Analysis of American’s Most Prominent Health Issues

Vincenzo Coppola, Ashish Sharma, Akshay Dwivedi

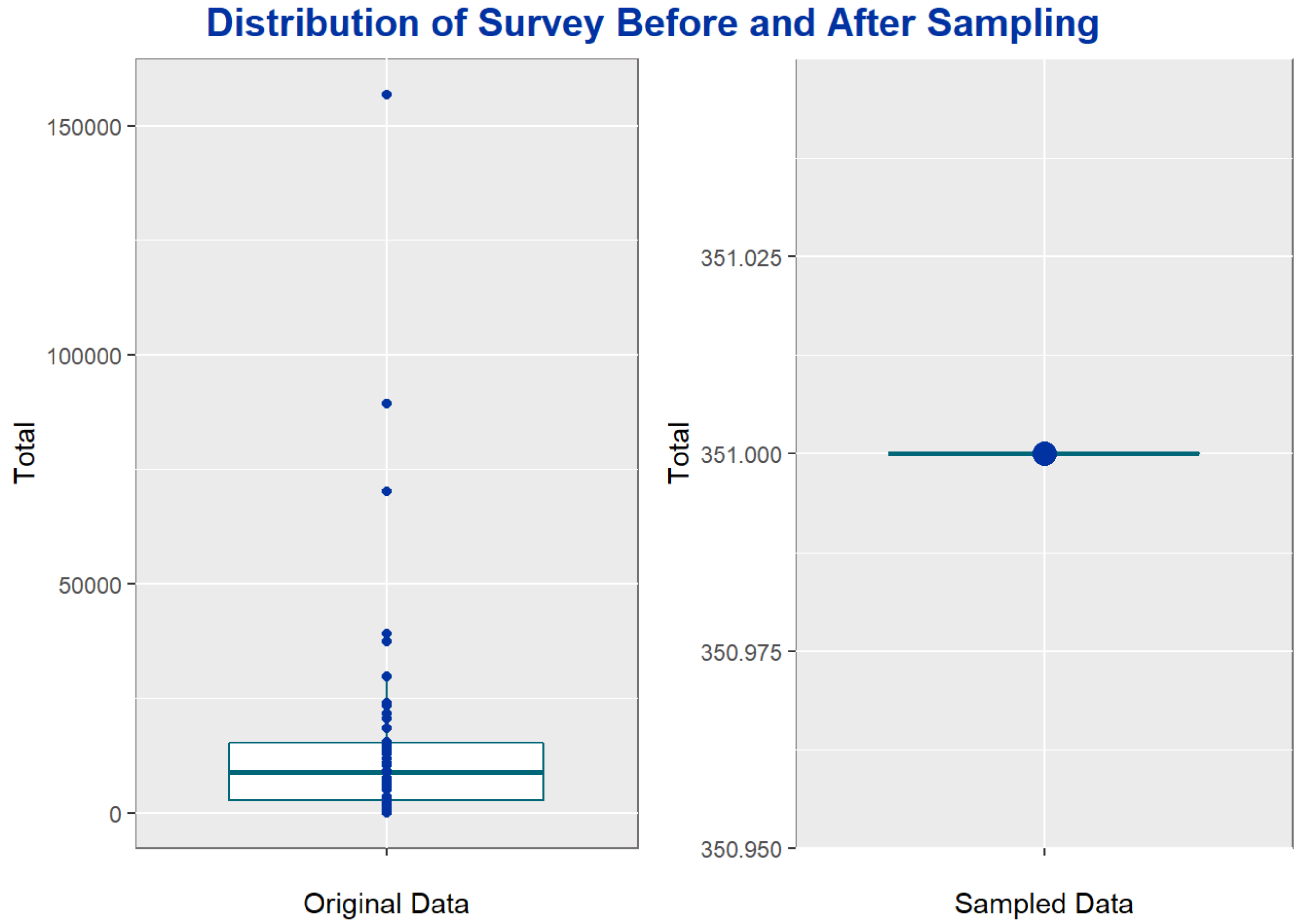
2/20/2022

Introduction

The “Centers for Disease Controls and Prevention” (CDC) is tasked with protecting America from “health, safety and security threats, both foreign and in the U.S.”. In recent years especially with COVID-19, the CDC has claimed to have America’s best interests in mind with their policies. Although it shouldn’t be, the CDC has been under scrutiny in many states and has been a political discussion. Regardless of peoples’ personal opinions, the health of our nation is at stake and the well being of its people is the most important thing. Therefore, the CDC “conducts critical science and provides health information that protects America against expensive and dangerous health threats” (source 5) responding when new ones arise. As the public’s health is absolutely a social issue, we as a group decided to take on the data surrounding America’s health so we could visualize any relationships in the data and see the facts at face value. In this report we will visualize some things to answer the questions we have about America its health and the CDC’s handling of the problems the US faces. As medical data is so important in the lives of the people it describes its important to use valid data and so first we will explain how we interpreted the to data to use.

There often exists a problem in surveyed data where the number of polled people in a specific location is directly proportional to the population of that location. In other words, there are far more people polled in areas of large population than in areas of small population. This can create bias in the data when looking at the actual statistics because the numbers in the densely surveyed locations will far outweigh the totals of the lesser surveyed ones. This is a problem because that simply might not be true. Looking at our data, we hypothesized this would be the case as some states, specifically California, Texas and New York would likely be far out surveyed due to their sheer overwhelming population over states like Delaware, Michigan and Alabama. Therefore this would uncover the common social issue that comes with surveying data. In order to look into this we analyzed the distribution of the data.

