STAT 231: Problem Set 7B

Brandon Kwon

due by 10 PM on Friday, April 16

This homework assignment is designed to help you further ingest, practice, and expand upon the material covered in class over the past week(s). You are encouraged to work with other students, but all code and text must be written by you, and you must indicate below who you discussed the assignment with (if anyone).

Steps to proceed:

1. In RStudio, go to File > Open Project, navigate to the folder with the course-content repo, select the course-content project (course-content.Rproj), and click "Open"
2. Pull the course-content repo (e.g. using the blue-ish down arrow in the Git tab in upper right window)
3. Copy ps7B.Rmd from the course repo to your repo (see page 6 of the GitHub Classroom Guide for Stat231 if needed)
4. Close the course-content repo project in RStudio
5. Open YOUR repo project in RStudio
6. In the ps7B.Rmd file in YOUR repo, replace "YOUR NAME HERE" with your name
7. Add in your responses, committing and pushing to YOUR repo in appropriate places along the way
8. Run "Knit PDF"
9. Upload the pdf to Gradescope. Don't forget to select which of your pages are associated with each problem. *You will not get credit for work on unassigned pages (e.g., if you only selected the first page but your solution spans two pages, you would lose points for any part on the second page that the grader can't see).*

If you discussed this assignment with any of your peers, please list who here:

