Data Visualization of Causes of Mortality in the United States

Yongtae Lee

10/31/2017

Background

- This post serves to learn deeply about data visualization, mainly using the most widely used R library: ggplot2. When confronted with enormous data with thousands of rows, it is necessary to import the data in an appropriate way and analyze this massive raw text file by using effective graphic tools. Throughout this blog post, we will employ multiple advanced data visualizing methods from ggplot2 to make the confusing data to be interpretable. So let's begin our journey to become a master of data visualization!

Getting started

Download required packages

At first, it is important to download required packages that assist with analyzing data. There are three packages needed to analyze the data: readr, dplyr, ggplot2.

- The readr package helps with importing data files and putting adjustments on them. To learn more details, please check the following link: https://cran.r-project.org/web/packages/readr/index.html
- The dplyr package allows us to do some data wrangling. To learn more details, please check the following link: https://cran.r-project.org/web/packages/dplyr/index.html
- The gglot2 package, the main package we will be exploring with, is essential to present data visualization. It has various features to produce graphics. To learn more details, please check the following link: https://cran.r-project.org/web/packages/ggplot2/index.html

Also, it is very crucial to load them by using library:

```
# loading packages by using library
library(readr)    # importing data
library(dplyr)    # data wrangling
```

```
## Warning: package 'dplyr' was built under R version 3.4.2
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)  # graphics
```

Importing the data

Now, let's see what our data is about and how it is structured. We will be using a data file 'leading-causes-of-death.csv' which contains some information about mortality from various diseases. You can download the data from the following link: https://catalog.data.gov/dataset?res_format=CSV&tags=mortality&page=2. Make sure to load the data to the current R session.

The data consists nine different columns:

1. year: years from 1999 to 2015
2. cause: name of the disease
3. state: states in the US
4. region: regions in the US
5. subregion: subregions in the US
6. deaths: number of people died
7. patients: number of patients who suffered from the particular disease
8. proportion: proportion of people died from the disease
9. spending: state healthcare spending amount (in dollars)

Here, we'll be reading the data by using read_csv from readr package.

```
# importing the data leading-causes-of-death.csv
base <- read_csv('leading-causes-of-death.csv',
                 # assigning column types for each column
                 col_types = cols(
                  year = col_integer(),
                  cause = col_character(),
                  state = col_character(),
                  region = col_character(),
                  subregion =  col_character(),
                  deaths = col_integer(),
                  patients = col_integer(),
                  proportion = col_double(),
                  spending = col_integer()
                  ))

base
```

```
## # A tibble: 9,490 x 9
##     year     cause state    region          subregion deaths patients
##    <int>     <chr> <chr>     <chr>              <chr>  <int>    <int>
##  1  1999 Alzheimer    NY Northeast    Middle Atlantic   1357    19386
##  2  1999 Alzheimer    HI      West            Pacific    109     1160
##  3  1999 Alzheimer    DC Northeast    Middle Atlantic     53      558
##  4  1999 Alzheimer    CT Northeast        New England    449     3939
##  5  1999 Alzheimer    AK      West            Pacific     24      202
##  6  1999 Alzheimer    NJ Northeast    Middle Atlantic   1041     8675
##  7  1999 Alzheimer    MS     South East South Central    356     2677
##  8  1999 Alzheimer    NV      West           Mountain    174     1279
##  9  1999 Alzheimer    FL     South     South Atlantic   3059    21392
## 10  1999 Alzheimer    PA Northeast    Middle Atlantic   2192    15222
## # ... with 9,480 more rows, and 2 more variables: proportion <dbl>,
## #   spending <int>
```

