Analytics Tech Mid-Term Project

# Overview

The opioid epidemic in the United States has been a growing problem for many years, with devastating consequences for individuals and communities. The issue has also placed significant strain on the healthcare system, particularly Medicare, which provides coverage for millions of older adults and people with disabilities. The widespread use and abuse of prescription painkillers and illegal opioids has resulted in a surge in opioid-related hospitalizations, overdoses, and deaths. Medicare has been grappling with how to address this public health crisis and ensure that patients receive effective and appropriate pain management while also reducing the risk of opioid misuse and addiction.

The knee-jerk response to curb the distribution of opioid painkillers has led to an unfortunate shift towards the use of heroin and synthetic opioids like fentanyl. The inflow of fentanyl into the United States from Mexico and China has become a major concern. It is alarming to know that there was a time when one could easily purchase a kilogram of fentanyl from overseas for as little as $2,500 using a Visa card and have it delivered through the US Postal Service.

In this assignment you are going to have several tasks:

1. Import several CSV files
   1. Physicians
   2. Scripts (files 1 – 10)
   3. Zip Codes
2. Create **physician summary** and **org summary** tables
3. There are **10 analysis questions** each requiring **1** **TABLE** and at least **1 CHART** (ex. barplot, count plot or boxplot) To get the results you’ll likely need to break things up into steps and perform various steps including:
   1. Query to filter data
   2. Merge to join data
   3. N-largest / N-smallest (top N, Bottom N)
   4. Concat to append datasets together
   5. agg (sum, count, mean, min, max)
   6. create barplots and count using seaborn
   7. Rename columns using df.column = [“col1”,”col2”,…]
4. Devise **5 of your own analysis questions** answered with tables and visualizations
5. **Executive summary** – Write a short (1 – 2 page) executive summary on your thoughts of the analysis and what kinds of conclusions if any can we draw/ derive from this data? What would likely make this analysis more interesting? In your opinion does it make sense that there can be drastically different rates of opiate prescriptions by city? What struck you the most about opiate and non-opiate prescriptions?

## About the data….

I’ve pulled the Medicare Provider Utilization and Payment Data, specifically the Part D Prescriber Public Use File (aka PUF data). The PUF data provides information on prescription drugs prescribed by individual physicians and other health care providers and paid for under the Medicare Part D Prescription Drug Program. The dataset identifies providers by their National Provider Identifier (NPI) and the specific prescriptions that were dispensed at their direction, listed by brand name (if applicable) and generic name. I’ve also pulled the Medicare part D Physician dataset as well as gathered some additional data to aide our analysis.

# TASKS

I’ve divided this project up into several tasks.

## TASK 1 – IMPORT

Import the following CSV files into panda’s data frames

* **Physician** there are **three** **CSV files** containing physician data you’ll need to import them and append them into a single Panda’s “physicians” data frame. The files can be found in the sample notebook:
  + When complete there should be **84037 rows and 73 columns**
* **Prescriptions** there are **ten (10) CSV files** containing prescription drug summary for each provider (MD, DDS, NP etc.). You’ll need to import them and append them into a single Panda’s “prescriptions” data frame. The files can be found in the sample notebook:
  + When complete there should be **1,975,658 rows and 9 columns.**
* **Zip** codes, finally there is a Semi-Colon delimited file of zip codes you’ll want to import that as well the file can be found in the sample notebook.
  + To read a semicolon (;) delimited file using Pandas, you can use the **read\_csv()** function with the **sep** parameter set to **';'**.