ANSWER: Alastair Poole

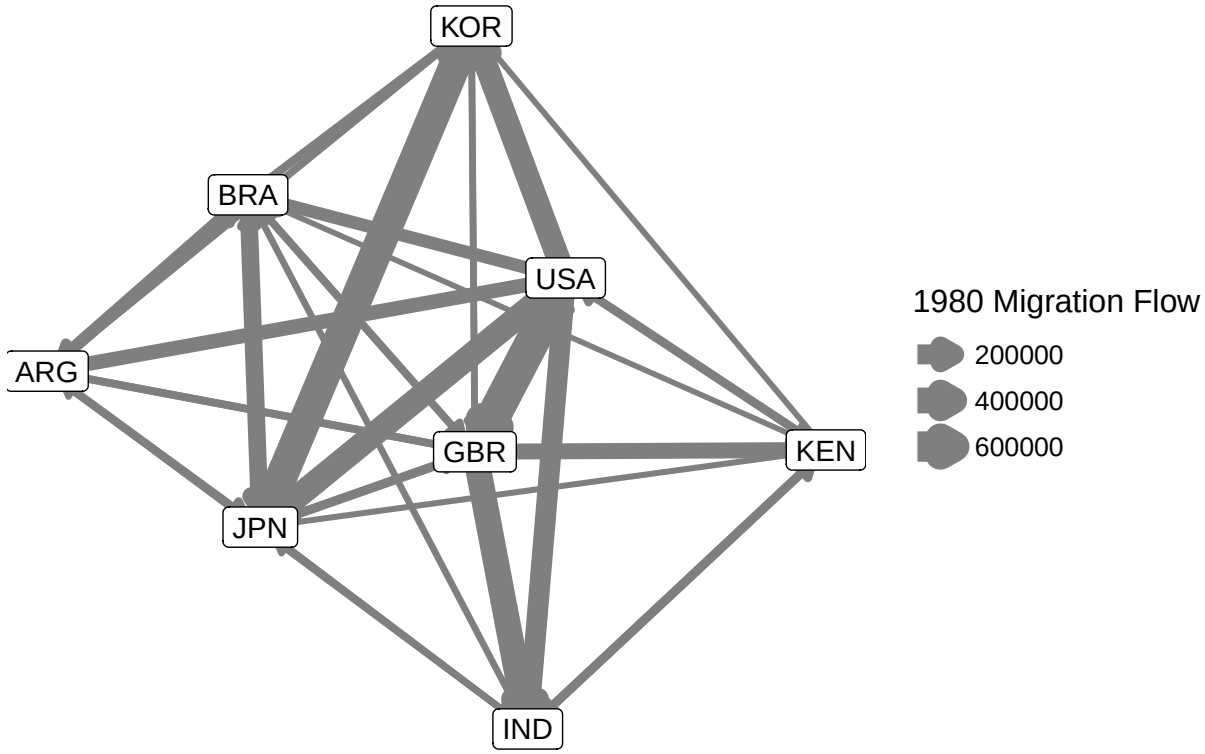
1. More migration

1a. Consider migration between the following countries: Argentina, Brazil, Japan, Kenya, Great Britain, India, South Korea, United States. Compare the TOTAL (males + females) migration between these countries over time. In separate (directed) graphs for 1980 and 2000, visualize the network for the these countries with edge width and/or edge color corresponding to migration flow size. Interpret the two graphs – what *information in context* do they convey?

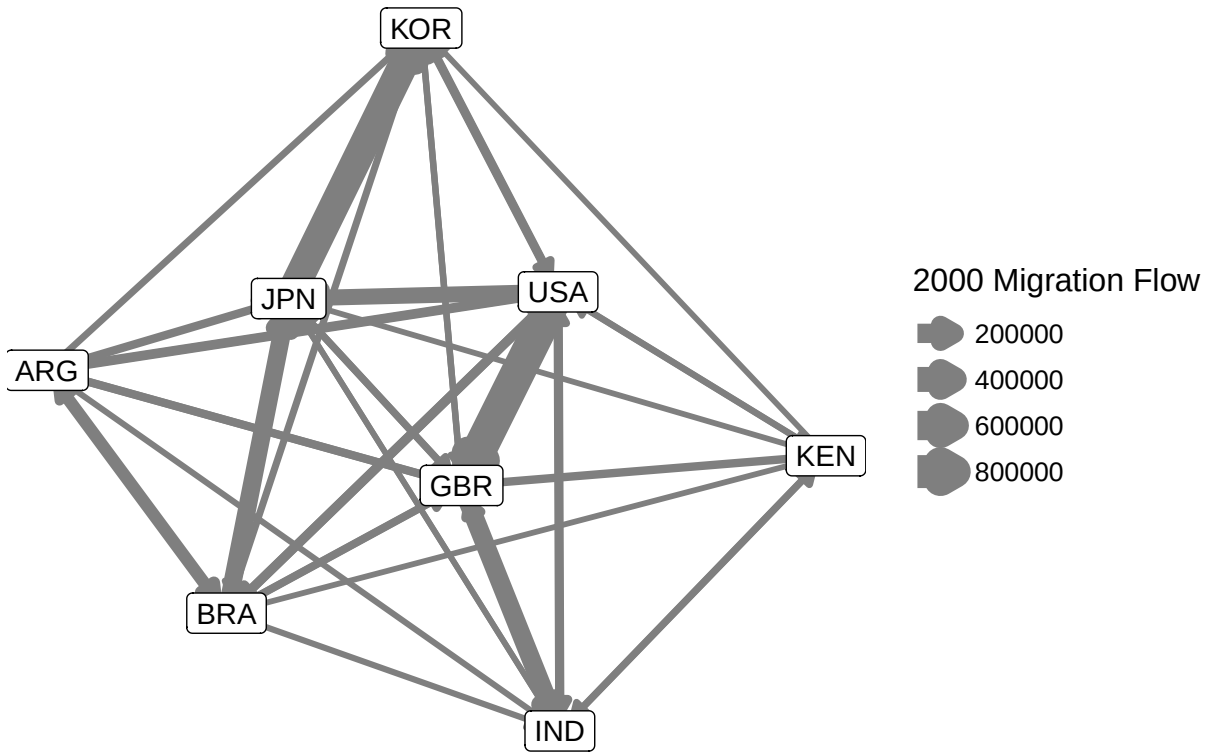
Don't forget to order the columns correctly and only keep relevant rows before transforming into a network object.