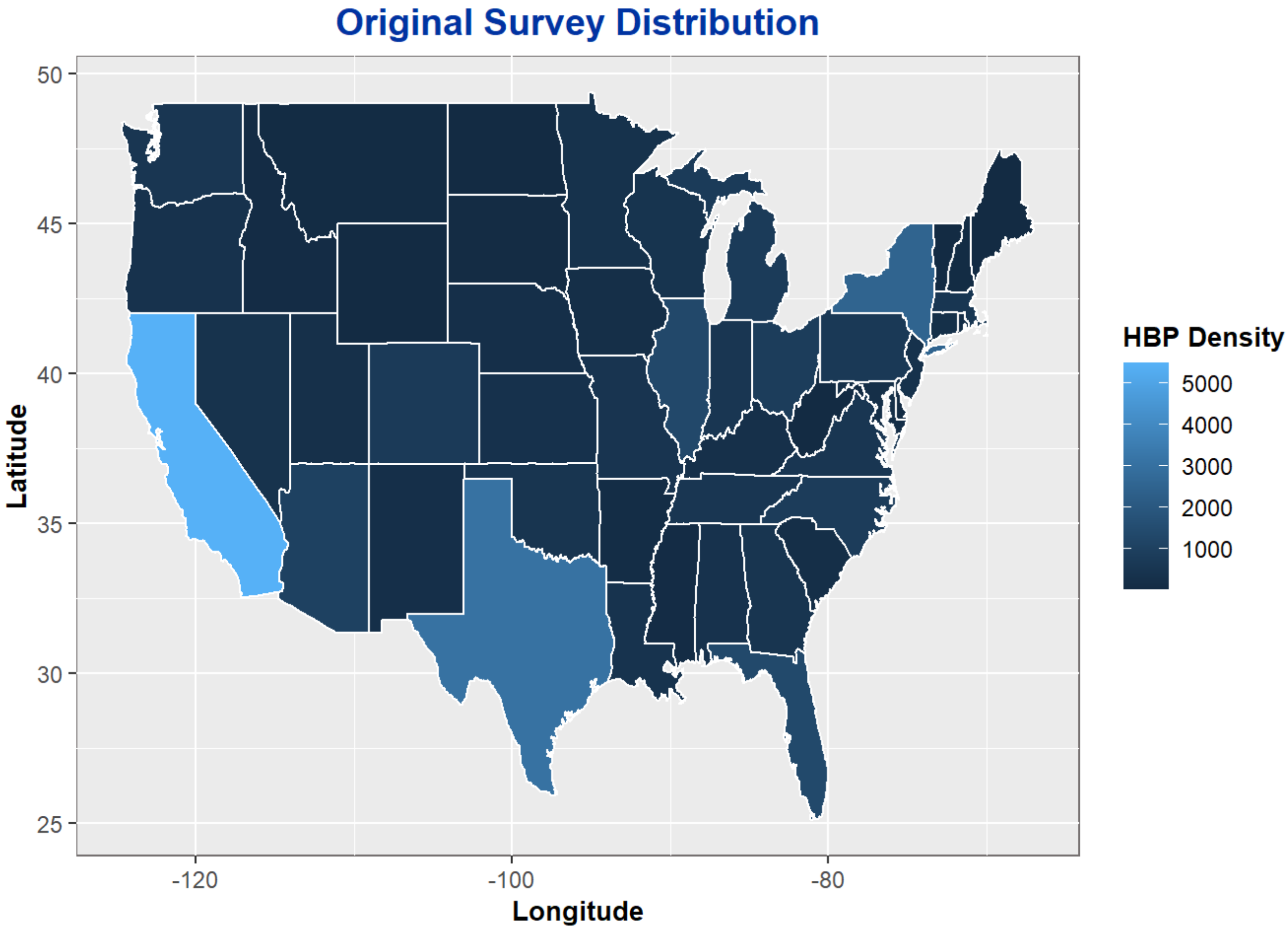
```
grid.arrange(bal_outliers_orig,bal_outliers,ncol=2,top =title1)
```



As we look at the distribution of the total data, it is clear that the original data has extreme outliers which are notably California, New York and Texas (the most populated states as hypothesized). On the right, the balanced data used in this analysis can be seen and all of the states have exact equal representation.

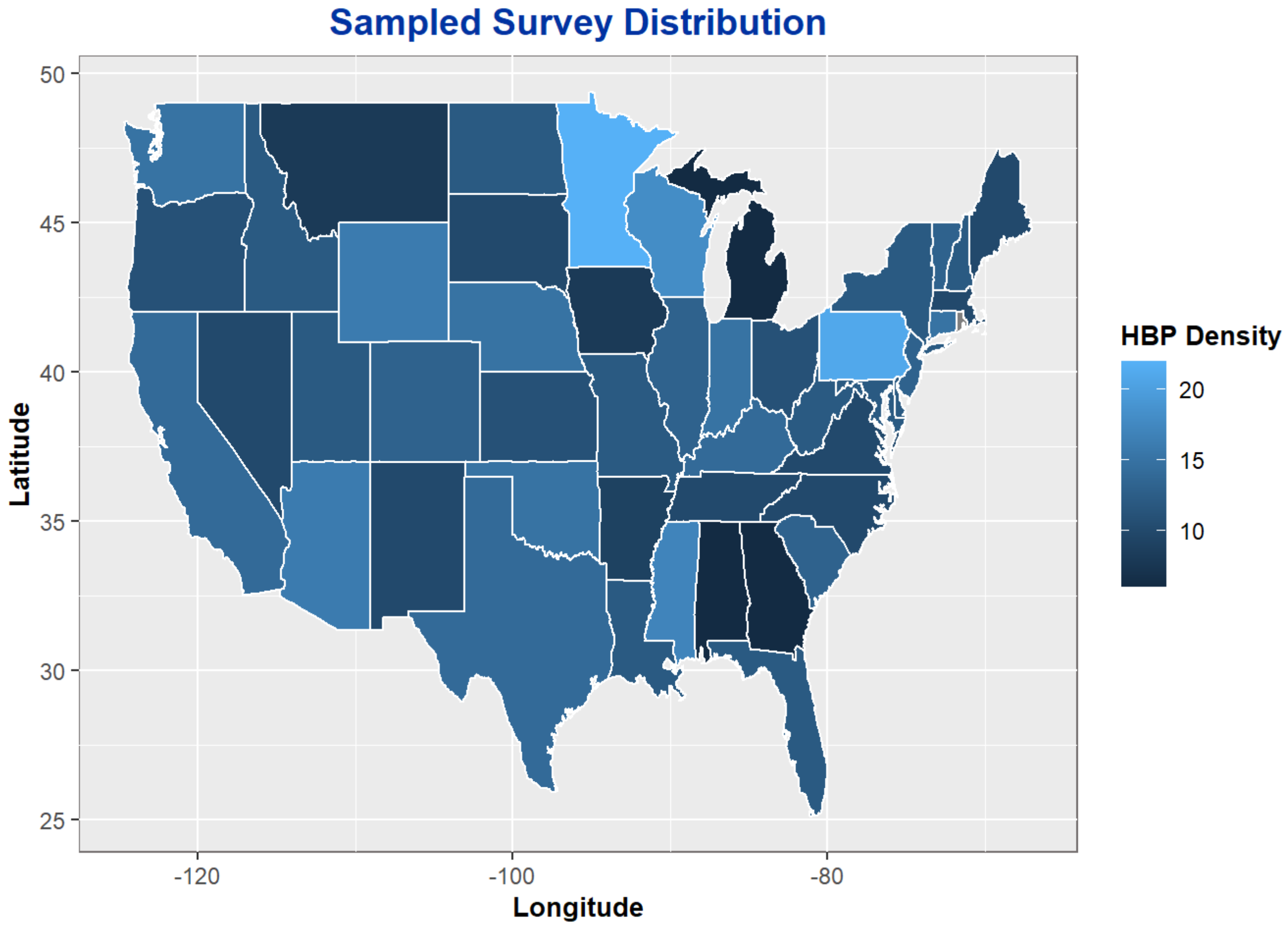
Sampling like this can also have its drawbacks but we wanted to ensure that it better represented the population of the U.S. Therefore, we created three visualizations which show similar information. The first, will show how the original dataset information is distributed on the US map using the total High Blood Pressure cases surveyed.

```
p1
```



