The Relationship Between Walkability and Health Outcomes of Counties in the US

1. Introduction

a. Set up

First load the packages you will use throughout the document. *Note: you can hide this step in the resulting output with the option* `echo=FALSE`.

b. Quick description of the dataset(s)

The first dataset details the walkability of locations across the US using a framework developed by the EPA (https://catalog.data.gov/dataset/walkability-index (https://catalog.data.gov/dataset/walkability-index)). The second dataset details various health outcomes such as poor mental health days a month, life expectancy, and diabetes prevalence (https://www.countyhealthrankings.org/explore-health-rankings/rankings-data-documentation (https://www.countyhealthrankings.org/explore-health-rankings/rankings-data-documentation)). Each row represents a county, given by FIPS codes. I am interested in how walkability of a city affects physical and mental health outcomes, such as the prevalence of diabetes, or the number of poor mental health days. Perhaps greater walkability enhances the time spent outdoors, which is linked to improved physical and mental health outcomes (https://www.apa.org/monitor/2020/04/nurtured-nature (https://www.apa.org/monitor/2020/04/nurtured-nature)). Additionally, greater walkability could reduce the overall commute time and instances of driving alone to work, which in turn can reduce each individual's carbon footprint.

For walkability, there are 220,740 observations of 117 variables. For health outcomes, there are 3195 observations of 690 variables.

```
# Create a dataset we can use to answer our research question

# Picking relevant information from the walkability index
walkability <- walkability_index %>%
  select(GEOID10, STATEFP, COUNTYFP, TRACTCE, NatWalkInd) %>%
  rename(GEOID = GEOID10,
         STATE = STATEFP,
         COUNTY = COUNTYFP,
         TRACT = TRACTCE)

# Picking relevant information from the health outcomes dataset
health <- health_index %>%
  select('State FIPS Code', 'County FIPS Code', '5-digit FIPS Code','Name', 'Poor m
ental health days raw value', 'Poor physical health days raw value', 'Adult obesity
raw value', 'Driving alone to work raw value', 'Life expectancy raw value', 'Freque
nt mental distress raw value', 'Diabetes prevalence raw value', 'Median household i
ncome raw value') %>%
  rename(STATE = 'State FIPS Code',
         COUNTY = 'County FIPS Code')
```

c. Define a research question

One thing we might want to explore is how the walkability of a city is related to health outcomes of its ceitizens. For example, we can look at the EPA's walkability index and calculate the average walkability per county. Then, we can compare walkability to measures of physical and mental health, such as prevalence of diabetes, life expectancy, or average number of poor mental health days. We might find that greater walkability is conducive to a decreased prevalence of diabetes, decreased number of poor mental health days, and an increased life expectancy.

Our datasets aren't tidy though! The variable names must be cleaned up to be easier to understand at a glance, and then matched so we can combine them with a key.

2,3,4. Make our data analyzable

Here is a giant piece of code that wrangles our data and results in a dataframe containing relevant health outcomes information, joined with walkability index score by county FIPs codes.

```
# WALKABILITY
new_walk <- walkability %>%
  group_by(STATE, COUNTY) %>%
  arrange(STATE, COUNTY) %>%
# find the mean walkability per state per county
  summarize(mean_walkind = mean(NatWalkInd, na.rm = T))

# HEALTH OUTCOMES
new_health <- health %>%
  mutate(STATE = as.numeric(STATE),
         COUNTY = as.numeric(COUNTY))

# join datasets with two keys (state and county)
my_data <- full_join(new_walk, new_health, by = c("STATE", "COUNTY"))
```

