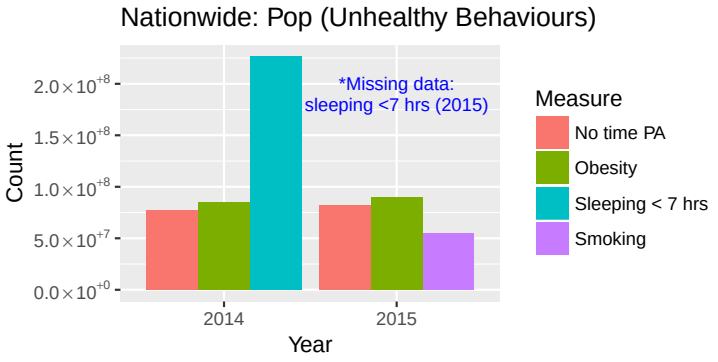
Unhealthy Behaviors and Health Issues in U.S.

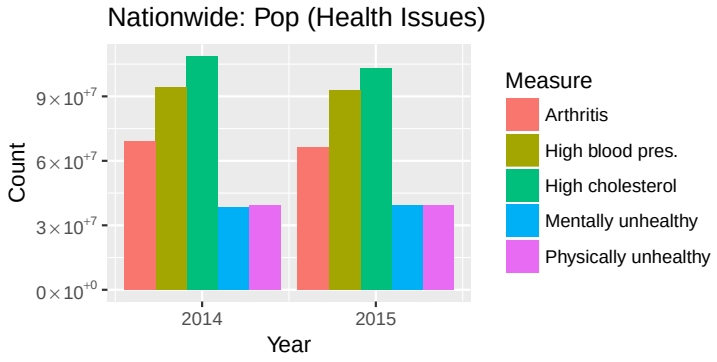
In Son, Zeng, 12/16/2017

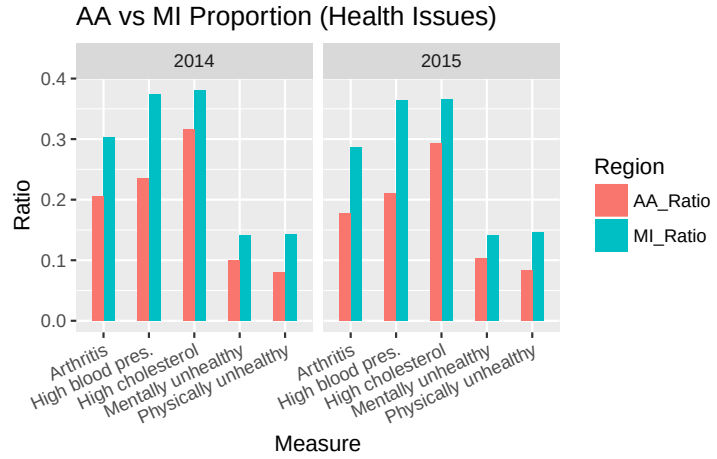
Introduction: Health condition is widely concerned by individuals as well as organizations; it is relevant to our lives. Researching the health problems such as unhealthy behavior and health outcomes (illnesses) in our place can raise public awareness about health and encourage individuals to take initiative in preventing health problems.

Research Question 2: The report focuses on **identifying** the top 3 nationwide unhealthy behaviors and top 5 health issues and **comparing** the level of unhealthy behavior and health issues of Ann Arbor compared to statewide level.
Methodology: In data cleaning part, I shorten the label of unhealthy behaviors and health issues in order to make nice graphics. To rank the top nationwide health problems, I use **population** as metric and use **slice** command from **dplyr** after sorting data. To compare the health condition between Ann Arbor and Michigan, I use **ratio** (proportion of population) as metric and use **gather** command from **tidyr** before graphing. Finally, to address the bias within dataset (underrepresentation in some states), I plot a representation heatmap (see Appendix) using packages **choroplethr** and **choroplethrMaps**.

Data description: The data has 1620206 rows containing the proportion of population in **500 cities** nationwide suffering 13 health issues, committing 5 unhealthy behaviors, and taking 9 prevention measures, in 2014 and 2015.
Data Source: The data is available at data







Findings:

- **California** and **Texas** are mostly represented states, in which 121 and 47 cities joined the study. However, there are 12 states where only 1 city is represented. (See Appendix)

- The top 3 nationwide unhealthy behaviors are **inadequate sleeping**, **obesity** and, **no time physical activity** in 2014, but the rank changed to **obesity**, **no time physical activity** and **smoking** in 2015 due to missing data. (Fig. 1)
- The top 5 health issues nationwide are **Arthritis**, **high blood pressure**, **high cholesterol**, **mentally and physically unhealthy** in both years. All the figures remain constant except for a drop in **high cholesterol** in 2015. (Fig. 2)
- The data shows that except for binge drinking, Ann Arbor has a lower overall rates in both unhealthy behavior and health issues compared to statewide level, implying that the health condition in Ann Arbor is **better** than state average. (Fig. 3,4)

Conclusion:

- Although the data is biased, the result is sufficient to show that inadequate sleep is worth concern nationwide.
- Residents in Ann Arbor should be aware of binge drinking since it is definitely an unhealthy behavior.