### task 1a – Clean up column names!

My advice is to make your life easier and clean up your column names using something like the following:

df**.**columns **=** ( df**.**columns

**.**str**.**strip()

**.**str**.**lower()

**.**str**.**replace(' ', '\_')

**.**str**.**replace('-', '\_')

**.**str**.**replace('(', '')

**.**str**.**replace(')', '')

**.**str**.**replace('?', '')

**.**str**.**replace('\'', '')

print(df**.**columns)

## TASK 2 – STRUCTURE

My advice is to look ahead to the tasks we’ll need to perform. You are going to create 2 data frames:

* physician\_summary
* org\_summary

### task 2a - create PHSYCIAN\_summary

To create the physician\_summary table you are going to need to perform the following:

* **Filter** “Rndrng\_Prvdr\_Ent\_Cd” == “I”
  + Rndrng\_Prvdr\_Ent\_Cd code of ‘I’ identifies referring providers registered as **individuals** and an Prscrbr\_Ent\_Cd of ‘O’ identifies referring providers registered as **organizations**.
* **Inner join** physician\_summary to the zip\_code dataframe on
  + left\_on = "Rndrng\_Prvdr\_Zip5"
  + right\_on = "zip\_code"
* **Create a new column** called “provider\_name” by combining the following

physician\_summary["provider\_name"] = physician\_summary["Rndrng\_NPI"].astype(str) + ": " + physician\_summary["Rndrng\_Prvdr\_First\_Name "] + " " + physician\_summary["Rndrng\_Prvdr\_Last\_Org\_Name "]

NOTE: When complete you should have around 72,658 records and 91 columns

### task 2b - create ORG\_summary

To create the org\_summary table you are going to need to perform the following:

* **filter** “Rndrng\_Prvdr\_Ent\_Cd” == “O”
  + Rndrng\_Prvdr\_Ent\_Cd code of ‘I’ identifies referring providers registered as individuals and an Prscrbr\_Ent\_Cd of ‘O’ identifies referring providers registered as organizations.
* **Inner join** it to the zip\_code dataframe on
  + left\_on = "Rndrng\_Prvdr\_Zip5"
  + right\_on = "zip\_code"
* **Create a new column** called “org\_name” by combining the following

org\_summary["rndrng\_npi"].astype(str) + ": " + org\_summary["rndrng\_prvdr\_last\_org\_name"] + " - " + org\_summary["official\_usps\_state\_code"]

it should produce a result like this “1003002254: Walgreen Co. – TN”

NOTE: When complete you should have around 4279 records and 91 columns

Check with org\_summary.shape

## TASK 3 Analysis

The analysis section consists of 10 questions you will need to produce a **result set** (table) named RESnn, where “nn” corresponds to the question number (ex. RES01) **and a chart**! You’ll want to output the table in one code cell and output the chart in another.

* Use print(“RESnn: question ”) before your output of your table
* Use **title** and **subtitle** function to identify the chart
* Ensure that your **X & Y axis** are labeled
* Ensure that the chart **figure size** is set so the chart makes sense.
* In some cases, you may need to create more than one chart or result to compare and contrast you can use the subplot function to accomplish this or simply create another set of results and charts just make sure your analysis makes sense
* Finally, you may be asked questions that require a response, please use a **text / markdown** **cell** to answer these not print.

**NOTE: The sample output provided is simply for reference your output likely will be different!**

**HINT: rename columns with the df.columns = [“new\_name”,”newname”] to make your life easier!**

### RES01: Title “Top 10 Medicare Providers by TOT\_MDCR\_PyMT\_AMT”

Using phsycian\_summary, create a table RES01 and a chart displaying the top 10 providers by tot\_mdcr\_pymt\_amt. Your table should look something like:

|  |  |  |
| --- | --- | --- |
|  | Provider\_Name | Sum total payment |
| 0 | 1851320691: Bradley Somer | 14030052.64 |
| 1 | 1366683823: Wanda Lowery | 11462667.6 |
| 2 | 1750541553: Shil Patel | 10526851.61 |
| 3 | 1023018728: Ross Kerns | 8945222.23 |
| 4 | 1801103643: Syed Sameer Nasir | 8347614.58 |

And chart:

