New York State Walkability and Chronic Health Indicators

Emma Mantel

2023-01-28

I just finished the Google Career Certificate in Data Analytics and wanted to create a small project to practice my newly acquired skills. This is by no means meant to be a real research project, but rather an exercise in finding data, figuring out what I can do with it, and putting everything into a presentable format.I have a basic interest in public health and one of my favorite pastimes is complaining about US car culture. I am a huge fan of being able to walk places (safely), and thus was curious about finding data about how walkable different areas are, and whether walkability might have any positive impact on the health of the people who live in those areas. That would give me just the right height of horse to sit on.

The walkability data came from the EPA Smart Location Database (SLD), and the Chronic Health Indicator (HD) data came from the CDC. It was myself-assigned job to mash them together and put them into some form of visualization.

Installing and loading libraries and data

First comes loading all the necessary libraries and data.

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0      v purrr   1.0.0
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.5.0
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(tigris)
```

```
## To enable caching of data, set `options(tigris_use_cache = TRUE)`
## in your R script or .Rprofile.
```

```
library(RColorBrewer)
library(awtools)

SLD <- read_csv("C:\\Users\\emmap\\Desktop\\Education & Training\\Coursera\\DataAnalytics\\Capstone\\Sma
```

```
## Rows: 220739 Columns: 182
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr   (8): GEOID10, GEOID20, STATEFP, COUNTYFP, TRACTCE, CSA_Name, CBSA_Name...
## dbl (174): OBJECTID, BLKGRPCE, CSA, CBSA, CBSA_POP, CBSA_EMP, CBSA_WRK, Ac_T...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# https://www.epa.gov/smartgrowth/smart-location-mapping#SLD
HD <- read_csv("C:\\Users\\emmap\\Desktop\\Education & Training\\Coursera\\DataAnalytics\\Capstone\\PLA
```

```
## Rows: 188456 Columns: 21
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (14): StateAbbr, StateDesc, LocationName, DataSource, Category, Measure,...
## dbl  (5): Year, Data_Value, Low_Confidence_Limit, High_Confidence_Limit, Tot...
## lgl  (2): Data_Value_Footnote_Symbol, Data_Value_Footnote
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# https://chronicdata.cdc.gov/500-Cities-Places/PLACES-Local-Data-for-Better-Health-County-Data-20/swc5
```

Cleaning the data

The data contain a lot of columns that I wasn't interested in. I narrowed down the scope of the project to New York State, because that is where I live. I could imagine this project being more interesting using other states that have more urban areas and smaller counties.

The SDL data contains a walkability index for each census block, but the HD data has county-level resolution. Thus, I aggregated the walkability indexes for each census block to form a county-level walkability index, taking into acocunt the population of each block, giving a kind of per capita walkability index, accounting for the fact that urban areas tend to be more walkable but might pack a lot more people into a single census block, compared to more rural areas.

```
ny_walkability <- SLD %>%
  filter(STATEFP== '36') %>%
  select('GEOID20', 'NatWalkInd', 'TotPop','COUNTYFP') %>%
  mutate(WalkIndScaled = NatWalkInd*TotPop) %>%
  group_by(COUNTYFP)%>%
  summarise(CountyPop=sum(TotPop), CountyWalkInd=sum(WalkIndScaled)/CountyPop)

head(ny_walkability)
```

```
## # A tibble: 6 x 3
##   COUNTYFP CountyPop CountyWalkInd
##   <chr>        <dbl>         <dbl>
## 1 001         307426         11.0
## 2 003          47025          5.57
## 3 005        1437872         13.1
## 4 007         194402          6.74
## 5 009          77686          6.09
## 6 011          77868          7.75
```

The counties data from the tigris packages contains the shape file for each counties which will be used in mapping later, and the FP code for the county, as well as its name, which will be necessary for joining all the data later.

