Walkability Index in US States

JonJeng and Given

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.2     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become error
```

```
library(ggthemes)
library(tidytext)
library(ggthemes)
library(lubridate)
library(kableExtra)
```

```
##
## Attaching package: 'kableExtra'
##
## The following object is masked from 'package:dplyr':
##
##     group_rows
```

Introduction:

Being students living in Northfield who do not own cars, we more often than not have to walk to our destinations. We recognize that Northfield is not a city, however, we were interested in how walkable cities are in the US. We decided to do our project on the walkability of different cities across the US. Walkability depends upon characteristics of the built environment that influence the likelihood of walking being used as a mode of travel.

Data: Where is it from

Our data is from the U.S. Environmental Protection Agency. The Walkability Index dataset characterizes every Census block group in the U.S. based on its relative walkability. According to the NationalWalkabilityIndex, Walkability depends upon characteristics of the built environment that influence the likelihood of walking being used as a mode of travel. The Walkability Index is based on the EPA's previous data product, the Smart Location Database (SLD). Block group data from the SLD was the only input into the Walkability Index, and consisted of four variables from the SLD weighted in a formula to create the new Walkability Index. The walkability index is based on measures of the built environment that affect the probability of whether people walk as a mode of transportation: street intersection density, proximity to transit stops, and diversity of land uses.This dataset shares the SLD's 2019 Census Block Groups.

We will analyze different factors that affect the walkability in US groups.

data: https://catalog.data.gov/dataset/walkability-index1

key1: https://geodata.epa.gov/arcgis/rest/services/OA/WalkabilityIndex/MapServer/0

key2: https://www.epa.gov/smartgrowth/smart-location-mapping

Body: Graphs

Question 1: Does the Count of workers in CBG (home location), 2017 per Total Population affect the National Walking Index?

Description: What is the relationship of the Walkability Index for the Number of Workers in CBG per Total Population in US States? What is the relationship between the total population in a CBG and the amount of workers in a CBG compared to the National Walking Index.

Graph 1:

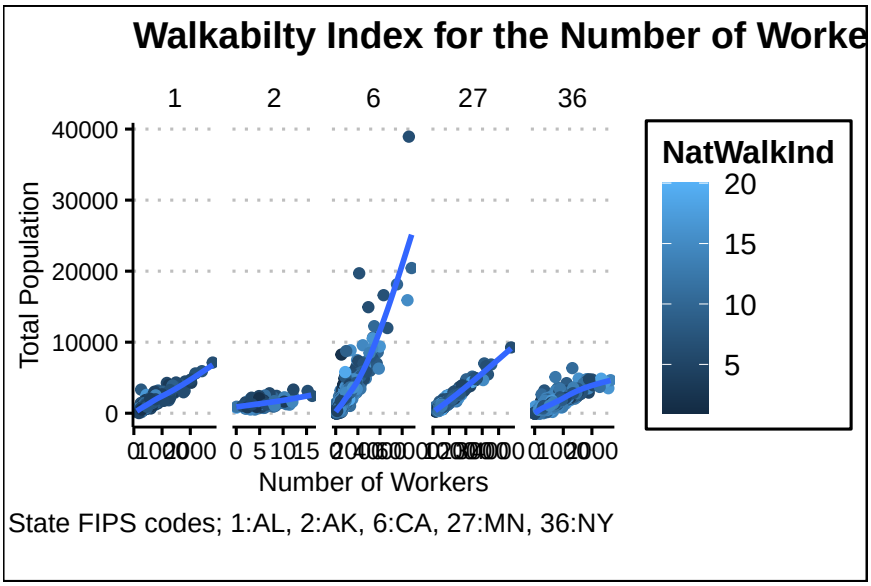
```
walkability |>
   filter(STATEFP == c(1,2,6,27,36),!is.na(CBSA)) |>
  select(STATEFP, COUNTYFP, CBSA, CBSA_Name, TotPop, NatWalkInd,Workers)|>

  group_by(CBSA_Name) |> # urban center with pop>10^4
  ggplot(aes(x=Workers,y=TotPop, color=NatWalkInd))+
  geom_jitter()+
  geom_smooth(se=FALSE)+
  facet_grid(~STATEFP, scales = "free")+
  labs(title="Walkabilty Index for the Number of Workers in CBG per Total Population in US States",capti
", x="Number of Workers", y="Total Population") +
  theme_clean()
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'