Exploring the data

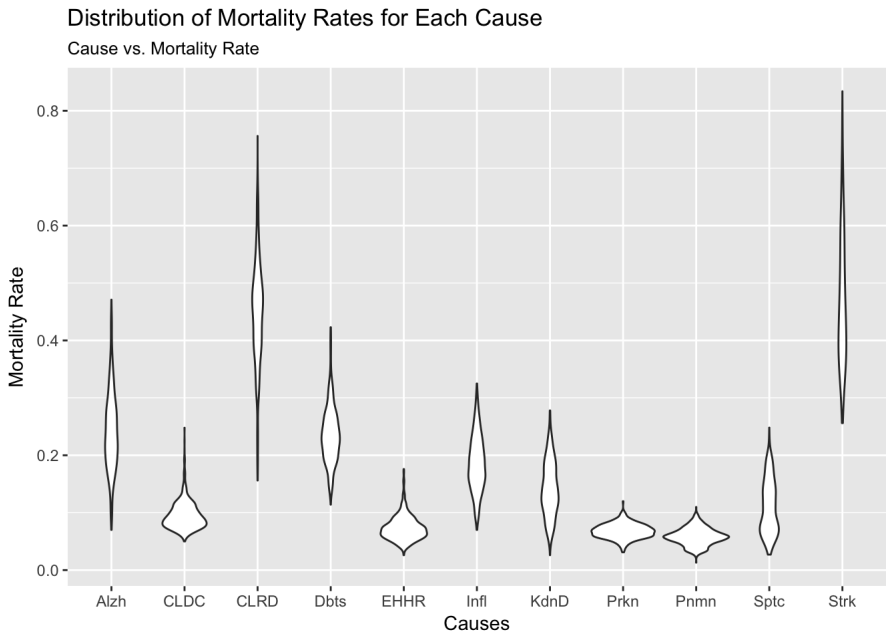
Yes! We have now successfully loaded our data. Today, we will be producing four meaningful observations by applying four distinct graphic features onto this data source:

1. Distribution of mortality rates for each cause
2. Time series of annual mortality rate for each cause
3. Mortality rate for each subregion and cause
4. Relationship between mortality rate and state healthcare spending amount

Distribution of mortality rates for each cause

To create the distribution of mortality rate for each cause, we will be using a violin plot from ggplot2: cause as x-variable and mortality rate (as known as proportion in the data) as y-variable.

```
# creating a violin plot
# distribution of mortality rates for each cause
g1 <- ggplot(base, aes(x=cause, y =proportion)) +
      # putting cause as x-variable and mortality rate (as known as proportion in the data) as y-variable
  geom_violin() +
      # applying violin plot
  labs(title="Distribution of Mortality Rates for Each Cause",
       subtitle="Cause vs. Mortality Rate",
       x="Causes",
       y="Mortality Rate") +
      # modifying axis, legend, and plot labels
  scale_x_discrete(labels = abbreviate)
      # shortening labels for disease names on the x-axis to prevent overlapping names

g1
```



From the violin plot, we can easily observe that stroke and CLRD (Chronic Lower Respiratory Disease) have relatively higher level of average mortality rates. We can also find that these two diseases have deviating mortality rates across states and time, as their distribution are more dispersed compared to those of other diseases.

More information about violin plot can be found from the following link: http://zevross.com/blog/2014/08/04/beautiful-plotting-in-r-a-ggplot2-cheatsheet-3/

Time series of annual mortality rate for each cause

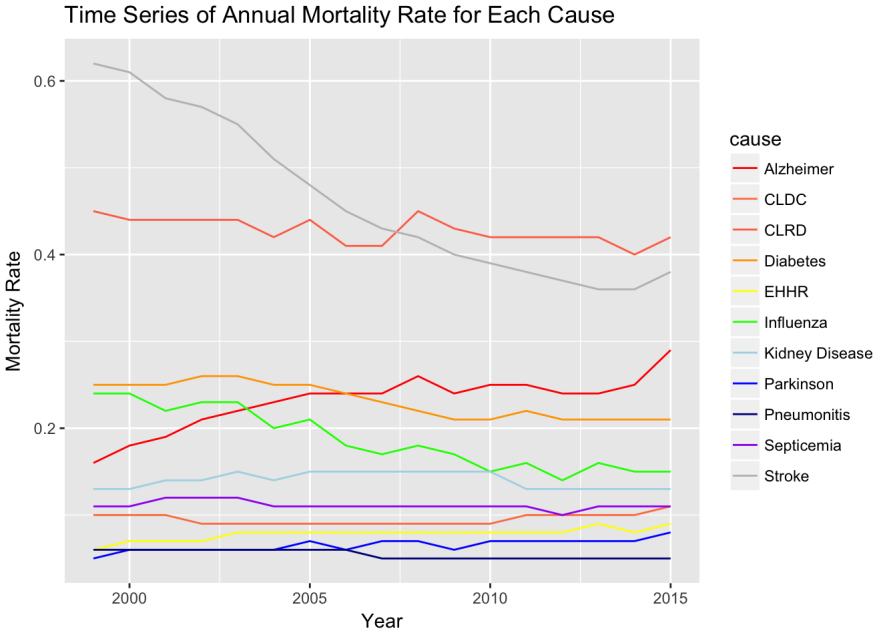
Now, we are going to create a time series of annual mortality rate. But before creating a graph, it is necessary to adjust the data. In order create a time series, we must obtain average mortality rate for each disease through grouping the data by cause and year columns.

```
# creating new data grouped by cause and year columns
group1 <- base %>%
  # selecting only four columns from base data
  select(cause, year, deaths, patients) %>%
  # grouping the data by cause and year columns
  group_by(cause, year) %>%
  # summarizing other columns(deaths and patients) to create average mortality rates for each disease
  summarise(deaths = sum(deaths),
            patients = sum(patients),
            mortality_rate = round(deaths/patients, digits =2)
            )
group1
```

```
## # A tibble: 187 x 5
## # Groups:   cause [?]
##        cause  year deaths patients mortality_rate
##        <chr> <int>  <int>    <int>          <dbl>
##  1 Alzheimer  1999  43780   267146           0.16
##  2 Alzheimer  2000  49558   274399           0.18
##  3 Alzheimer  2001  53852   279173           0.19
##  4 Alzheimer  2002  58866   282448           0.21
##  5 Alzheimer  2003  63457   287866           0.22
##  6 Alzheimer  2004  65965   291658           0.23
##  7 Alzheimer  2005  71599   298445           0.24
##  8 Alzheimer  2006  72432   305787           0.24
##  9 Alzheimer  2007  74632   313176           0.24
## 10 Alzheimer  2008  82435   319818           0.26
## # ... with 177 more rows
```