> ANSWER: In general, there seems to be a greater migration flow amongst these countries in 1980 as compared to 2000. In 1980, the US and Great Britain seemed to be more centralized in terms of migration flow into and out of the countries. In 2000, Japan seems to be a part of this elite group. This definitely makes because Japan has become more industrialized as time continues to progress. Therefore, it makes sense that more people are moving in and out of this country more often.

```
options(scipen = 999)

path_in <- "~/Spring 2021/Data Science - STAT 231/course-content/data/"
MigrationFlows <- read_csv(paste0(path_in,"/MigrationFlows.csv"))

# Argentina, Brazil, Great Britain, Japan, Kenya,  India, South Korea, United States
countries <- c("ARG", "BRA", "GBR", "JPN", "KEN", "IND","KOR", "USA")

# need migration overall:

# do some prelim data wrangling to combine numbers for males + females

MigrationFlows1 <- MigrationFlows %>%
  filter(destcode %in% countries & origincode %in% countries) %>%
  group_by(destcode, origincode) %>%
  summarise(total1980 = sum(Y1980), total2000 = sum(Y2000)) %>%
  filter(total1980 != 0 & total2000 != 0)

migration_flow1980 <- graph_from_data_frame(MigrationFlows1 %>%
                      select(-total2000) %>%
                        filter(total1980 != 0), directed = TRUE)

ggplot(data = migration_flow1980
       , aes(x = x, y = y, xend = xend, yend = yend)) +
  geom_edges(arrow=arrow(type="closed", length=unit(6,"pt"))
             , color = "gray50"
             , aes(size = total1980)) +
  geom_nodes() +
  geom_nodelabel(aes(label = name)) +
  theme_blank() +
  ggtitle("Selected Countries") +
  labs(size = "1980 Migration Flow")
```




```
migration_flow2000 <- graph_from_data_frame(MigrationFlows1 %>%
                      select(-total1980) %>%
                        filter(total2000 != 0), directed = TRUE)

ggplot(data = migration_flow2000
       , aes(x = x, y = y, xend = xend, yend = yend)) +
  geom_edges(arrow=arrow(type="closed", length=unit(6,"pt"))
             , color = "gray50"
             , aes(size = total2000)) +
  geom_nodes() +
  geom_nodelabel(aes(label = name)) +
  theme_blank() +
  ggtitle("Selected Countries") +
  labs(size = "2000 Migration Flow")
```



1b. Compute the *unweighted* in-degree for Japan in this network from 2000, and the *weighted* in-degree for Japan in this network from 2000. In 1-2 sentences, interpret these numbers in context (i.e., without using the terms "in-degree" or "weighted").

> ANSWER: These numbers suggest that people migrating to Japan have come from 7 other countries, with a total of 931,809 people moving into Japan in 2000.

```
igraph::degree(migration_flow2000, mode = "in")
```

```
## ARG BRA GBR IND JPN KEN KOR USA
##   4   4   7   6   7   4   6   7
```

```
strength(migration_flow2000, weights = E(migration_flow2000)$total2000)
```

```
##     ARG     BRA     GBR     IND     JPN     KEN     KOR     USA
##  114798  272796 1231228  229552  931809   13588  648652 1209469
```

1c. Among these same countries, identify the top 5 countries *of origin* and *of destination* (separately) in 1980 using (weighted) degree centrality. Interpret this information.

> ANSWER: The top 5 countries of origin are USA, Japan, Great Britain, Brazil, and Argentina. This suggests that in 1980, more people migrated out of these countries than the others. The top 5 countries of destination are Korea, Great Britain, India, Japan, and USA. This suggests that in 1980, more people migrated into these countries than the others.

```
sort(strength(migration_flow1980, weights=E(migration_flow1980)$total1980, mode="in"))
```

```
##    BRA    ARG    KEN    USA    JPN    IND    GBR    KOR
## 105472 107789 122617 180296 502540 643586 832184 993074
```

```
sort(strength(migration_flow1980, weights=E(migration_flow1980)$total1980, mode="out"))
```

```
##     KOR     IND     KEN     ARG     BRA     GBR     JPN     USA
##    8270   22709   27882   69756  194211  647261  705932 1811537
```

1d. Among these same countries, identify the top 5 countries *of origin* and *of destination* (separately) in 2000 using (weighted) degree centrality. Interpret this information.

