DASC 1104 Example Project Proposal

John Tipton

1 My blog link

My blog is available at https://dasc-1104-blog.netlify.app.

```
library(here)
library(ggplot2)
library(tidyverse)
library(readxl)
knitr::opts_chunk$set(echo = FALSE, tidy = TRUE)
dat1 <- read.csv(here::here("data", "tidytuesday", "data", "2018", "2018-11-13",
    "malaria_deaths.csv"))
dat2 <- read.csv(here::here("data", "tidytuesday", "data", "2018", "2018-11-13",
    "malaria_deaths_age.csv"))
dat3 <- read.csv(here::here("data", "tidytuesday", "data", "2018", "2018-11-13",
    "malaria_inc.csv"))

## rename deaths from malaria per 100K
dat1 <- rename(dat1, deaths_per_100K = Deaths...Malaria...Sex..Both...Age..Age.standardized..Rate...per

## rename the incidence of malaria per 1000 population at risk
dat3 <- rename(dat3, incidence_per_1000_at_risk = Incidence.of.malaria..per.1.000.population.at.risk...
```

Explore the data – show this for the proposal but not the final report!

```
glimpse(dat1)
```

```
## Rows: 6,156
## Columns: 4
## $ Entity          <fct> Afghanistan, Afghanistan, Afghanistan, Afghanistan,...
## $ Code            <fct> AFG, AFG, AFG, AFG, AFG, AFG, AFG, AFG, AFG, AFG, A...
## $ Year            <int> 1990, 1991, 1992, 1993, 1994, 1995, 1996, 1997, 199...
## $ deaths_per_100K <dbl> 6.802930, 6.973494, 6.989882, 7.088983, 7.392472, 7...
```

```
glimpse(dat2)
```

```
## Rows: 30,780
## Columns: 6
## $ X         <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17...
## $ entity    <fct> Afghanistan, Afghanistan, Afghanistan, Afghanistan, Afgha...
## $ code      <fct> AFG, AFG, AFG, AFG, AFG, AFG, AFG, AFG, AFG, AFG, AFG, AF...
## $ year      <int> 1990, 1991, 1992, 1993, 1994, 1995, 1996, 1997, 1998, 199...
## $ age_group <fct> Under 5, Under 5, Under 5, Under 5, Under 5, Under 5, Und...
## $ deaths    <dbl> 184.6064, 191.6582, 197.1402, 207.3578, 226.2094, 236.328...
```

```
glimpse(dat3)
```

```
## Rows: 508
## Columns: 4
## $ Entity                     <fct> Afghanistan, Afghanistan, Afghanistan, A...
## $ Code                       <fct> AFG, AFG, AFG, AFG, DZA, DZA, DZA, DZA, ...
## $ Year                       <int> 2000, 2005, 2010, 2015, 2000, 2005, 2010...
## $ incidence_per_1000_at_risk <dbl> 1.071000e+02, 4.650000e+01, 2.390000e+01...
```

2 Malaria deaths

For this project, I am examining the malaria deaths by age dataset contained in the `malaria_deaths_age.csv` file on the Tidy Tuesday website. The data consists of 30780 observations of 5 variables. The variable `entity` is a factor with 228 levels that represents the region for the deaths. The variable `code` is a factor with 196 levels which represents a three letter code for the country/region. The variable `year` is a discrete integer variable that records the years between 1990 and 2016. The variable `age_group` is a discrete ordered factor with 5 levels that groups the population into age groups under 5, 5-14, 15-49, 50-69 and 70 or older. The variable `deaths` records the number of deaths from malaria for each region, year, and age group. Initial exploration shows that the largest variability in the number of deaths occurs across age groups with the largest number of deaths occurring in children under 5. Other variability in deaths occurs when examining death rates by region.

The second dataset we use to understand malaria is the incidence rate of malaria given in the `malaria_inc.csv` file on the Tidy Tuesday website. The malaria incidence rate is the average number of people who contract malaria per 1000 people at risk. The malaria incidence dataset has 508 observations of 4 variables: the variable `Entity`, which is a factor with 127 levels, that represents the region for the malaria cases, the variable `Code`, which is a factor with 101 levels, that represents a three letter code for the country/region, the variable `year`, which is a discrete integer variable that records the years between 2000 and 2015 in an interval of 5 years, and the variable `Incidence` which records the number of malaria cases per 1000 people at risk.

- Question 1: First, is there a significant difference in malaria deaths by age group. If so, which groups are different. To test this, I will generate different data visualizations (likely boxplots) as well as calculate group-level statistics like means, medians, and standard deviations.
- Question 2: Second, how has the death rate from malaria changed over time? To investigate this, I will generate data visualizations of the death rate over time. In the visualization I will add model smooths and explore facets and groupings of other variables to see if the time trends vary with other variables.
- Question 3: Third, has the incidence of malaria per 100K people at risk changed over time? First, I will mutate the data to calculate the incidence rate per 100K. Then, I will produce visualizations of the incidence over time while exploring other grouping variables and facetes.
- Question 4: To be determined.

3 Covid-19 data analysis

```r
dat_full <- read.csv(file = here::here("data", "covid-data", "us-states.csv"))

dat_pop <- read.csv(file = here::here("data", "covid-data", "nst-est2019-alldata.csv"))
```

```r
glimpse(dat_full)
```

```
## Rows: 14,094
## Columns: 5
## $ date   <fct> 2020-01-21, 2020-01-22, 2020-01-23, 2020-01-24, 2020-01-24, ...
## $ state  <fct> Washington, Washington, Washington, Illinois, Washington, Ca...
## $ fips   <int> 53, 53, 53, 17, 53, 6, 17, 53, 4, 6, 17, 53, 4, 6, 17, 53, 4...
## $ cases  <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 1, 1, 1, 2, 1, 1, 1, 2, 1, 1, ...
## $ deaths <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
```

```
## there are 151 variables -- not printing them all for this proposal
## glimpse(dat_pop)
```

3.1 Data

Using data available on GitHub from the NY Times, I am going to investigate the ongoing Covid-19 pandemic in the United States. The variables in the Covid-19 data include the date, the state, the federal id code (fips), the number of cases recorded in the state for the given date, and the number of deaths recorded in the state for the given date. The US population data is from the US Census Bureau website. There are 151 varaibles in the US Census data – we will use the `NAME` variable (state variable) and the estimated population in 2019 `POPESTIMATE2019`.

- Question 1: Does the proportion of people being infected each day follow an exponential growth curve?
  - I will test this by making two graphics, one of the number of cases as a function of time and the other as the logarithm of the number of cases over time. By adding a smoother to the visualization, we can detect exponential growth with a linear trend in the log-transformed data. I will also explore plotting this data with other facets and groupings (states, regions, ages, ethnicities) to see if there is evidence of exponential growth.
- Question 2: Has the average rate of cases in the US grown faster/slower/the same in the last 3 weeks than the 3 weeks prior?
  - To explore this, I will define a new variable as the ratio of new cases from data to day. Using this new variable, I will visualize the distributions to see if there is a difference and explore if there are difference by regions/states/other groupings. I will also calculate summary statistics like the mean/median/standard deviation of the ratio of new daily cases for the two time periods.
- Question 3: Using the the week April 10-17, has the number of new cases per 100K people (using population data available here) over the this week, is there a difference in infection rate based on spatial region in the US (northwest, south, northeast, etc.).
  - This question can be answered using a visual representation of the number of cases per 100K (a mutated variable) over a spatial map.
- Question 4: to be determined
  -