Homework 9

Due: Wednesday April 24 at 11:59pm

Use RStudio for all exercises. Provide one code file that contains all the code and includes comments noting which code is for which exercises. You will also need to comment on the results, so place them in a Word (or Open Office or HTML or PDF) document and add your comments to answer the questions. You can use knitr to programmatically create your document or use an R notebook if you like. Be sure to include your name in the file name for all files submitted.

Any code based on code from elsewhere must reference in comments the source of the original code. **Code files must be the actual code files**, not code pasted into some other document.

Exercise 1 is based on the **`Grunfeld`** data set included in the **`Ecdat`** package. This package will need to be installed.

Exercise 2 is based on the **`ggplot2movies::movies`** dataset from R used in streamgraph examples.

Exercises 3 and 4 use the **`LocationData.csv`** data file attached to this assignment on compass. This data set is a simulated data set about location migration by region in the United States. The data is for migration from one region to another. The first column contains the initial region of residence and the second column contains final region of residence.

Exercise 1:

The **Grunfeld** data set includes investment and valuation numbers for 10 companies over a range of years. The data set and variables are described in R documentation.

From the data set, construct the following 3 streamgraphs:

- Investment by firm over time
- Value by firm over time
- Capital by firm over time

Interpret what the plots tell us about changes in investment, value, and capital over the time frame of the data. Also comment on magnitudes and trends for those measures for the firms (e.g. firms with higher or lower values in general, firms with relatively flat values over time, firms with larger increases or decreases over the timespan or during specific periods, etc.)

Exercise 2:

In this exercise, we will construct a circular graphic for a few characteristics of the movies data. Everyone will do the first few steps, and graduate students will add an extra layer. Note that there will be a few thousand observations so the graphic may be slow to render.

Starting from the **`ggplot2movies::movies`** data frame, construct a new data set that only contains observations with known mpaa values. Omit any observations for which mpaa is not known.

Use the year variable as the x axis, and create a circular graphic with mpaa values as the sectors and the following tracks starting with the outermost track:

- Scatter plot with movie length as the y coordinate and year as the x coordinate
- A label for the mpaa rating
- Track lines with year as the x coordinate and budget as the y coordinate
- Histograms for number of movies made (**graduate students only**)

Based on the graphic, comment on any interesting features in the variables plotted across mpaa ratings.

Exercise 3:

In this exercise, we consider migration patterns within and between regions.

Construct a chord diagram for the data in **`LocationData.csv`**. The data will need to be processed into the appropriate form for the necessary plotting function. Comment on trends across regions (e.g. which regions have higher and lower migration numbers out of region, and which out-of-region migration paths are more and less common).

Exercise 4 (graduate students only):

Aggregate the regions from exercise 3 into three groups based on the names of the regions. Those with North in their name should be aggregated into one North group, those with Midwest into one Midwest group, and those with South into one South group.

After making this aggregation, create a chord diagram for this new data and answer the same questions as in exercise 3.