Homework 1

Instructions

This assignment is entirely about GitHub and data management. The goal is to give you a chance to practice wrangling and tidying data. We do this very early in the class because we will start doing some empirical analysis using real data soon. The faster you are comfortable with the datasets, the better. For more detailed instructions on how to submit your homework answers, please see the overview page here.

The due date for initial submission is **1/29**, the revision due date is **1/31**, and the final due date is Friday, **2/2**.

Building the data

Most of your professional lives will likely involve managing data. It can be tedious but also extremely rewarding when you finally get to find out what's going on in the analysis stage. Anyway, let's get to work! All of these questions require you to use the Medicare Advantage GitHub Repo.

Enrollment Data

Run the `R` code to organize the Monthly Plan Enrollment Data focusing only on 2010 through 2015. Once you've created your final dataset (it's called *full_ma_data* in my code), answer the following:

1. How many observations exist in your current dataset?
2. How many different *plan_types* exist in the data?
3. Provide a table of the count of plans under each plan type in each year. Your table should look something like Table 1.

Table 1: Plan Count by Year

	2010	2011	2012	2013	2014	2015
Type 1	22	25	10	31	45	22
Type 2	37	39	16	27	39	37
Type 3	12	36	47	39	18	12

4. Remove all special needs plans (SNP), employer group plans (eghp), and all "800-series" plans. Provide an updated version of Table 1 after making these exclusions.

5. Merge the contract service area data to the enrollment data, and restrict the data only to contracts that are approved in their respective counties. The `R` script to create the service area dataset is here: Contract Service Area. And you can follow the _BuildFinalData.R script to see where/how I join the datasets. Limiting your dataset only to plans with non-missing enrollment data, provide a graph showing the average number of Medicare Advantage enrollees per county from 2010 to 2015. Be sure to format your graph in a meaningful way.

Premium Data

Now we're going to incorporate the plan premium information. This is part of the "Plan Characteristics" data, and the underlying `R` scripts for these files can be found here: Plan Characteristics.

6. Merge the plan characteristics data to the dataset you created in Step 5 above. Note that you'll need to join the Market Penetration Data in order to get the information you need to merge the plan characteristics. This is because the plan characteristics data only have state name and county (not FIPS codes). The penetration files have both FIPS codes and state/county names, so that dataset serves as a good crosswalk file. Provide a graph showing the average premium over time. Don't forget about formatting!

7. Provide a graph showing the percentage of $0 premium plans over time. Also…remember to format things.

Summary Questions

With all of this data work and these great summaries, let's take a step back and think about what all this means.

8. Why did we drop the "800-series" plans?

9. Why do so many plans charge a $0 premium? What does that really mean to a beneficiary?

10. Briefly describe your experience working with these data (just a few sentences). Tell me one thing you learned and one thing that really aggravated you.