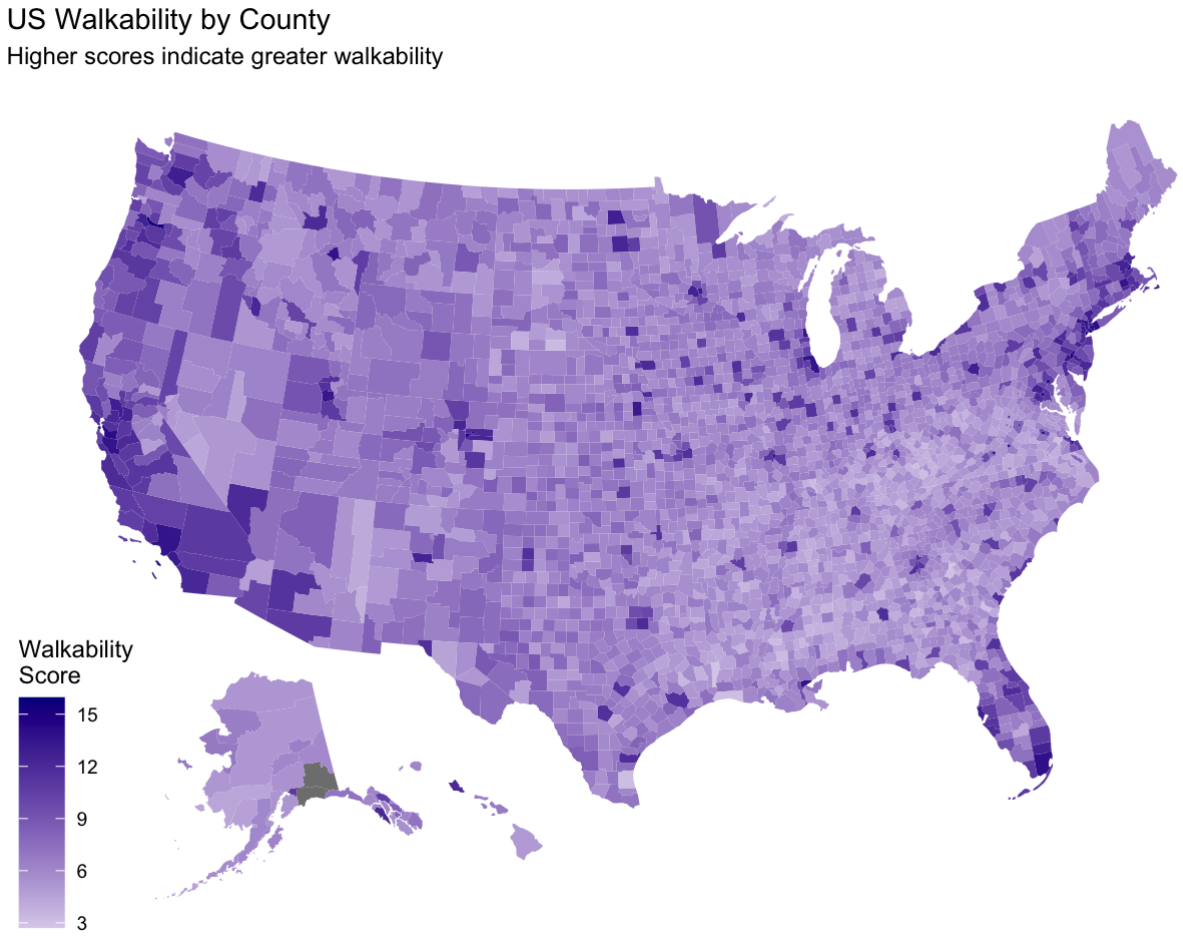
5. Visualize and analyze our data

Visualize our predictor: walkability (by county). The walkability index is described by a variety of variables, such as the prevalence of mixed-use blocks, proximity to transit stops, and greater intersection density.

```
# Visualizing walkability
# install.packages("usmap")
library(usmap)
library(ggplot2)
library(tidyverse)

my_data <- my_data %>%
  rename(fips = '5-digit FIPS Code')

plot_usmap(data = my_data, values = "mean_walkind", color = "transparent") +
  labs(title = "US Walkability by County", subtitle = "Higher scores indicate great
er walkability", fill = "Walkability
Score") +
  scale_fill_gradient2(low = "white", high = "dark blue")
```

