Analysis of Recent Diabetes and Obesity Trends

Alex Norris
Department of Mathematics and Statistics, Auburn University, Auburn, AL
oan0003@auburn.edu

Abstract

Within the United States, diabetes and obesity rates have significantly increased over the last two decades. Diabetes affects nearly 35 million Americans today, with about 90% of those cases being type II cases [1]. Data has been accumulated over the years from the CDC, NIH, and other health organizations in order to find patterns or underlying causes for the health crisis. With this data we can apply various statistical methods in order to discover trends and relationships inside the aggregated diabetes and obesity data. Several relationships and trends were identified when analyzing discrepancies across race, age, and gender. By utilizing various statistical and visualization packages in R, these discrepancies are shown throughout the analysis. In the United States, diabetes disproportionately affects the lives of minorities, specifically black and american indian populations. Much of the root cause for diabetes is directly linked with higher rates of obesity. Throughout analysis it can be seen that numerous factors have a strong correlation to the growing diabetes problem, mainly the also growing obesity factor.

I. Introduction

Currently in the United States, chronic health conditions are on the rise. Of the many chronic health conditions, the Centers for Disease Control marks diabetes as a major problem for the United States population. This problem lands close to home, as Alabama lands itself on the list as one of the most diabetes-heavy populations in the entire country. When briefly looking at diabetes rates, it can be noted that along with Alabama, many other southern states also hold higher rates of diabetes than the other sectors of the country. Along with the elevated rates of diabetes comes elevated rates of obesity and an overweight population. Many factors are tracked when finding the cause for diabetes, but obesity is currently the leading risk factor for type II diabetes [2]. For describing medical conditions like diabetes, we first need to provide a proper background to the disease.

When analyzing diabetes, there first must be a line drawn between type I and type II diabetes. Type I diabetes can be seen in a much smaller percentage of total cases, about 5% and is often seen as more severe and uncontrollable. Here, the immune system mistakenly attacks the body's pancreas, the organ responsible for producing insulin. Insulin is the hormone required for controlling glucose in the blood [3]. This type of diabetes is usually developed early on in life and requires lifetime close monitoring of the person's insulin levels. Moving on to the much more common type II diabetes. Type II diabetes usually develops later in a person's life, typically after 45 years old. Unlike type I, the pancreas is still making insulin, but the cells are becoming resistant to its glucose managing ability [3]. However, as the years go by the pancreas does become less effective and stop producing insulin since the body cannot do much with the hormone. Treatment usually starts with diet changes and healthier lifestyles, but can turn into medications for blood sugar regulation and even insulin injections. One very important point to mention is the fact that type II diabetes is typically diagnosed in adults that are obese or overweight. This has led us to believe with a healthy diet and exercise, type II diabetes can be prevented in some ways. With obesity being the leading risk factor for type II diabetes diagnosis,

it is worth investigating why obesity rates are increasing, leading to an increase in diabetes rates as well.

Moving forward, with the knowledge that it is possible to cut down on the type II rates across the county, studies have shown that even with this information, rates are still increasing, especially in the south. The goal of this analysis is to figure out why diabetes rates are going up, specifically within racial groups and southern states. We will conduct further analysis into this problem when comparing rates of diabetes and obesity in states like Alabama to the national average.

The CDC has hundreds of options for acquiring public health data. Most of the data used in this analysis is aggregated time series data collected by the CDC through surveys like the National Health Interview Survey. These surveys consist of numerous questions that require simple yes or no responses. We will be focusing on the aggregated race, gender, and state responses in the National Health Interview Surveys for diabetes and obesity rates, along with exercise amounts.

II. Methods & Analysis

Once all data has been collected, we can check the integrity of the information and see what cleaning procedures must be taken to ensure the elimination of noise and incorrect responses or data entry. Starting with the CDC sponsored U.S Diabetes Surveillance System data [9], we can filter out invalid responses and select the crucial attributes for analysis. This dataset contained 17 attributes, each with a name, type, description, and valid contents. A description document and lookup table are included with the archived data that ranges from 1980 all the way to 2017. First filter out any non useful information before cleaning. Throughout analysis, the Tidyverse and ggpubr packages were utilized for preparing and applying statistical methods to the data.

