

HEALTH CARE DATA ANALYTICS AND DATA MINING

Group 1



January 24, 2025

GROUP 1

Doreen Accorley, Peter Mukiibi, Biniyam Abebe, Yash Puri

**Introduction**

The National Plan and Provider Enumeration System (NPPES) is a comprehensive registry of individual and institutional healthcare providers, along with their credentials. Maintained by the Center for Medicare and Medicaid Services (CMS) under the Department of Health and Human Services (HHS), this extensive database offers valuable insights into various aspects of healthcare providers and their services across the United States. As one of the most detailed resources on healthcare providers, NPPES holds significant potential to reveal trends that can shape healthcare policy and decision-making.

This project focuses on utilizing advanced data analytics techniques to explore the NPPES dataset and uncover meaningful patterns and correlations within the healthcare provider landscape. By delving into this complex dataset, the aim is to generate insights that can enhance the understanding of the provider network within the healthcare industry. The final report is a collaborative effort by the group to analyze the data and address the questions outlined in the assignment.

**TOOLS USED**

* SAS Enterprise Guide
* Excel
* Tableau

**OBJECTIVES:**

1. Leverage statistical knowledge and tools to explore and analyze the NPPES repository, addressing key questions related to healthcare providers.
2. Examine the techniques and processes of data mining, cleaning, and analysis, using the NPPES repository as a practical application.

Q1.

|  |  |  |
| --- | --- | --- |
| **Student’s First Name** | **Student’s Last Name** | **State where the doctor was first licensed** |
| Biniyam | Abebe | DC |
| Doreen | Accorley | New York |
| Peter | Mukiibi | Texas |
| Yash | Puri | Massachusetts |

In our analysis we use Sas Enterprise Guide we selected Provider Last Name (Legal Name), Provider First Name and Provider License Number state co (label : Provider License Number State code\_1). We filter by first name and last name to filter.

Q2. Table Analysis

Results

The FREQ Procedure

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  | | --- | | **Frequency**  **Percent**  **Row Pct**  **Col Pct**  **Cumulative Col%** | | |  |  |  |  | | --- | --- | --- | --- | | Table of Provider Gender Code by Is Sole Proprietor | | | | | Provider Gender Code | **Is Sole Proprietor** | | | | **N** | **Y** | **Total** | | F | 777095  45.79  66.75  69.20  69.20 | 387134  22.81  33.25  67.45  67.45 | 1164229  68.61      68.61 | | M | 345848  20.38  64.92  30.80  100.00 | 186845  11.01  35.08  32.55  100.00 | 532693  31.39      100.00 | | Total | 1122943  66.18 | 573979  33.82 | 1696922  100.00 | |

Statistics for Table of Provider Gender Code by Is Sole Proprietor

| **Statistic** | | **DF** | **Value** | | **Prob** |
| --- | --- | --- | --- | --- | --- |
| **Chi-Square** | | 1 | 542.7150 | | <.0001 |
| **Likelihood Ratio Chi-Square** | | 1 | 541.0097 | | <.0001 |
| **Continuity Adj. Chi-Square** | | 1 | 542.6336 | | <.0001 |
| **Mantel-Haenszel Chi-Square** | | 1 | 542.7147 | | <.0001 |
| **Phi Coefficient** | |  | 0.0179 | |  |
| **Contingency Coefficient** | |  | 0.0179 | |  |
| **Cramer&apos;s V** | |  | 0.0179 | |  |
| **Pearson Chi-Square Test** | | | |
| **Chi-Square** | 542.7150 | | |
| **DF** | 1 | | |
| **Asymptotic Pr > ChiSq** | <.0001 | | |
| **Exact Pr >= ChiSq** | <.0001 | | |

| **Likelihood Ratio Chi-Square Test** | |
| --- | --- |
| **Chi-Square** | 541.0097 |
| **DF** | 1 |
| **Asymptotic Pr > ChiSq** | <.0001 |
| **Exact Pr >= ChiSq** | <.0001 |
| **Mantel-Haenszel Chi-Square Test** | |
| **Chi-Square** | 542.7147 |
| **DF** | 1 |
| **Asymptotic Pr > ChiSq** | <.0001 |
| **Exact Pr >= ChiSq** | <.0001 |

