Project proposal

Avatar: Xavier Ghazaleh, Abigail Phillips, Saajan Patel, Harriet Wiser

10/09/2020

Section 1. Introduction

Death is a certainty, but what we die of depends on the conditions (environmental, economic, political, etc.) under which we live. A significant factor that determines these conditions is our position in the world, both geographically and socially - so from where we live to who has power over us. This means that causes of death and their prevalence can vary based on our country and our region. This data set puts numbers to those disparities. Understanding how someone's situation affects their risks is broadly important because it determines what interventions are put in place, governmentally or otherwise.

We have many avenues we would like to explore with this dataset. The immediate thing that comes to mind is tracking how the impact of certain afflictions changes over time in different locations - as well as how the impact of certain afflictions varies based upon location in the first place. We are also interested in how different issues relate to others. An obvious example of this is alcohol disorders and liver disease, but we also want to trace more subtle potential links like perhaps HIV/AIDS and suicide. Although we, of course, probably won't be able to establish causation without broader understanding of the context these factors exist in, correlation can still lead to some interesting conclusions. The conflict and terrorism columns also merit examine, although unfortunately there are many places that don't have data for those areas, so it would have to be a more place-limited analysis.

Research Question: How are homicide and suicide mortality rates related to the mortality rates of other diseases and issues such as dementia, aid, and liver disease?

Hypothesis: Homicide and suicide will be have a positive correlation with at least one other disease or issue such as dementia, aids, and liver disease.

Section 2. Data description

The numbers are drawn from a study called the Global Burden of Disease done by the medical journal The Lancet. It is a world-wide observational study that collects data starting in 1990 and updated up to 2017, although we are using a version that ends in 2016. The data include mortality rates by the percent of deaths caused by 32 issues every year in 228 regions/countries. The causes of death (observations) seem to be roughly organized by prevalence, although this, of course, varies. The first three are cardiovascular disease, cancer, and respiratory disease. The countries and regions (variables) are ordered alphabetically, with each location having all 27 years they were tracked represented right after each other (i.e. it is ordered by location not by timeline). The first country is Afghanistan, and the fifth location gives us our first region, Andean Latin America.

Section 3. Glimpse of data

```r
library(tidyverse)
library(readxl)
global_mortality <- read_excel("data/global_mortality.xlsx")

glimpse(global_mortality)
```

```
## Rows: 6,156
## Columns: 35
## $ country                                      <chr> "Afghanistan", "Afghanis...
## $ country_code                                 <chr> "AFG", "AFG", "AFG", "AF...
## $ year                                         <dbl> 1990, 1991, 1992, 1993, ...
## $ `Cardiovascular diseases (%)`                <dbl> 17.61040, 17.80181, 18.3...
## $ `Cancers (%)`                                <dbl> 4.025975, 4.054145, 4.17...
## $ `Respiratory diseases (%)`                   <dbl> 2.106626, 2.134176, 2.20...
## $ `Diabetes (%)`                               <dbl> 3.832555, 3.822228, 3.90...
## $ `Dementia (%)`                               <dbl> 0.5314287, 0.5324973, 0....
## $ `Lower respiratory infections (%)`           <dbl> 10.886362, 10.356968, 10...
## $ `Neonatal deaths (%)`                         <dbl> 9.184653, 8.938897, 8.84...
## $ `Diarrheal diseases (%)`                      <dbl> 2.497141, 2.572228, 2.70...
## $ `Road accidents (%)`                          <dbl> 3.715944, 3.729142, 3.81...
## $ `Liver disease (%)`                           <dbl> 0.8369093, 0.8455159, 0....
## $ `Tuberculosis (%)`                            <dbl> 5.877075, 5.891704, 6.03...
## $ `Kidney disease (%)`                          <dbl> 1.680611, 1.671115, 1.70...
## $ `Digestive diseases (%)`                      <dbl> 1.058771, 1.049322, 1.06...
## $ `HIV/AIDS (%)`                                <dbl> 0.01301948, 0.01451458, ...
## $ `Suicide (%)`                                 <dbl> 0.4366105, 0.4422802, 0....
## $ `Malaria (%)`                                 <dbl> 0.4488863, 0.4550191, 0....
## $ `Homicide (%)`                                <dbl> 1.287020, 1.290991, 1.32...
## $ `Nutritional deficiencies (%)`               <dbl> 0.3505045, 0.3432123, 0....
## $ `Meningitis (%)`                              <dbl> 3.037603, 2.903202, 2.84...
## $ `Protein-energy malnutrition (%)`            <dbl> 0.3297599, 0.3221711, 0....
## $ `Drowning (%)`                                <dbl> 0.9838624, 0.9545860, 0....
## $ `Maternal deaths (%)`                         <dbl> 1.769213, 1.749264, 1.76...
## $ `Parkinson disease (%)`                       <dbl> 0.02515859, 0.02545063, ...
## $ `Alcohol disorders (%)`                       <dbl> 0.02899828, 0.02917152, ...
## $ `Intestinal infectious diseases (%)`         <dbl> 0.1833303, 0.1781074, 0....
## $ `Drug disorders (%)`                          <dbl> 0.04120540, 0.04203340, ...
## $ `Hepatitis (%)`                               <dbl> 0.1387378, 0.1350081, 0....
## $ `Fire (%)`                                    <dbl> 0.1741567, 0.1706712, 0....
## $ `Heat-related (hot and cold exposure) (%)`   <dbl> 0.1378229, 0.1348266, 0....
## $ `Natural disasters (%)`                       <dbl> 0.00000000, 0.79760256, ...
## $ `Conflict (%)`                                <dbl> 0.932, 2.044, 2.408, NA,...
## $ `Terrorism (%)`                               <dbl> 0.007, 0.040, 0.027, NA,...
```