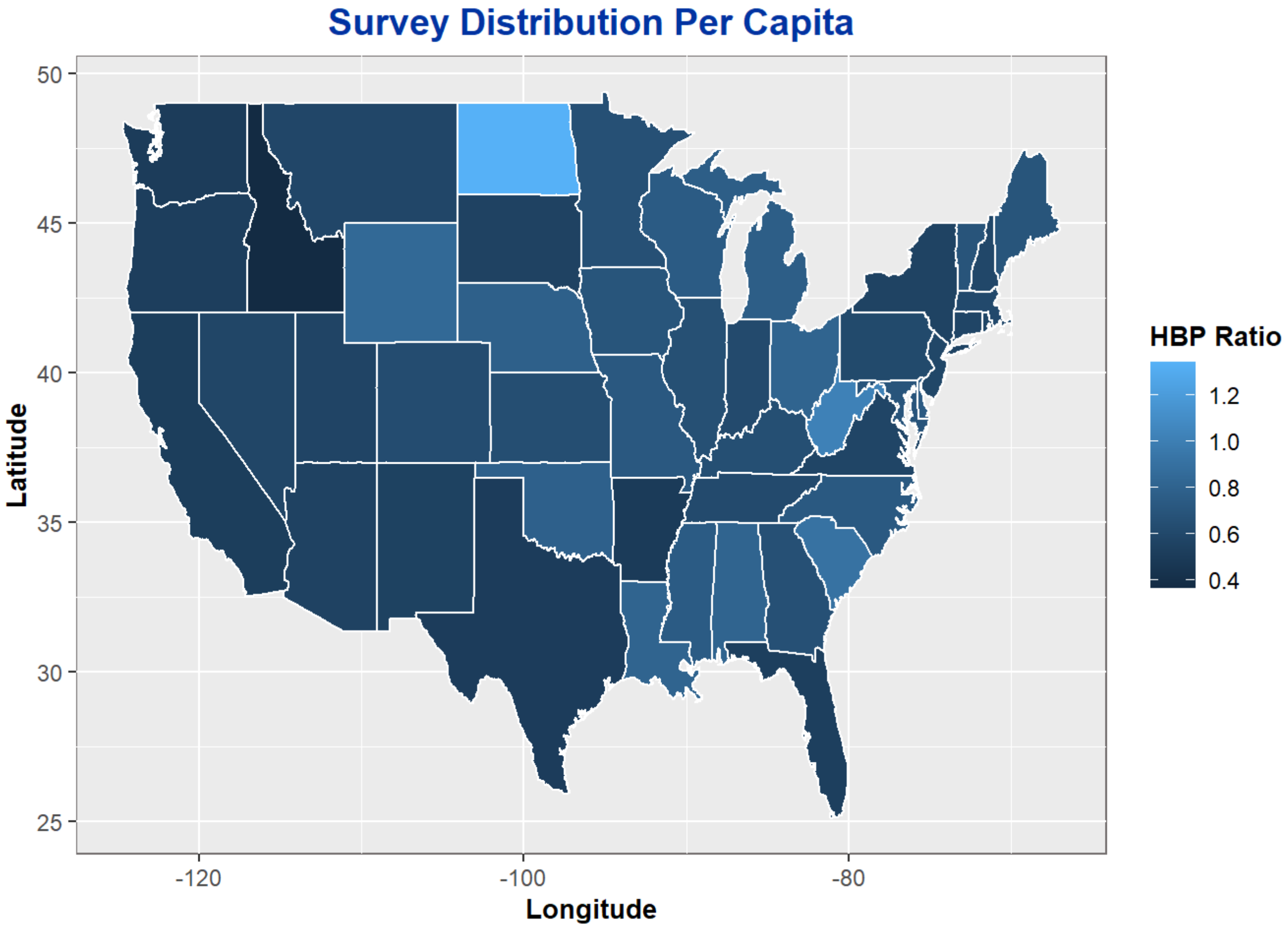
As can be seen, California, Texas and New York are far and away the brightest states and all other states are relatively the same shades of blue. This would be fine if the facts are that way. Looking at the aforementioned down sampled data we see the following plot.

```
p2
```



This shows those same the issues we hypothesized, as with out the bias in the data all the states have a much more equal representation and some other states such as Pennsylvania and Minnesota actually represent the higher High Blood Pressure regions. In order to prove that this data is more representative we found the total number of diseases per capita and plotted on the same map. In order to do this we found the number of unique cities for each state and summed up the total populations to get a state population estimate and divided the total count of diseases per state by that number respectively. With this ratio we were able to plot the per capita diseases as follows:

```
p3
```



To conclude this data verification, our hypothesis was confirmed and as can be seen the per capita plot is much more represented by the down sampled data than by the original data collected by the CDC. This shows that in order to visualize the data accurately we needed to preprocess it and give a fair representation. This is an important conclusion as it clearly verifies the social issues that deal with public health survey accuracy/representation. This data is not the greatest measure of the public health of our country but in the following sections we plan to visualize some other conclusions to our hypotheses and better understand this CDC data.