Appendix: Graphs, Tables and Code Chunk

Heatmap: Number of cities represented each State

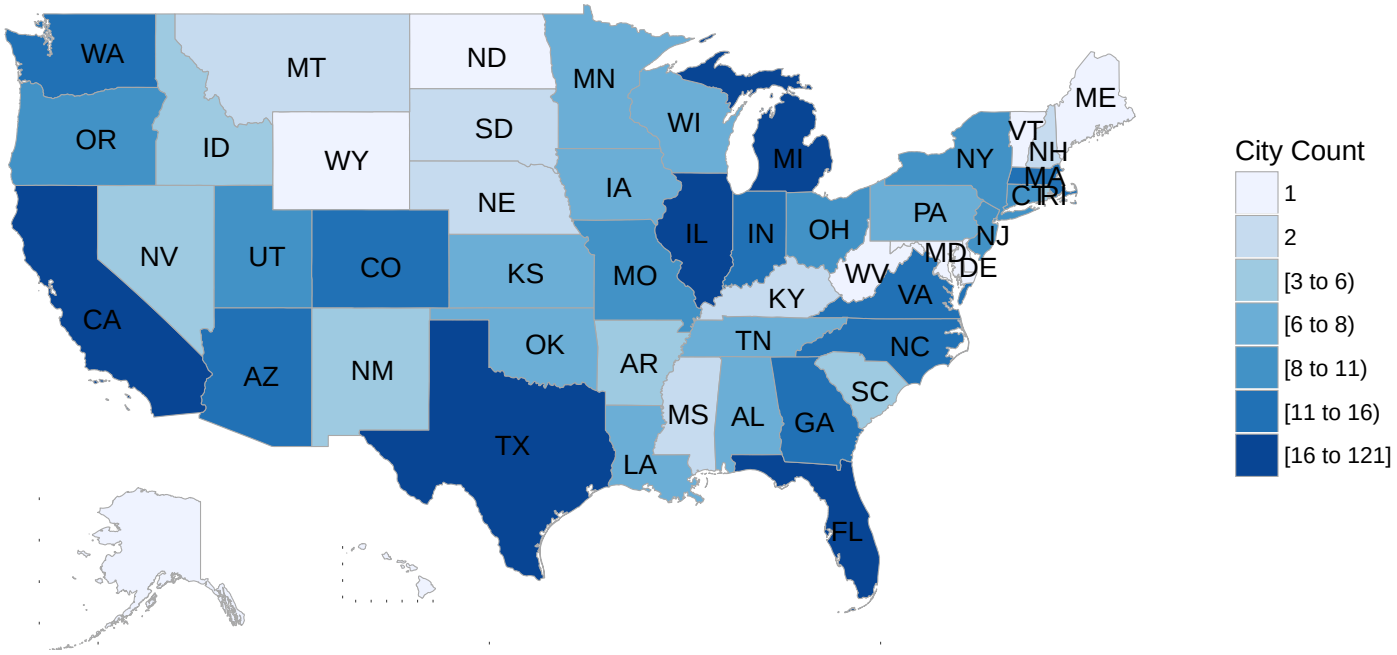


Table 1: Mostly represented states

region	value
california	121
texas	47
florida	33
illinois	18
michigan	16
colorado	14
north carolina	14
washington	14

