WHO), some leisure-time physical activities include the following but not limited to these: “walking, cycling, dancing, gardening, hiking, swimming etc.” (*WHO | Physical Activity and Adults*, n.d.). Engaging in at least 30 to 60 minutes of leisure-time physical activities will help improve cardiorespiratory and muscular fitness, bone health, reduce the risk of cardiovascular diseases, and depression. In order to answer the question posed, a bar plot was used to show relationships between the attributes and the height of the bars simply describes how the question was answered. Based on the results of the bar plot below, it can be concluded that individuals with less than high school do not engage in leisure-time physical activities while college graduates engage in more leisure-time physical activities.

A screenshot of a cell phone

Description automatically generated

A screenshot of a cell phone

Description automatically generated

Figure 9: R code and Bar plot

Furthermore, another important analysis for this project is basically focusing on the physical activity levels of young adults age 18 to 44 years. it is important to focus on this age group because there is a rise in the percentage of physical inactivity among young adults. This can be as a result of various socioeconomic factors such as: economic stability, physical environment, level of education, etc. According to the CDC, the percentage of physical inactivity in the U.S. for young adults age 18 to 44 years is about 19.3% (*CDC, 2019*). This project focuses primarily on young adults in Florida and based on the analysis it can be concluded that percentage of physical inactivity for this same age group is about 22.3%. Based on this result, Florida ranks as the fourth state with the highest percentage of physical inactivity in young adults in the U.S. The figure below shows a visual representation of the percentage of physical activity in young adults in Florida based on different sample sizes using the geom\_point function in R.

A screenshot of a cell phone

Description automatically generated

A screenshot of a computer screen

Description automatically generated

Figure 10: R code and Jitter plot

# **CONCLUSION**

It is imperative that the research question is fully understood in order to effectively interpret the solution and analysis. The research question simply asks, “To what extent should physical activities be increased in order to reduce chronic diseases or hospitalization?” A simple answer to this question is that physical activity levels should be increased by at least 60 minutes weekly. By increasing the weekly levels of physical activities, it can be concluded that the rate of chronic and preventable diseases will greatly decrease. In addition to reducing the rate of chronic diseases, another goal of this project is to focus on increasing physical activity levels in young adults in order to help reduce the cost of healthcare in general and promote a healthier lifestyle.

# 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

*Explore Physical Inactivity in Florida | 2019 Annual Report*. (n.d.). America’s Health Rankings. Retrieved April 25, 2020, from https://www.americashealthrankings.org/explore/annual/measure/Sedentary/population/Sedentary\_18\_44/state/FL

Partnership, G. S. (2015, July 14). *Aggregate Data Definition*. The Glossary of Education Reform. https://www.edglossary.org/aggregate-data/

*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 | Physical Activity and Adults*. (n.d.). WHO; World Health Organization. Retrieved April 23, 2020, from https://www.who.int/dietphysicalactivity/factsheet\_adults/en/

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

Wickham, H., Francois, R., Henry, L., & Müller, K. (2015). dplyr: *A grammar of data manipulation*. R package version 0.4, 3.

Wickham, H. (2016). ggplot2: *Elegant graphics for data analysis*. Springer.

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