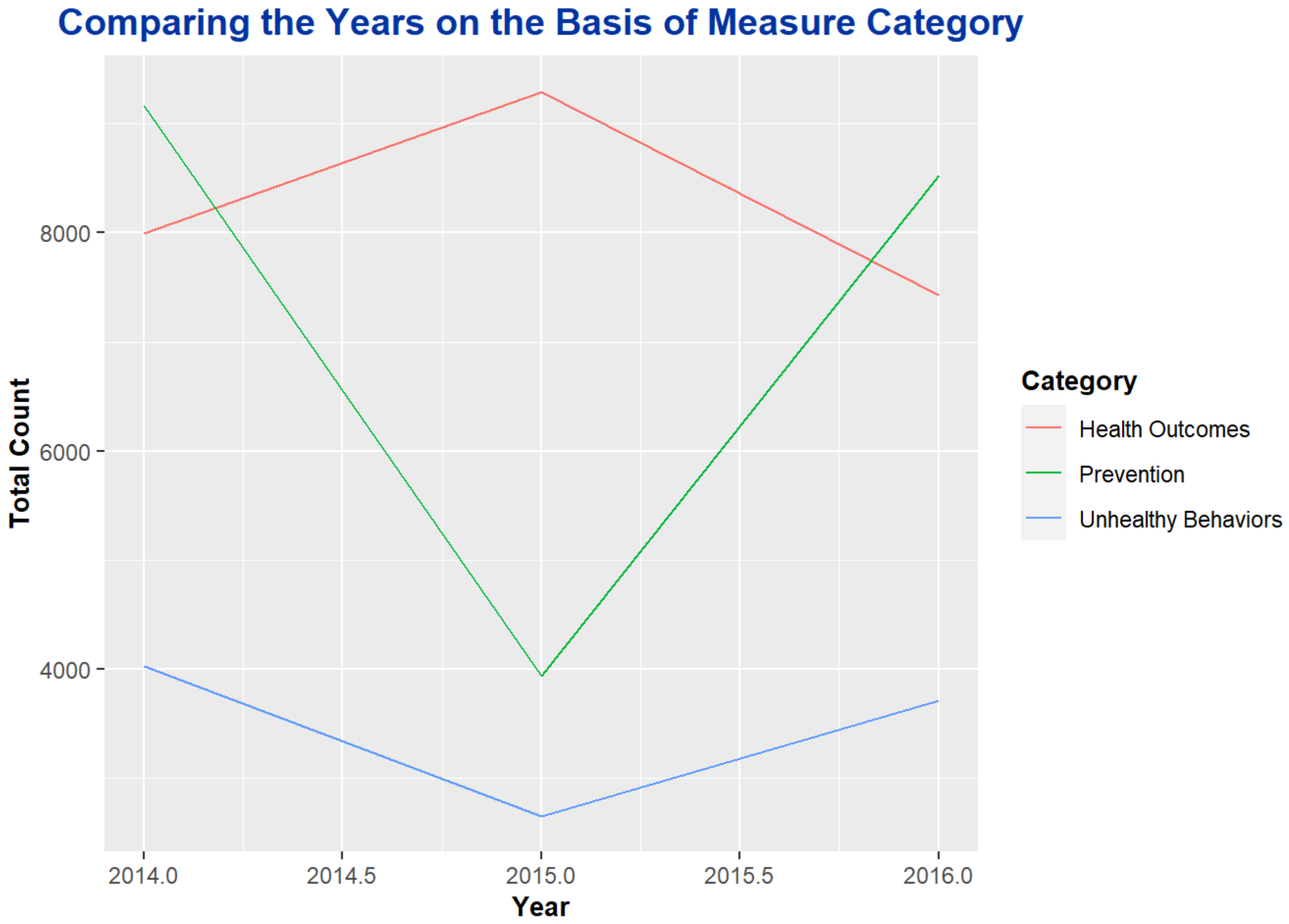
Section 1: Health Issue Relationships

First, we looked to explore the relationships between different health issues in America. It is easy to assume that health issues such as diabetes are related to other issues such as high cholesterol. We shared this hypothesis and believed that there is likely many relationships in the issues within our data. This hypothesis raised the question:

Q1. How do the CDC health measures relate to each other, specifically how do the unhealthy habits relate to preventative measures, as well as a lack of health insurance?

Therefore, we created the following visualizations first importing additional yearly data so we could see a yearly comparison of preventative measures, unhealthy behaviors and health outcomes.

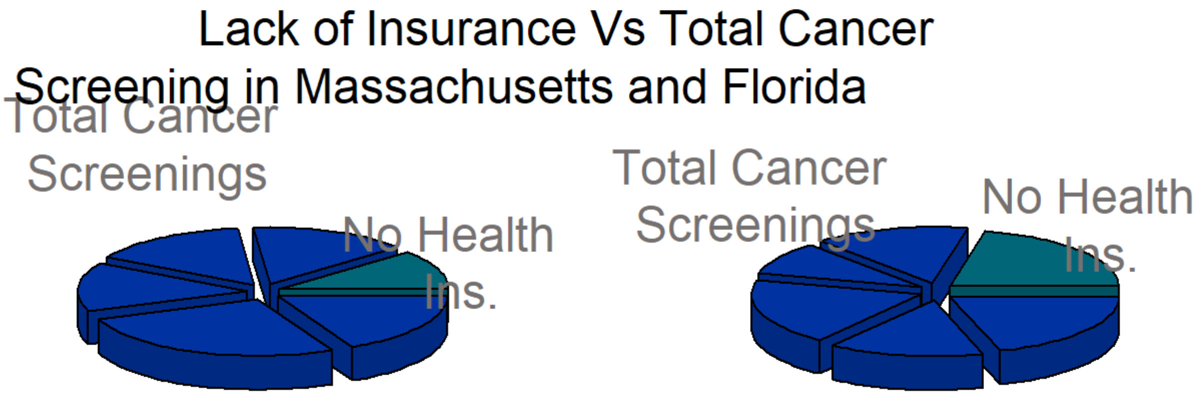
```
line_p
```


Comparing the Years on the Basis of Measure Category

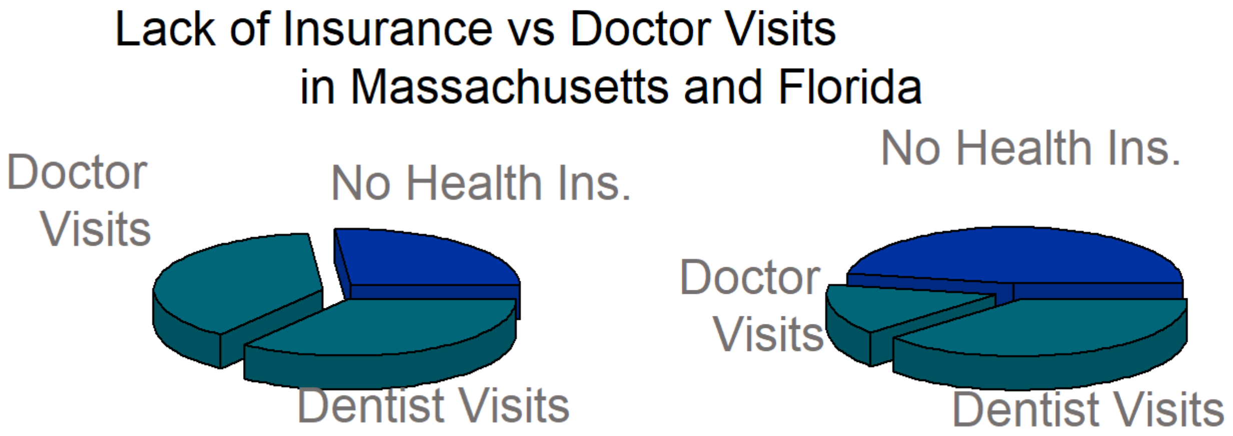
As can be seen in the line graph, as preventative measures decreased over the years, health outcomes increased and vice versa. The data clearly shows this trend and while it could be simply due to the survey it makes sense that as health prevention is not done, worse health issues arise.