Starting with the 7th attribute of the data, ‘indicatorid’ has 97 possible values, so we must first narrow down what type of indicator we should analyze. Lets focus on newly diagnosed cases of diabetes as our ‘indicatorid’ of 2. By narrowing down new cases, we can more easily track increasing numbers of diabetes across the country. To filter out everything but newly diagnosed cases from the large dataset, we can use the filter function within the dplyr package that is included with Tidyverse. After the filter we are left with about 800 of our original 29,000 entries. When viewing our new dataset, we can see that there are multiple values used for the ‘estimateid’ attribute. Here we will focus on the estimated rate of diabetes per 1,000 people. After filtering further we can view that there are several null entries in the ‘Estimate’ attribute marked with a “.”. Once the data has been narrowed down and cleaned, let's take a look at diabetes estimates by race nationally. From this survey’s newly diagnosed numbers, only three races are included, so keep that in mind later. We can group all ages together by setting the ‘ageid’ to 99 as well as gender with a ‘genderid’ of 0. The new dataset does not appear to have any null or false values, so we can start by making a scatterplot to view the data.

To begin visualizing the data, we can use the excellent visualization package ggplot that is included within Tidyverse. Start by calling the ggplot function and passing our filtered data archive and the aesthetic settings. Here we bind the years to the x axis, the numeric estimate per 1,000 to the y axis, and the columns to each of the three listed races. In order to plot the data, we add the ‘geom_point()’ argument to get our first scatterplot. The plot shows that the black and hispanic populations have had significantly higher rates of diabetes over the last 20 years. Next we add a regression line for each race with the ‘stat_smooth(method=”lm”)’ argument. With this

linear regression line we can better visualize the fit of the data by compressing the y component down. Finally lets add one more thing to the plot. With the 'stat_cor()' arg we can get our correlation coefficient and p-value from the lines. The r values or correlation coefficients show that there is a weak positive relationship at 0.3 for the black population estimates and the p-value corresponds to this with a high value of 0.18 meaning that the probability of type I error is 18%. Type I error is defined as the probability that a false positive occurs, meaning the null hypothesis is true but rejected. The hispanic and white populations show moderate positive relationships with a hispanic R of 0.65 and the white population with an R of 0.53. Both have small p-values of 0.013 and 0.0015 so the probability of type I error is low here. Furthermore, it appears that even with high amounts of variance for the black population, they are disproportionately affected by diabetes in this country, while hispanics also are experiencing much higher rates and look to surpass rates in the black population with much less variance and chance of type I error.

Figure 1. National Diabetes Estimates by Race

Moving forward, now that we have identified a significant discrepancy within race, let us now observe any obvious trends by gender. Long time research of the human body has shown that with males, rates of diabetes are typically higher due to the storage of fat in a typical male. Males usually store fat in the abdomen, stressing internal organs, while women tend to store

extra weight on their hips, not causing as much harm to internal organs. Like we have done above, filter the data again within newly diagnosed cases of diabetes, but this time group race and keep the genders separate in individual columns. We can simply copy the filtering and plotting steps already done and just swap out values and labels. The final plot can be seen below with somewhat surprising results. Although males are more likely biologically, socioeconomic factors seem to have had a commanding hold on the female population. Females show a higher estimated rate of diabetes than males do from the start of the data around 1980 all the way to the early 2010s. Perhaps with more research and data, we could find correlation with general healthcare and living standards of the female population nationally. Back to the plot, we see that both male and female populations show a strong positive relationship with R values of 0.86 and 0.88. Both have remarkably low p-values as well, showing very little chance of type I error. In conclusion, diabetes rates are increasing among men and women, but the male population is seeing larger growth.

