**Healthcare Data Analytics and Data Mining**

**Assignment 1: Analyzing the database “National Plan and Provider Enumeration System (NPPES)”**

In this assignment we mine and do analytics on a large publicly available database. The National Plan and Provider Enumeration System (NPPES) database reports the status of the US healthcare providers. This is an example dataset for the “Supply/Provider” side of the market in our triangular model. It is updated periodically to reflect the most recent status of healthcare providers in terms of address, nature of practice, credentials, organizational relationship, licensure status and so on. The database comes as part of a full downloadable package where you will also find the supplementary documents on the variables’ definition, version control, and other useful information. The database is maintained by a US federal agency named Centers for Medicare and Medicaid Services CMS. The address of the data warehouse is:

[http://download.cms.gov/nppes/NPI\_Files.html](http://download.cms.gov/nppes/NPI_Files.html%20%20)

You need to download the **Full Replacement Monthly NPI File** at the link NPPES Data Dissemination (October 15, 2019) - ZIP format (706 MB). The zipped package is 706 MB and when unzipped the database file named “npidata\_pfile\_20050523-20191013.csv” is 7.2 GB which is obviously a very large file. You will have to use a proper platform for super large databases or a SQL engine of your choice that can handle a database this size. The first 1,000 rows of the data are imported into an Excel file named npidata\_pfile\_20050523-20191013\_1000Rows.xlsx on LATTE so you can get a sense of the file structure and variables as you write your codes to import the file.

**♫♪♫►Hint:** The data set npidata\_pfile\_20050523-20191013.csv has close to 5.9 Million rows and 329 variables for QC of your download before you start mining! Also note that this dataset has many variable names that are repeated with sequential numbers such as “*Provider License Number\_1*” thru “*Provider License Number\_15*”. These variables are there to let providers with multiple answers having enough columns to feel in their data in the national database. Of course a big number of those repeated variables are empty for most of the providers that have only one or a few items to report to the national database. So when you read this homework always double-check to make sure you are picking the right variable name with right variable sequential number.

This warm-up practice is only meant to guide you through some basic analytics on the healthcare provider profile database. We will use an extended version of this database with more data content for each provider in future case studies, so this simpler database should help you laying the groundwork for future undertaking.

**♫♪♫►Hint:** As the data expert members of the team are having fun wrapping their heads around the monster database, the rest of the team can plan almost all analysis on the Excel mini version of the data which is a true mini copy of the real giant file.

There are some questions in this homework that apply to entire database (nationwide analytics) while some other questions pertain to only your specific set of US states according to your team number. Appendix 1 lists the assignment of US states to each submission team. For instance where the question asks for team specific set of states, the submission team 2 will have to filter the national database for only 8 team-specific states listed below and then run the analysis on only 8 states. To filter the national database for your submission team specific list found in Appendix 1, please use the variable “*Provider Business Practice Location Address State Name*” and filter the national file for your own team specific state subset.

This data segmentation in each homework will insure that at least for some of the homework questions, different teams are working on different subsets of a larger national database and hence the answers are different, and in turn fairness in rewarding the original work is achieved ☺

|  |  |  |
| --- | --- | --- |
| **Submission Team 2** | FL | Florida |
| **Submission Team 2** | GA | Georgia |
| **Submission Team 2** | AZ | Arizona |
| **Submission Team 2** | CO | Colorado |
| **Submission Team 2** | CT | Connecticut |
| **Submission Team 2** | MS | Mississippi |
| **Submission Team 2** | HI | Hawaii |
| **Submission Team 2** | RI | Rhode Island |

**♫♪♫►Hint:** This database profiles both individual/person providers and healthcare facilities and clinics. Please use the filter on the database after you have downloaded to limit your analytics on only “individual or person providers” using *“Entity Type Code”* =1 where the homework question is talking about individual providers and use *“Entity Type Code”* =2 when the subject of the question is a healthcare facility and not an individual.

**Question 1—Find your own doctor!**

Please find your own healthcare provider (PCP or GP) and report the state in which she/he was first licensed using the variable “*Provider License Number State Code\_1*” . For a team of 6 students, list all six 6 answers, one per student, ordered by students’ last name. If you do not have a doctor, find one for a close family member or friend.