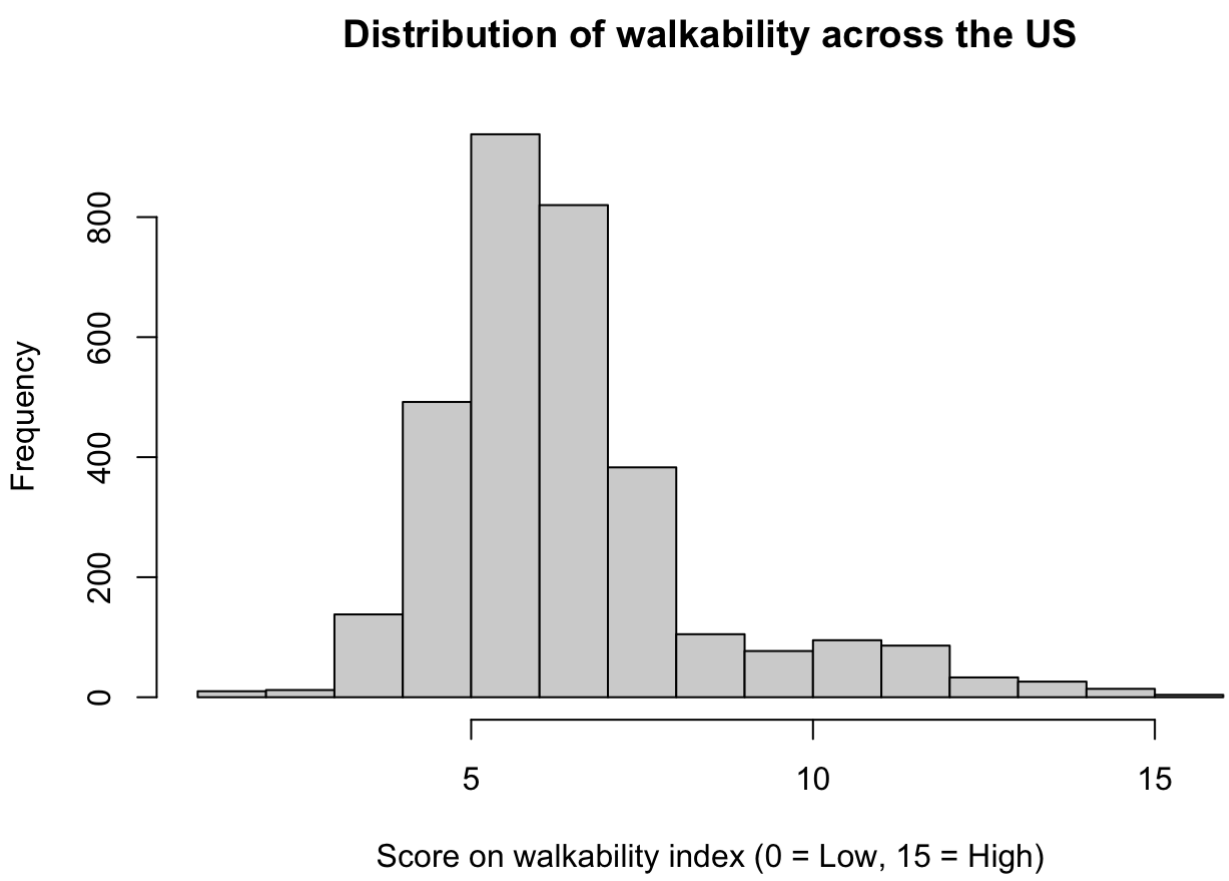


This is interesting! Walkability seems especially prevalent around major east and west coast cities. There also appears to be a clear disparity in walkability between urban and rural areas. Let's run some summary statistics to explore this distribution further.

```
# Summary statistics about walkability across US counties
summary(my_data$mean_walkind)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   1.000   5.200   6.036   6.426   7.018  15.957      53
```

```
hist(my_data$mean_walkind,
     xlab = "Score on walkability index (0 = Low, 15 = High)",
     main = "Distribution of walkability across the US")
```