Chart

Description automatically generated

### RES02: Title “Top 10 Medicare ORGANIzations by TOT\_MDCR\_PyMT\_AMT”

Using org\_summary, create a table RES02 and a chart displaying the top 10 organizations by tot\_mdcr\_pymt\_amt. Your table should look something like:

|  |  |
| --- | --- |
| org\_name | tot\_mdcr\_pymt\_amt |
| 1538144910: Laboratory Corporation Of America ... | 234670845.50 |
| 1174009245: Vikor Scientific, Llc - SC | 42198509.44 |
| 1699782722: Aegis Sciences Corporation - TN | 33066425.94 |
| 1134538713: Mako Medical Laboratories, Llc - NC | 31287531.70 |
| 1235186800: Pathgroup Labs, Llc - TN | 21797255.78 |

Chart

Description automatically generated with low confidence

### RES03: Title “Nurse Practitioners by State”

Using physician\_summary, create a table and chart of “Nurse Practitioner” using rndrng\_prvdr\_type and rndrng\_prvdr\_state\_abrvtn .

|  |  |
| --- | --- |
| **rndrng\_prvdr\_state\_abrvtn** | **count** |
| TN | 5314 |
| NC | 4997 |
| SC | 2552 |

Chart:

Chart

Description automatically generated with medium confidence

### RES04: Title “Organization Total Spending by State

Using org\_summary, summarize tot\_mdcr\_pymt\_amt by rndrng\_prvdr\_state\_abrvtn (state) you should have a table that looks like this:

|  |  |
| --- | --- |
| **state** | **sum** |
| NC | 641174881.70 |
| TN | 408660971.70 |
| SC | 289073137.80 |

And chart:

Chart

Description automatically generated

### RES05: Title “Top 5 ORG spending by STATE and Specialty”

Use org\_summary, sum tot\_mdcr\_pymt\_amt by rndrng\_prvdr\_state\_abrvtn and rndrng\_prvdr\_type you should get the top 5 provider types by the sum of tot\_mdcr\_pymt\_amt for each state. Your output table should look something like this:

|  |  |  |
| --- | --- | --- |
| state | provider\_type | sum |
| TN | Clinical Laboratory | 128751819.8 |
| TN | Ambulance Service Provider | 128737375.5 |
| TN | Ambulatory Surgical Center | 108401378.1 |
| TN | Mass Immunizer Roster Biller | 10849333.97 |
| TN | Centralized Flu | 7241423.04 |
| SC | Ambulance Service Provider | 109359849 |
| SC | Ambulatory Surgical Center | 78569205.43 |
| SC | Clinical Laboratory | 71811426.38 |
| SC | Mass Immunizer Roster Biller | 10384709.16 |
| SC | Independent Diagnostic Testing Facility (IDTF) | 9208319.55 |
| NC | Clinical Laboratory | 357244114.2 |
| NC | Ambulance Service Provider | 144528807.8 |
| NC | Ambulatory Surgical Center | 78164524.07 |
| NC | Mass Immunizer Roster Biller | 16004206.92 |
| NC | Pharmacy | 13438189.16 |

Hint: break this up into multiple steps here are some snips of code you might find helpful:

RES05 = (org\_summary

.groupby(["rndrng\_prvdr\_state\_abrvtn","rndrng\_prvdr\_type"])

.agg({"tot\_mdcr\_pymt\_amt":["sum", “count”]})

.reset\_index()

)

RES05.columns = ["state","provider\_type","sum"]

(RES05

.sort\_values(by=['state', 'sum'],ascending=False)

.groupby(["state"])

.head(5)

)

And chart:

Chart, bar chart

Description automatically generated

### RES06: Title “Top 5 PROVIDER spending Specialty”

Use phys\_summary, sum tot\_mdcr\_pymt\_amt and rndrng\_prvdr\_type you should get the top 5 provider types by the sum of tot\_mdcr\_pymt\_amt Your output table should look something like this

|  |  |
| --- | --- |
| rndrng\_prvdr\_type | tot\_mdcr\_pymt\_amt |
|  | **sum** |
| Ophthalmology | 505082507.8 |
| Internal Medicine | 485139913.6 |
| Hematology-Oncology | 450390569.9 |
| Nurse Practitioner | 443327249.8 |
| Family Practice | 381226996.7 |

