**HW3: Insurance Claims Data & Analytics**

**Due: 12:00 pm, Monday Feb 25**

**Context:**

This homework lays out some analytics on the most important database in market for health, the insurance claim data. We use the publicly available All Payer Claim Data published by the State of Vermont. The databases available to us are named Vermont Uniform Hospital Discharge Data Sets. These databases consist of inpatient discharge data, outpatient procedures and services data, and emergency department data. Each data set according to the website includes:

* Case-specific diagnostic discharge data
* Some socio-demographic characteristics of the patient
* Medical reason for the admission
* Treatment and services provided to the patient
* Duration and status of the patient's stay in the hospital
* Full, undiscounted total and service-specific charges billed by the hospital

They are found in the link provided below. We want to use 2016 databases and codebooks since the 2017 upload is still too fresh and probably subject to revision.

<http://www.healthvermont.gov/health-statistics-vital-records/health-care-systems-reporting/hospital-discharge-data>

**Important reminder: Use 2016 data since the 2017 is not finalized yet.**

**Question 1: Clustering Costs – data used: Inp and RevCode**

In this question we will do a simple cluster analysis. Clustering is a big part of AI and for the most part the technique follows the same basic ideas as explained in this question. We plan to look into clusters of certain costs categories of the inpatient hospital DRGs and try to interpret the clusters or “to make sense of them” so to speak. Understanding and making sense of the clusters is an art and could be mastered only by training and of course having a rich domain knowledge because there is no science to tell us how to understand the properties of the clusters that a mathematical model has created for us.

Inpatient hospitalizations are identified and priced and paid by DRGs. So each hospital admission has a DRG. You can see the list of DRGs in the Excel file “*HS\_2016VT\_PUF\_FILE\_LAYOUT\_and\_CODES*” from the website. While the inpatient file provides you with a single summary dollar value for hospital charges, you can see all the details of those charges in the Revenue Code file VTREVCODE16. In the Revenue file there are more than one standard to classify the detailed services. The most common coding system is Revenue Code which is shown the variable RevCode in the file. There is another coding system that is known as Primary Cost Center or PCCR which tells you the department from which the bill is initiated. You do have the PCCR lists in the Excel file HS\_2016VT\_REVCODE\_FILE\_LAYOUT\_and\_CODES.

Start from the Inpatient file VTINP16. Filter your hospital admissions to only DRGs between 20 and 977. We do not need other DRGs for this analysis. Then you link your filtered Inpatient data to the Revenue Code file using the UNIQ variable. In your Revenue file exclude the low dollar value services (less than $100) by dropping the REVCHRGS <100. Then sum all the charges by the PCCR category because some hospital admissions may have multiple charges submitted from the same PCCR department so you want to add them all. (In your SQL you can use something like this syntax: Select UNIQ, DRG, PCCR, Sum(REVCHRGS) Group By UNIQ, DRG, PCCR From myIP\_RevCode\_Merged\_Table).

Now that you have summed together all dollars for PCCR per hospital admission you need to cross tabulate your list of selected DRGs (in the row) and the mean value of the PCCRs. This cross-tabulation will have 687 rows (one per DRG) and 54 columns, one column per PCCR for all your PCCRs

For this exercise we want to focus on the two specific PCCR cost centers:

**PCCR 3700 Operating Room**

**PCCR 4000 Anesthesiology**

Now create a new cost category by combining the **PCCR 3700 Operating Room + PCCR 4000 Anesthesiology.** Name this new cost category as **PCCR\_OR\_and\_Anesth\_Costs**. You now have 54+1 PCCR categories as column headings.

While your column headings can have the PCCR code in them you want to plug in the PCCR real name to your Revenue File and use the PCCR names instead otherwise your cross-tabulation column headings are a bunch of PCCR numbers and that file would not serve you well in moving to cluster analysis. So you want your column heading to be for example “Physical Therapy” instead of PCCR 5000 or Ultra Sound instead of PCCR 3630. Also make sure you plug in the DRG names to your cross-tabulated file so you will have the RDG number and the name both as your row headings. So far your x-tab will have DRG names as the rowheadings, the name of all + 1 PCCRs as the column heading, and the average costs ($) of each PCCR category as the cell values.

You will soon notice that many cells in your x-tab are empty. This is because for example pregnancy DRG does not use any Nuclear Medicine-Diagnostic service so that cell will be empty. In order to avoid any problem with your cluster analysis software you want to preferably turn all those empty cells to zero dollars.

With the x-tab cleaned and formatted neatly you are ready to move to the cluster analysis. We want to find meaningful clusters of DRGs based on the dollar charges of our newly created variable PCCR\_OR\_and\_Anesth\_Costs. So we have 687 average costs in dollars and want to use them to create meaningful clusters. Among many methods of clustering that your software can perform use k-Means algorithm.

Obviously you can ask your software to identify any number of clusters in your data. The optimal number of clusters could be done by tracing the value of the Calinski-Harabasz f-statistics. So ask your software to report the Calinski-Harabasz f-statistics every time you run your cluster analysis. Try to cluster your cost data into 2,3,4, and 5 clusters and examine the f-stat each time to see where you have the best clustering of costs (i.e. where your f-stat is the highest). Take a note of your f-stats for this section however in the next section we only focus on solutions with 3 clusters.

Now consider your solution with only 3 clusters. Can you graph your costs and somehow color code your cost points to reflect the 3 clusters? You can use any tool for graphing your clusters but I just used Excel and did it for 4 clusters just as an example. You can get some ideas as how to do with 3 clusters. Or you can use any other methods as long as you can visually differentiate between your own 3 clusters. I am sure you will do prettier than mine ☺

**Interpretation and making sense of your clusters:**

This is where the things get exciting, sort of, or maybe not! First a bit of bad news ☹ more often than not machine suggested clusters are not intuitive enough to understand. Here are some typical reactions from others who have tried the AI: I wonder why these X number of hospital DRGs are clustered together and the other Y hospital DRGs in another clusters and the other Z number of them in other and the rest in a 4th cluster? Why did the computer decide to group them this way and why not in another way? It looks stupid and I cannot easily see any similarity/pattern within clusters and dissimilarity across the clusters. I do not like what I see and cannot wrap my mind around these nonsense clusters. I hate this whole AI thing since it does not make any sense to me! This is in particular the case when you want to cluster the data based on more than one variable. Here I simplified the world for you to cluster your DRGs using only and only one single cost variable. I could have asked for more than one variable to ruin your weekend!

So at this point your entire team need to use its rich domain knowledge and come up with some good theories as what are the common properties of the hospital inpatient DRGs of each cluster that make them relatively similar or closely related to one another, and on the other hand, relatively dissimilar to other clusters’ DRGs. And beware not to make the very common mistake that most people do: do NOT focus on the costs and their magnitude within and between clusters more than what you have to, rather spend most of your brain on understanding the common properties of the DRGs. You are clustering and classifying inpatient DRG admissions and not costs. You only use certain cost categories to help you make a meaningful classification of the hospital DRGs so stay focus on studying and exploring the DRGs based on many properties that are available to you through the Excel code lists and other resources.

So in your team class presentation spend most of your time and effort on a good interpretation and characterization of the DRG clusters since rest of the technical stuff are all easily done by your laptop! Human brain and thought process is vital to make sense of AI results, and that is a very important point to consider.

Graph 1: An example of graphic presentation of the clusters (4 clusters here)