Sample Size = 1696922

Based on the 2x2 cross-tabulation of gender (Male/Female) and sole proprietor status (Yes/No), the Chi-Square test reveals a substantial correlation between gender and the likelihood of being a sole proprietor (χ² = 542.7150, p < 0.0001). Notably, females constitute a larger percentage of sole proprietors (33.25%) compared to males (35.08%). The column percentages from the cross-tabulation further demonstrate that 67.45% of sole proprietors are female, while 32.55% are male. The Chi-Square test (χ² = 542.7150, p < 0.0001) corroborates that this disparity is statistically significant. This finding suggests that females are more probable than males to engage in sole proprietorship. While the observed difference is substantial, the effect size (Cramér’s V = 0.0179) indicates that the association is relatively weak. Consequently, while gender plays a role, it is not a strong determinant of sole proprietor status.

Q3. Table Analysis

Results

The FREQ Procedure

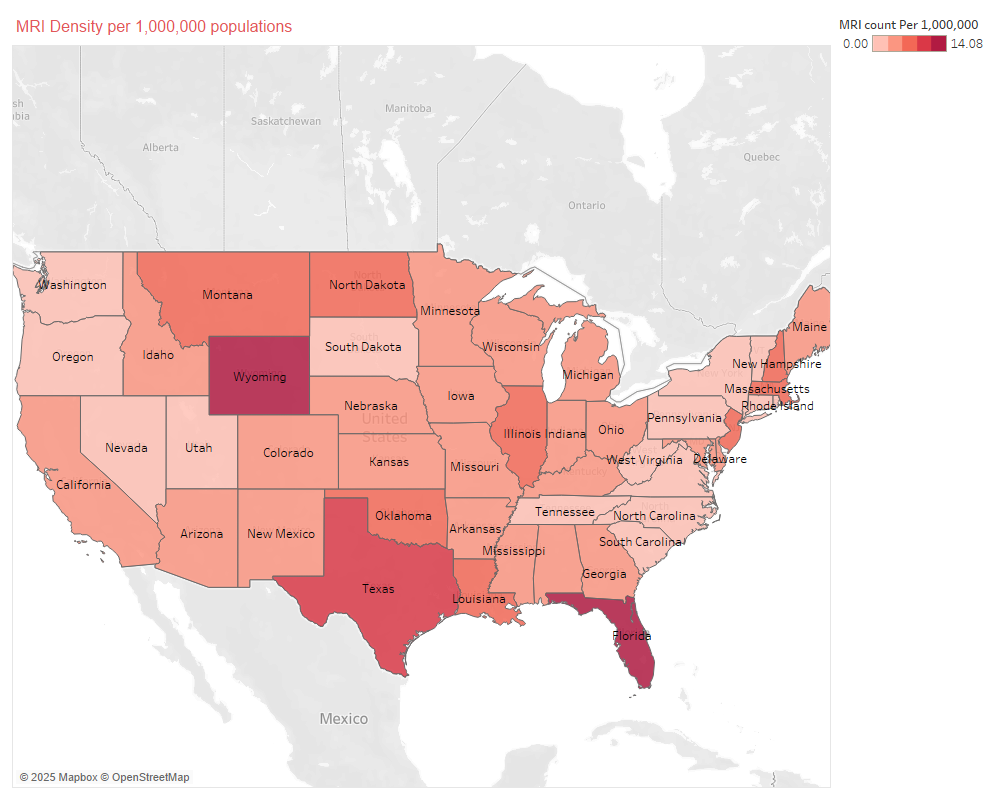
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  | | --- | | **Frequency**  **Percent**  **Row Pct**  **Col Pct**  **Cumulative Col%** | | | **Table of Provider Gender Code by Risk\_Category** | | | | | --- | --- | --- | --- | | **Provider Gender Code** | **Risk\_Category** | | | | **High Risk** | **Low Risk** | **Total** | | **F** | 2465  7.70  16.26  19.07  19.07 | 12692  39.63  83.74  66.44  66.44 | 15157  47.32      47.32 | | **M** | 10460  32.66  62.00  80.93  100.00 | 6412  20.02  38.00  33.56  100.00 | 16872  52.68      100.00 | | **Total** | 12925  40.35 | 19104  59.65 | 32029  100.00 | |