## Warning: The following aesthetics were dropped during statistical transformation:
## colour.
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
## The following aesthetics were dropped during statistical transformation:
## colour.
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
## The following aesthetics were dropped during statistical transformation:
## colour.
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
## The following aesthetics were dropped during statistical transformation:
## colour.
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a `group` aesthetic or to convert a numerical
```

```
##   variable into a factor?
## The following aesthetics were dropped during statistical transformation:
## colour.
## i This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## i Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
```



Analysis:

From our graph we observed that Count of workers in CBG was directly proportional with the total population in the observed area. Generally, we observed that the national walking index was the highest in states with total population less than 5000 people. States with lower populations had lower number of workers in the CBG, which correlated wit a higher national walking index. CA has the highest average National Walking Index compared to the other states. CA also has the highest number of workers per CBG, which may contribute to this trend.

Question: Does the the number of households with cars for each major Core-Based Statistical Area (CBSA) in the United States, affect the Mean Walk Index?

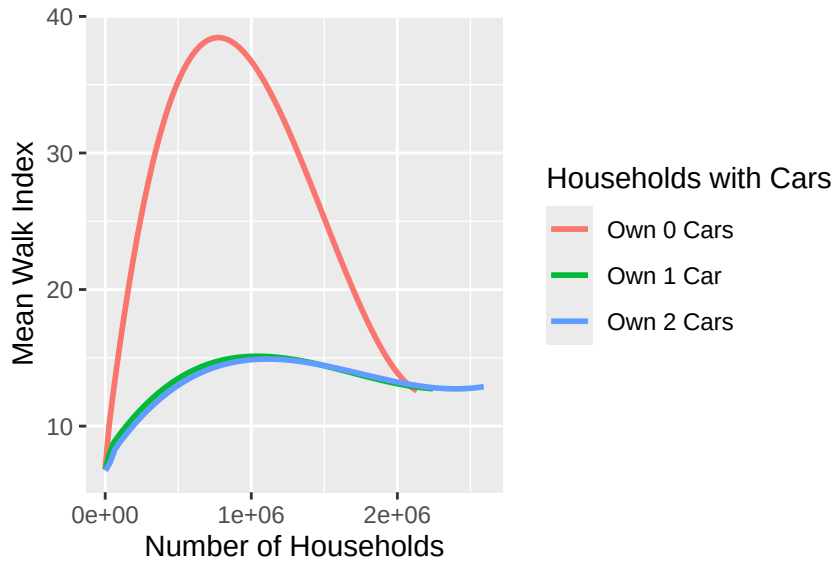
Description: In this line graph, we have the Number of Households on the X axis, and the Mean Walk Index on the Y axis. The Mean walk index ranges from 0 to 40, and the range for number of residents is from 0 to roughly 2,500,000.

Graph 2:

```
walkability |>
  filter(!is.na(CBSA)) |>
  select(STATEFP, COUNTYFP, CBSA, CBSA_Name, TotPop, D5AR,D5BE,D5CE, Workers, AutoOwn0, AutoOwn1, AutoO
  group_by(CBSA_Name) |>
  summarize("subareas" = n(),
            Population = sum(TotPop),
            "Own 0 Cars" = sum(AutoOwn0),
            "Own 1 Car" = sum(AutoOwn1),
            "Own 2 Cars" = sum(AutoOwn2p),
            MeanWalkInd = mean(NatWalkInd)) |>
  pivot_longer("Own 0 Cars":"Own 2 Cars",
```