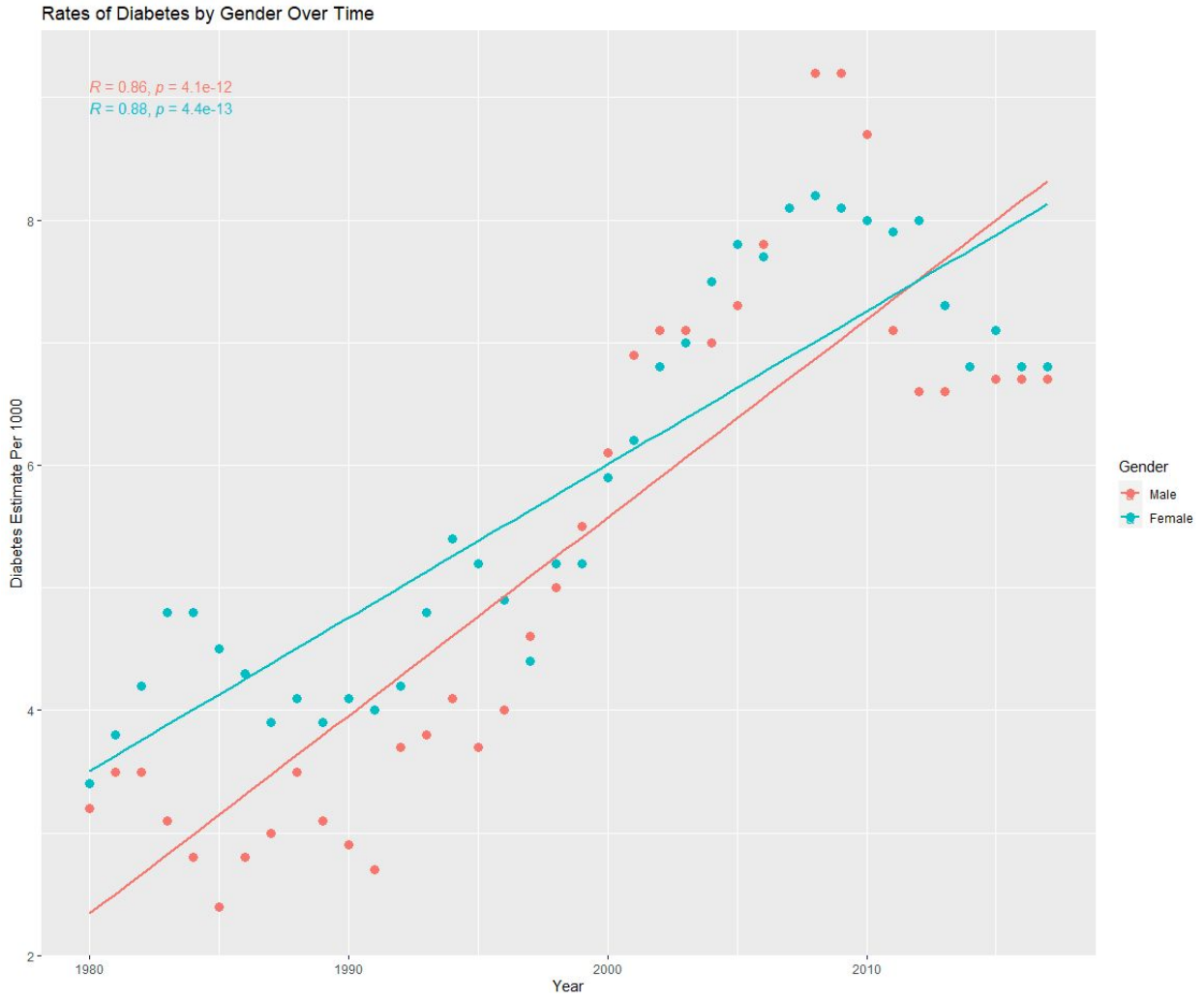

Figure 2. National Diabetes Estimates by Gender

Now that we have conducted analysis on diabetes estimates across the country, we should move on to what may be causing these increased rates. As mentioned before, obesity is the leading risk factor for developing the most common form of diabetes, type II diabetes. In order to

