DSE6630 Final Project

Merrimack College DSE6630: Healthcare & Life Sciences Analytics

Ryan Canfield, and Malek Sabri

30 June 2024

Table of Contents

# Loading in the data file.  
OriginDF <- read.csv("MA\_COUNTY\_CANCER\_MORTALITY\_RATES\_1980\_2014.CSV")  
  
# \*\*Note\*\* all the heads should be commented out when knitting to remove extra tables this is just a way to confirm and see if the code is working.  
#head(OriginDF)

# 1 Introduction

## 1.1 Background

This data set shows the mortality rate of each cancer in Massachusetts. It is a record showing the number of deaths by county gender and year. Some of the important variable information is:

**location\_name** – Is the county that is being measured/recorded. **FIPS** – Is the unique value assigned by the government to identify geographic entities within the U.S.. **cause\_name** – Is the type of cancer that caused the mortality. **sex** – Is the sex of the patient. (Both is the combination of both the male and female rates). **year\_id** – Is the year that is being recorded. the years go from 1980 - 2014. **mx** – Is the number of deaths caused by the cancer in that year. **upper** and **lower** – Are the upper and lower bounds of mx.

To read this data, we will use an example which can be found on line 106. On that line it says: In the year 1980 there were an estimation of 305 male deaths due to Neoplasms cancer in Barnstable County

The first 100 or so rows show the number for the entire state not, not by county so we removed the rose containing data that we didn’t need.

# Removes data that we didn't need.  
df <- OriginDF[-(1:105), ]  
  
#head(df)

Here we are reading in shape files found on the website below to get the county borders for massachusetts. [here](https://www.mass.gov/info-details/massgis-data-counties?_gl=1*1atxnmf*_ga*MTAyOTc5ODQ4LjE3MTk2MTk0NDU.*_ga_MCLPEGW7WM*MTcxOTYxOTU2Ny4xLjEuMTcxOTYxOTU3Ni4wLjAuMA..#downloads-)

ma\_sf <- read\_sf("C:/Users/RRC/Downloads/counties/COUNTIESSURVEY\_POLYM.shp")   
ma\_sf$geometry

## Geometry set for 14 features   
## Geometry type: MULTIPOLYGON  
## Dimension: XY  
## Bounding box: xmin: 33863.73 ymin: 777606.4 xmax: 330837 ymax: 959743  
## Projected CRS: NAD83 / Massachusetts Mainland  
## First 5 geometries:

## MULTIPOLYGON (((308049.3 871046.7, 308304.8 871...

## MULTIPOLYGON (((75164.18 940237.2, 77276.85 940...

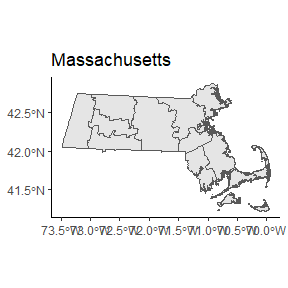
## MULTIPOLYGON (((241907.9 851743.9, 241923.5 851...

## MULTIPOLYGON (((274929.9 804009.5, 274934.5 804...

## MULTIPOLYGON (((250195 959315.3, 253308.7 95689...

Reading in the shape file and looking at the geometry of it as well.

# Initiate the plot and tell it the data set  
ggplot(ma\_sf) +  
 # Use the geom\_sf function to tell it how to make the map  
 geom\_sf() +  
 # Set the theme to classic  
 theme\_classic() +  
 # Let's give it a title  
 ggtitle("Massachusetts")

 ## 1.2 Questions to examine

1. Based on the data we have we can look at most most common cancer mortalities in the three counties with the lowest income compared to the highest income (look at incomes online ) and map then to compare. after finding results we can come pair either hospital numbers or poor insurance due to low income

top 3 highest 1. Nantucket 2. Norfolk 3. Middlesex

top 3 lowest 1. Hampden 2. Berkshire 3. Franklin

1. We can use bar graphs and other visuals and aggregate the data to look at highest mortality cancer types between males females and both combined. From here we can come up with a few recommendations like creating more tests to watch out for these types of cancer, and seeing trends in mortalities and seeing if any new cancer arises as highest deaths.

# 2 Methods and Visualizations

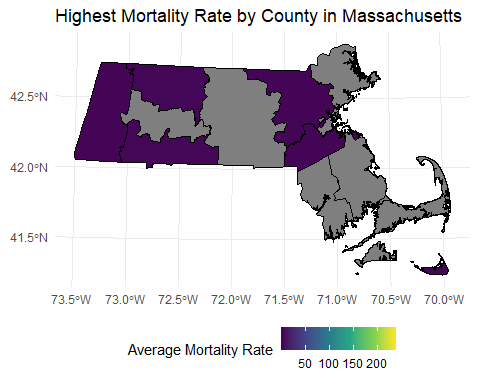
## 2.1 Question 1

### Wrangling, Preprocessing, and Mapping

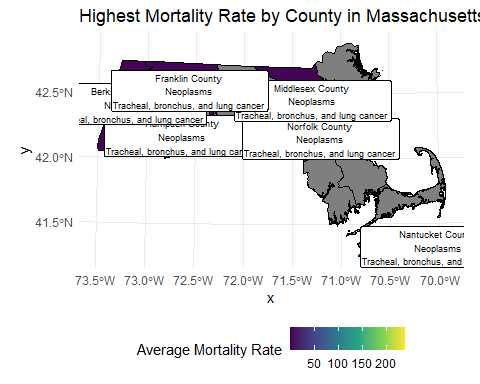
# Filtering data to include only the top three highest and top three lowest counties based on income. (We found these by researching)  
filtered\_data1 <- df %>%  
 filter(location\_name %in% c('Nantucket County', 'Norfolk County', 'Middlesex County',   
 'Hampden County', 'Berkshire County', 'Franklin County')   
 & sex %in% c('Both'))  
  