Chart

Description automatically generated with low confidence

### RES07: Title “top 10 Generic Opiates by Total Day Supply”

Use prescriptions, filter for is\_opiate == “Yes”, group by gnrc\_name sum tot\_day\_suply select the top 10 opiates your table should look like this

|  |  |
| --- | --- |
| gnrc\_name | tot\_day\_suply |
|  | **sum** |
| Hydrocodone/Acetaminophen | 776370204 |
| Oxycodone Hcl/Acetaminophen | 177779252 |
| Tramadol Hcl | 144511176 |
| Oxycodone Hcl | 97625860 |
| Morphine Sulfate | 35075620 |
| Fentanyl | 7821429 |
| Methadone Hcl | 6650875 |
| Acetaminophen With Codeine | 4832944 |
| Hydromorphone Hcl | 4366616 |
| Oxymorphone Hcl | 1930748 |

And chart:

Chart

Description automatically generated

### RES08: Title “Top 10 Most Expensive Drugs”

Use prescriptions, create a new column called claim\_cost by dividing tot\_drug\_cst by tot\_clms then simply group by brnd\_name and take the mean of top 10 brnd\_name(brand name) drugs by claim cost.

|  |  |
| --- | --- |
| brnd\_name | mean |
| Myalept | 165046.884 |
| Strensiq | 112137.005 |
| Berinert | 107622.893 |
| Ravicti | 95992.306 |
| Firazyr | 75005.836 |
| Lumizyme | 73466.83 |
| Chenodal | 65809.59 |
| Aldurazyme | 63541.299 |
| Ayvakit | 61090 |
| Procysbi | 57488.641 |

And chart

Chart, funnel chart

Description automatically generated

### RES09: Title “Top 2 Cities in SC, NC, TN by sum of Total Day Supply for Hydrocodone and Oxycodone”

Take the prescriptions table

* filter gnrc\_name for “Hydrocodone/Acetaminophen” and “Oxycodone Hcl/Acetaminophen” (our top 2 opiates prescribed).
* Join result physican\_summary,
* Filter physican\_summary for official\_usps\_state\_code for SC, NC, and TN
* Group by ['official\_usps\_city\_name', 'official\_usps\_state\_code']
* Get the sum and mean of tot\_day\_suply and tot\_drug\_cost
* Sort values by Sum of tot\_day\_suply largest to smallest (ascending = False)
* group by official\_usps\_state\_code
* Get the top 2 Cities for each state

Your table should look something like this:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| official\_usps\_state\_code | 0fficial\_usps\_city\_name | tot\_day\_suply |  | tot\_drug\_cst |
|  |  | sum | mean | sum | mean |
| TN | Knoxville | 3648035 | 1678.801 | 5628765.79 | 2590.32 |
| TN | Chattanooga | 2702112 | 1898.884 | 4483955.44 | 3151.06 |
| SC | Spartanburg | 2222916 | 2751.134 | 2892592.43 | 3579.94 |
| SC | Greenville | 1712100 | 1188.958 | 2720653 | 1889.34 |
| NC | Charlotte | 3669324 | 1277.62 | 7970929.61 | 2775.39 |
| NC | Raleigh | 2246097 | 1184.65 | 7169602.51 | 3781.44 |

And I want you to create two graphs (1. Sum Total day supply, and 2. Mean day suply)

Chart, bar chart, funnel chart

Description automatically generated

Chart, bar chart

Description automatically generated

NOTE: of course, break this up into multiple steps and take creative liberty on how to produce this analysis. For example, you might want to do this for each state separately (hint) and concatenate (pd.concat) them together. Whatever path you decide to do this your results should be like the table above and your graphs should look like those above.