> ANSWER: The top 5 countries of origin are Great Britain, Korea, Japan, India, and USA This suggests that in 2000, more people migrated out of these countries than the others. The top 5 countries of destination are USA, Japan, Great Britain, Brazil, and Argentina. This suggests that in 1980, more people migrated into these countries than the others.

```
sort(strength(migration_flow2000, weights=E(migration_flow2000)$total2000, mode="in"))
```

```
##    KEN    ARG    BRA    USA    IND    JPN    KOR    GBR
##  11699  41082  67542 163747 206519 294863 639215 901279
```

```
sort(strength(migration_flow2000, weights=E(migration_flow2000)$total2000, mode="out"))
```

```
##     KEN     KOR     IND     ARG     BRA     GBR     JPN     USA
##    1889    9437   23033   73716  205254  329949  636946 1045722
```

1e. What is the diameter of this network in 2000? In 1-2 sentences, interpret this value.

> ANSWER: This suggests that the length of the longest of all computed shortest paths is 2 in terms of migration flow. In other words, when computing lengths of paths between two countries, the longest of the shortest ones has a length of 2.

```
diameter(migration_flow2000)
```

```
## [1] 2
```

1f. What is the density of this network in 2000? In 1-2 sentences, interpret this value.

> ANSWER: This suggests that the portion of potential connections that are actual connections is very high. This suggests that these countries are all relatively connected when it comes to migration flow.

```
graph.density(migration_flow2000)
```

```
## [1] 0.8035714
```

2. Mapping spatial data

Reproduce the map you created for Lab08-spatial (and finish it if you didn't in class). In 2-4 sentences, interpret the visualization. What stands out as the central message?

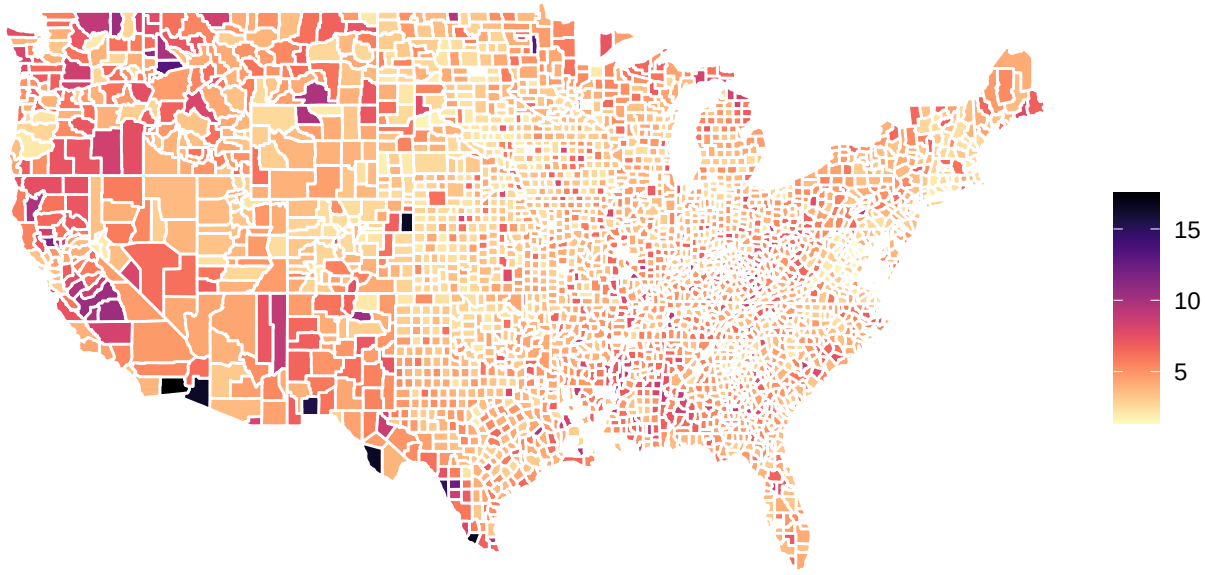
NOTE: you do NOT need to say what colors are representing what feature (e.g, NOT: "In this map, I've colored the countries by GDP, with green representing low values and red representing high values") – this is obvious to the viewer, assuming there's an appropriate legend and title. Rather, what *information* do you extract from the visualization? (e.g., "From the choropleth below, we can see that the percent change in GDP per capita between 1957-2007 varies greatly across countries in Central America. In particular, Panama and Costa Rica stand out as having GDPs per capita that increased by over 200% across those 50 years. In contrast, Nicaragua's GDP per capita decreased by a small percentage during that same time span.")

