Overview

Our project is an analysis of the leading causes of death in the United States. It compares deaths in different US regions and New York City. This is done using an interactive stacked area chart comparing causes of death in each region per year. The stacked chart can be used to create a line graph that provides more information about individual causes of death and a bar graph compares gender and sex.

Data

NYC Data

Link: https://data.cityofnewyork.us/Health/New-York-City-Leading-Causes-of-Death/jb7j-dtam#column-menu

This data was obtained from the NYC Open Data site. It contains information about causes of death from 2007 to 2011 for different races and sexes. The data was already in an easy-to-use format, so no processing was done (besides removing some unnecessary data columns) outside of the javascript code. The final table contained the following values:

1. Year
2. Ethnicity
3. Sex
4. Cause of Death
5. Count

US Data

Link: http://webappa.cdc.gov/sasweb/ncipc/leadcaus10_us.html

The dataset was obtained from the CDC database. It contains death reports per region from 1999-2013 for different races and sexes. In order to remain consistent with the NYC dataset only the data from 2007 to 2011 was used.

The datasets could only be accessed separately for each year, gender, and race combination. Because of this each set had to be downloaded separately and combined manually. There are also several gaps in the data, but we decided to include causes of death that were missing gender or ethnicity data since it still gave a more complete view of the story. The site also warns “Year-to-year death data for a given state can sometimes be affected by unexpectedly large numbers of death certificates with the underlying cause coded as “other ill-defined and unspecified causes of mortality.” The final table contained the following values:

1. Year
2. Ethnicity
3. Sex
4. Cause of Death
5. Region
6. Count

Processing in Code
In order to use multiple graphs to visualize the data a lot of processing needed to be done in the code to combine the correct fields for each graph. More information will be provided in the following section for each visualization.

Visualization

Map
The graphic starts with a map selectable by four US regions or NYC. Clicking on a region creates a stacked area chart for each region.

Stacked Area Chart:
The stacked area chart shows the number of deaths vs. year for all causes of death.

The data that is passed into the function to create the chart is filtered for only the selected region. However the data is separated for all values, so we created a json variable “tmp” to formalize the dataset to a json structure based on “Year” and “Cause”, and accumulated the “Count” of each cause.

Then we created an array “keypoolarr” based on this json variable. And since not all the causes are in every year, we use a keypool variable to store all the causes, and set the count of cause to 0 if this year doesn’t have the specific cause.

Finally, we found the largest value of the count. Comparing the largest count value of each year, and find the bigger one as the max. Then use the max value as 100% percent.

The x scale is simply a linear scale showing the percentage of deaths, and the y scale is categorical for each year.

We expected this graph to show more of a story about the changes in causes of death per year, but its actually very hard to see this in the final product. However, it did succeed in providing a relatively good comparison between each cause of death. One problem we originally ran into was that some of the causes consist of a very small percentage of the graph and are hard to see and select in the area chart. In order to fix this we provided a large legend that is also selectable and interacts with the graph. The legend also allows the viewer to see a list of all of the leading causes of death.

Clicking on a cause creates a line graph and bar graph based on that cause’s data.

Line Graph
The line graph provides a more detailed view of a the number of deaths for a single cause vs. year. This graph did allow for a better view of the changes in the numbers of deaths over the years, which was hard to visualize with the above chart.

The data used to create this graph is a subset of the list used to create the stacked area chart. The data filtered for region above is also filtered by the selected cause. The deaths for each ethnicity and gender are then combined by year.

Hovering over the bubbles for a year will change the information in the bar chart to show information for different ethnicities and sex.

Bar Chart
In order to get an even more in depth view of the data we used a bar graph the shows the number of deaths vs. ethnicity and sex.

The data used is a subset of the line graph data. The data filtered by region and by cause (used in the line graph) is further filtered by year. The values are then combined for each ethnicity (with separate counts for each sex).

This graph is the most detailed view of our datasets, and provides information to compare deaths between different demographics.

Analysis

The data confirms some common knowledge such as the fact that heart disease and cancer are among the highest causes of death, but it also provides information on some of the more specific causes that are less known such as atelectasis.

The changes of each cause over the years don't give us much information since the time scale is so small. It would have been nice to look at the full set of US data (from 1999 - 2013), but because of the hassle of downloading that data set and for the sake of staying consistent with the NYC dataset we limited ourselves to just the 5 years shown. Small trends can be observed in some cases (Ex. general upward trend in cancer deaths); however, again since it's such a small time scale, these must be taken with a grain of salt.

The bar graph definitely seems to be skewed by the population. The non-hispanic white demographic clearly has many more deaths than the other ethnicities in almost all cases, most likely because they also constitute the majority of population. The male and female data does seem to be comparable. Male deaths are expectedly greater in cases such as "unexpected injury", but trends are less predictable in the cases of medical causes.