Part 1

This is an exercise that will test your ability to work with large datasets to draw interesting conclusions and present results in a compelling way.

You will be working with the Social Connectedness Index (SCI), a dataset built from an anonymized snapshot of Facebook users and their friendship networks. The data measure the intensity of social connections between counties. (For one overview, see this coverage in the New York Times).

Relevant data: "county_county_sci.tsv", "sf12010countydistancemiles.csv"

Our team has a presentation coming up to a set of policymakers in Washtenaw County, Michigan. Using the SCI and the county distance datasets, make a set of exploratory plots describing the social connections of Washtenaw County. Specifically:

(a) Summarize the distribution of Washtenaw's Social Connectedness Index to other counties
(b) Which counties are most strongly connected to Washtenaw?
(c) Merge in the distance data and describe the relationship between distance to Washtenaw and connectedness to Washtenaw

The team is also interested in exploring the relationship between a county's *network concentration* – e.g., the share of a county's Facebook friends that are located nearby – and other important socio-economic measures. For this part:

(d) Using the "county_county_sci.tsv" and "sf12010countydistancemiles.csv", construct a county-level measure of *network concentration*. Briefly justify your measure (there is no single "right" answer). Remember that the social connectedness index is defined as:

$$Social\ Connectedness\ Index_{i,j} = \frac{FB\_Connections_{i,j}}{FB\_Users_i * FB\_Users_j}$$

(e) Merge in the county_demographics dataset and describe relationships between network concentration and 2-3 other county level measures. Suggest possible explanations of why these relationships might exist. Discuss any ideas you have on how your explanations could be tested (perhaps using other data or in other contexts).

Data Dictionary

1. county_county_sci.tsv
   - This is the Facebook Social Connectedness Index (SCI). A full description of the data is available here. We include the "US Counties – US Counties" data, a symmetric measure between every pair of US counties. The original data can be found here.
   - The columns are:
     - user_loc = The FIPS code of the first county (the user's county)
     - fr_loc = The FIPS code of the second county (the friend's county)
     - scaled_sci = The (symmetric) Social Connectedness Index between counties, as detailed in the full description document linked to above.

2. sf12010countydistancemiles.csv
   - This is the distance between every county pair (from the NBER)
   - ***Important:*** the county pairs between a county and *itself* are excluded
   - The columns are:
     - county1 = The FIPS code of the first county
     - mi_to_county = Miles between the centers county1 and county2
     - county2 = The FIPS code of the second county

3. county_description.csv
   - This is a set of simple county descriptors
   - The columns are:
     - county_fips = County FIPS code
     - county_name = County name
     - state_fips = State FIPS code
     - state_name = State name
     - state_abrev = State abbreviation

4. county_demographics.csv
   - This is a set of county-level demographics and socio-economic outcomes from Bailey et al., 2018
   - The columns are county_fips, measure, and value
   - "measure" is one of:
     a. no_highschool = The share of the population that did not attend high school
     b. total_population = The total population
     c. male_population = The male population
     d. median_age = The median age
     e. pct_white_alone = The share of the population that is White alone
     f. median_hh_income = The median household income
     g. mean_hh_income = The mean household income
     h. share_below_povline = The share of the population below the poverty line
     i. obama_share_vs_mccain = Of those who voted for Obama or McCain for president in 2008, the share that voted for Obama
     j. cz = The commuting zone this county falls within
     k. e_rank_b = A measure of upward mobility (i.e., how likely it is for you to have a higher income than your parents). Higher values mean more upward mobility. From Chetty et al. 2020.
     l. frac_kteenbirthfem = The teen birth rate
     m. sk97 = A measure of social capital (i.e., the general strength of the relationships and trust of people in a county). From Rupasingha et al. 2006.
     n. le_agg_q[X]_[Y] = For people of gender Y in this county with incomes in quarter X of the national income distribution, their life expectancy.

Part 2

This is an exercise that will test your ability to work with cleaning and processing data. You will work with Twitter data attributed to Russia's Internet Research Agency (IRA) to investigate Russian trolls' tweeting patterns in response to different political events.

Since 2016, many journalistic accounts have reported that Russia aimed to stoke polarization on issues related to police violence and race relations. For this exercise, we will explore these claims quantitatively and see whether IRA tweeting patterns really are affected by BLM-related events in the US. Specifically, we'll look at the following events: the death of Freddie Grey on August 19, 2015, Sandra Bland on July 13, 2015, and the Alon Sterling shooting on July 5, 2016.

Relevant data: "ira_tweets_csv_hashed.dta"

(a) Create a panel of the following variables at the day level:

- Number of tweets
- Average engagement metrics (replies, likes, quotes, and retweets)
- Number of tweets mentioning either "BLM" or "Black Lives Matter"

(b) Make a graph that shows the yearly evolution of the 6 variables in your tweets dataset (tweet count, four engagement metrics, tweets about BLM).

(c) Run the following regression:

$$Y_t = B_0 + \beta_1 X_t + \varepsilon_t \quad (1)$$

where Y_t is the outcome variable at time t, X_t is an indicator that equals 0 during the pre-treatment period and 1 post-treatment. Set the pre/post window as 30 days from the event. Your code should loop over each of the 3 events as well as each of the 6 variables in your tweets dataset (tweet count, four engagement metrics, tweets about BLM).

(d) What is the interpretation of each coefficient?