track this major risk factor, we need to analyze the rates of obesity and the total overweight population within the country as a whole and also take a look at rates in particularly unhealthy southern states like Alabama. We should start by analyzing chronic CDC data [10] on national and statewide obesity rates. The data needed to be aggregated by hand when combining all years. All of this survey data covers years 2011 to 2018. This gives us an appropriate range for discovering trends that may be causing overall increased rates of diabetes. By starting with the national aggregated data, we can use the filtering functionality used earlier within the Tidyverse package. Since the data is already cleaned and rather minimal, only containing a few attributes, we do not need to do much prep before plotting. We can start by removing the "2 or more races" as well as the "other" entries for the race attribute. Like earlier let us make a scatter plot comparing the races as well as a regression line overlaid. Although not shown in the plot below, the R values for each of the six listed races, all showed a strong positive relationship with values greater than 0.80 and all with p-values less than 0.015. The first thing that stands out when analyzing the plot below, is the incredibly low rate of diabetes for the asian population. That was the major outlier as typically minorities show greater health disparity in these studies. With further research we could track the cultural difference in diets to explain this. Black, American Indian and Hispanic populations show much higher rates of obesity than other races.

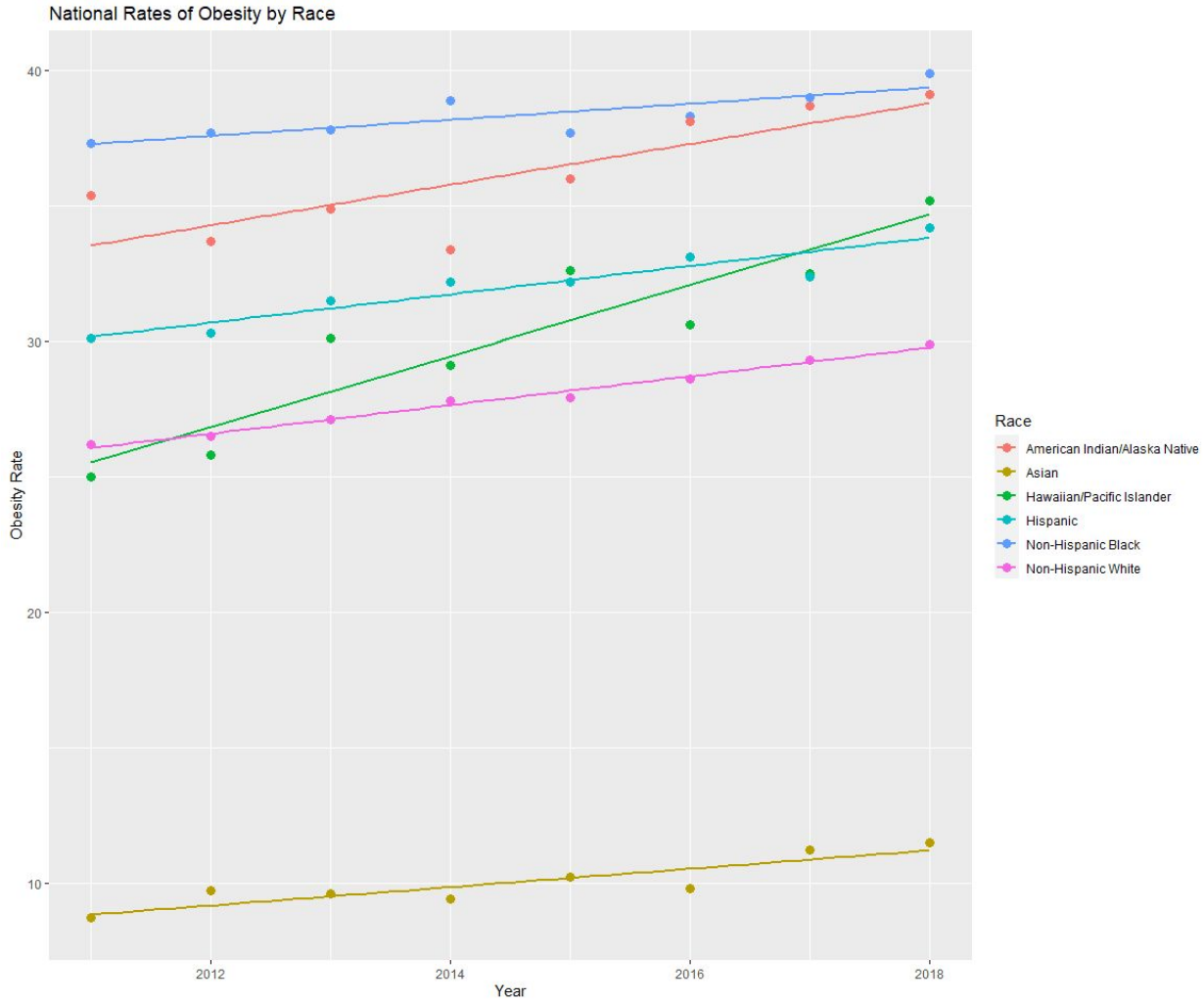

Figure 3. National Obesity Rates by Race

Since the southern states in the United States typically show greater health disparity than other parts of the country, we should visualize the rates of obesity in the state of Alabama. Once we have analyzed Alabama, we can compare to the national averages and see if there is any significant difference. Using the same dataset as above but only using data from the state of Alabama, we found some holes in the data. Earlier we filtered out the "2 or more races" and the "other" entries but the state data requires more cleaning. The sample sizes are too small and some races are even void of data. We need to filter out "Asian" and "Hawaiian/Pacific Islander" to properly plot the alabama data. Next, follow the same plotting steps to analyze the data. Immediately, I noticed the extremely low R value of 0.045 and high p-value of 0.92 for the American Indian population. Although nationally, the American Indian population shows a high rate of obesity, the data in Alabama is insignificant and must be discarded. Formerly, we display obesity rates by the three largest races in Alabama. The black and hispanic populations have moderate positive relationships with R values of 0.58 and 0.62, and p-values of 0.1 and 0.13 signalling a significant probability of type I error. With this plot, we see that in Alabama, the black population reaches a nearly 45% rate of obesity, while the white population follows with close to 35% obesity rates, and the hispanic lowest at around 30% obesity rates. The plot below compares the state's rates to the national ones, showing a significant difference.