To further see relations of health issues, we created another visualization as we hypothesized that health insurance is related to these useful preventive measure such as regular doctor and dentist visits as well as cancer screenings. In the below pie charts, we visualize this hypothesis.

```
line = -1
cex = 1.5
side = 3
adj=1.5
par(mfcol=c(1,2))
pie3D(df_for_pie1$Totals_for_pie1, labels = pielabels, explode = 0.1, col = c("#016678", "#0033A1","#0033A1","#0033A1","#0033A1","#0033A1"), labelcol = "#767171", col.main = "Black")

pie3D(df_for_pie2$Totals_for_pie2, labels = pielabels, explode = 0.1, col = c("#016678", "#0033A1","#0033A1","#0033A1","#0033A1","#0033A1"), labelcol = "#767171", col.main = "Black")
mtext("Lack of Insurance Vs Total Cancer\n Screening in Massachusetts and Florida", side=side, line=line, cex=cex, adj=adj)
```



```
line = -1
cex = 1.5
side = 3
adj=2
par(mfcol=c(1,2))

pie3D(df_for_pie3$Totals_for_pie3, labels = pielabels2, explode = 0.1, col = c("#0033A1","#016678","#016678"), labelcol = "#767171", col.main = "Black")

pie3D(df_for_pie4$Totals_for_pie4, labels = pielabels2, explode = 0.1, col=c("#0033A1", "#016678","#016678"), labelcol = "#767171", col.main = "Black")
mtext("Lack of Insurance vs Doctor Visits\n in Massachusetts and Florida", side=side, line=line, cex=cex, adj=adj)
```



Conclusion

The data seen in this section clearly show there are relationships in the preventative health measures and the outcomes. As could be seen with the line chart these were inversely proportionate and as more clearly could be seen in the pie charts as the Lack of Access slice increased (In low income states such as Florida), the preventative measures such as doctor visit and cancer screenings both decreased. This is important data to see as it promotes the ideas of universal free healthcare.

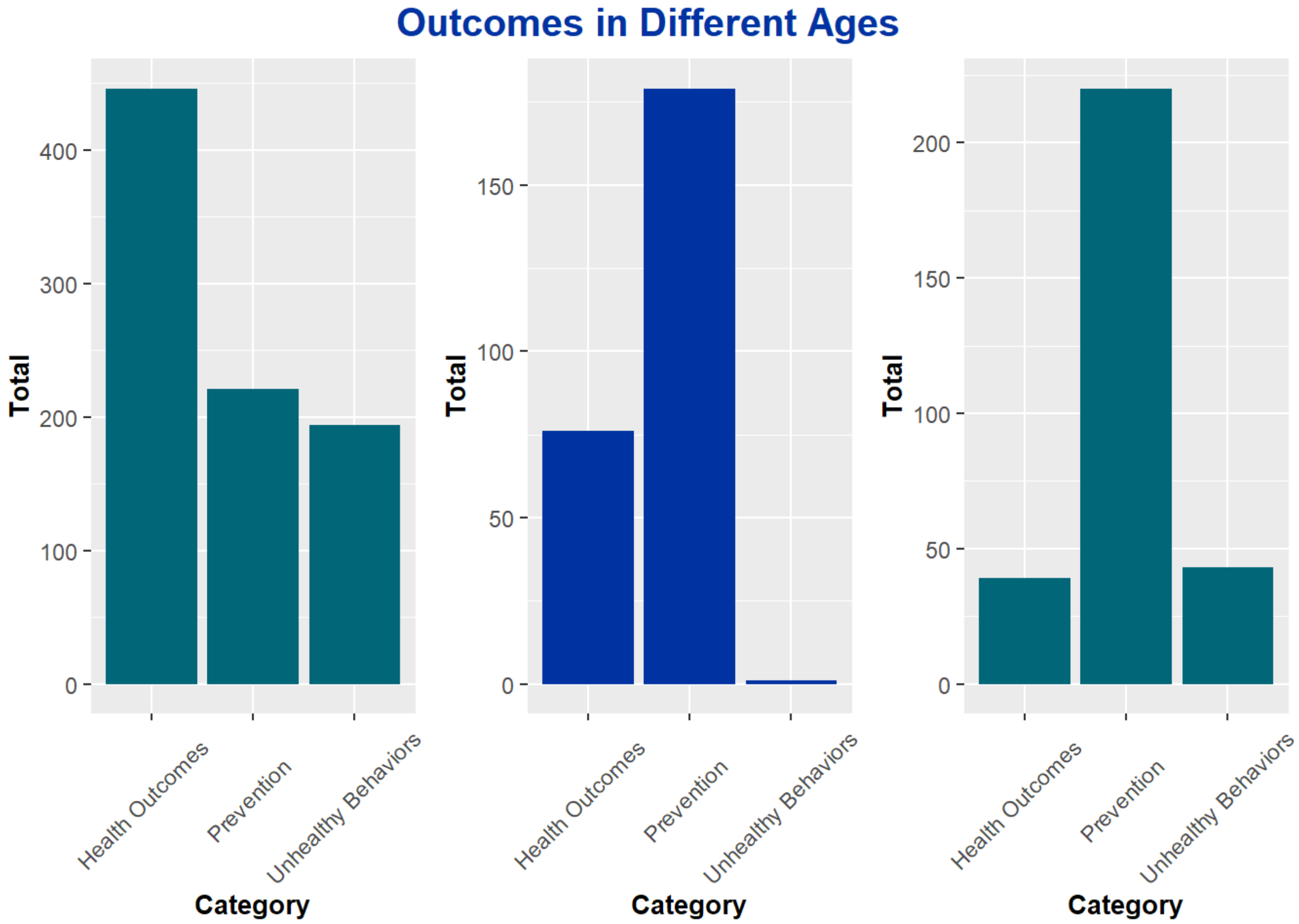
Section 2: Demographics

Next, we will introduce how ages and genders may relate to health issues and preventative measures. First we will address the following question as we hypothesize there is some relation to out declining health as we age as well as depending on various genders.

Q2. How are different types of diseases and preventives related to the demographics of America?

First, we aimed to visualize a trend in age and in order to do so we extracted the ages present in the health types of the data set. Plotting the totals for these age ranges and looking at the categories of health (Health Outcomes, preventative, and Unhealthy Behaviors). These categorize the data as either a health consequence/developed disease, preventative/cautious measure, or an unhealthy behavior respectively. After plotting the number of cases for each category respectively and separating them by age we can see the below plot.