> ANSWER: I can see that moving from west to east in the United States, the unemployment rate seems to decrease gradually. I am not sure why this is the case, but I can assume that because it is very expensive to live in the west coast, it is more difficult to obtain a sufficient job.

```
county_employment <- readxl::read_excel(paste0("~/Spring 2021/Data Science - STAT 231/course-content/da
                                        , sheet = 1
                                        , skip = 7) %>%
  janitor::clean_names()

usa_counties <- map_data(map = "county", region = ".")

county_employment <- county_employment %>%
  mutate(area_name = gsub(pattern = " County.*", replacement = "", x = area_name),
         area_name = gsub(pattern = " Borough.*", replacement = "", x = area_name),
         area_name = gsub(pattern = " Area.*", replacement = "", x = area_name),
         area_name = gsub(pattern = " Municipality.*", replacement = "", x = area_name),
         area_name = str_to_lower(area_name))

final_df <- usa_counties %>%
  inner_join(county_employment, by = c("subregion" = "area_name"))


ggplot(final_df, aes(x = long, y = lat, group = group, fill = unemployment_rate_2000)) +
  geom_polygon(colour = "white") +
  theme_void() +
  coord_fixed(ratio = 1.3) +
  labs(title = "Unemployment Rate by Counties in the United States"
       , subtitle = "in the Year 2000"
       , caption = "* Hawaii and Alaska not shown above"
       , fill = "") +
  scale_fill_viridis(option = "magma", direction = -1)
```

Unemployment Rate by Counties in the United States

in the Year 2000



* Hawaii and Alaska not shown above

3. Mapping spatial data at a different level

Create a map at the world, country, or county level based on the choices provided in lab08-spatial, that is at a DIFFERENT level than the map you created for the lab (and included above). For instance, if you created a map of US counties for the lab, then choose a country or world map to create here.

Note: While I recommend using one of the datasets provided in the lab so you don't spend a lot of time searching for data, you are not strictly required to use one of those datasets.

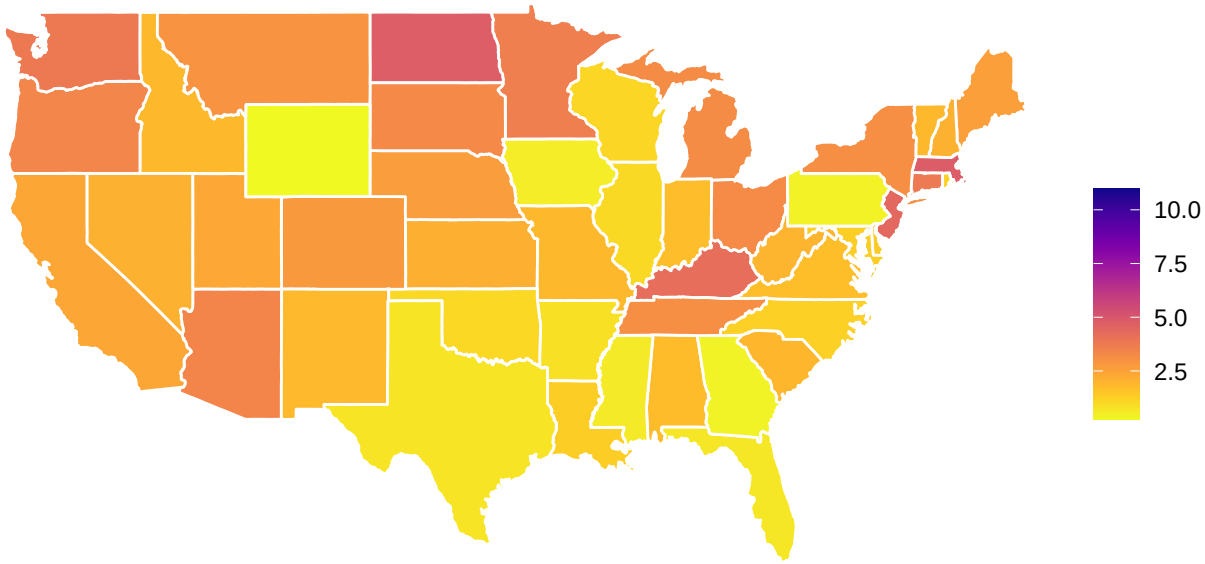
Describe one challenge you encountered (if any) while creating this map.