```
ny_counties <- counties("NY") %>%
  select('COUNTYFP','NAME','geometry')
```

```
## Retrieving data for the year 2021
```

```
## Simple feature collection with 6 features and 2 fields
## Geometry type: MULTIPOLYGON
```

```
## Dimension:     XY
## Bounding box:  xmin: -78.30932 ymin: 41.58061 xmax: -73.57334 ymax: 45.01586
## Geodetic CRS:  NAD83
##     COUNTYFP         NAME                       geometry
## 47       101      Steuben MULTIPOLYGON (((-77.20041 4...
## 167      091     Saratoga MULTIPOLYGON (((-73.67891 4...
## 175      003     Allegany MULTIPOLYGON (((-78.04342 4...
## 205      075       Oswego MULTIPOLYGON (((-75.90417 4...
## 212      111       Ulster MULTIPOLYGON (((-73.94659 4...
## 324      089 St. Lawrence MULTIPOLYGON (((-74.64908 4...
```

The HD data is collected for each county and contins a number of measures for different health outcomes. Here, I selected coronary heart disease (CHD), depression, Obesity, and leisure time physical activity as indicators to correlate with walkability. Again, I am filtering for New York states. We get the county name and coordinates for the county's centerpoint from this dataset as well. Those coordinates came in a string that I split up into longitude and latitude numeric values for later plotting.

```
ny_HD <- HD %>%
  filter(StateAbbr=='NY' & DataValueTypeID=='AgeAdjPrv') %>%
  select('LocationName','MeasureId','Data_Value','Data_Value_Unit','TotalPopulation','Geolocation') %>%
  filter(MeasureId %in% c('CHD','DEPRESSION','OBESITY','LPA') )

ny_HD_loc <- ny_HD %>%
  mutate(Geo_clean=substr(Geolocation,9,nchar(Geolocation)-1)) %>%
  mutate(Geo_split=strsplit(Geo_clean, split=" ")) %>%
  unnest_wider(Geo_split,names_sep ="_LOC", transform=as.numeric) %>%
  select(-Geo_clean) %>%
  rename(LON=Geo_split_LOC1, LAT=Geo_split_LOC2) %>%
  mutate(LON=-1*LON)
```

The walkability data and county data are merged by the county FP code, and that new dataset is then merged with the HD data by county name. ny_data_loc is going to be used fro plotting the New York State map later on, the ny_data will be used to plot the variables in scatterplots. In retrospect, I could probably have used just one dataframe for both purposes, but this script unsurprisingly went through many iterations and this is how it turned out.

```
ny_data_loc <- ny_counties %>% merge(ny_walkability) %>%
  merge(ny_HD_loc, by.x='NAME', by.y='LocationName')

ny_data <- ny_counties %>%
  select('COUNTYFP', 'NAME') %>%
  merge(ny_walkability) %>%
  merge(ny_HD, by.x='NAME', by.y='LocationName')
```

```
## Simple feature collection with 6 features and 11 fields
## Geometry type: MULTIPOLYGON
## Dimension:     XY
## Bounding box:  xmin: -78.30932 ymin: 41.99821 xmax: -73.67676 ymax: 42.82258
## Geodetic CRS:  NAD83
##       NAME COUNTYFP CountyPop CountyWalkInd  MeasureId Data_Value
## 1   Albany      001    307426     10.959696 DEPRESSION       19.4
## 2   Albany      001    307426     10.959696        CHD        5.0
## 3   Albany      001    307426     10.959696        LPA       19.7
## 4   Albany      001    307426     10.959696    OBESITY       26.7
## 5 Allegany      003     47025      5.566061        CHD        6.1
## 6 Allegany      003     47025      5.566061 DEPRESSION       22.0
```