# Only including variables we need for the visual we don't need gender since it is all the same and we do not need year since we will be grouping the data next.  
dfBoth <- filtered\_data1[, c('FIPS', 'location\_name', 'cause\_name', 'mx')]   
  
# Grouping by cancer type and getting the average of each mortality rate from each year.  
dfBothFinal <- dfBoth %>%  
 group\_by(FIPS, location\_name, cause\_name) %>%  
 summarise(averge\_mortality\_rate = mean(mx))

## `summarise()` has grouped output by 'FIPS', 'location\_name'. You can override  
## using the `.groups` argument.

# Merge summarized data with spatial data (if necessary)  
dfBothFinalMerged <- merge(ma\_sf, dfBothFinal, by.x = "FIPS\_ID", by.y = "FIPS", all.x = TRUE)  
  
# Plot the map with ggplot2, labeling the top causes  
ggplot(dfBothFinalMerged) +  
 geom\_sf(aes(fill = averge\_mortality\_rate), color = "black") +  
 scale\_fill\_viridis\_c(name = "Average Mortality Rate", label = scales::comma) +  
 theme\_minimal() +  
 ggtitle("Highest Mortality Rate by County in Massachusetts") +  
 labs(fill = "Average Mortality Rate") +  
 theme(legend.position = "bottom")



# Sort and select top six counties by average mortality rate  
top\_causes <- dfBothFinalMerged %>%  
 arrange(desc(averge\_mortality\_rate)) %>%  
 slice\_head(n = 6) %>%  
 select(FIPS\_ID, location\_name, cause\_name)  
top\_causes2 <- dfBothFinalMerged %>%  
 arrange(desc(averge\_mortality\_rate)) %>%  
 slice(7:12) %>%  
 select(FIPS\_ID, location\_name, cause\_name)  
  
# Plot the map with ggplot2, labeling the top causes  
ggplot(dfBothFinalMerged) +  
 geom\_sf(aes(fill = averge\_mortality\_rate), color = "black") +  
 geom\_sf\_label(data = top\_causes, aes(label = paste(location\_name, cause\_name, top\_causes2$cause\_name, sep = "\n")), size = 2.5) +   
  
 scale\_fill\_viridis\_c(name = "Average Mortality Rate", label = scales::comma) +  
 theme\_minimal() +  
 ggtitle("Highest Mortality Rate by County in Massachusetts") +  
 labs(fill = "Average Mortality Rate") +  
 theme(legend.position = "bottom")



## Grouping by cancer type and getting the average of each mortality rate from each year.  
#dfBothFinal <- dfBothFinal %>%  
# arrange(desc(averge\_mortality\_rate))  
#  
#dict <- tribble(  
# ~ Variable, ~ Description,  
# "Nantucket County (Highest income) -- Neoplasms", "Average Mortality Rate -- 239.00577",  
# "Hampden County (Lowest income) -- Neoplasms", "Average Mortality Rate -- 230.91646",  
# "Norfolk County (2nd Highest income) -- Neoplasms", "Average Mortality Rate -- 220.48933",  
# "Middlesex County (3rd highest income) -- Neoplasms", "Average Mortality Rate -- 217.44186",  
# "Berkshire County (2nd lowest income) -- Neoplasms", "Average Mortality Rate -- 217.25613",  
# "Franklin County (3rd lowest income) -- Neoplasms", "Average Mortality Rate -- 209.75202",  
#   
# "Nantucket County (Highest income) -- Tracheal, bronchus, and lung cancer ", "Average Mortality Rate -- 67.78133",  
# "Hampden County (Lowest income) -- Tracheal, bronchus, and lung cancer ", "Average Mortality Rate -- 65.18665",  
# "Norfolk County (2nd Highest income) -- Tracheal, bronchus, and lung cancer ", "Average Mortality Rate -- 61.35590",  
# "Berkshire County (2nd lowest income) -- Tracheal, bronchus, and lung cancer ", "Average Mortality Rate -- 60.89512",  
# "Middlesex County (3rd highest income) -- Tracheal, bronchus, and lung cancer ", "Average Mortality Rate -- 59.44581",  
# "Franklin County (3rd lowest income) -- Tracheal, bronchus, and lung cancer ", "Average Mortality Rate -- 56.63687",  
#   
# "Nantucket County (Highest income) -- Colon and rectum cancer ", "Average Mortality Rate -- 32.66684",  
# "Hampden County (Lowest income) -- Colon and rectum cancer ", "Average Mortality Rate -- 31.56224",  
# "Norfolk County (2nd Highest income) -- Colon and rectum cancer ", "Average Mortality Rate -- 30.52043",  
# "Berkshire County (2nd lowest income) -- Colon and rectum cancer ", "Average Mortality Rate -- 30.41453",  
# "Middlesex County (3rd highest income) -- Colon and rectum cancer ", "Average Mortality Rate -- 29.84633",  
# "Franklin County (3rd lowest income) -- Colon and rectum cancer ", "Average Mortality Rate -- 29.50000",  
#)  
#  
#kable(  
# dict,  
# format = "html",  
# col.names = c("", ""),  
# caption = "Table 1. Data dictionary of explaining County Cancer type and Rates.") %>%  
#   
# group\_rows(index = c("1st Highest Mortality Rate" = 6, "2nd Highest Mortality Rate" = 6, "3rd Highest Mortality Rate" = 6)) %>%  
# kable\_styling(bootstrap\_options = c("hover"),  
#)  
#  
#

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

## 2.2 Question 2

### Wrangling, Preprocessing, and Mapping