> ANSWER: One challenge I had when creating this map was making the states lowercase to better make the map.

```
hate_crimes <- fivethirtyeight::hate_crimes

# states in the US
usa_states <- map_data(map = "state", region = ".")
#change all states to lowercase
hate_crimes <- hate_crimes %>%
  mutate("state" = tolower(state))
#creating new datset
hate_crimes_map <- hate_crimes %>%
  inner_join(usa_states, by = c("state" = "region"))

ggplot(hate_crimes_map, aes(x = long, y = lat, group = group
                      , fill = avg_hatecrimes_per_100k_fbi)) +
  geom_polygon(color = "white") +
  theme_void() +
  coord_fixed(ratio = 1.3) +
  labs(title = "Average Hate Crimes per 100k, by state*"
       , subtitle = "as Reported by the FBI"
       , caption = "* Hawaii and Alaska not shown above"
       , fill = "") +
  scale_fill_viridis(option = "plasma", direction = -1)
```

Average Hate Crimes per 100k, by state*

as Reported by the FBI



* Hawaii and Alaska not shown above

4. Love Actually (OPTIONAL PRACTICE)

This problem is *optional* and will not be graded, but is given to provide additional practice interpreting networks and as another real-world example of network analysis that might be intriguing to film buffs.

Consider the figure "The Two Londons of 'Love Actually' " in this FiveThirtyEight article.

2a. Based on this figure, is the network connected? In 1-2 sentences, please explain.

ANSWER:

2b. Based on the figure, what is the (unweighted) degree for Emma Thompson? What is the (unweighted) degree for Keira Knightley? Explain what these values mean for these characters.

ANSWER:

2c. Based on the figure, for whom would the (unweighted) betweenness centrality measure be higher: Colin Firth or Hugh Grant? Explain what this implies.

ANSWER:

5. Migration network on a world map! (OPTIONAL PRACTICE)

This problem is *optional* and will not be graded, but is given to provide additional coding practice and as a challenge to incorporate networks and mapping techniques together.

Create a world map that visualizes the network of countries we examined in #1 for the year 2000. For example, arrows to and from each of countries on the world map could have edge widths relative to their weighted degree centrality to represent migration to and from the countries.

Code to get you started is provided below.

```
# from mdsr package
# should see 'world_cities` df in your environment after running
data(world_cities)

# two-letter country codes
# Argentina, Brazil, Great Britain, Japan, Kenya
# India, South Korea, United States
countries2 <- data.frame(country3=countries
                         , country2 = c("AR", "BR", "GB", "JP"
                                        , "KE", "IN","KR", "US"))

# find capitals for anchoring points; can't find D.C., use Boston
cities <- c("Buenos Aires", "Brasilia", "London", "Tokyo", "Nairobi"
             , "New Delhi", "Seoul", "Boston")

anchors <- world_cities %>%
  right_join(countries2, by = c("country" = "country2")) %>%
  filter(name %in% cities) %>%
  select(name, country, country3, latitude, longitude)

# one suggested path:
# 1. based on the anchors dataset above and your Migration 2000 dataset created for # 1,
#    create dataframe that would supply geom_curve with the relevant arrow locations
#    (start points and end points)
# 2. create world map dataset using `map_data` function
# 3. use geom_polygon to create world map, geom_point and/or geom_text to add
#    city points, and geom_curve to add weighted/colored arrows
```