```
  names_to = "Cars",
  values_to = "NumCars") |>
  ggplot(aes(x = NumCars, y = MeanWalkInd, color = Cars)) +
    geom_smooth(se = FALSE) +
  labs(y = "Mean Walk Index",
       x = "Number of Households",
       color = "Households with Cars")
```

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



Analysis: The graph shows that in CBSAs with around 750,000 households, there is a slight increase in the mean walk index. In the CBSAs, Households with one or two cars have similar walk indices. Households with no cars show a significant increase in walk index. This data makes sense because households with no cars probably do more walking than households with cars due to transportation needs.

Question: Does the Mean Walk Index for each major Core-Based Statistical Area (CBSA) affect the wage range of each household in the United States?

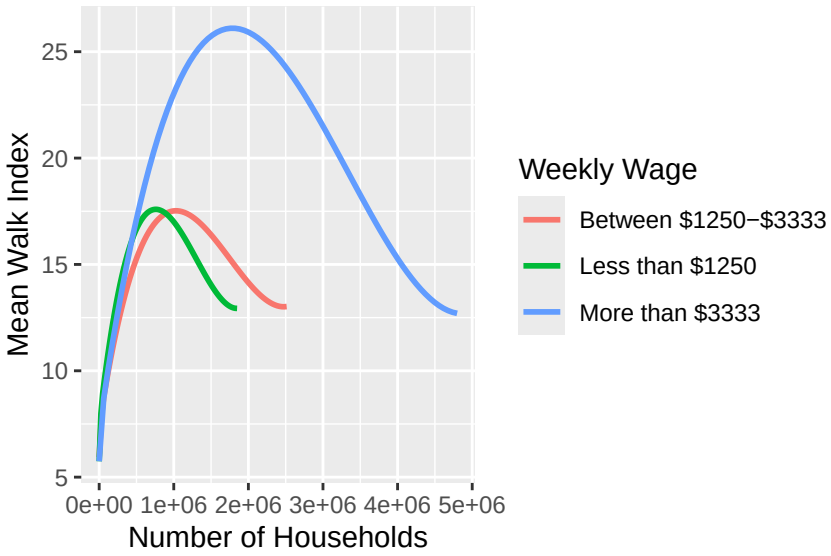
Description: In the following line, and scatter plot, we have the Mean Walk Index on the X axis, and Number of People on the Y axis. The Mean walk index ranges from 0 to 14, and the range for number of residents is from 0 to roughly 1,200,000. The graph generally shows that as the Walk index increases, the number of people with higher wages increases.

Graph 3:

```
walkability |>
  filter(!is.na(CBSA)) |>
  select(STATEFP, COUNTYFP, CBSA, CBSA_Name, TotPop, D5AR,D5BE,D5CE, Workers, AutoOwn0, AutoOwn1, AutoO
  group_by(CBSA_Name) |>
  summarize("subareas" = n(),
            Population = sum(TotPop),
            "Less than $1250" = sum(E_LowWageWk),
            "Between $1250-$3333" = sum(E_MedWageWk),
            "More than $3333" = sum(E_HiWageWk),
            MeanWalkInd = mean(NatWalkInd)) |>
  pivot_longer("Less than $1250":"More than $3333",
```

```
  names_to = "Wage",
  values_to = "NumPeople") |>
  mutate(Wage = fct_reorder2(Wage, NumPeople, MeanWalkInd)) |>
  ggplot(aes(x = NumPeople, y = MeanWalkInd, color = Wage)) +
  geom_smooth(se = FALSE) +
    labs(y = "Mean Walk Index",
         x = "Number of Households",
         color = "Weekly Wage")
```

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



Analysis: Households that make more than 3,333 dollars per week tend to have higher walk indices. The weekly wages of household between 1,250-3,333 dollars and 1,250 dollars have similar walk indices. A population of around 1,750,000, and weekly wage of over 3333 dollars has the highest Mean Walk Index with a value of around 27.

Conclusion:

In this project, we explored the walkability of different cities across the United States using the Walkability Index provided by the U.S. Environmental Protection Agency. Our analyses focused on understanding how various factors such as the number of workers in a Census Block Group (CBG), the number of households with cars, and household wage ranges influence the walkability of each different area of the US.

Our analysis showed a direct relationship between the number of workers in a CBG and the area's total population. The data indicated that CBSAs with more households owning no cars show a significant increase in the Mean Walking Index. We reasoned this thinking households without cars relied more on walking than any other mode of transportation. Last we found that households with higher wages (more than $3,333 per week) had higher walkability index. Logically, wealthier areas might invest more in walkable infrastructure, and higher-wage earners can afford to live in areas with better walking infrastructure.

In the future it would be interesting to compare the walkability of US state cities with that of European cities such as Milan Italy which has been rated the most walkable city in world as of March 2020.