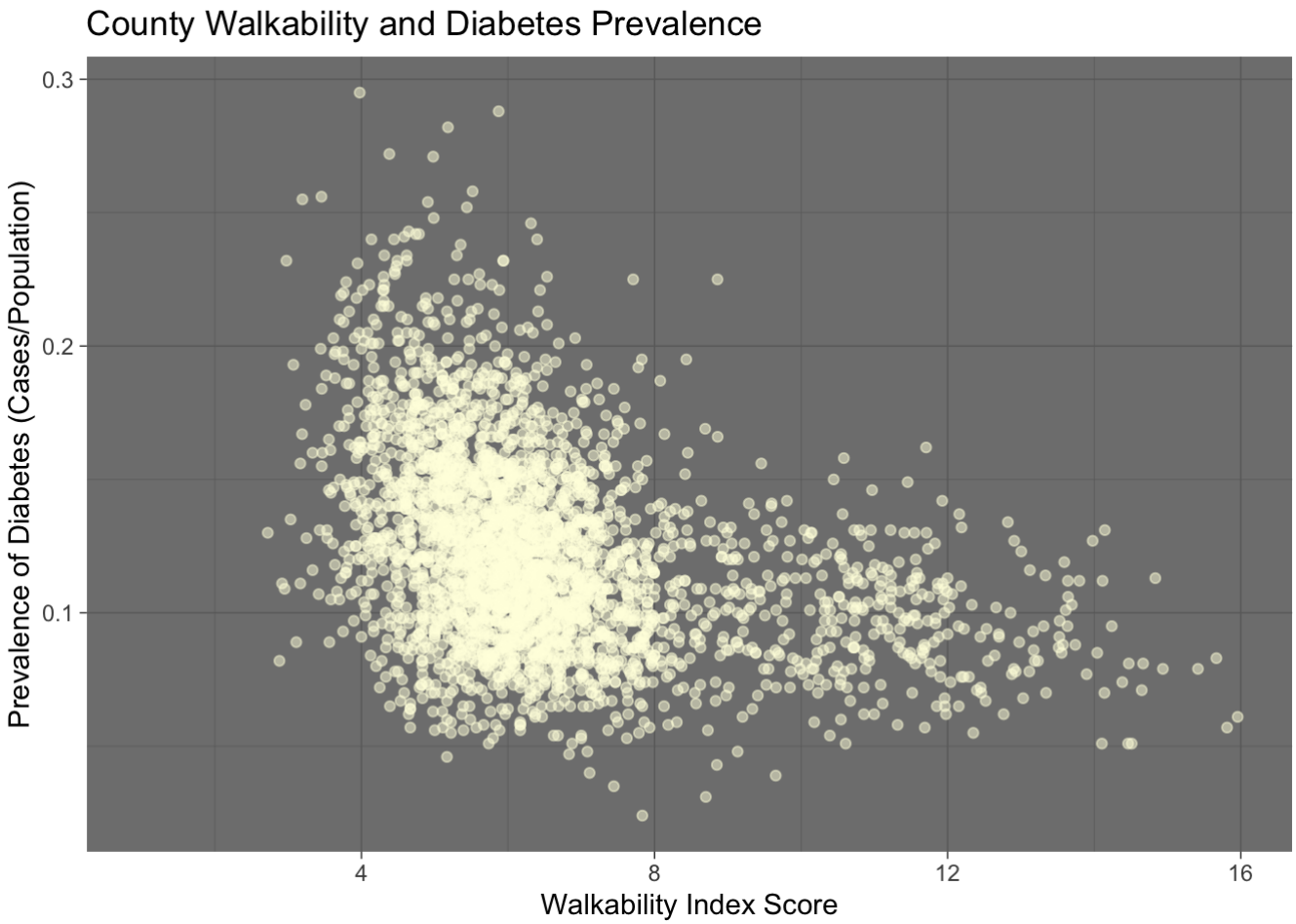
The walkability distribution is skewed right with a median of 6.036. The EPA interprets scores of 1.2 - 8.3 as Rural - Suburban. So, you would need to rely on a car to get around the typical US county.