Let's use this grouped data to produce a time series. A line graphic would be appropriate since we want to visualize trend of annual mortality rate by each cause.

```
# Time series of annual mortality rate for each cause

g2 <- ggplot(group1, aes(x=year, y=mortality_rate, col=cause)) +
  # assigning year as x-axis, mortality rate as y-axis, cause as color
  geom_line() +
  # applying line plot
  labs(title="Time Series of Annual Mortality Rate for Each Cause",
       x = "Year",
       y="Mortality Rate") +
  # modifying axis, legend, and plot labels
  scale_color_manual(labels = c("Alzheimer", "CLDC", "CLRD", "Diabetes", "EHHR", "Influenza", "Kidney Disease",
                                "Parkinson", "Pneumonitis", "Septicemia", "Stroke"),
                     values = c("Alzheimer" = "red",
                                "CLDC" = "coral",
                                "CLRD" = "coral1",
                                "Diabetes" = "orange",
                                "EHHR" = "yellow",
                                "Influenza" = "green",
                                "Kidney Disease" = "light blue",
                                "Parkinson" = "blue",
                                "Pneumonitis" = "dark blue",
                                "Septicemia" = "purple",
                                "Stroke" = "gray"))
  # matching diseases with distinct colors

g2
```



From the time series, we can observe that the mortality rates are quite homogenous across different years, except for the three diseases: Alzheimer, Stroke, and Influenza. While the mortality rates for stroke and influenza have significantly decreased over time, Alzheimer's mortality rate has much increased. Thus, it seems very important for healthcare industry to take a careful look on those diseases and analyze what factors contribute such changes.