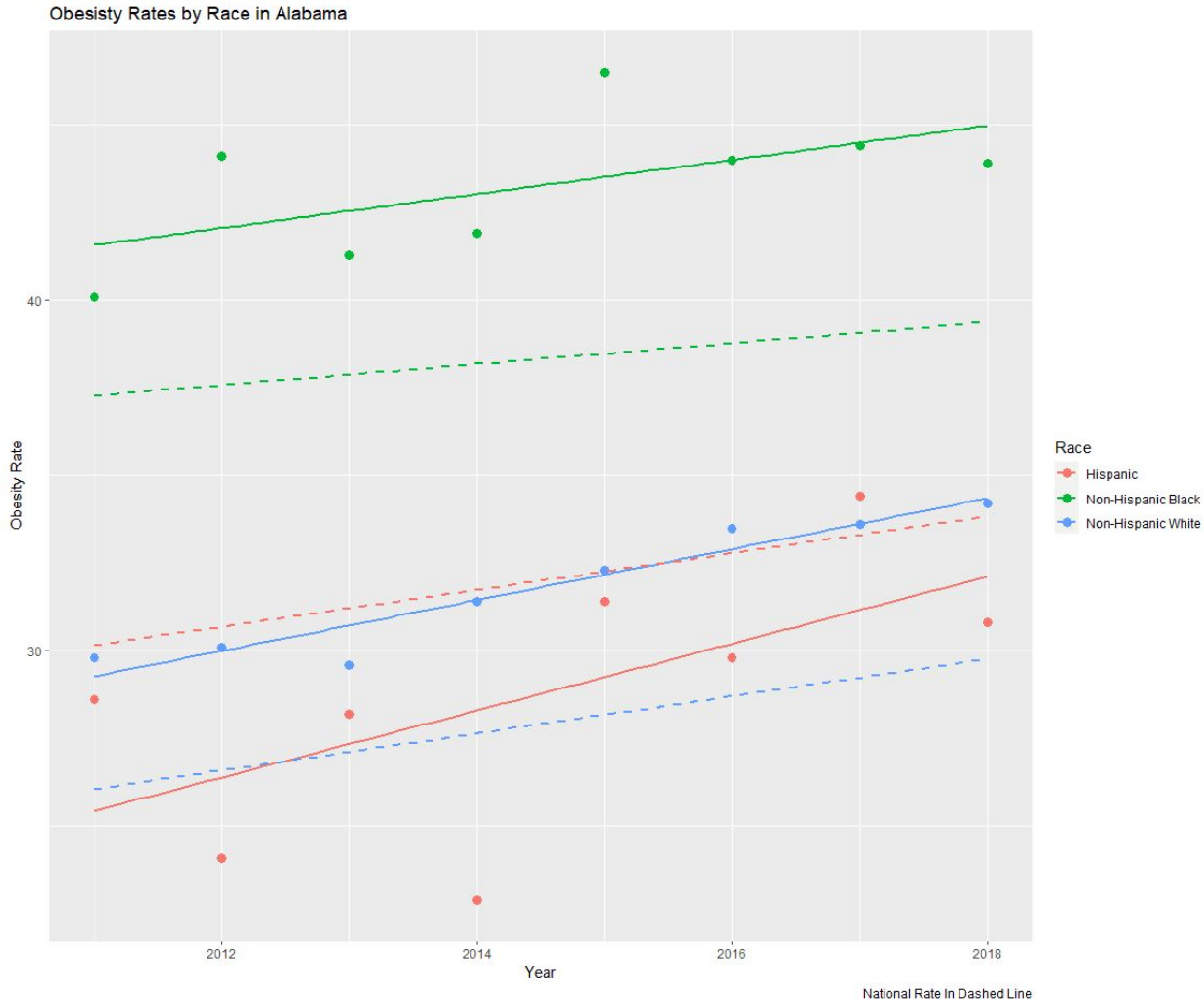

Figure 4. Alabama Obesity Rates by Race Vs National Rates(DashedLine)

Now that we have plotted the rates of obesity across various races and geographic locations spanning within the last decade, we have a good reason to believe that just as obesity may be causing diabetes, there must be some underlying cause for increasing rates of obesity as well. A future goal for this study is to take a look at exercise rates nationally and within Alabama to find the root cause for increased obesity while also taking into account overweight status as well. While there are many factors for diabetes, exercise is certainly one of the most crucial, and with simple methods like the ones used above, trends and relationships could be easily identified.

III. Conclusion

In conclusion, the United States has appalling rates of diabetes when compared to the rest of the world and the problem is just growing every year. Diabetes is one of the most researched diseases in the world today, leaving the country and people little room to ignore facts and numbers. We have applied various statistical methods to discover the underlying trends and relationships within racial and gender groups, as well as regional groups. From our data analysis, we have first found that minority groups such as the black and hispanic populations in this country suffer from higher rates of diabetes than the white population. Other groups like the American Indian population have shown to suffer the worse in recent studies. The data used did not include this group and the study definitely suffers by not representing this minority. Along with these groups, asians also have shown higher rates of diabetes than the white population nationally. When analyzing the differences in gender, it seemed that the gap was relatively small, as males and females suffered similar rates, with men recently surpassing women in the last 15 years. There are numerous socio-economic reasons that may be the cause of this discrepancy, but within this analysis, I focused on the leading risk factor for type II diabetes.