```
##   Data_Value_Unit TotalPopulation                    Geolocation       LON
## 1               %          303654 POINT (-73.9740095 42.5882401) -73.97401
## 2               %          303654 POINT (-73.9740095 42.5882401) -73.97401
## 3               %          303654 POINT (-73.9740095 42.5882401) -73.97401
## 4               %          303654 POINT (-73.9740095 42.5882401) -73.97401
## 5               %           45587 POINT (-78.0261531 42.2478532) -78.02615
## 6               %           45587 POINT (-78.0261531 42.2478532) -78.02615
##       LAT                         geometry
## 1 42.58824 MULTIPOLYGON (((-73.96379 4...
## 2 42.58824 MULTIPOLYGON (((-73.96379 4...
## 3 42.58824 MULTIPOLYGON (((-73.96379 4...
## 4 42.58824 MULTIPOLYGON (((-73.96379 4...
## 5 42.24785 MULTIPOLYGON (((-78.04342 4...
## 6 42.24785 MULTIPOLYGON (((-78.04342 4...
```

```
## Simple feature collection with 6 features and 9 fields
## Geometry type: MULTIPOLYGON
## Dimension:     XY
## Bounding box:  xmin: -78.30932 ymin: 41.99821 xmax: -73.67676 ymax: 42.82258
## Geodetic CRS:  NAD83
##       NAME COUNTYFP CountyPop CountyWalkInd  MeasureId Data_Value
## 1   Albany      001    307426     10.959696 DEPRESSION       19.4
## 2   Albany      001    307426     10.959696        CHD        5.0
## 3   Albany      001    307426     10.959696        LPA       19.7
## 4   Albany      001    307426     10.959696    OBESITY       26.7
## 5 Allegany      003     47025      5.566061        CHD        6.1
## 6 Allegany      003     47025      5.566061 DEPRESSION       22.0
##   Data_Value_Unit TotalPopulation                    Geolocation
## 1               %          303654 POINT (-73.9740095 42.5882401)
## 2               %          303654 POINT (-73.9740095 42.5882401)
## 3               %          303654 POINT (-73.9740095 42.5882401)
## 4               %          303654 POINT (-73.9740095 42.5882401)
## 5               %           45587 POINT (-78.0261531 42.2478532)
## 6               %           45587 POINT (-78.0261531 42.2478532)
##                         geometry
## 1 MULTIPOLYGON (((-73.96379 4...
## 2 MULTIPOLYGON (((-73.96379 4...
## 3 MULTIPOLYGON (((-73.96379 4...
## 4 MULTIPOLYGON (((-73.96379 4...
## 5 MULTIPOLYGON (((-78.04342 4...
## 6 MULTIPOLYGON (((-78.04342 4...
```

Calculating some values

Lastly, I am calculating correlation coefficients between the health indicators and walkability index for each county. I am also loading/creating color palettes and a data frame to insert the correlation values as labels later during plotting.