Hint: I created a new column of State + City and a palette to color the bars. Similar to the code below:

#RES09.columns = ["city","state","sum\_day\_suply","mean\_day\_suply","sum\_cost","mean\_cost"]

#RES09["state\_city"] = RES09["state"] + " - " + RES09["city"]

my\_palette = {"TN - Knoxville": "orange",

"TN - Chattanooga": "orange",

"SC - Spartanburg": "red",

"SC - Greenville":"red",

"NC - Charlotte" : "lightblue",

"NC - Raleigh" : "lightblue"

}

sns.barplot(data=RES09,

y="state\_city",

x="sum\_day\_suply",

palette=my\_palette

);

Part 3 of RES09: ANSWER in MARKDOWN: What if anything does this tell you about hydrocodone & oxycodone prescriptions to Medicare Part D by city and state? What additional info would make for more interesting analysis?

### RES10: Title “An OPIATE for everyone?”

My Recommendation is to break this up into a few steps:

Using the Zips table filter for official\_usps\_state\_code == TN, SC, NC and official\_usps\_city\_name equal to your cities from RES09. Group by official\_usps\_state\_code and official\_usps\_city\_name then SUM the population you should end up with a table like this:

|  |  |  |
| --- | --- | --- |
| **official\_usps\_city\_name** | **official\_usps\_state\_code** | **population** |
|  |  | **sum** |
| Charlotte | NC | 923655 |
| Chattanooga | TN | 210667 |
| Greenville | SC | 221595 |
| Knoxville | TN | 391360 |
| Raleigh | NC | 555041 |
| Spartanburg | SC | 112001 |

NEXT join this table to RES09 on State and City,Create a new column “Day Supply Per Person” by Dividing sum of **tot\_day\_suply by sum of population your table should look like this:**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| state\_city | city | state | population | sum\_day\_suply | mean\_day\_suply | sum\_cost | mean\_cost | **day\_suply\_per\_person** |
| NC - Charlotte | Charlotte | NC | 923655 | 3669324 | 1277.62 | 7970929.61 | 2775.393 | **3.973** |
| TN - Chattanooga | Chattanooga | TN | 210667 | 2702112 | 1898.884 | 4483955.44 | 3151.058 | **12.826** |
| SC - Greenville | Greenville | SC | 221595 | 1712100 | 1188.958 | 2720653 | 1889.342 | **7.726** |
| TN - Knoxville | Knoxville | TN | 391360 | 3648035 | 1678.801 | 5628765.79 | 2590.32 | **9.321** |
| NC - Raleigh | Raleigh | NC | 555041 | 2246097 | 1184.65 | 7169602.51 | 3781.436 | **4.047** |
| SC - Spartanburg | Spartanburg | SC | 112001 | 2222916 | 2751.134 | 2892592.43 | 3579.941 | **19.847** |

Make a barplot like this:

Chart, bar chart, funnel chart

Description automatically generated

Part 3 of RES10: ANSWER in MARKDOWN: What if anything does this tell you about hydrocodone & oxycodone prescriptions to Medicare Part D patients by city population? what additional info would make for more interesting analysis?

## TASK 4 Your Analysis

### RES11: Title “your analysis”

### RES12: Title “your analysis”

### RES13: Title “your analysis”

### RES14: Title “your analysis”

### RES15: Title “your analysis”

## Executive Summary

Write a short (1 – 2 page) executive summary on your thoughts of the analysis and what kinds of conclusions if any can we draw/ derive from this data? What would likely make this analysis more interesting? In your opinion does it make sense that there can be drastically different rates of opiate prescriptions by city? What struck you the most about opiate and non-opiate prescriptions?

# TURN IN

You should submit the following:

1. **Notebook (in HTML)** appropriately formatted and organized so that I can follow it
   1. Each question should have a markdown cell with ## RES01 for example
   2. Each question should produce a TABLE and at least 1 CHART
2. **Word doc** with your executive summary