Kevin Garner
Srinivas Havanur
Prasanna Sajjan
CS 595
Professor Cartledge

Project Questions

1. **What is the question?**

   What is the correlation between the average Medicare billings for the "Cardiovascular stress test" procedure by address and the total cost of pharmaceutical payments made to each address?

2. **Why is it important?**

   It is important because the professor wants to know what factors might affect different costs of her medical procedure throughout the country. She feels that some pharmaceutical companies pay the physicians to use their products with some procedures that have the highest profit margin. She thinks there is a correlation between the average Medicare billings of her procedure "Cardiovascular stress test" with that of the payments made by the pharmaceutical companies to the physicians' offices. Hence she is interested to find out the correlation between the average cost of all Medicare billings made for cardiovascular stress test procedures and the total of all payments made to a physicians' office.

3. **What have others done to try and solve the question?**

   We found a similar case where attorneys were interested in finding out whether or not pharmaceutical companies were paying doctors to use their drugs for specific procedures. Attorneys took the Medicare data provided by the CMS (Center for Medicare and Medicaid Services) and attempted to find the correlation between it and pharmaceutical company payments (provided by the CMS under the Sunshine Act) to detect fraud among doctors receiving payments.

   For example, the Medicare database provides the information pertaining to the number of times the doctor prescribed a particular drug in a year. The sunshine database provides the information about the drug manufacturer paying the prescribing doctor a certain amount in meals or other transfers of value during that year. The attorneys used this information to analyze whether pharmaceutical companies providing kickbacks to a doctor effected that doctor's prescribing practices or not, which is essentially what the English professor was concerned with in our project in terms of the Cardiovascular Stress Test.

Reference:

Sullivan, T. (2014, April 29). *Qui Tam lawyers mining the medicare database for potential fraud*. Retrieved from http://www.policymed.com/2014/04/qui-tam-lawyers-mining-the-medicare-database-for-potential-fraud.html

4. What will I do to solve the problem?

a) Login to fast.cs.odu.edu

b) Secure shell to the Hadoop server

c) Make sure these files are available in my working directory: project_update_93015.pig, project_update_93010.pig, makeFile.sh,project_count_93015.pig, project_count_93010.pig . The pig scripts will be made to use Pig Latin commands to perform necessary procedures to the Medicare and Pharmaceutical databases (such as normalization, filtering to find only the addresses that provide the specific procedure codes, and joining the two databases).

d) Run the makeFile.sh bash script. It will ask the user whether to extract the data with unique code 93015(Cardiovascular stress test) or to extract the data for unique code 93010(Electrocardiogram report). Further it will ask the user to input the directory name where output files will be stored.

e) Once the output file is generated for both the parts, I will use the output files to generate scatter plots and find the correlation between average Medicare billings for "Cardiovascular stress test" procedure by address and sum of pharmaceutical payments made to each address. In addition, the correlation value and p value is calculated after inserting the output into a program called "RGui." Its mathematical functions for correlation and p-value will be used for the calculations. Likewise, it will be done for unique code 93010. After finding these correlation values, I will analyze these values and the scatter plots to determine whether the professor's hypothesis of there being a correlation between the average Medicare billings of her procedure "Cardiovascular stress test" with that of the payments made by the pharmaceutical companies is valid or invalid.

f) The Pig Latin script will also keep track of and output the count of both Medicare and pharmaceutical data in the databases before the join and after the join to generate the Venn diagram for both the unique codes.

5. What will I do to prove that I have solved the problem?

I will extract only a small number of rows from each database into new files and test my script using these small databases. Before running my script, I will edit the rows to ensure that there are similar addresses and correct procedure codes to properly test my script. I will also add redundancies (such as white space, multiple abbreviations and full address names, and hashes) to ensure that my normalization procedures work correctly. Every possible scenario found in the original databases will be duplicated in these "practice" databases to ensure that my script handles all of these scenarios in the best and most efficient way possible. I will verify that the output generated is actually correct and consistent with the “practice” databases and that everything was calculated correctly and joined correctly. After doing all of this and finding a correlation with the real databases, I will repeat the procedure using a different CPT code and compare the results to see if there is a similar correlation between both results or if they are completely different. If they are similar or the same, then my conclusion for the first code would be valid. If they are not similar, then I would go through my script again and re-check how I go about filtering the data, normalizing it, and joining the databases because there would be some problem somewhere in my procedure that I would need to alter to make my data analysis more efficient.

6. What is the conclusion?

We found the correlation value to be 0.003328825 for the CPT code 93015. This value is very close to 0. From this, we can conclude that both the Medicare billings and the pharmaceutical costs are not closely related to each other. There is a huge difference in the cost between these two and hence the correlation value is almost 0. The scatterplot also did not show any relevant correlation. The Venn diagram shows how the number of records in each data set are on par with each other and how many common records exist in between these two. This is further verified by the correlation 0.01871026 which was calculated for the CPT code 93010. This value is also close to 0, showing no real correlation between average Medicare Billings and Pharmaceutical payments. The professor’s hypothesis is invalid. Pharmaceutical companies don’t necessarily pay physicians to use their products with procedures that have the highest profit margin.