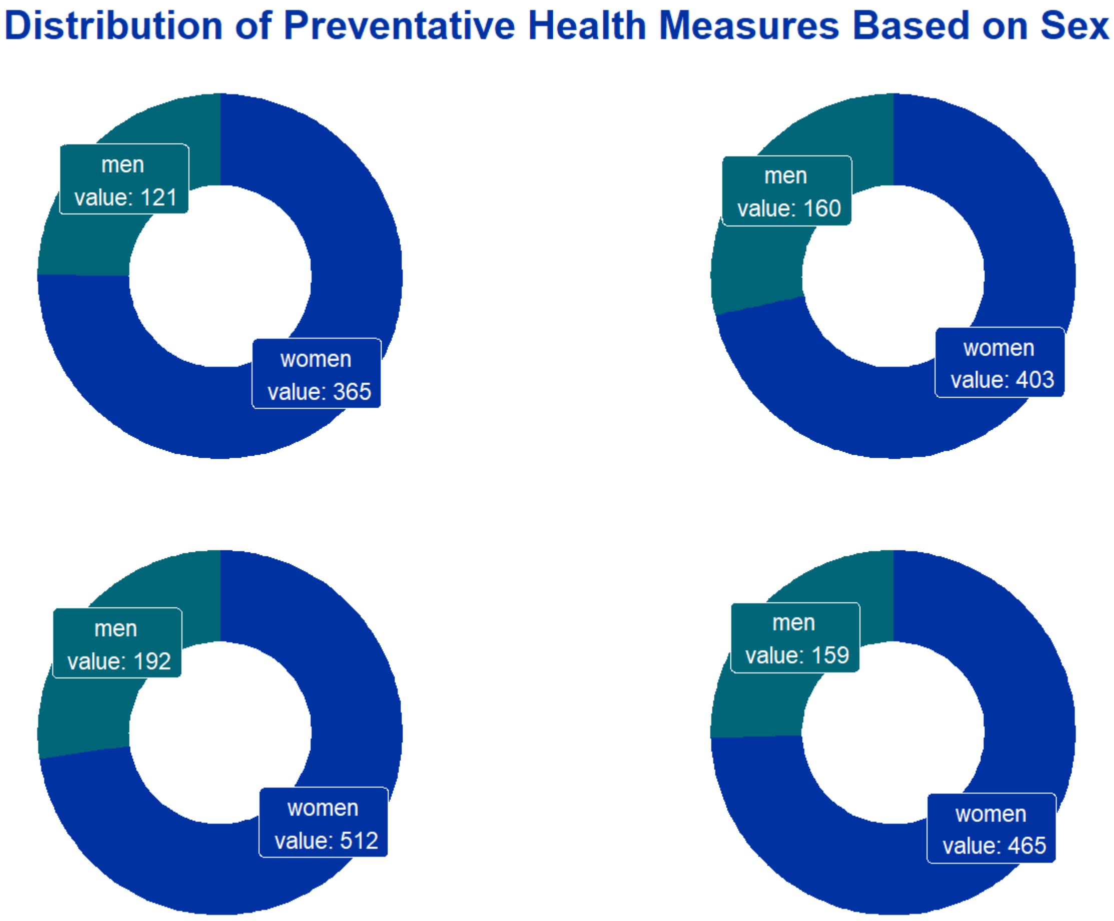
```
grid.arrange(eighteens,fifties,old,ncol = 3,top=title_bar1)
```



This tells us some useful information which may have been intuitive to hypothesize. As age increases and the plots move left to right unhealthy behaviors such as binge drinking, sleeping less than 7 hours and smoking tend to decrease while preventative measures such as vaccines and cancer screenings increase.

Next, we aimed to explore if a trend in sex exists and in order to do so we extracted the occurrences of men and women present in the health types of the dataset. In doing so it was discovered that these were mainly occurring in the preventative health concerns such as cancer screening, mammography, Papanicolaou test and colonoscopies. In order to actually visualize this we created four pie charts as seen below. We first assigned all states to a respective region based on their location. We then plotted the totals for all male preventative cases vs all female preventative cases for each region.

```
grid.arrange(west_pie, East_pie, midwest_pie, south_pie,nrow=2,ncol=2,top= title_pie2d)
```



Distribution of Preventative Health Measures Based on Sex

From the data visualizations, it is easily noticeable that the preventative measures surveyed for females far outweigh the measures taken for men. The pie charts are very distinctly separated and the trend is consistent for all regions regardless of location in the country.

Conclusion

Are these trends truly the case in the U.S.? Do we as a country truly increase our preventative measures and decrease our unhealthy behaviors as we get older? Does sex have an impact on the preventative measures taken for adults in that women are more likely to be screened for health concerns? The data suggests both to be true and it intuitively makes sense based on real life experiences but to truly confirm the trend more information should be collected by the CDC regarding the demographics of the U.S. and the people who they survey individually.