Now, let's see if walkable cities actually lead to more walking, and perhaps a decreased prevalence of diabetes:

```
# Relationship between walkability and prevalence of diabetes

# Renaming variables for easier manpulation
my_data <- my_data %>%
  rename(diabetes = 'Diabetes prevalence raw value') %>%
  rename(poor_mental = 'Poor mental health days raw value')
# Changing diabetes prevalence & mental health to a numeric variable
  class(my_data$diabetes) = "numeric"
  class(my_data$poor_mental) = "numeric"

# Visualizing the association between walkability and prevalence of diabetes
my_data %>%
  ggplot(aes(x = mean_walkind, y = diabetes)) +
  geom_point(size = 1.5, alpha = 0.5, color = "light yellow") +
  scale_y_continuous() %>%
  labs(title = "County Walkability and Diabetes Prevalence", x = "Walkability Index
Score", y = "Prevalence of Diabetes (Cases/Population)") +
  theme_dark()
```

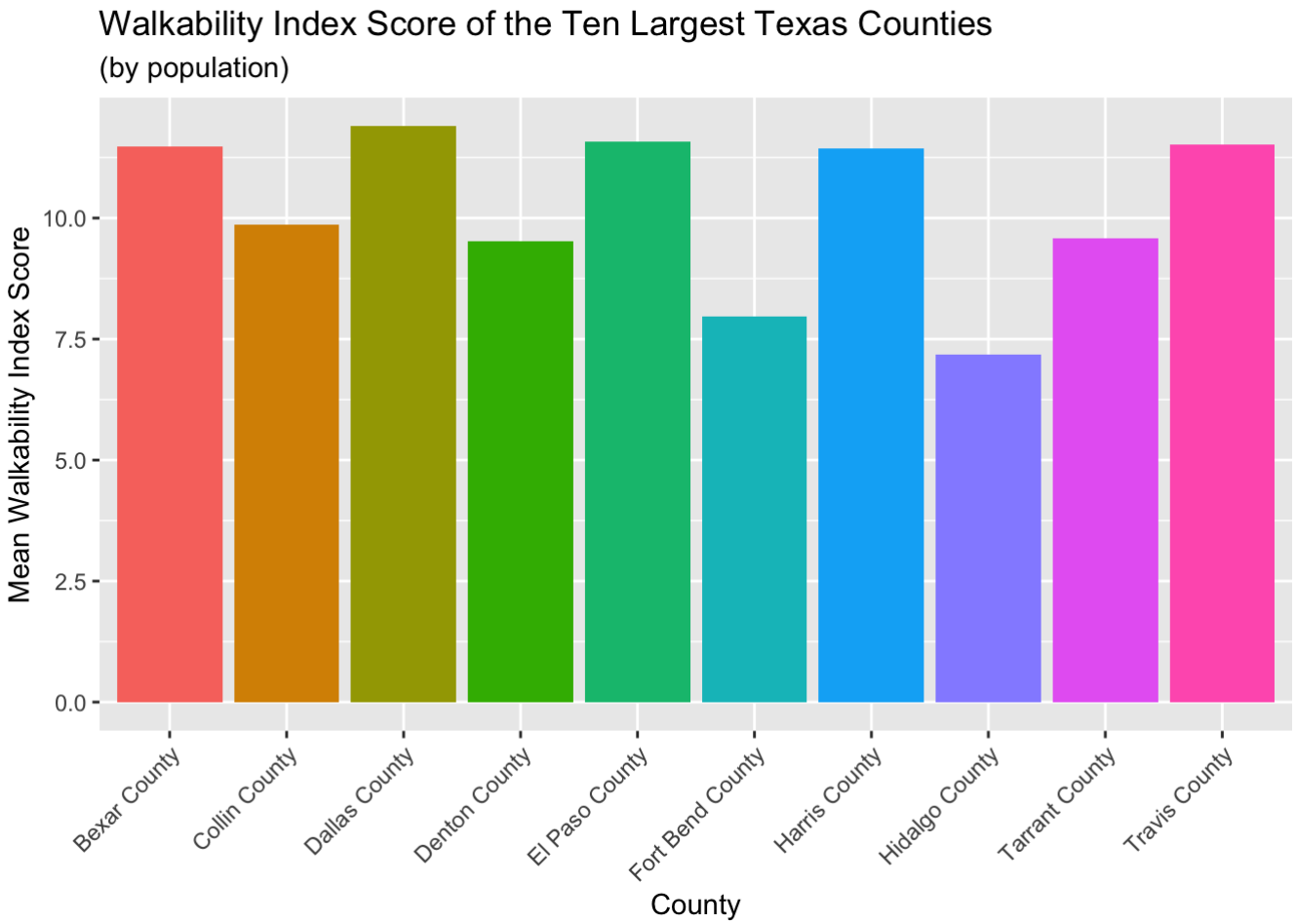


In terms of health outcomes, there seems to be a negative, weak-to-moderate association between the walkability of a county and the prevalence of diabetes. This means that counties scoring higher in walkability exhibit lower rates of diabetes. However, walkability seems to have less of an impact on diabetes prevalence among Walkability Index Scores from 4-8 as the points are scattered widely across the range of percentages. What if we took a look at walkability index scores across the largest Texas counties?

```
# A dataset within library(usmap)
view(countypop)

my_data <- full_join(my_data, countypop, by = c("fips"))

# Filter data to include Texas counties only; then find top 10 populations
my_data %>%
  filter(abbr == "TX") %>%
  top_n(10, pop_2015) %>%
  arrange(desc(pop_2015)) %>%
# Visualize the walkability of the ten counties
  ggplot(aes(x = Name, y = mean_walkind, fill = Name)) +
  geom_bar(stat = "summary", fun = "mean") +
   theme(axis.text.x = element_text(angle = 45, hjust = 1),
    legend.position = "none") +
  labs(title = "Walkability Index Score of the Ten Largest Texas Counties", subtitl
e = "(by population)", x = "County", y = "Mean Walkability Index Score")
```