Obesity is perhaps the largest reason to blame for the increasing rates of diabetes in this country. The Mayo Clinic identifies obesity and overweight status as the leading risk factor for developing type II diabetes [4]. From this observation, I compared rates of obesity by race again to see if there was correlation to the diabetes rates across the white, black, hispanic, and other populations in this country. From my analysis, nationally black, indian, and hispanic groups all showed obesity rates much higher than the white and asian populations. The asian population, specifically, showed remarkably low rates of obesity. The asain population stands out because even though they experience low rates of obesity, they still show much higher rates of diabetes than the general white population. This visualization showed that obesity rates for the most part had a strong relationship with diabetes rates in the country. Southern states like Alabama have shown even higher rates of obesity, with the black population approaching nearly 45% of the group being obese. This is conclusive that often poorer states' populations' health drastically suffers from lower living conditions and poverty. Further research is required here to analyze the factors more. Exercise was the other major factor for developing a overweight and obese state that I did not get to.

Minorities in this country suffer greatly from diabetes and leading factors such as obesity are the current leading cause. Exercise data could further improve this study along with diet data to figure out what exactly is causing the increased rates of diabetes and obesity. Finally, a strong connection can be made that some states such as Alabama suffer from as close to a 15% increase in rates of obesity. Solving the obesity crisis is perhaps the first step to stopping the increase in type II diabetes.

References

[1] https://www.cdc.gov/diabetes/basics/type2.html
[2] Barnes A. S. (2011). The epidemic of obesity and diabetes: trends and treatments. *Texas Heart Institute journal*, *38*(2), 142–144.
[3] https://hopkinsdiabetesinfo.org/diagnosis-of-diabetes/
[4] https://www.mayoclinic.org/diseases-conditions/diabetes/symptoms-causes/syc-20371444

[9] Diagnosed Diabetes NHIS
https://gis.cdc.gov/grasp/diabetes/DiabetesAtlas.html
[10] National/Alabama Obesity Data
https://chronicdata.cdc.gov/Nutrition-Physical-Activity-and-Obesity/Nutrition-Physical-Activity-and-Obesity-Behavioral/hn4x-zwk7
[11] Nutrition, Physical Activity, and Obesity: Data, Trends and Maps
https://nccd.cdc.gov/dnpao_dtm/rdPage.aspx?rdReport=DNPAO_DTM.ExploreByLocation&rdRequestForwarding=Form

