STAT 231: Problem Set 7B

Kevin Ma

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: TA hours with Andrea

1. More migration

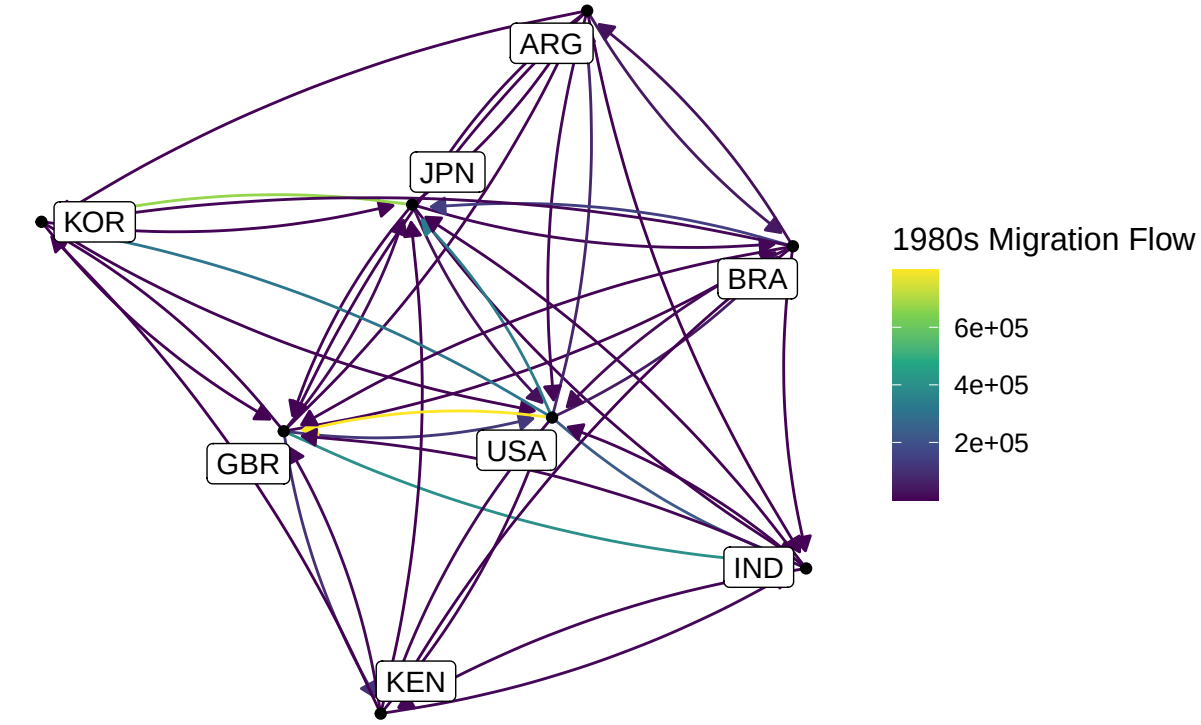
1a. Consider migration between the following countries: Argentina, Brazil, Japan, Kenya, Great Britain, India, South Korea, United States. Compare the TOTAL (males + females) migration between these countries over time. In separate (directed) graphs for 1980 and 2000, visualize the network for the these 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.