Important privacy hint: To protect your privacy **PLEASE DO NOT** reveal or report anything else besides the state about your doctor in your homework submission!

**Question 2—Only for the states assigned to your submission team:**

Run a statistical test exploring gender difference in practicing as a “*Sole Proprietor*”.

**♫♪♫►Hint:** Run 2x2 cross-tab for male/female for Gender by Yes/No for Sole Proprietor.

To report the statistical significance on your x-tabulation please use the Fisher’s Exact Test for a 2x2 table to test if the gender preference differences that you observe is significant or not (p-value <0.05 is significant). Note: depending on your programming languages some may conduct the Fisher’s Exact test under the Chi-Sqr test category. **Clean up notes:** Please remove any value besides Male/Female M/F from the gender variable and any value besides Sole Proprietor Y/N. Only then you will have a 2x2 x-tab. After all, do you see females being less or more likely to establish a solo practice office?

**Question 3—Only for the states assigned to your submission team:**

Please test this hypothesis: male doctors are more likely than their female peers to choose the practices that are associated with higher risk for a higher reward.

To test the hypothesis use the variable *Healthcare Provider Taxonomy Code\_1*. The codebook “*Health Care Provider Taxonomy Code Set*” is available for public use in the website:

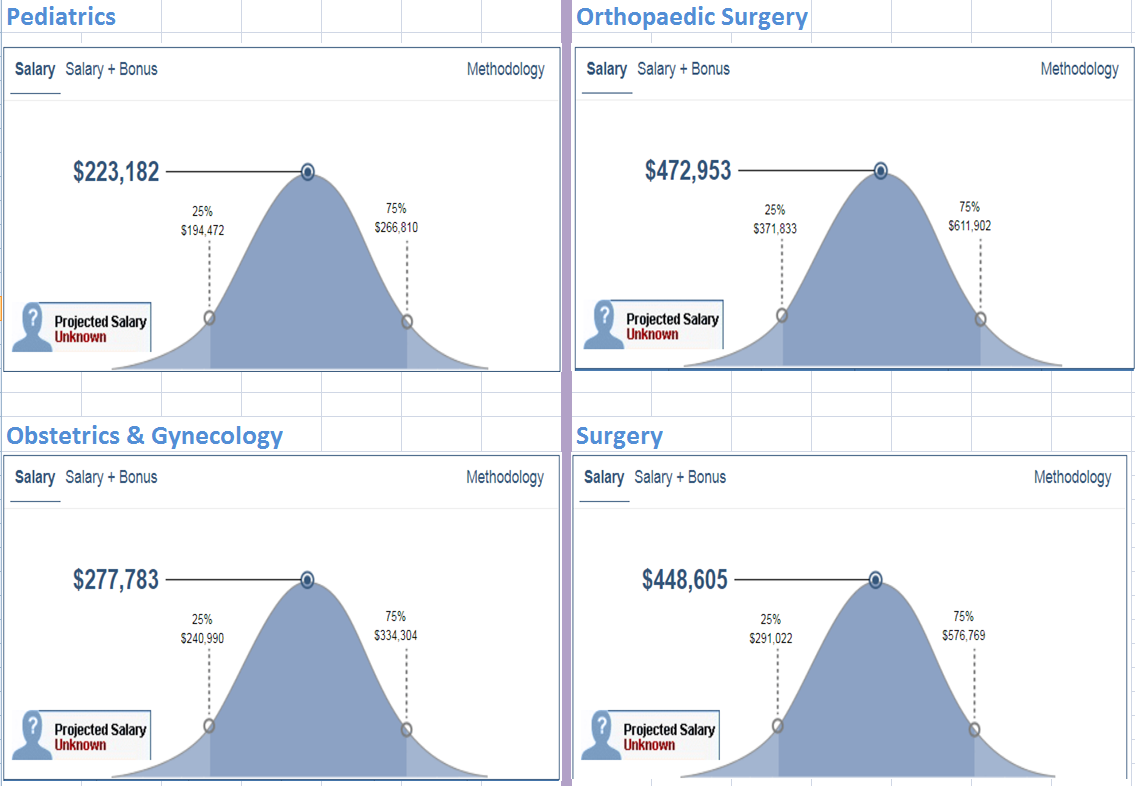
<http://www.wpc-edi.com/reference/>

For this question please cross-tabulate male and female providers (drop other gender values) by low and high risk-reward categories. Let your low risk/reward category to include “Obstetrics & Gynecology” and “Pediatrics” and your high risk/reward include “Surgery” and “Orthopaedic Surgery”. Example: in the *Health Care Provider Taxonomy Code Set* the taxonomy code for Obstetrics & Gynecology is **207V00000X**. You need to find the other 3 codes there too. Note that once again you need 2x2 cross-tab (2 practice categories X 2 genders) followed by a Fisher’s Exact Test with p-value.

Just FYI, I did high/low reward classification using this website (graphs below):

<https://swz.salary.com/SalaryWizard/Pediatric-Physician-Salary-Details-Boston-MA.aspx>

Your Category 1 (High reward) appears in right and Category 2 (low reward) in left.



**Question 4: National heat map of MRI centers – Applies to all states – Applies to healthcare facilities and not individual providers**

Go back to your original national database download and use “*Entity Type Code*” =2. We are now working on healthcare facilities and not the individual providers. From the taxonomy code set you can see that the Magnetic Resonance Imaging MRI centers are identified by the taxonomy code 261QM1200X*.