**Final Project Initial proposal**

**Group Members:**

Levey, Trevor (PID: 720492717)

Peng, Qianhe (PID: 730003665)

Schulz, Kirsten(PID: 720471345)

**1.What questions will you try to answer? List 4-7 possible questions.**

For the final project, we will try to answer the following questions:

1. How is our data set organized and what information does it include?
2. What is the average, lowest and highest inpatient medical cost in each state and region?
3. How the cost of inpatient medical care varies across states and regions?
4. What is the relationship between Average Covered Charges and Average Total Payments across states and regions?
5. Are there certain patterns about costs in big/small cities (urban-rural difference) within the same state?
6. Are there any relations between the hospital size (by the number of hospital’s total discharges) and the inpatient medical costs?
7. What are the concerns and limitations of the dataset?

**2.What datasets will you use? You should have already found and taken a first look at the datasets. Make sure the data is clean enough to reasonably use and actually has the information content to answer your questions.**

We will use the data.gov dataset [Inpatient Prospective Payment System (IPPS) Provider Summary for the Top 100 Diagnosis-Related Groups (DRG) - FY2011](https://data.cms.gov/Medicare-Inpatient/Inpatient-Prospective-Payment-System-IPPS-Provider/97k6-zzx3). This data set recorded costs for different procedures, and payments after insurance. The data has been aggregated by groups defined by procedure, healthcare service provider, and location (city, state, and region). Included in each group are, number of discharges, average covered charges, average total payments, and average medicare payments. There are 163,065 total entries in this data set. This data set is from 2011, and any conclusions drawn will reflect that this is not very recent, but is the most recent data available.

**3.What are some things you will do with the data to get at your questions? For example, what are some plots you might make?**

We will be able to compare costs of common procedures using the Diagnosis-Related Groups (DRG) definition variable. This categorizes similar diagnoses/ procedures during the hospital stay. This variable will allow us to compare medical costs across regions and states in the United States. Provider State, Provider Zip Code, and Provider City will give our geographical data for the observations. In addition to Average Total Payments, this data set provides the Average Covered Charges (medicare coverage) for the patients. We will be able to extract average covered charges and average total payments by area and treatment. We will calculate the distribution of treatment types across regions and states so we can account for cost discrepancies due to extreme data points. Our main objective is to create a map of the United States that would give average medical costs by region. We would clearly display our findings using Shiny and a color-scaled map.

**On a second page include a list of 3 other ideas you brainstormed with a couple bullet points of detail.**

1. Look at the relationship between fruit and vegetable prices and health and nutrition.
   1. Using <https://catalog.data.gov/dataset/fruit-and-vegetable-prices>
   2. We would make an app that could look at key statistics and plots for the different categories of fruits and vegetables, as well as individual fruits and vegetables.
2. Look at the relationship between nutrient intake and demographic characteristics
   1. Using <https://catalog.data.gov/dataset/food-consumption-and-nutrient-intakes>
   2. We would also make an app for this. It would be similar to the idea above, but would explore different relationships.
3. Look at the US electric system operating data
   1. Using <https://www.eia.gov/beta/realtime_grid/#/status?end=20171108T19>
   2. We would explore the use of electricity by regions in the United States.