Mortality rates for each subregion and cause

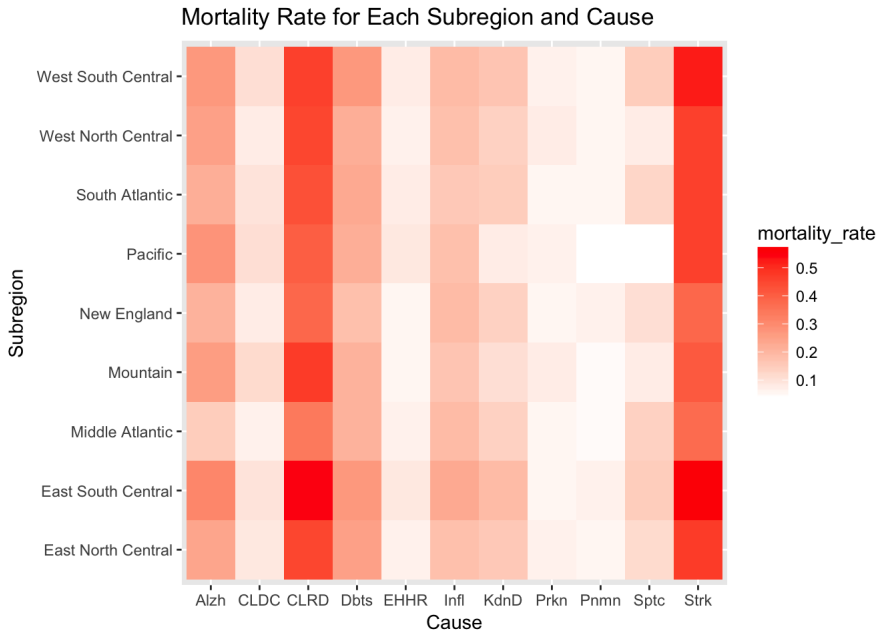
Now, we are going to create a plot of three variables using heat map. In this part, mortality rates for each cause and subregion in the US will be analyzed. But first, let's create another grouped data in order to apply it on the heat map.

```
# creating a new data grouped by subregion and cause columns
group2 <- base %>%
  # using the original data 'base'
  select(subregion, cause, deaths, patients) %>%
  # selecting four columns from 'base': subregion, cause, deaths, patients
  group_by(subregion, cause) %>%
  # grouping by two columns: subregion and cause
  summarise(deaths = sum(deaths),
            patients = sum(patients),
            mortality_rate = round(deaths/patients, digits =2)
            )
  # summarizing other columns(deaths and patients) to create average mortality rates for each disease

group2
```

```
## # A tibble: 99 x 5
## # Groups:   subregion [?]
##              subregion          cause deaths patients mortality_rate
##                  <chr>          <chr>  <int>    <int>          <dbl>
##  1 East North Central      Alzheimer 204666   863484           0.24
##  2 East North Central           CLDC  75933   847159           0.09
##  3 East North Central           CLRD 381998   842440           0.45
##  4 East North Central       Diabetes 208498   846366           0.25
##  5 East North Central           EHHR  64274   859269           0.07
##  6 East North Central      Influenza 154055   855346           0.18
##  7 East North Central Kidney Disease 139417   849114           0.16
##  8 East North Central      Parkinson  58759   837921           0.07
##  9 East North Central    Pneumonitis  50266   856712           0.06
## 10 East North Central     Septicemia  98707   846066           0.12
## # ... with 89 more rows
```

Now, let's apply heat map on the newly grouped data 'group2'. We will be using a 'geom_tile' graphic feature from ggplot2 to observe the result: cause as x-axis, subregion as y-axis, and mortality rate as gradient of color for each box.

```r
# heat map of mortality rate for each subregion and cause
g3 <- ggplot(group2, aes(x = cause, y = subregion)) +
    # assigning cause as x-axis and subregion as y-axis
  geom_tile(aes(fill = mortality_rate)) +
    # producing a heat map by using geom_tile from ggplot2
    # filling each box by mortatlity rate
  scale_fill_gradient(low = "white", high = "red") +
    # assigning colors to fill each box
  labs(title="Mortality Rate for Each Subregion and Cause",
       x = "Cause",
       y = "Subregion") +
    # modifying axis, legend, and plot labels
  scale_x_discrete(labels = abbreviate)
    # shortening labels for disease names on the x-axis to prevent overlapping names

g3
```



From this heat map, we can easily find that CLRD and Stroke are densely colored compared to other diseases, as we can expect from the previous graphics. Also, it is observed that two regions have higher mortality rates: West South Central and East South Central. Based on this heat map, healthcare industry should also analyze why some regions have high mortality rates than the other regions.

More information about heat map can be found from the following link: https://learnr.wordpress.com/2010/01/26/ggplot2-quick-heatmap-plotting/

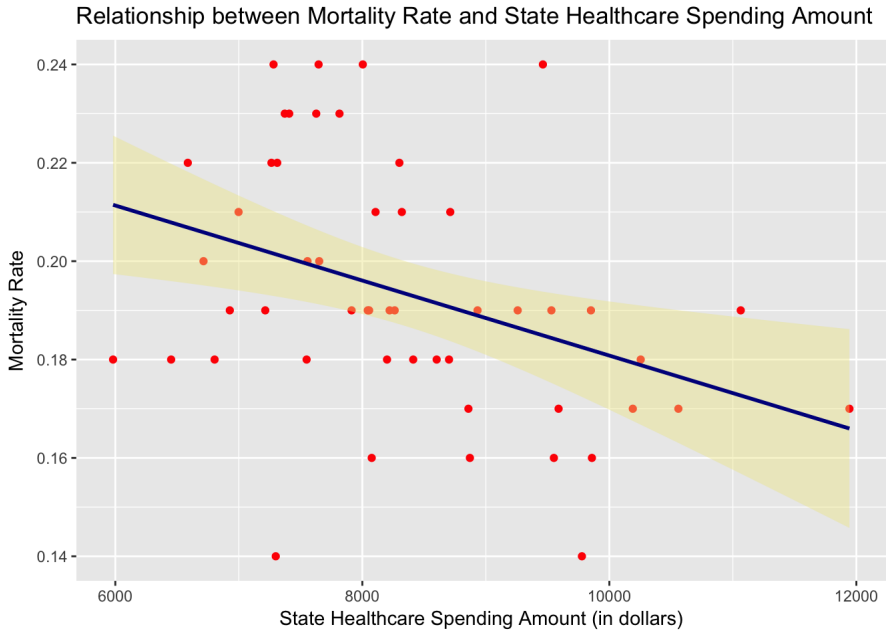
Relationship between state healthcare spending amount and mortality rate