R Code

```
##########################
#
#       Diabetes
#
##########################


# Diabetes data located here https://gis.cdc.gov/grasp/diabetes/DiabetesAtlas.html#
# Within the compressed directory, there are 3 files
# 1. National_Data.csv
# 2. Data_Dictionary_National_1980_2017.docx
# 3. Lookup_Table.xlsx

# Read in the national diabetes data set from the CDC, covers 1980 to 2017
archive = read.csv("C:/Users/anorr/Desktop/Stat6000/FinalProject/Diabetes_Statistical_

# Throughout this project the Tidyverse and ggpubr packages for R are utilized
#install.packages("tidyverse")
#install.packages("ggpubr")
library(tidyverse)
library(ggpubr)
# Start by narrowing down to just new diagnoses
# Fetch all rows that are classified as "newly diagnosed"
# 'indicatorid' of 2 represents a group that is newly diagnosed with diabetes
# Also filter by 'estimateid' of 72 to get rate per 1000 people
newly_diagnosed = dplyr::filter(archive, indicatorid == 2 & estimateid == 72)

# select just the year, race, and estimate attributes from the data
year_race_rate_estimate = dplyr::select(newly_diagnosed, yearid, raceid, Estimate)

# Select just the races from the newly_diagnosed table)
race = dplyr::select(newly_diagnosed, raceid)
# find unique entries in the race table
unique(race)
# unique races: 0(total), 1(white), 2(black), and 7(Hispanic)


# First visualize diabetes rates across the 3 included races
# Tidyverse + ggplot scatter plot
# Filter out total race and total age, but keep total gender

race_archive <- archive %>%
  filter(indicatorid == 2 &
           estimateid == 72 &
           raceid != 0 &
           genderid == 0 &
           ageid == 99 & Estimate != ".")

# Plot newly diagnosed diabetes estimates across races
ggplot(race_archive, aes(x = yearid,
                         y = as.numeric(Estimate),
                         col = as.factor(raceid))) +
        geom_point(size = 3) +
        xlab("Year") + ylab("Diabetes Estimate Per 1000") +
        labs(title = "Rates of Diabetes by Race Over Time", col="Race") +
        scale_color_hue(labels = c("White", "Black", "Hispanic")) +
        stat_smooth(method = "lm", se = FALSE) +
        stat_cor()


# Next visualize the diabetes rates across gender
# Filter out total age and total gender, but keep total race
gender_archive <- archive %>%
  filter(indicatorid == 2 &
           estimateid == 72 &
           raceid == 0 &
           genderid != 0 &
           ageid == 99 & Estimate != ".")

# Plot newly diagnosed diabetes estimates across gender
ggplot(gender_archive, aes(x = yearid,
                         y = as.numeric(Estimate),
                         col = as.factor(genderid))) +
      geom_point(size = 3) +
      xlab("Year") + ylab("Diabetes Estimate Per 1000") +
      labs(title = "Rates of Diabetes by Gender Over Time", col="Gender") +
      scale_color_hue(labels = c("Male", "Female")) +
      stat_smooth(method = "lm", se = FALSE) +
      stat_cor()
```

```r
# Alex Norris


##########################
#
#       Obesity
#
##########################

library(tidyverse)
library(ggpubr)

nationalObesity = read.csv("C:/Users/anorr/Desktop/Stat6000/FinalProject/Diab
alabamaObesity = read.csv("C:/Users/anorr/Desktop/Stat6000/FinalProject/Diabe


nationalO <- nationalObesity %>%
  filter(Race.Ethnicity != "2 or more races" &
           Race.Ethnicity != "Other")

ggplot(nationalO, aes(x = YearStart,
                      y = as.numeric(Data_Value),
                      col = as.factor(Race.Ethnicity))) +
  geom_point(size = 3) +
  xlab("Year") + ylab("Obesity Rate") +
  labs(title = "National Rates of Obesity by Race", col="Race") +
  stat_smooth(method = "lm", se = FALSE) +
  stat_cor()


# Alabama data lacks values for certain races, so we filter them out
alabamaO <- alabamaObesity %>%
  filter(Race.Ethnicity != "2 or more races" &
           Race.Ethnicity != "Other" &
           Race.Ethnicity != "Asian" &
           Race.Ethnicity != "Hawaiian/Pacific Islander")

ggplot(alabamaO, aes(x = YearStart,
                     y = as.numeric(Data_Value),
                     col = as.factor(Race.Ethnicity))) +
  geom_point(size = 3) +
  xlab("Year") + ylab("Obesity Rate") +
  labs(title = "Rates of Obesity in Alabama by Race", col="Race") +
  stat_smooth(method = "lm", se = FALSE) +
  stat_cor()




nationalComp <- nationalObesity %>%
  filter(Race.Ethnicity != "2 or more races" &
           Race.Ethnicity != "Other" &
           Race.Ethnicity != "Asian" &
           Race.Ethnicity != "Hawaiian/Pacific Islander" &
           Race.Ethnicity != "American Indian/Alaska Native")

alabamaComp <- alabamaObesity %>%
  filter(Race.Ethnicity != "2 or more races" &
           Race.Ethnicity != "Other" &
           Race.Ethnicity != "Asian" &
           Race.Ethnicity != "Hawaiian/Pacific Islander" &
           Race.Ethnicity != "American Indian/Alaska Native")


ggplot(alabamaComp, aes(x = YearStart,
                        y = as.numeric(Data_Value),
                        col = as.factor(Race.Ethnicity))) +
  geom_point(size = 3) +
  xlab("Year") + ylab("Obesity Rate") +
  labs(title = "Obesisty Rates by Race in Alabama", col="Race",
       caption = "National Rate In Dashed Line") +
  stat_smooth(method = "lm", se = FALSE) +
  stat_smooth(method = "lm",linetype = "dashed", se = FALSE,
              data = nationalComp, aes(x = YearStart,
                                       y = as.numeric(Data_Value),
                                       col = as.factor(Race.Ethnicity))) +
```