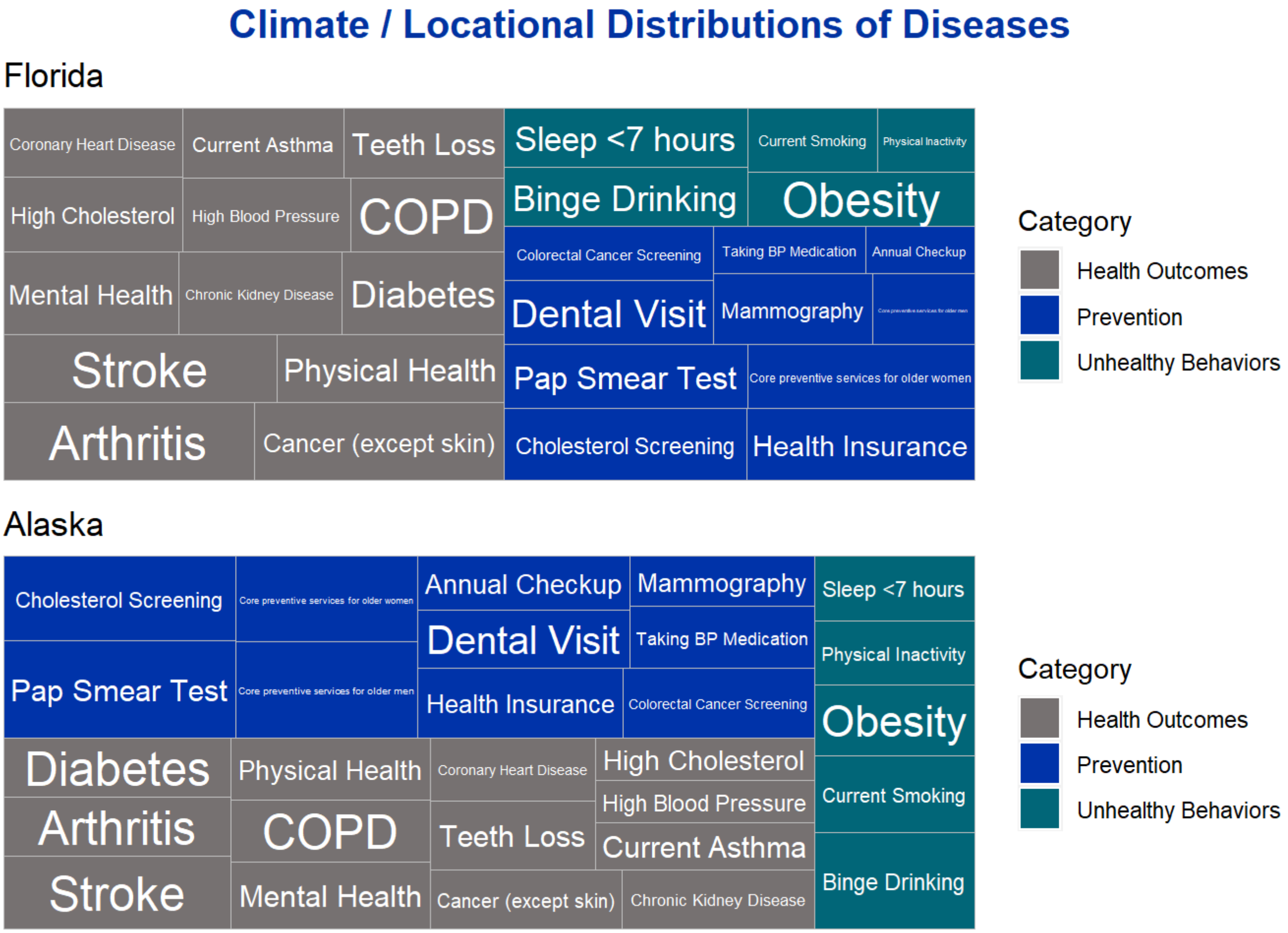
Section 3 - Location

Finally, as the name of the data set suggests, we wanted to look at just how the 500 cities, or the locations for that matter play a role in American health. To do this, we explored with two different visualizations to answer the following question.

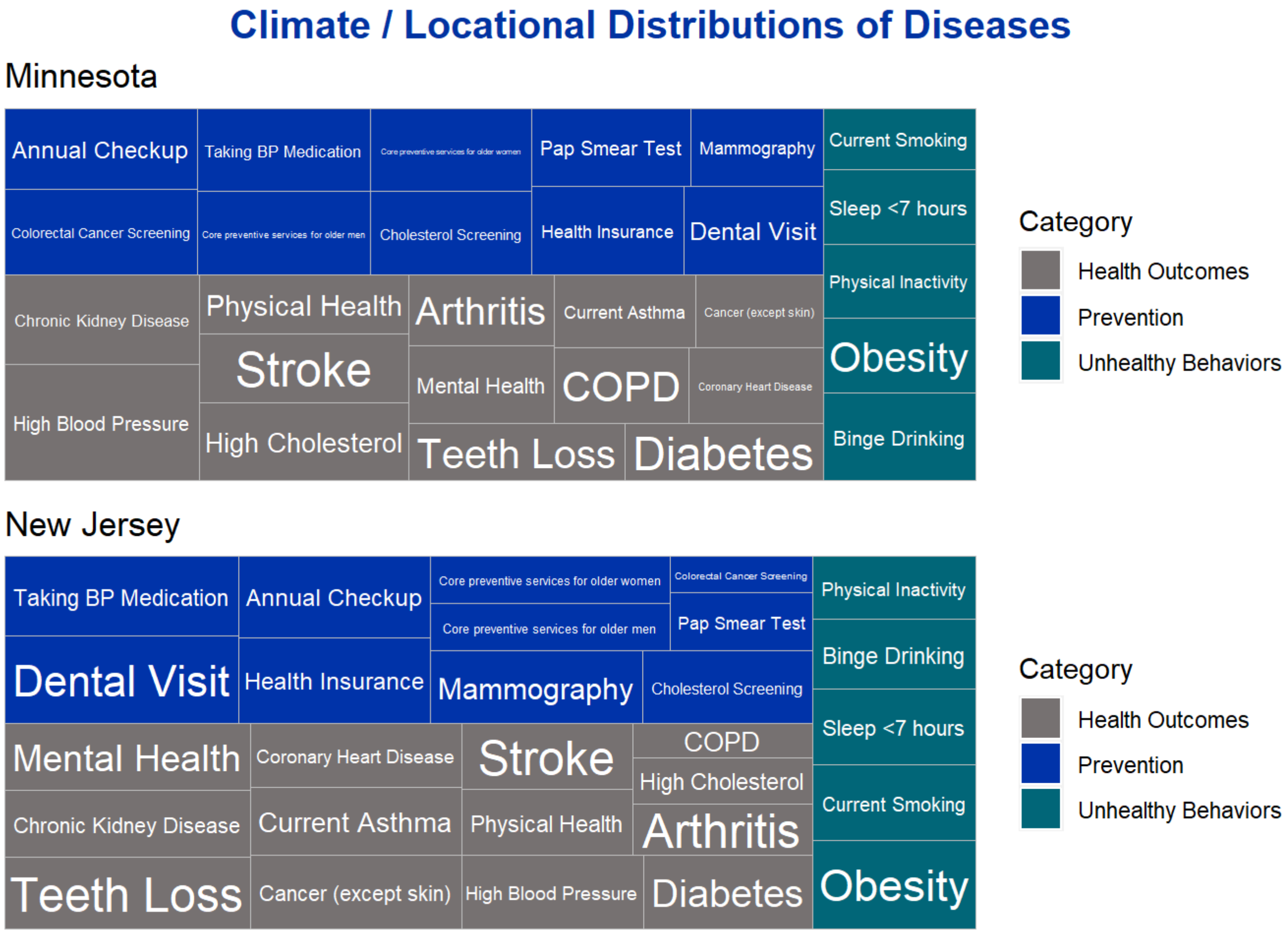
Q3. How do location factors effect the different types of health issues in an area?