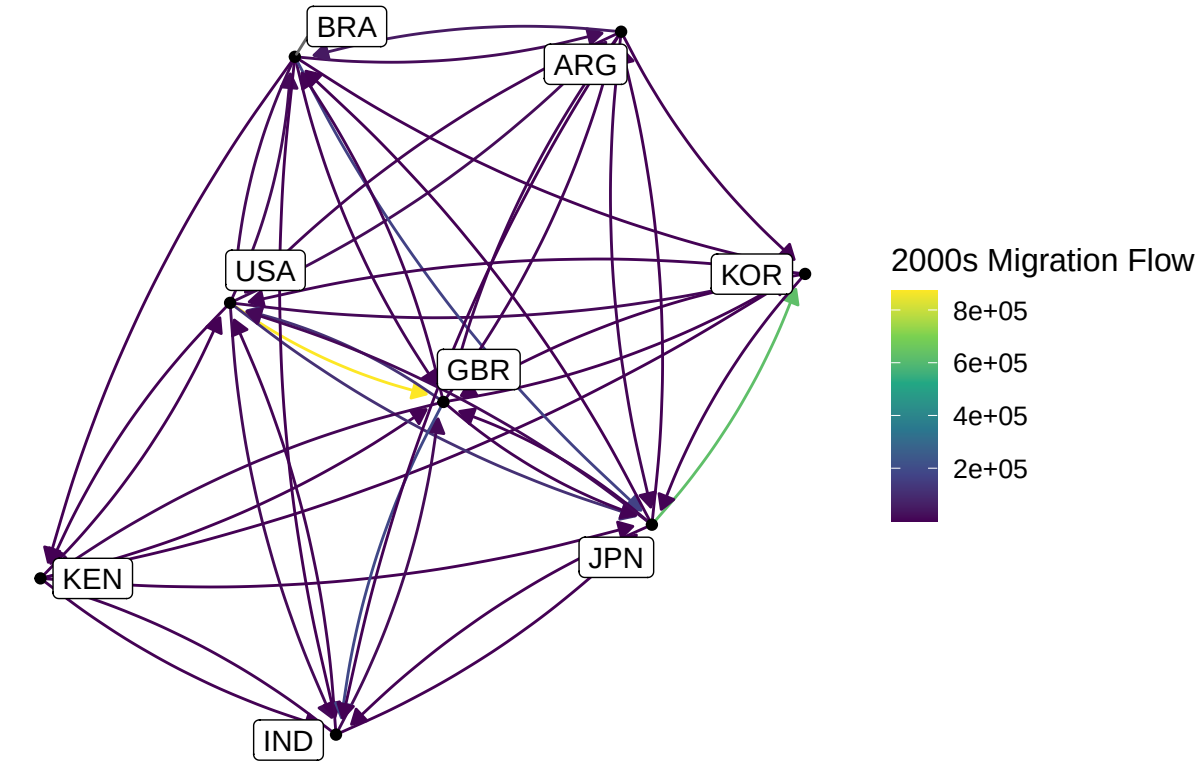
> ANSWER: I think that migration was more evenly spread across countries in the 1980s. Migration in and out of India and Argentina have declined. On the other hand, migration across countries that already had strong migration seem to have increased, for example between GBR and USA as well as Japan and Korea.

```
path_in <- "~/Github/Stat231/Homework/BPsets"
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")

total_migration <- MigrationFlows %>%
  filter(Y1980 > 0, Y2000 > 0) %>%
  select(origincode, destcode, Y1980, Y2000) %>%
  filter(destcode %in% countries & origincode %in% countries) %>%
  group_by(destcode, origincode)%>%
  summarise(Migration2000 = sum(Y2000), Migration1980 = sum(Y1980))

# need migration overall:
# do some prelim data wrangling to combine numbers for males + females
migration1980 <- total_migration%>%
  select(destcode, origincode, Migration1980)

migration2000 <- total_migration%>%
  select(destcode, origincode, Migration2000)

mig_1980 <- graph_from_data_frame(migration1980
                                  , directed = TRUE)
mig_2000 <- graph_from_data_frame(migration2000
                                  , directed = TRUE)
```

```
ggplot(data = ggnetwork(mig_1980)
       , aes(x = x, y = y, xend = xend, yend = yend)) +
  geom_edges(curvature = 0.1
             , arrow=arrow(type="closed", length=unit(6,"pt"))
             , aes(color = Migration1980)) +
  geom_nodes() +
  geom_nodelabel_repel(aes(label = name)) +
  theme_blank() +
  ggtitle("Migration Across A Set Of Countries") +
  labs(color = "1980s Migration Flow") +
  scale_color_continuous(type = "viridis")
```




```
ggplot(data = ggnetwork(mig_2000)
       , aes(x = x, y = y, xend = xend, yend = yend)) +
  geom_edges(curvature = 0.1
             , arrow=arrow(type="closed", length=unit(6,"pt"))
             , aes(color = Migration2000)) +
  geom_nodes() +
  geom_nodelabel_repel(aes(label = name)) +
  theme_blank() +
  ggtitle("Migration Across A Set Of Countries") +
  labs(color = "2000s Migration Flow") +
  scale_color_continuous(type = "viridis")
```