Statistics for Table of Provider Gender Code by Risk\_Category

| **Statistic** | | **DF** | **Value** | | **Prob** |
| --- | --- | --- | --- | --- | --- |
| **Chi-Square** | | 1 | 6937.9234 | | <.0001 |
| **Likelihood Ratio Chi-Square** | | 1 | 7333.6275 | |  |
| **Mantel-Haenszel Chi-Square** | | <.0001 | 6937.7068 | | <.0001 |
| **Phi Coefficient** | |  | -0.4654 | |  |
| **Contingency Coefficient** | |  | 0.4220 | |  |
| **Chi-Square** | 6937.9234 | | |
| **DF** | 1 | | |
| **Asymptotic Pr > ChiSq** | <.0001 | | |
| **Exact Pr >= ChiSq** |  | | |

Sample Size = 32029

We conducted a 2x2 cross-tabulation to analyze the gender distribution among high-risk and low-risk providers in relation to their risk-reward categories. The results revealed a significant association between gender and practice risk-reward category. Specifically, 80.93% of high-risk providers are male, compared to 19.07% female, while 66.44% of low-risk providers are female and 33.56% male. The Chi-Square test (χ² = 6937.9234, p < 0.0001) demonstrated a highly significant relationship between gender and practice risk-reward category. Moreover, the negative Phi coefficient (-0.4654) further confirmed the inverse relationship between female gender and high-risk practices. These findings strongly support the hypothesis that female gender is associated with low-risk practices.

Q4. 

The heatmap depicts the concentration of MRI centers per 1,000,000 population across the United States. It was calculated using the taxonomy code 261QM1200X for MRI facilities. The analysis reveals significant variations in MRI center density. Florida stands out with the highest density, significantly surpassing other states. Wyoming, despite its smaller population, also exhibits a high density, possibly due to centralized healthcare resources serving broader rural areas. Conversely, states like Idaho and Alaska have notably low densities. Florida’s elevated density could be attributed to its large population of retirees and elderly residents, who necessitate more diagnostic imaging services, and its extensive healthcare infrastructure accommodating seasonal visitors and residents.

Q5. 

The chart illustrates concentration of plastic surgery centers per 100,000 population across U.S. states. Washington, D.C., Rhode Island, and New York lead the nation in the demand for cosmetic surgeries, closely followed by California, Utah, and Maryland. Several factors contribute to this demand. Firstly, these states boast urban populations with higher disposable incomes, creating a market for cosmetic procedures. Secondly, social pressures and cultural acceptance of beauty standards play a significant role in driving the demand. Moreover, D.C.’s high concentration of plastic surgery centers could be attributed to the professional environment where appearances may hold greater importance. In contrast, states like California and New York are driven by the entertainment and fashion industries, which prioritize beauty standards. Lastly, the absence of insurance coverage for most cosmetic procedures suggests that these demand-heavy states likely have affluent populations willing to pay out-of-pocket for such services.

**ANNEX**

**SAS ENTERPRISE GUIDE CODE AND STEP BY STEP ANSWER**

  Source Column: t1.Provider First Name

Column Name: Provider First Name

Type:  Basic

Filter:

   WHERE t1.'Provider First Name'n = 'SEIFE'

AND

Source Column: t1.Provider Last Name (Legal Name)

Column Name: Provider Last Name (Legal Name)

Type:  Basic

Filter:

   WHERE t1.'Provider Last Name (Legal Name)'n = 'YOHANNES'

OR

Source Column: t1.Provider First Name

Column Name: Provider First Name

Type:  Basic

Filter:

   WHERE t1.'Provider First Name'n = 'LYNNE'

AND

Source Column: t1.Provider Last Name (Legal Name)

Column Name: Provider Last Name (Legal Name)

Type:  Basic

Filter:

   WHERE t1.'Provider Last Name (Legal Name)'n = 'LAZARUS'

OR

Source Column: t1.Provider First Name

Column Name: Provider First Name

Type:  Basic

Filter:

   WHERE t1.'Provider First Name'n = 'KRISTINA'

AND

Source Column: t1.Provider Last Name (Legal Name)

Column Name: Provider Last Name (Legal Name)

Type:  Basic

Filter:

   WHERE t1.'Provider Last Name (Legal Name)'n = 'ORIO'

OR

Source Column: t1.Provider First Name

Column Name: Provider First Name

Type:  Basic

Filter:

   WHERE t1.'Provider First Name'n = 'TRACY'

AND

Source Column: t1.Provider Last Name (Legal Name)

Column Name: Provider Last Name (Legal Name)

Type:  Basic

Filter:

   WHERE t1.'Provider Last Name (Legal Name)'n = 'HARADON'

2.

