STAT 3622
Project Proposal

October 24, 2016

Gordon Hart
3035338963

gordonhart@hku.hk

University of Hong Kong
Department of Statistics and Actuarial Science

1 Introduction

For the `STAT 3622` final project I propose a data visualization of mortality data gathered from the United States' Center for Disease Control and Prevention (CDC). The CDC provides extensive records of deaths in the US from 1968 up to 2014: the dataset includes cause of death, location of death, age at time of death, sex, race, pre-existing conditions, and many more elements. Access to data this detailed provides countless opportunities for analysis and visualization of mortality in the United States.

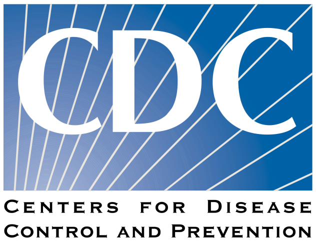

Figure 1: The kind providers of the data utilized for this project.

The data is released in a raw format that would require a non-trivial amount of processing to prepare it for analysis. As a computer science student with nearly 10 years of programming experience and data manipulation, I believe that I am capable of handling this task and that the format of the data will not prove an insurmountable hinderance for the project.

However, given that the data is provided in year-by-year releases, all with slightly different schema, it is not realistic to analyze each of the years in the hopes of producing time-series data for the last four decades. Rather, focus will be placed on the 1968 and 2014 datasets, examining state- and country-wide differences in cause, age, location, etc. of death for Americans.

Various visualizations focused on different characteristics from the dataset will be produced. See the following section for a more detailed outline of the proposed graphics.

2 Proposed Graphics

The goal of this visualization project is to provide a cross-section of American life and well-being through a juxtaposition of citizens in two very different years, 1968 and 2014, at time of death.

2A Cause of Death

Cause of death can offer clues as to how a person lived; on a mass scale, cause of death statistics can reflect the overall health of a nation.

1. *Cause of Death by Location*
   A map is proposed that highlights the leading cause of death by US state, with an animated function switching between 1968 and 2014.

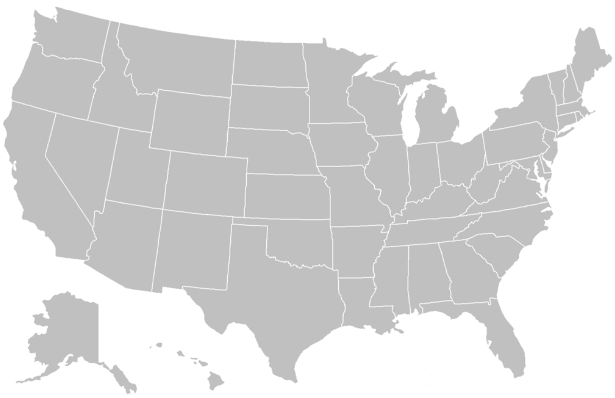

Figure 2: A blank map of states similar to the one that would be color-coded with cause of death in this proposed visualization.

2. *Cause of Death vs. Age of Death*
   A plot is proposed that identifies the three leading causes of death for each age group, grouped together in five-year intervals, in hopes of identifying the *risk periods*: specific ages at which a person may be threatened by a specific disease/factor.
3. *Cause of Death by Race*
   What are the leading killers of people of different ethnic backgrounds? A plot is proposed to display the handful of leading causes of death for each race identified in the dataset.
4. *Cause of Death by Gender*
   Are there diseases that primarily target on sex over the other? A plot is proposed to show the discrepancies between deaths of citizens of different genders from the same ailment.

2B Age of Death

Age at time of death is a key statistic for the evaluation of the health of a populace. Life expectancy statistics are extremely widespread, but do not offer the full picture. This section of visualizations aims to clear up some of the ambiguity surrounding that particular statistic.

1. *Age of Death by Race*
   Is there a difference in life expectancy for different ethnic groups? A simple bar plot is proposed to show the average life expectancy based on deaths in 1968 and 2014 for the different groups identified in the CDC dataset.
2. *Age of Death vs. Gender*
   Most of us know that women are expected to live longer than men; if this is indeed true, what do the death rates by age look like for the two different genders? Is there a particular period in life in which many men die but many women live through?
3. *Age of Death by Location*
   Which states die young? Are there states that are substantially more healthy (in that they live longer) than others? How do the trends identified in the *race* and *gender* plots hold up when grouped by state boundaries?

Additional plots may be included as familiarity with the dataset and its offerings increases.

3 Presentation and Technologies

As an experienced web developer the findings from this project will be organized an presented as a single, interactive web page.

The forecasted technologies for presentation include `HTML/CSS` (of course), `JavaScript`, `plotly` (either `js` or `R`), and likely `ggplot2`.

For data processing before visualization, I will write processing scripts in the functional programming language `OCaml`. This language is quite powerful and is well-suited for sifting and sorting through large datasets such as those provided by the CDC.

4 Final Remarks

Mortality is a topic that every one of us must grapple with throughout our lives. Through a data-driven inspection of this inevitable event I hope to remove some of the mystery surrounding death, providing viewers with a more complete idea of the methods in which they may meet their demise.

The visualizations discussed in Section 2 may not comprise the exact set of visualizations presented along with the final submission; during the data preparation process I will use my discretion to decide which proposed figures should be removed and which additional figures should be created and included.