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: The unweighted in-degree for Japan in this network from 2000 is 7. This means that there were 7 other countries that people left from to migrate to Japan. In addition, the weighted in-degree for Japan in this network from 2000 is 294,863, which means that 294,863 people migrated to Japan in 2000.

```
V(mig_2000)$degree <- igraph::degree(mig_2000, mode = "in")
V(mig_2000)$wtdegree <- strength(mig_2000
                                 , weights=E(mig_2000)$Migration2000, mode = "in")

stats <- data_frame(name = V(mig_2000)$name,
                    degree = V(mig_2000)$degree,
                    wtdegree = V(mig_2000)$wtdegree )
```

```
## Warning: `data_frame()` was deprecated in tibble 1.1.0.
## Please use `tibble()` instead.
```

```
statsJapan <- stats %>%
  filter(name == "JPN")
statsJapan
```

```
## # A tibble: 1 x 3
##   name  degree wtdegree
##   <chr>  <dbl>    <dbl>
## 1 JPN        7   294863
```

```
#Weighted in-degree
```

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 in 1980 were USA, JPN, GBR, BRA, and ARG. These countries were the top 5 most popular countries from the list of selected countries to migrate from in 1980 to the other selected countries. The top 5 countries of destination in 1980 were KOR, GBR, IND, JPN, and USA. These countries were the top 5 most popular countries from the list of selected countries to migrate to in 1980 from the list of selected countries.

```
head(sort.int(strength(mig_1980, weights = E(mig_1980)$Migration1980, mode = "out"),
              decreasing = TRUE), 5)
```

```
##     USA     JPN     GBR     BRA     ARG
## 1811537  705932  647261  194211   69756
```

```
# of destination
head(sort.int(strength(mig_1980, weights = E(mig_1980)$Migration1980, mode = "in"),
              decreasing = TRUE), 5)
```

```
##    KOR    GBR    IND    JPN    USA
## 993074 832184 643586 502540 180296
```

1d. Among these same countries, idenitify 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 in 2000 were USA, JPN, GBR, BRA, and ARG. These countries were the top 5 most popular countries from the list of selected countries to migrate from in 1980 to the other selected countries. The top 5 countries of destination in 1980 were GBR, KOR, IND, JPN, and USA. These countries were the top 5 most popular countries from the list of selected countries to migrate to in 1980 from the list of selected countries. There is not much change between 1980 and 2000 in terms of order.

```
head(sort.int(strength(mig_2000, weights = E(mig_2000)$Migration2000, mode = "out"),
              decreasing = TRUE), 5)
```

```
##     USA     JPN     GBR     BRA     ARG
## 1045722  636946  329949  205254   73716
```

```
# of destination
head(sort.int(strength(mig_2000, weights = E(mig_2000)$Migration2000, mode = "in"),
              decreasing = TRUE), 5)
```

```
##    GBR    KOR    JPN    IND    USA
## 901279 639215 294863 206519 163747
```

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

> ANSWER: The diameter of this network in 2000 is 2. This means that the shortest path between any two countries in this network is two edges.

```
diameter(mig_2000, directed = TRUE)
```