FIRST SELECTED BY OUR GIVEN STATE(QUIRY ONE)

Source Column: t1.Provider License Number State Co

Column Name: Provider License Number State Co

Type:  Basic

Filter:

   WHERE t1.'Provider License Number State Co'n = 'CA'

OR

Source Column: t1.Provider License Number State Co

Column Name: Provider License Number State Co

Type:  Basic

Filter:

   WHERE t1.'Provider License Number State Co'n = 'DC'

OR

Source Column: t1.Provider License Number State Co

Column Name: Provider License Number State Co

Type:  Basic

Filter:

   WHERE t1.'Provider License Number State Co'n = 'IA'

OR

Source Column: t1.Provider License Number State Co

Column Name: Provider License Number State Co

Type:  Basic

Filter:

   WHERE t1.'Provider License Number State Co'n = 'LA'

OR

Source Column: t1.Provider License Number State Co

Column Name: Provider License Number State Co

Type:  Basic

Filter:

   WHERE t1.'Provider License Number State Co'n = 'MD'

OR

Source Column: t1.Provider License Number State Co

Column Name: Provider License Number State Co

Type:  Basic

Filter:

   WHERE t1.'Provider License Number State Co'n = 'MT'

OR

Source Column: t1.Provider License Number State Co

Column Name: Provider License Number State Co

Type:  Basic

Filter:

   WHERE t1.'Provider License Number State Co'n = 'NE'

OR

Source Column: t1.Provider License Number State Co

Column Name: Provider License Number State Co

Type:  Basic

Filter:

   WHERE t1.'Provider License Number State Co'n = 'OH'

OR

Source Column: t1.Provider License Number State Co

Column Name: Provider License Number State Co

Type:  Basic

Filter:

   WHERE t1.'Provider License Number State Co'n = 'WA'

THEN BY GENDER(ONLY F AND M) QUIRY 2

   Source Column: t1.Provider Gender Code

Column Name: Provider Gender Code

Type:  Basic

Filter:

   WHERE t1.'Provider Gender Code'n = 'M'

THEN BY SOLE PRIORITY (Y/N ONLY) QUIRY THREE

Source Column: t1.Is Sole Proprietor

Column Name: Is Sole Proprietor

Type:  Basic

Filter:

   WHERE t1.'Is Sole Proprietor'n = 'Y'

OR

Source Column: t1.Is Sole Proprietor

Column Name: Is Sole Proprietor

Type:  Basic

Filter:

   WHERE t1.'Is Sole Proprietor'n = 'N'

QUESTION 3- SELECTED LICENSE NUMBER STATE CODE, HEALTHCARE PROVIDER TAXONOMY CODE, PROVIDER GENDER CODE

THEN FILTERED BY STATE

Source Column: t1.Provider License Number State Co

Column Name: Provider License Number State Co

Type:  Basic

Filter:

   WHERE t1.'Provider License Number State Co'n = 'CA'

OR

Source Column: t1.Provider License Number State Co

Column Name: Provider License Number State Co

Type:  Basic

Filter:

   WHERE t1.'Provider License Number State Co'n = 'DC'

OR

Source Column: t1.Provider License Number State Co

Column Name: Provider License Number State Co

Type:  Basic

Filter:

   WHERE t1.'Provider License Number State Co'n = 'IA'

                        OR

Source Column: t1.Provider License Number State Co

Column Name: Provider License Number State Co

Type:  Basic

Filter:

   WHERE t1.'Provider License Number State Co'n = 'LA'

  OR

Source Column: t1.Provider License Number State Co

Column Name: Provider License Number State Co

Type:  Basic

Filter:

   WHERE t1.'Provider License Number State Co'n = 'MD'

OR

Source Column: t1.Provider License Number State Co

Column Name: Provider License Number State Co

Type:  Basic

Filter:

