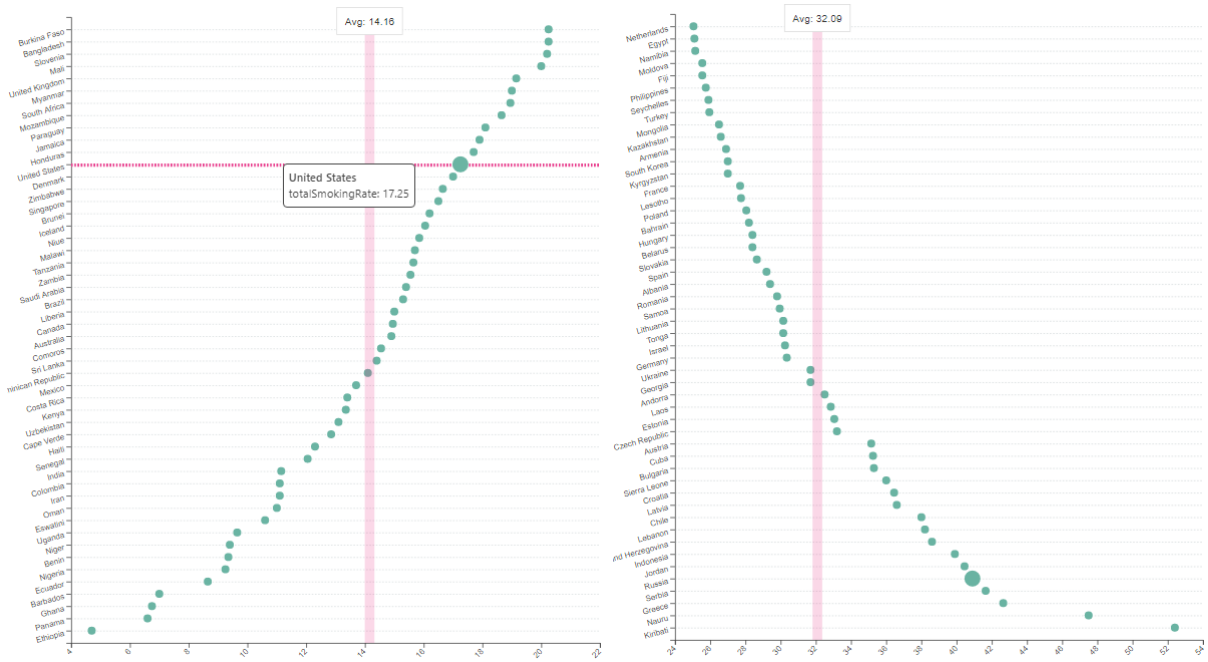
Midterm Project (Rough Draft)

Hansol Kim

Overview

Original visualization:



Claim:

> The tobacco prevalence is affected by several factors. Properity is the first of these. Wealthier countries tend to smoke more. Certain religions such as Christianity and Judaism, are anti-tobacco.

Design

It seems like a scatter plot was used in this original visualization. Since there are over a hundred countries listed in the data set, the author may have thought the scatter plot is the best plot to represent numerous rows. The author could have labeled the country names on the x-axis, but to make it a easier to view, I think he or she labeled the country names on the y-axis and the total smoking rate on the x-axis. The dots or the points are presented in green and there are no specific size is assigned. When the viewer moves the cursor on a certain point, the total rate of that country shows up. The author also used lines in the background so that it is easy to follow which point belongs to which country. The author used all the country names in the plot. I think there could be an adjustment here. Although these country names are appropriate variables to use, there are too many to of them to show in one plot. Thus, I think we can reduce some countries, for example, filtering only the OECD countries. The claim is clear and straightforward, but the visualization is not very effective. We can barely find any association between its claim and its visualization. For example, the original visualization does not really show how economic status or the religion of a country has affect on the smoking rate.

Data

Data Overview

I was able to find the original data, and found couple other data sets to join gdp and religion into the original data set. The data sets of smoking rate came from the website which can be downloaded as a csv file: https://worldpopulationreview.com/country-rankings/smoking-rates-by-country (https://worldpopulationreview.com/country-rankings/smoking-rates-by-country) This website allows anyone to download the data because this website provides open databases of various research on the world population. The data seems to be a single file with no aggregation with other files. Furthermore, no personal information was stored in the data set. We may need to include all the data. There should not be any biased action during the data collection process. After all the data is collected, then we can manipulate the data. For example, using left_join or inner_join to filter any unmatching or missing data.

Data Details

For smoking and gdp, each row represents country names in their initials. For religion, each row represents country names in their full names. nrow(smoking) = 115, nrow(gdp) = 115, nrow(religion) = 228 The religion data set may be interesting to know whether or not the dominant religion of a certain country has affect on the smoking rate.

```
# your code to load the data here
smoking <- read_csv('world_smoking.csv')
```

```
##
## ── Column specification ────────────────────────────────────────────────────────
## cols(
##   LOCATION = col_character(),
##   INDICATOR = col_character(),
##   SUBJECT = col_character(),
##   MEASURE = col_character(),
##   FREQUENCY = col_character(),
##   TIME = col_double(),
##   Value = col_double(),
##   `Flag Codes` = col_character()
## )
```

```
nrow(smoking)
```

```
## [1] 115
```

```
gdp <- read_csv('world_smoking.csv')
```

```
##
## ── Column specification ────────────────────────────────────────────────────────
## cols(
##   LOCATION = col_character(),
##   INDICATOR = col_character(),
##   SUBJECT = col_character(),
##   MEASURE = col_character(),
##   FREQUENCY = col_character(),
##   TIME = col_double(),
##   Value = col_double(),
##   `Flag Codes` = col_character()
## )
```

```
nrow(gdp)
```

```
## [1] 115
```

```
religion <- read_csv('world_religion.csv')
```

```
##
## ── Column specification ────────────────────────────────────────────────────────
## cols(
##   country = col_character(),
##   chistians = col_double(),
##   muslims = col_double(),
##   unaffiliated = col_double(),
##   hindus = col_double(),
##   buddhists = col_double(),
##   folkReligions = col_double(),
##   other = col_double(),
##   jews = col_double()
## )
```

```
nrow(religion)
```

```
## [1] 228
```

Wrangling

Replication

Alternatives

Design

Implementation

Summary