Project Proposal

due October 11, 2021 by 11:59 PM

The Gre8est Team: Roni Ochakovski, Matthew Wang, Judy Zhong

10/11/2021

Load Packages

```
library(tidyverse)
```

Load Data

```
library(readr)
h_spend <- read_csv("../data/Health-Spending.CSV")
wb_tot_lifeexp <- read_csv("../data/World Bank Life Exp.csv")
wb_f_lifeexp <- read_csv("../data/World Bank Female Life Exp.csv")
wb_m_lifeexp <- read_csv("../data/World Bank Male Life Exp.csv")
```

Introduction and Data, including Research Questions

Our general research question is to investigate the relationship between a country's healthcare spending and the life expectancies of their population. We are using health spending data from the Global Health Data Exchange [1] and life expectancy data from the World Bank Open Data [2-4]. In addition to this, we are using world econometric data from the World Bank [6]. For these 5 data sets, we will be analyzing data from all countries that are included in both databases. The health spending data was collected from a wide variety of sources that included program reports, budget data, national estimates, and National Health Accounts (NHAs). The variables that we are most concerned with are location ID, location name, location category, year, and total health spending (2020 USD). This spending estimate is in constant thousands of 2020 United States Dollars. The life expectancy data were collected by the United Nations Statistics Division through the Demographic Yearbook vital statistics questionnaire. The variables that we are most concerned with are country or area, total life expectancy at birth (years), male life expectancy at birth (years), and female life expectancy at birth (years). Some existing literature has already explored the association between healthcare spending and health outcomes [5]. However, we would like to explore this relationship ourselves as we are all interested in seeing whether the common-sense notion that increased spending on healthcare should increase life expectancies holds true across different countries. And if not, delving deeper into those variations.

References

[1] Global Burden of Disease Collaborative Network. Global Health Spending 1995-2018. Seattle, United States of America: Institute for Health Metrics and Evaluation (IHME), 2021.
[2] World Bank. "Life expectancy at birth, total (years)" World Development Indicators, The World Bank Group, 2021, https://data.worldbank.org/indicator/SP.DYN.LE00.IN
[3] World Bank. "Life expectancy at birth, male (years)" World Development Indicators, The World Bank Group, 2021, https://data.worldbank.org/indicator/SP.DYN.LE00.MA.IN
[4] World Bank. "Life expectancy at birth, female (years)" World Development Indicators, The World Bank

Group, 2021, https://data.worldbank.org/indicator/SP.DYN.LE00.FE.IN
[5] Gallet CA, Doucouliagos H. The impact of healthcare spending on health outcomes: A meta-regression analysis. Soc Sci Med. 2017 Apr;179:9-17. doi: 10.1016/j.socscimed.2017.02.024. Epub 2017 Feb 20. PMID: 28237460.

Glimpse

*Note: If glimpse() were to be used, our document would be over the length criteria. Please visit our github repo to see full datasets.

```
## # A tibble: 6 x 33
##   location_id location_name iso3  level    year the_total_mean the_total_ppp_me~
##         <dbl> <chr>         <chr> <chr>   <dbl>          <dbl>             <dbl>
## 1         160 Afghanistan   AFG   Country  1995         528409           2193179
## 2         160 Afghanistan   AFG   Country  1996         516915           2145473
## 3         160 Afghanistan   AFG   Country  1997         509874           2116248
## 4         160 Afghanistan   AFG   Country  1998         485561           2015335
## 5         160 Afghanistan   AFG   Country  1999         463720           1924685
## 6         160 Afghanistan   AFG   Country  2000         446201           1851971
## # ... with 26 more variables: ghes_total_mean <dbl>, ghes_total_ppp_mean <dbl>,
## #   ppp_total_mean <dbl>, ppp_total_ppp_mean <dbl>, oop_total_mean <dbl>,
## #   oop_total_ppp_mean <dbl>, dah_total_mean <dbl>, dah_total_ppp_mean <dbl>,
## #   the_per_cap_mean <dbl>, the_per_cap_ppp_mean <dbl>,
## #   ghes_per_cap_mean <dbl>, ghes_per_cap_ppp_mean <dbl>,
## #   ppp_per_cap_mean <dbl>, ppp_per_cap_ppp_mean <dbl>, oop_per_cap_mean <dbl>,
## #   oop_per_cap_ppp_mean <dbl>, dah_per_cap_mean <dbl>, ...
```