* Use “*Healthcare Provider Taxonomy Code\_1*” variable in your database to find the MRI centers. You need to calculate the density of MRI centers by state population. You will first need to calculate the number of MRI clinics per 1000,000 population for each state. For population statistics please use any reliable source on the web that reports the US population by states. Once you have calculated the MRI density per 1000,000 population, draw the heat map for the 50 states using any software of your choice. Any idea as why the density of MRI per population in Florida is so high? (Bear in mind that we do not know how many MRI machines is in each facility so our speculation about Florida may not be valid enough w/o detailed information)

**Submission Format:** Your homework submission needs to be a well written, well edited, well organized, technical report. Given the relatively generous submission team sizes in this class you will have to plan for division of both analytical tasks, and the reporting tasks. The quality of final report therefore is part of your final grade insuring your technical report writing skills are also improving in this module course, which is an extremely important qualification in a rather competitive job market!

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

**Appendix 1:** Assignment of US states to each submission team

|  |  |  |
| --- | --- | --- |
| **Submission Team Number** | State Code | State Name |
| **1** | TX | Texas |
| **1** | NC | North Carolina |
| **1** | IN | Indiana |
| **1** | MD | Maryland |
| **1** | LA | Louisiana |
| **1** | UT | Utah |
| **1** | NE | Nebraska |
| **1** | MT | Montana |
| **1** | VT | Vermont |
| **2** | FL | Florida |
| **2** | GA | Georgia |
| **2** | AZ | Arizona |
| **2** | CO | Colorado |
| **2** | CT | Connecticut |
| **2** | MS | Mississippi |
| **2** | HI | Hawaii |
| **2** | RI | Rhode Island |
| **3** | CA | California |
| **3** | OH | Ohio |
| **3** | TN | Tennessee |
| **3** | WI | Wisconsin |
| **3** | KY | Kentucky |
| **3** | IA | Iowa |
| **3** | NM | New Mexico |
| **3** | ME | Maine |
| **3** | WY | Wyoming |
| **4** | NY | New York |
| **4** | NJ | New Jersey |
| **4** | MA | Massachusetts |
| **4** | AL | Alabama |
| **4** | OK | Oklahoma |
| **4** | KS | Kansas |
| **4** | ID | Idaho |
| **4** | DE | Delaware |
| **4** | DC | District of Columbia |
| **5** | PA | Pennsylvania |
| **5** | VA | Virginia |
| **5** | MO | Missouri |
| **5** | MN | Minnesota |
| **5** | OR | Oregon |
| **5** | AR | Arkansas |
| **5** | WV | West Virginia |
| **5** | SD | South Dakota |
| **6** | IL | Illinois |
| **6** | MI | Michigan |
| **6** | WA | Washington |
| **6** | SC | South Carolina |
| **6** | PR | Puerto Rico |
| **6** | NV | Nevada |
| **6** | NH | New Hampshire |
| **6** | ND | North Dakota |