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

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

> ANSWER: The density of this network in 2000 is 80.36%. Thus, around 80% of the total possible edges are there, meaning that this networks is well-connected. Many countries have edges to the majority of the other countries.

```
graph.density(mig_2000)
```

```
## [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: In the map below we can observe the number of hate crimes per 100 thousand across different states in the United States from 2013 to 2014. In particular, North Dakota, Kentucky, New Jersey, and Massachusetts report the highest number of hate crimes per 100,000. Hate crimes is spread out geographically and politically. For example Georgia and Mississippi report low numbers of hate crimes. Thus, this data may run into self-reporting bias from states.

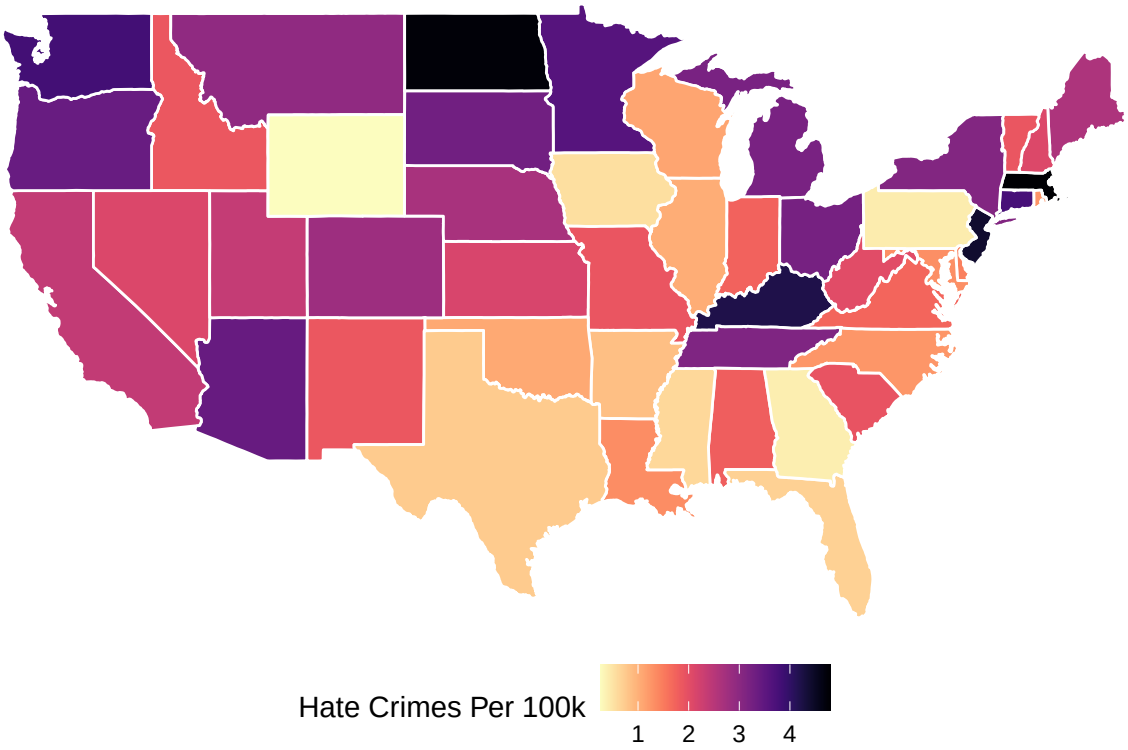
```
library(fivethirtyeight)
```

```
## Some larger datasets need to be installed separately, like senators and
## house_district_forecast. To install these, we recommend you install the
## fivethirtyeightdata package by running:
## install.packages('fivethirtyeightdata', repos =
## 'https://fivethirtyeightdata.github.io/drat/', type = 'source')
```

```
data(state)

# creates a data frame with state info
state_info <- data.frame(state_full = tolower(state.name)
                         , State = state.abb
                         , Region = state.region)
usa_states <- map_data(map = "state"
                       , region = ".")
hate_crimes <- fivethirtyeight::hate_crimes

hate_crimes <- hate_crimes %>%
  rename(State = state_abbrev)

hatecrimes_map <- hate_crimes %>%
  left_join(state_info, by = "State") %>%
  right_join(usa_states, by = c("state_full" = "region"))