```
corr_data <- pivot_wider(ny_data_loc, names_from="MeasureId", values_from="Data_Value")
corr_OBESITY <- cor(corr_data$CountyWalkInd, corr_data$OBESITY)
corr_DEPRESSION <- cor(corr_data$CountyWalkInd, corr_data$DEPRESSION)
corr_LPA <- cor(corr_data$CountyWalkInd, corr_data$LPA)
corr_CHD <- cor(corr_data$CountyWalkInd, corr_data$CHD)
graphLabels <- data.frame(MeasureId = c("CHD","DEPRESSION", "LPA","OBESITY"),
                          corcoefs= paste("correlation coefficient: ", c(format(corr_CHD, digits=2),for
```

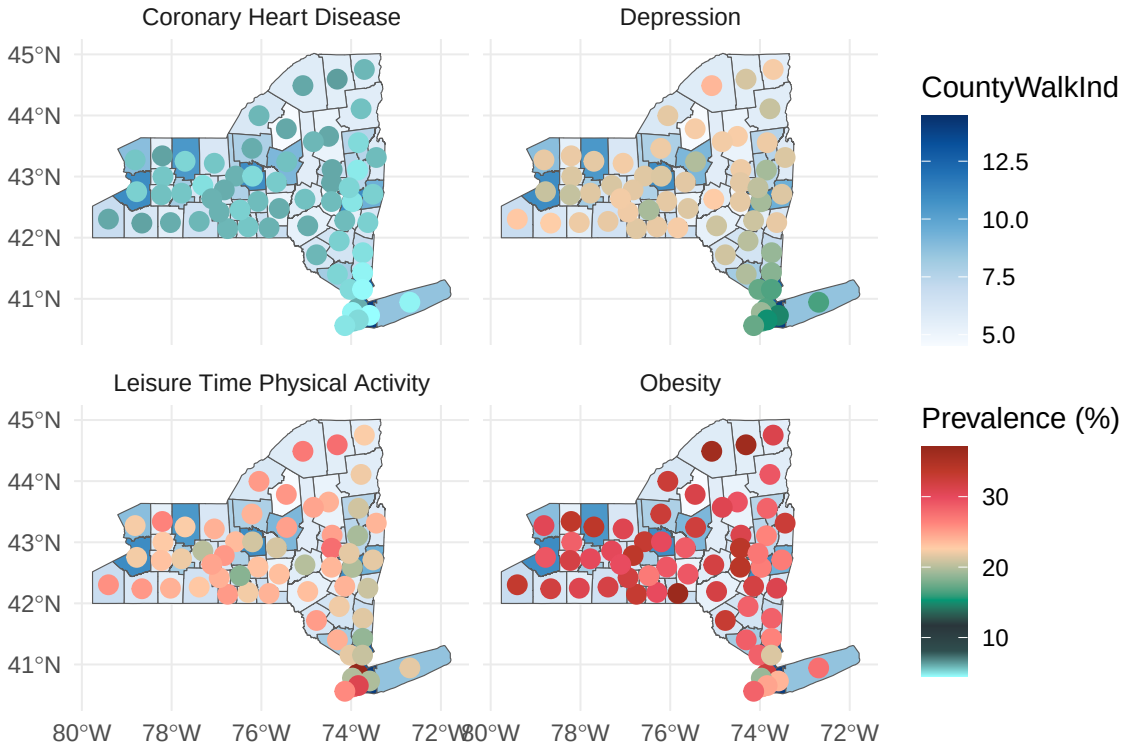
```
                          format(corr_LPA, digits=2),format(corr_OBESITY, digits=2))),
                          x     = c(11, 10.5, 10,11),
                          y     = c(6, 22,30,35))