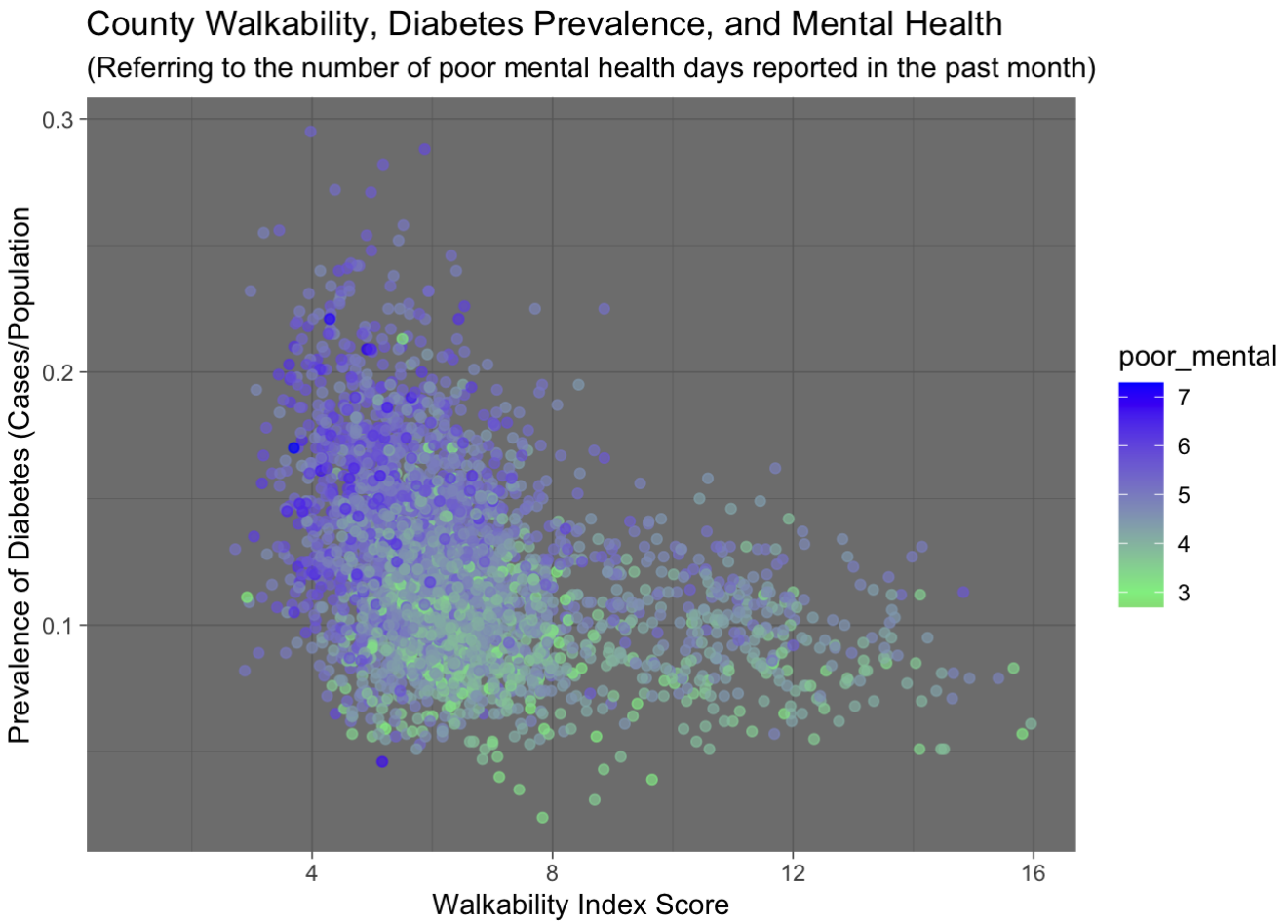


Walkability Index Score across the most populous Texas counties seems to hover around 8-12. This is still quite far away from our highest possible score of 16. We would expect to see higher walkability index scores within cities, but as Texas is one of the largest and most spread out states, it makes sense that our walkability index scores peak around 12.

Let's investigate walkability further through health outcomes, such as the prevalence of diabetes or the average number of poor mental health days:

```
# Relationship between walkability, the prevalence of diabetes, and the number of r
eported poor mental health days in the last month

my_data %>%
  ggplot(aes(x = mean_walkind, y = diabetes, color = poor_mental)) +
  geom_point(size = 1.5, alpha = 0.7) +
  scale_y_continuous(name = "Prevalence of Diabetes (Cases/Population") +
  labs(title = "County Walkability, Diabetes Prevalence, and Mental Health", subtit
le = "(Referring to the number of poor mental health days reported in the past mont
h)", x = "Walkability Index Score", y = "Prevalence of Diabetes (Cases/Populatio
n)") +
  scale_color_gradient2(mid  = "light green", high = "blue", midpoint = 3) +
  theme_dark()
```



Adding the number of poor mental health days to our previous visualization, we observe that there does seem to be an association between Walkbility Index Score and the number of poor mental health days! The association between walkabilidy index score and number of poor mental health days appears to be negative and weak. Higher numbers of poor mental health days seem to be associated with Walkability Index Scores of less than 8.

6. Discussion

While we observed an association between greater walkability and reduced cases of diabetes and poor mental health days, we should be careful about our interpretation as “walkability” isn’t the most robust measure of how livable or happy a city is. According to the EPA’s data, walkability in the US hinges heavily on whether the area is urban or not, and in these instances, the urban/rural divide brings confounding variables such as income, housing, and industry. A county could have blocks dedicated to be mixed-use and pedestrian-friendly, but actually getting from one block to another requires driving (Houston).

Developing a robust measure of walkability is challenging because there isn’t a one-size-fits-all measure for how walkable a city is. To further explore this topic, we should incorporate more variables such as median income or number of pedestrian-involved car accidents to develop a more holistic view of each county.

Acknowledgements: thanks to the EPA and the CHR&R for providing these datasets!