In order to address this, we first wanted to look into how types of health issues are distributed for individual states in various areas of the U.S. or areas of differing climates and political views. In order to do so we created the following four tree maps showing the differences in health issues for Alaska and Florida as well as New Jersey and Minnesota. These four states show varying climates and regions from very cold in Alaska to slightly warmer Minnesota to an Atlantic New Jersey and finally to a more topical Florida. The following are those results.

```
grid.arrange(florida_tree,alaska_tree,nrow=2,top=title_tree)
```



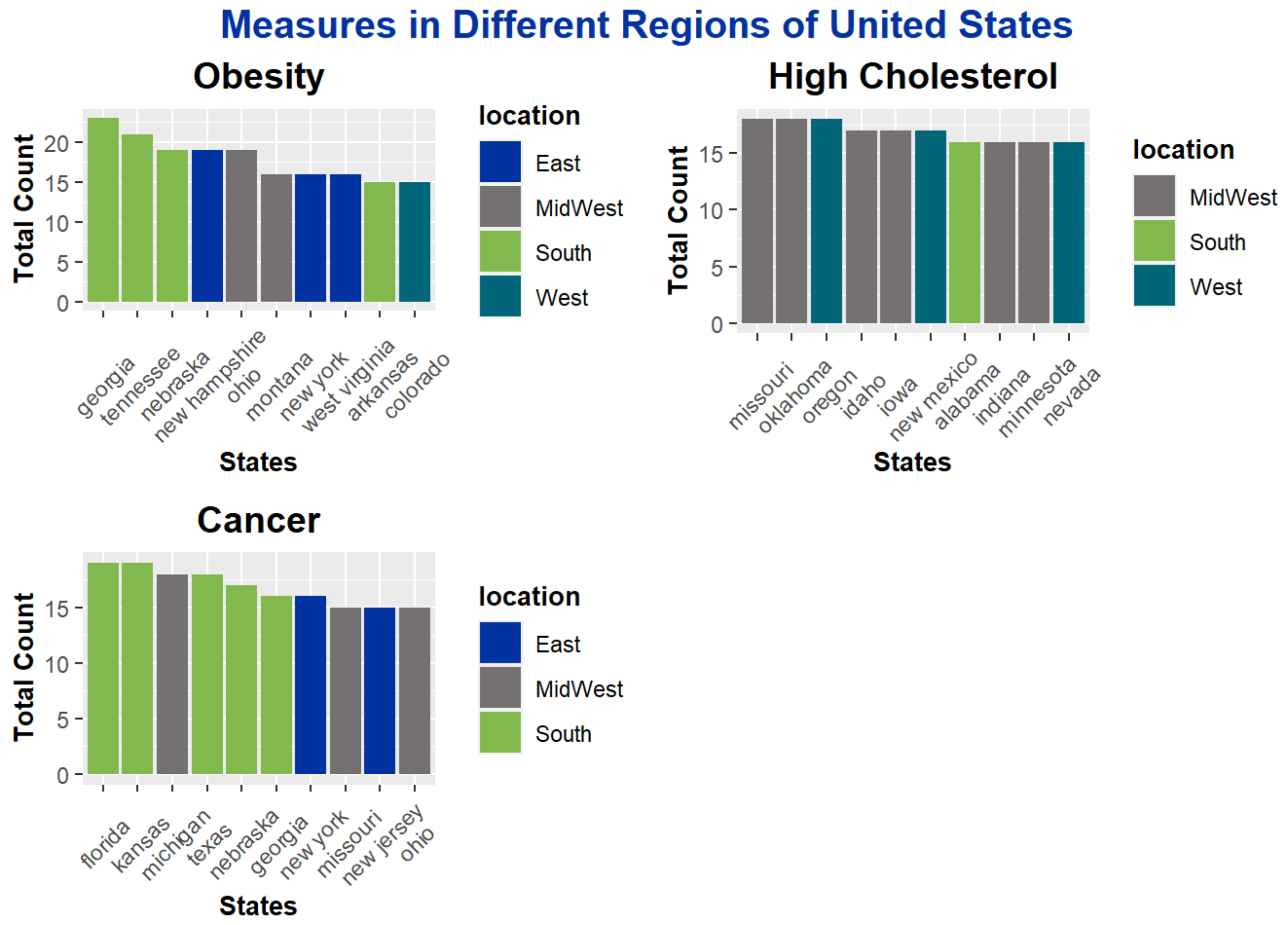
```
grid.arrange(minn_tree,nj_tree,nrow=2,top=title_tree)
```



As can be seen, there are many variances in the data from the tree maps. Namely the warmer state in Florida has far more frequency of Binge Drinking and unhealthy behaviors as well as a greater lack of access to health insurance. Whereas, Minnesota representing mid-western climates shows a higher frequency of high blood pressure population. Alaska represents a greater frequency of preventative measures for women and finally new jersey shows a great difference from the others in yearly dental visits, mental health issues and Asthma.

We then decided to look into the regions more specifically as East, Midwest, South and West to see how the differences would again appear. To do this we created the three following visualizations containing the top 10 states in frequency of health issues. The three health issues analyzed were obesity, high cholesterol, and cancer as these are three of the most major issues that Americans face. Below the top ten states can be seen plotted as bars with color correlating to one of the four regions.

```
grid.arrange(barp1,barp2,barp3,nrow=2,ncol=2,top=title_bar2)
```



Conclusion

In conclusion, location definitely does show a difference for health issues. As can be seen, warmer locations like Florida have more unhealthy habits possibly due to the party culture. It also has less access to health insurance compared to New Jersey which has a great deal of insurance access. This could be attributed to the incomes of each state as New Jersey is a high income state and Florida is not. The bar charts show a different story but also confirm the location effect. In the south, Obesity and Cancer are very dominant compared to the other regions. Meanwhile, high cholesterol while also prominent in the South is a major issue in the Mid West. Overall, regions like the East and West appear healthier in the data which can possibly be attributed to the high incomes there as well as healthier habits.

Final Conclusion

Overall, the U.S. City Health data provided by the CDC was quite revealing but to ensure that the facts were accurate it did take a great deal of pre-processing. Overall, we examined the social issues that come with biased data from government surveys. We then managed to explore a wide array of factors such as income, location, prevention vs outcome, age, gender and even climate. Through a variety of different visualization methods we were able to wrangle the data and provide meaningful answers to our questions as well as prove/disprove our hypotheses. The data speaks for itself and although it may not be the best accumulation or representation of the U.S. in data form, we were able to tell a story about the health of the United States. The CDC claims to have America’s citizens’ well being as the forefront of there mission. We discovered that their data holds so much information to understanding the facts regarding the Nation's health and the trends that lie beneath. In visualizing the data, we can better understand the facts and if we can best understand the facts we can better prepare for the future of a Nation's health.

Sources

1. https://www.kaggle.com/jennifersantiago/500-cities-local-data-for-better-health-2018 (https://www.kaggle.com/jennifersantiago/500-cities-local-data-for-better-health-2018)
2. https://chronicdata.cdc.gov/500-Cities-Places/500-Cities-Local-Data-for-Better-Health-2018-relea/rja3-32tc (https://chronicdata.cdc.gov/500-Cities-Places/500-Cities-Local-Data-for-Better-Health-2018-relea/rja3-32tc)
3. https://www.youtube.com/watch?v=AgWgPSZ7Gp0 (https://www.youtube.com/watch?v=AgWgPSZ7Gp0)
4. https://cran.r-project.org/web/packages/plotrix/plotrix.pdf (https://cran.r-project.org/web/packages/plotrix/plotrix.pdf)
5. https://worldpopulationreview.com/state-rankings/average-income-by-state (https://worldpopulationreview.com/state-rankings/average-income-by-state)