Table 2: Mostly underrepresented states

region	value
alaska	1
delaware	1
district of columbia	1
hawaii	1
maine	1
maryland	1
north dakota	1
vermont	1
west virginia	1
wyoming	1

```r
library(dplyr)
library(tidyr)
library(ggplot2)
library(maps)
library(maptools)
library(SpatialEpi)
library(gridExtra)
library(data.table)
library(choroplethr)
library(choroplethrMaps)
library(reshape2)

# Data cleaning
Measure <- c("Current lack of health insurance among adults aged 18\342\200\22364 Years",
             "Arthritis among adults aged >=18 Years",
             "Binge drinking among adults aged >=18 Years",
             "High blood pressure among adults aged >=18 Years",
             "Taking medicine for high blood pressure control among adults aged >=18 Years wit
             "Cancer (excluding skin cancer) among adults aged >=18 Years",
             "Current asthma among adults aged >=18 Years",
             "Coronary heart disease among adults aged >=18 Years",
             "Visits to doctor for routine checkup within the past Year among adults aged >=18
             "Cholesterol screening among adults aged >=18 Years",
             "Fecal occult blood test, sigmoidoscopy, or colonoscopy among adults aged 50\342\
             "Chronic obstructive pulmonary disease among adults aged >=18 Years",
             "Physical health not good for >=14 days among adults aged >=18 Years",
             "Older adult men aged >=65 Years who are up to date on a core set of clinical pre
             "Older adult women aged >=65 Years who are up to date on a core set of clinical p
             "Current smoking among adults aged >=18 Years",
             "Visits to dentist or dental clinic among adults aged >=18 Years",
             "Diagnosed diabetes among adults aged >=18 Years",
             "High cholesterol among adults aged >=18 Years who have been screened in the past
             "Chronic kidney disease among adults aged >=18 Years",
             "No leisure-time physical activity among adults aged >=18 Years",
             "Mammography use among women aged 50\342\200\22374 Years",
             "Mental health not good for >=14 days among adults aged >=18 Years",
             "Obesity among adults aged >=18 Years",
             "Papanicolaou smear use among adult women aged 21\342\200\22365 Years",
             "Sleeping less than 7 hours among adults aged >=18 Years",
             "Stroke among adults aged >=18 Years",
             "All teeth lost among adults aged >=65 Years")


Measure_renamed <- c("No health insurance",
             "Arthritis",
             "Binge drinking",
             "High blood pres.",
             "High blood pres. contr.",
             "Cancer",
```

```
              "Asthma",
              "CHD",
              "Routine checkup",
              "Cholesterol screening",
              "FOBT, sigmoidoscopy, colonoscopy",
              "COPD",
              "Physically unhealthy",
              "Men: Flu & PPV shot, Cancer screening",
              "Women: Flu & PPV shot, Cancer screening, Mammogram",
              "Smoking",
              "Dental clinic",
              "Diabetes",
              "High cholesterol",
              "CKD",
              "No time PA",
              "Mammography use Woman",
              "Mentally unhealthy",
              "Obesity",
              "Papanicolaou smear use",
              "Sleeping < 7 hrs",
              "Stroke",
              "All teeth lost")

matched <- data.frame(Measure, Measure_renamed) %>%
  mutate_if(is.factor, as.character)

# Import dataset
df2 = fread("/Users/son520804/Desktop/STATS506/Individual Project/500_Cities__Local_Data_for_B
df3 = fread("/Users/son520804/Desktop/STATS506/Individual Project/500_Cities__Local_Data_for_B
Health_500 <- rbind(df2, df3) %>%
  mutate(Measure = matched$Measure_renamed[match(Measure, matched$Measure)]) %>%
  filter(Measure != "All teeth lost")

# Representation of each state (Motivation: Underrepresentation is always an issue worth con
state_rep <- Health_500 %>%
  group_by(StateDesc,CityName) %>%
  summarize(frequency=n()) %>%
  group_by(StateDesc) %>%
  summarize(CityCount=n()) %>%
  filter(StateDesc != "United States") %>% # Delete the row to fit the map
  mutate(StateDesc = lapply(StateDesc,tolower)) %>% # Change the state name to lowercase
  rename(region = StateDesc, value = CityCount)%>%
  mutate(region = unlist(region))

# Import the U.S.Map and evaluate (heat map)
staterep_heatmap = state_choropleth(state_rep,
                title = "Heatmap: Number of cities represented each State",
                legend = "City Count")
staterep_heatmap
```

```r
state_rep_table1 <- state_rep %>%