   WHERE t1.'Provider License Number State Co'n = 'MT'

OR

            Source Column: t1.Provider License Number State Co

Column Name: Provider License Number State Co

Type:  Basic

Filter:

   WHERE t1.'Provider License Number State Co'n = 'NE'

OR

Source Column: t1.Provider License Number State Co

Column Name: Provider License Number State Co

Type:  Basic

Filter:

   WHERE t1.'Provider License Number State Co'n = 'OH'

    OR

Source Column: t1.Provider License Number State Co

Column Name: Provider License Number State Co

Type:  Basic

Filter:

   WHERE t1.'Provider License Number State Co'n = 'WA'

THEN FILERED BY HEALTHCARE PROVIDER TAXONOMY COD

Source Column: t1.Healthcare Provider Taxonomy Cod

Column Name: Healthcare Provider Taxonomy Cod

Type:  Basic

Filter:

   WHERE t1.'Healthcare Provider Taxonomy Cod'n = '208600000X'

OR

Source Column: t1.Healthcare Provider Taxonomy Cod

Column Name: Healthcare Provider Taxonomy Cod

Type:  Basic

Filter:

   WHERE t1.'Healthcare Provider Taxonomy Cod'n = '207X00000X'

OR

Source Column: t1.Healthcare Provider Taxonomy Cod

Column Name: Healthcare Provider Taxonomy Cod

Type:  Basic

Filter:

   WHERE t1.'Healthcare Provider Taxonomy Cod'n = '208000000X'

OR

Source Column: t1.Healthcare Provider Taxonomy Cod

Column Name: Healthcare Provider Taxonomy Cod

Type:  Basic

Filter:

   WHERE t1.'Healthcare Provider Taxonomy Cod'n = '207V00000X'

THEN FILTERED BY GENDER

Source Column: t1.Provider Gender Code

Column Name: Provider Gender Code

Type:  Basic

Filter:

   WHERE t1.'Provider Gender Code'n = 'M'

OR

Source Column: t1.Provider Gender Code

Column Name: Provider Gender Code

Type:  Basic

Filter:

   WHERE t1.'Provider Gender Code'n = 'F'

THEN WE USED COMPUTED COLUMNS  THEN NEW THEN IN ADVANCED EXPRESSION WE PUT CASE

    WHEN Healthcare\_Provider\_Taxonomy\_Cod IN ('207X00000X', '208600000X') THEN 'High Risk'

    WHEN Healthcare\_Provider\_Taxonomy\_Cod IN ('208000000X', '207V00000X') THEN 'Low Risk'

    ELSE 'Unknown'

END

 ALTERNATIVELY WE EXPORTED IN EXCEL WE CREATED A NEW COLUMN AND PUT =IF(OR(B2="208000000X", B2="207V00000X"), "Low Risk", IF(OR(B2="208600000X", B2="207X00000X"), "High Risk", "Unknown"))

THEN WE USED TABLE ANALYSIS UNDER DESCRIBE WE CHOOSE PROVIDER GENDER CODE AND RISK FACTOR FOR TABLE ANALYSIS,CELL STATISTICS WE SELECTED ROW PERCENTAGE, CELL FREQUENCY , COLUMN PERCENTAGE AND CUMULATIVE AND FROM TABLE STATISTICS WE SELECTED CHI SQUARE.

QUESTION 4.

SELECTED Entity Type Code, Healthcare Provider Taxonomy Cod AND Provider License Number State Co.

Source Column: t1.Entity Type Code

Column Name: Entity Type Code

Type:  Basic

Filter:

   WHERE t1.'Entity Type Code'n = 2

 THEN FILTERED BY

Source Column: t1.Healthcare Provider Taxonomy Cod

Type:  Basic

Filter:

   WHERE t1.'Healthcare Provider Taxonomy Cod'n = '261QM1200X'

THEN EXPORT TO EXCEL

number 5

Selected : **Healthcare Provider Taxonomy Code\_1and : Provider License Number State**

Source Column: t1.Healthcare Provider Taxonomy Cod

Column Name: Healthcare Provider Taxonomy Cod

Type:  Basic

Filter:

   WHERE t1.'Healthcare Provider Taxonomy Cod'n = '208200000X'