Finally, we will be creating a dot plot and a linear trend line on state healthcare spending amount and mortality rate. As previously done, let's create another grouped data to use for the new graph.

```
# creating a new data grouped by state
group3 <- base %>%
    # based on the original data 'base'
  select(state, spending, deaths, patients) %>%
    # selecting four columns: state, spending, deaths, patients
  group_by(state) %>%
    # grouping by state
  summarise(spending = max(spending),
            deaths = sum(deaths),
            patients = sum(patients),
            mortality_rate = round(deaths/patients, digits =2)
            )
    # summarizing other columns(spending, deaths, and patients)

group3
```

```
## # A tibble: 51 x 5
##    state spending  deaths patients mortality_rate
##    <chr>    <dbl>   <int>    <int>          <dbl>
##  1    AK    11064   12235    63354           0.19
##  2    AL     7281  217036   920531           0.24
##  3    AR     7408  136215   589309           0.23
##  4    AZ     6452  213256  1172222           0.18
##  5    CA     7549 1120005  6365511           0.18
##  6    CO     6804  141508   781820           0.18
##  7    CT     9859  128047   806703           0.16
##  8    DC    11944   18432   108035           0.17
##  9    DE    10254   31809   175122           0.18
## 10    FL     8076  727980  4571496           0.16
## # ... with 41 more rows
```

```
# relationship between state healthcare spending amount and mortality rate
g4 <- ggplot(group3, aes(spending, mortality_rate)) +
    # assigning spending as x-axis and mortality rate as y-axis
  geom_point(color = "red") +
    # creating a dot plot
  geom_smooth(method = "glm", color = "dark blue", fill = "khaki") +
    # creating a linear trend line
  labs(title="Relationship between Mortality Rate and State Healthcare Spending Amount",
       x = "State Healthcare Spending Amount (in dollars)",
       y = "Mortality Rate")
    # modifying axis, legend, and plot labels

g4
```



From the graphic, we can easily find that mortality rate and state healthcare spending amount is negatively correlated. This makes a lot of sense: the higher the state spending on healthcare is, the less is the mortality rate. This graphic is meaningful because it proves that healthcare spending is effective in reducing mortality rate.

Putting all together

From these four graphics, we have learned that some meaningful observations can be successfully made by using various features of ggplot2. In fact, there are more tools available in the package. Many researchers and students produce advanced graphics with ggplot2 and support their arguments. The following link is a cheat sheet for ggplot2 provided by RStudio: https://www.rstudio.com/wp-content/uploads/2015/03/ggplot2-cheatsheet.pdf

I hope this post helped you better understand graphic features available in R. You can also apply your own data and produce many other graphics! Take advantage of this amazing toolkits provided by R and conduct your own experiments. Data visualization is not only efficient, but also very fun!

References

- image: https://www.sas.com/en_us/insights/big-data/data-visualization/_jcr_content/socialShareImage.img.png
- data: https://catalog.data.gov/dataset?res_format=CSV&tags=mortality&page=2
- readr package: https://cran.r-project.org/web/packages/readr/index.html
- dplyr package: https://cran.r-project.org/web/packages/dplyr/index.html
- gglot2 package: https://cran.r-project.org/web/packages/ggplot2/index.html
- ggplot2 violin plot: http://zevross.com/blog/2014/08/04/beautiful-plotting-in-r-a-ggplot2-cheatsheet-3/
- ggplot2 heatmap: https://learnr.wordpress.com/2010/01/26/ggplot2-quick-heatmap-plotting/
- ggplot2 cheatsheet: https://www.rstudio.com/wp-content/uploads/2015/03/ggplot2-cheatsheet.pdf