ggplot(hatecrimes_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(fill = "Hate Crimes Per 100k") +
  theme(legend.position="bottom") +
  scale_fill_viridis(option = "magma", direction = -1) +
  labs(title = "Hate Crimes per 100k by State 2013-2014",
       subtitle = "Kevin Ma")
```

Hate Crimes per 100k by State 2013−2014

Kevin Ma



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.

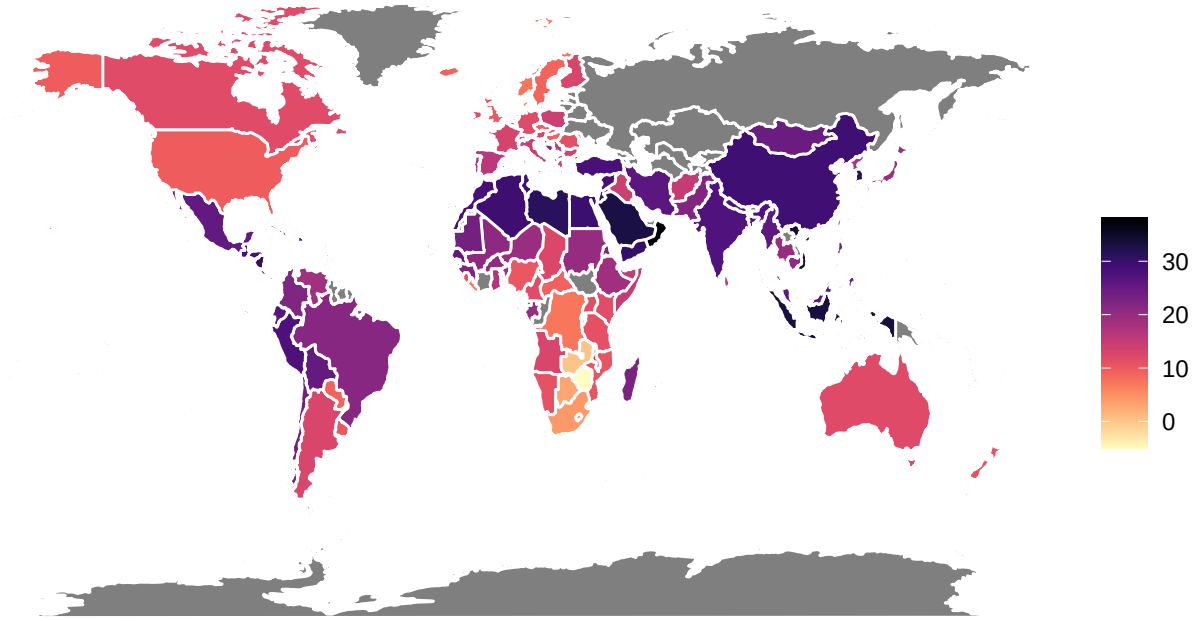
Note: While I recommend using one of the datasets provided in the lab so you don't spend a lot of time searching for data, you are not strictly required to use one of those datasets.

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

ANSWER:

```
library(gapminder)

world_map <- map_data(map = "world"
                      , region = ".")
world_map$region[world_map$region == "USA"] <- "United States"
world_map$region[world_map$region == "UK"] <- "United Kingdom"
world_map$region[world_map$region == "Democratic Republic of the Congo"] <- "Congo, Dem. Rep."
world_map$region[world_map$region == "Yemen"] <- "Yemen, Rep."
world_map$region[world_map$region == "North Korea"] <- "Korea, Dem. Rep."
world_map$region[world_map$region == "South Korea"] <- "Korea, Rep."
world_map$region[world_map$region == "Slovakia"] <- "Slovak Republic"

gapminder <- gapminder::gapminder

gapminder1 <- gapminder %>%
  filter(year == 1952 | year == 2007) %>%
  select(country, year, lifeExp) %>%
  mutate(lifeExp = as.numeric(as.character(lifeExp))) %>%
  group_by(country) %>%
  mutate(diff = lifeExp - lag(lifeExp)) %>%
  filter(year == 2007) %>%
  rename(region = country)

mapping <- gapminder1 %>%
  right_join(world_map, by = "region")

ggplot(mapping, aes(x = long, y = lat, group = group
                      , fill = diff)) +
  geom_polygon(color = "white") +
  theme_void() +
  coord_fixed(ratio = 1.3) +
  labs(title = "Change in Life Expectancy between 1952 and 2007"
       , caption = "Countries in grey have no data"
       , fill = "") +
  scale_fill_viridis(option = "magma", direction = -1)
```

Change in Life Expectancy between 1952 and 2007



Countries in grey have no data

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
```