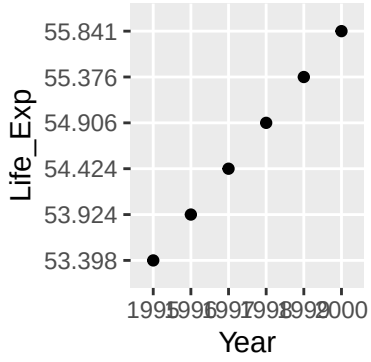
```
## # A tibble: 6 x 33
##   `Country Name` `Country Code` `1990` `1991` `1992` `1993` `1994` `1995` `1996`
##   <chr>          <chr>          <chr>  <chr>  <chr>  <chr>  <chr>  <chr>  <chr>
## 1 Afghanistan    AFG            50.331 50.999 51.641 52.256 52.842 53.398 53.924
## 2 Albania        ALB            71.836 71.803 71.802 71.86  71.992 72.205 72.495
## 3 Algeria        DZA            66.938 67.27  67.575 67.877 68.194 68.54  68.919
## 4 American Samoa ASM            ..     ..     ..     ..     ..     ..     ..
## 5 Andorra        AND            ..     ..     ..     ..     ..     ..     ..
## 6 Angola         AGO            45.306 45.271 45.23  45.201 45.201 45.246 45.35
## # ... with 24 more variables: 1997 <chr>, 1998 <chr>, 1999 <chr>, 2000 <chr>,
## #   2001 <chr>, 2002 <chr>, 2003 <chr>, 2004 <chr>, 2005 <chr>, 2006 <chr>,
## #   2007 <chr>, 2008 <chr>, 2009 <chr>, 2010 <chr>, 2011 <chr>, 2012 <chr>,
## #   2013 <chr>, 2014 <chr>, 2015 <chr>, 2016 <chr>, 2017 <chr>, 2018 <chr>,
## #   2019 <chr>, 2020 <chr>
```

Data Analysis Plan

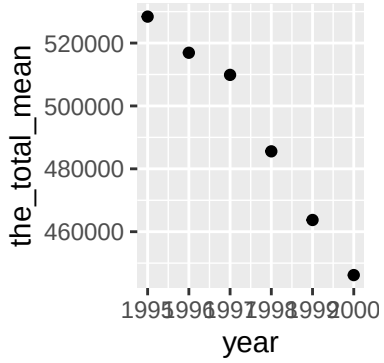
The predictor would be health spending within a particular country. The outcome would be life expectancy (whether total or per age group). We may also explore differences in life expectancies between genders as a function of health spending within the country or the association between the geography of a country and its global health spending. We will be comparing spending across countries and life expectancies within those countries so no control group would be necessary.

```
pivot_longer(
  wb_tot_lifeexp,
  cols = "1995" : "2020",
  names_to = "Year" ,
```

```
  values_to = "Life_Exp"
  ) %>%
  head() %>%
  ggplot(aes(x = Year, y = Life_Exp)) +
  geom_point()
```




```
h_spend %>%
  head() %>%
   ggplot(aes(x = year, y = the_total_mean)) +
  geom_point()
```



Above we output two very simple and preliminary scatter plots for the trend in life expectancy over 5 years and health spending (labeled the_total_mean) over the same 5 years in Afghanistan. As we can see, there is a very clear and opposite trend in the data (which is contradictory to what we expect). Therefore, this is a good indication that patterns are present across the two datasets and that furthr exploration is necessary. Over the course of the project, we will combine the two datasets and create similar scatter plots with mean spending on the x-axis and life expectancy on the y-axis with possible facets or color-coordination by gender.

A regression analysis with focus on R and R^2 will be useful to analyze the relationship between health spending and life expectancies. We can also categorize the health spending into low and high categories and do the same for life expectancies. Then, we can run a Chi-Squared test for Independence to find if the two variables (now categorical) are independent or not. Our hypothesis is that higher health spending is associated with higher life expectancy. To support this, we would expect to see a positive and strong correlation (with a high positive R value and high R^2 value). We would also expect to reject the null hypothesis for independence of health spending and life expectancy through a low p-value in the Chi-Squared test.