  arrange(-value) %>%
  slice(1:8)
kable(state_rep_table1, caption = "Mostly represented states")
state_rep_table2 <- state_rep %>%
  filter(value == 1)
kable(state_rep_table2, caption = "Mostly underrepresented states")

# Change of scientific notation
fancy_scientific <- function(l) {
     # turn in to character string in scientific notation
     l <- format(l, scientific = TRUE)
     # quote the part before the exponent to keep all the digits
     l <- gsub("^(.*)e", "'\\1'e", l)
     # turn the 'e+' into plotmath format
     l <- gsub("e", "%*%10^", l)
     # return this as an expression
     parse(text=l)
}

# Top 3 nationwide behaviors
behaviors <- Health_500 %>%
  filter(Category == "Unhealthy Behaviors", StateDesc != "United States") %>%
  mutate(Data_Value = Data_Value / 100,
         Pop = PopulationCount * Data_Value,
         Year = as.factor(Year)) %>%
  group_by(Year, Measure) %>%
  summarize(Count = sum(Pop, na.rm=TRUE)) %>%
  arrange(-Count)%>%
  slice(1:3)

# Barplot (GGplot)
barplot_behaviors <- ggplot(behaviors, aes(x=Year, y=Count))+
                     geom_bar(stat="identity", aes(fill = Measure), position = "dodge") +
                     labs(title="Nationwide: Pop (Unhealthy Behaviours)") +
                     annotate("text", label="*Missing data:",
                              x = as.factor(2015), y = 200000000, size = 3, colour="blue")+
                     annotate("text", label="sleeping <7 hrs (2015)",
                              x = as.factor(2015), y = 180000000, size = 3, colour="blue")+
                     scale_y_continuous(labels=fancy_scientific)


# Top 5 illnesses
illnesses <- Health_500 %>%
  filter(Category == "Health Outcomes", StateDesc != "United States") %>%
  mutate(Data_Value = Data_Value / 100,
         Pop = PopulationCount * Data_Value) %>%
  group_by(Year, Measure) %>%
  summarize(Count = sum(Pop, na.rm=TRUE)) %>%
```

```
  ungroup() %>%
  mutate(Year = ifelse(Year =="2013",2014,
                       ifelse(Year =="2014", 2014, 2015))) %>%
  group_by(Year) %>%
  arrange(-Count)%>%
  slice(1:5)

#dput(unique(illnesses$Measure))

# Barplot (GGplot)
barplot_illnesses <- ggplot(illnesses, aes(x=Year, y=Count))+
                     geom_bar(stat="identity", aes(fill = Measure), position = "dodge") +
                     labs(title="Nationwide: Pop (Health Issues)") +
                     scale_x_continuous(breaks= c(2014,2015)) +
                     scale_y_continuous(labels=fancy_scientific)

grid.arrange(barplot_behaviors, barplot_illnesses, ncol=2)

# Comparison between Ann Arbor and the State standard
MI <- Health_500 %>%
  filter(StateAbbr=="MI") %>%
  mutate(Data_Value = Data_Value / 100,
         Pop = PopulationCount * Data_Value) %>%
  group_by(Year, Category, Measure,CityName) %>%
  summarize(Population=sum(PopulationCount,na.rm=TRUE), Count = sum(Pop, na.rm=TRUE))%>%
  ungroup()%>%
  group_by(Year, Category, Measure) %>%
  summarize(AA_Ratio = Count[CityName %in% "Ann Arbor"] / Population[CityName %in% "Ann Arbor"


# Two graphs (will be changed to one)

MI_unhealthy <- MI %>%
  tidyr::gather(Region, Ratio, c(`AA_Ratio`,`MI_Ratio`)) %>%
  filter(Category =="Unhealthy Behaviors")

# This data.frame is for the sake of annotation (specific)
annotation_text1 <- data.frame(Measure = 3,Ratio = 0.41, Region=as.factor(MI_unhealthy$Region)
annotation_text2 <- data.frame(Measure = 3,Ratio = 0.37, Region=as.factor(MI_unhealthy$Region)

Unhealthy<- ggplot(MI_unhealthy, aes(x = Measure, y= Ratio, fill = Region), xlab="") +
  geom_bar(stat="identity", width=.5, position = "dodge")+
  labs(title="AA vs MI Proportion (Unhealthy Behaviours)") +
  theme(axis.text.x = element_text(angle = 25, hjust = 1)) +
  facet_wrap(~Year) +
  geom_text(data=annotation_text1, size=3, colour="blue", label="*Missing data:") +
  geom_text(data=annotation_text2, size=3, colour="blue", label="sleeping < 7 hrs(2015)")

MI_illnesses <- MI %>%
```

```
  filter(Category =="Health Outcomes") %>%
  ungroup() %>%
  mutate(Year = ifelse(Year =="2013",2014,
                       ifelse(Year =="2014", 2014, 2015))) %>%
  mutate(Total_Ratio = AA_Ratio + MI_Ratio) %>%
  group_by(Year) %>%
  arrange(-Total_Ratio) %>%
  slice(1:5) %>%
  ungroup() %>%
  dplyr::select(-`Total_Ratio`) %>%
  tidyr::gather(Region, Ratio, c(`AA_Ratio`,`MI_Ratio`))

Illnesses<- ggplot(MI_illnesses, aes(x = Measure, y= Ratio, fill = Region), xlab="") +
  geom_bar(stat="identity", width=.5, position = "dodge") +
  labs(title="AA vs MI Proportion (Health Issues)") +
  theme(axis.text.x = element_text(angle = 25, hjust = 1)) +
  facet_wrap(~Year)

grid.arrange(Unhealthy, Illnesses, ncol=2)
```