Data Report

A. Description of the Data:

We acquired our data set from openpaymentsdata.cms.gov, a government website which reports on physician payments by pharmaceutical companies per Section 6002 of the Affordable Care Act. Our data set includes data from the first reporting period of the act, August to December 2013. Since this was the first period, CMS did not yet have a full year's worth of data.

We used the "General Payment Data with Identifying Recipient Information – Detailed Dataset 2013 Reporting Year" data set, and leveraged the website's online data manipulation tool to filter for the "Recepient_Country" to the United States so that we only get data on United States physicians. Next, we used the site's "Filter" tool and its "Sort & Roll-Up Tool" feature to group the data by the "Applicable_Manufacturer_or_Applicable_GPO_Making_Payment_Name" (the name of the pharma company making the payment) and "Recepient_State" (the physician receiving the payment) variables. We then rolled up the data in these variables by count of the "Physician_Profile_ID" (a unique identifier of the receiving physician) data and the sum of the "Total_Amount_of_Payment_USDollars" (amount of money received) data. We then used the websites data export tool to create a CSV file of the resulting filtered data. The CSV file created included following data variables:

Data Variable	**Definition**
Recepient_State	The state where payments are made
Applicable_Manufacturer_or_Applicable_GPO_Making_Payment_Name	The pharma company that made the payment
Physician_Profile_ID*	The COUNT of unique physicians that have received a payment from said pharma company in said state
Total_Amount_of_Payment_USDollars	The SUM of money received by the physicians by said pharma company in said state

* Lines of data with a Physician_Profile_ID of "0" indicated that a physician profile ID was not present for a specific monetary contribution in the data set, since the contribution was made to a hospital instead of an individual physician. Given that our report visualization shows the amount of money given to individual physicians, not hospitals, we created a line of code in our visualization that sets the radius of the associated circle to 0 so that circles are not generated for money given to hospitals.

Existing CSV to JSON converters converted Physician_Profile_ID and Total_Amount_of_Payment_USDollars into strings. As such, we developed and used a separate code to convert the CSV data into a JSON string using the JSON.stringify function. We converted the mentioned variables into JSON numbers instead of strings. We then copied and pasted the generated JSON data into a JSON file, which we used in our project.

We created two visualizations. The first included all states and all pharma contributions to individual physicians. In the second visualization, we only showed the top 5 states by population according to the US Census population estimate for 2014.
(source from Wikipedia:
http://en.wikipedia.org/wiki/List_of_U.S._states_and_territories_by_population).
We also only showed monetary payments of more than $250,000 dollars in total amount made to less than 500 physicians.

B. Data Mapping to Visual Elements
Directions for this section: A description of the mapping from data to visual elements. Describe the scales you used, such as position, color, or shape. Mention any transformations you performed, such as log scales. (10 pts)

For each state where money received, we created circles of with radiuses represented by the “Total_Amount_of_Payment_USDollars” spent variable. Each circle represents a pharma company, with the circle’s size representing the amount spent on physicians in that state. We arranged the circles on the x axis according to their respective “Physician_Profile_ID” variable, which shows the number of physician the money was spent on by the company in that state.

We used x-scale to represent the number of doctors/physicians being paid by each company. The scale is linear and maps number of doctors from 0 to the max doctor number to range [150,1170] on the canvas. So the y axis was translated to point(150,0).
We used y-scale to represent different states where money or donation received from different companies. The scale is linear and maps states of 1-50 to range [1900, 110]. So the x axis was translated to point(0,1900).
We use colors and radius to represent the amount of money paid by the company.
The color scale is linear and maps to 3 colors: yellow, pink and purple.
The radius scale is square root and map the amount of money to range of [0,200].

C. The Data Story
Directions for this section: The story. What does your visualization tell us? What was surprising about it? (5 pts)

The visualization shows us the amount of money that pharmaceutical companies spent on groups of physicians by state. What was surprising about the visualization is that the pharma company Genentech, Inc. spent over one hundred million dollars on a few hundred physicians in CA over the course of five months. Given that law requiring pharma companies to submit payments data is fairly new, and that the end of 2013 is the first reporting cycle, it is possible that such a large amount could be the result of data quality issues that occurred in the reporting process.

Also, there is an intriguing story involving the type of trend that companies follow when donating to different doctors. The x-scale range on our largest graph can be deceiving, but the numbers on the bottom signify a range of thousands of doctors, and it appears that company donations focus on both a large pool of doctors, and other times a significant few; that is, not specifically one or the other.

There isn't as strong of evidence of corruption (large amounts of money to only one-two doctors) as we had hoped, in the sense that we expected a large amount of the large circles to be focused on the y axis, and not spread out along the x axis as they were. Although there are still a significant amount of payments to a few concentrated group of doctors, there are also many payments that are given to a multitude of doctors at once. Although our data cannot tell everything, we think we handled our incredibly large data sample size very well, and it definitely corrected our views about pharmaceutical donations.