pal1 <- brewer.pal(9,'Blues')
pal2 <- a_palette
pal2 <- c("#97FFFF","#2F4F4F",pal2[1:8])
```

Plotting

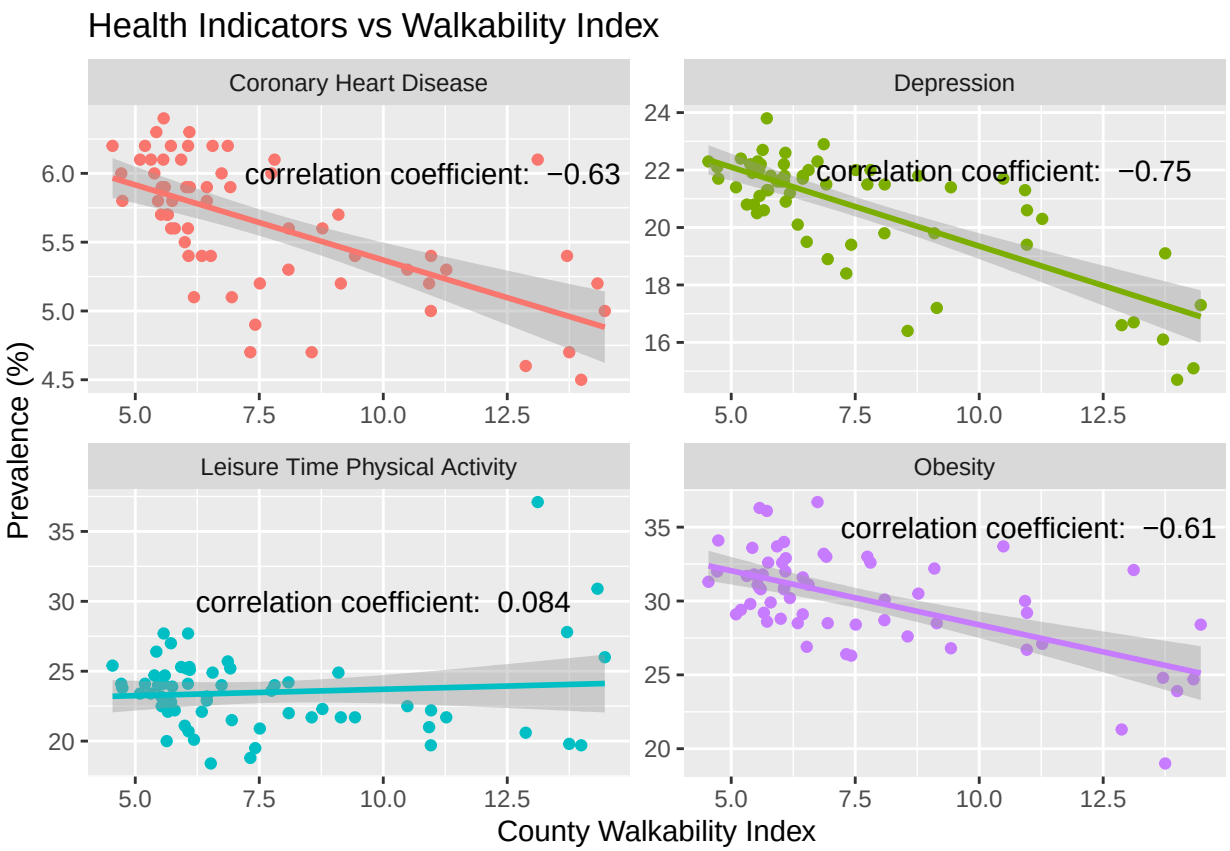
Finally, we get to make the plots. This code is porbably pretty straighforward, but trust me, it took me some time to get here! It is very obvious that the walkability is much higher in the counties that contain the few cities that exist in NYS. The New York City area is made up of several counties, and Rochester, Albany, and Buffalo each pull up the walkability index for their respective counties.Just from this plot, there seems to be some kind of relationship between walkability and at depression, obesity, and CHD. This is further supported by the next plot.

```
ggplot(ny_data_loc) +
  geom_sf(aes(fill=CountyWalkInd)) +
  theme_minimal() +
  scale_fill_gradientn(colors=pal1) +
  geom_point(mapping=aes(x=LON,y=LAT, color=Data_Value), size=3) +
  scale_color_gradientn(colors=pal2) +
  labs(title='',x='', y='', color='Prevalence (%)') +
  facet_wrap(~MeasureId,labeller = as_labeller(c(`CHD`="Coronary Heart Disease",`DEPRESSION` = "Depressi
  guides(size='none')
```



```
ggplot(ny_data) +
  geom_point(aes(x=CountyWalkInd, y=Data_Value, color=MeasureId), show.legend = FALSE) +
  facet_wrap(~MeasureId, scale='free', labeller = as_labeller(c(`CHD`="Coronary Heart Disease",`DEPRESSI
  geom_smooth(aes(x=CountyWalkInd, y=Data_Value, color=MeasureId), method='lm',show.legend = FALSE) +
  labs(title='Health Indicators vs Walkability Index',x='County Walkability Index', y='Prevalence (%)',
  geom_text(data=graphLabels,aes(x=x, y=y,label=corcoefs))
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Discussion

Plotting the variables against one another with their respective correlation coefficient shows that there seems to be an inverse relationship between Depression, Obesity, and CHD, and the walkability index in a given county. Do people who live in walkable areas actually move more and thus avoid certain chronic diseases? Is preventative healthcare more accessible in more urban areas? Is the population in cities simply younger and thus less likely to have developed these conditions? It would take some more investigation to tell which factors drive this relationship, but it was certainly a fun foray into publicly available data and analytics!