**Capstone Project Preliminary Analysis Template**

**Name: Viridiana Ramirez**

**Project Title: Medicare enrollment and usage in NEW YORK 2022 (might change later)**

**Project Objective: Provide insights to Medicare Enrollment and Identify demographic attributes of beneficiaries.**

|  |
| --- |
| **- Take a closer look at how Medicare enrollment changes month to month**  **- Dive into the demographics of beneficiaries like age, gender, race, and ethnicity in different parts of New York State by Hospital discharge analysis**  **- Explore how different demographic factors might impact things like hospital stays, admission types, and severity of illness**  **- Map out areas with high Medicare enrollment and see if there are any connections to regions with lots of hospital visits**  **Audience: healthcare providers and administrators, policymakers, insurance companies, patients** |
|
|
|
|

**Data Collection Source:**  *Downloaded data from CMS.gov* ***https://data.cms.gov/summary-statistics-on-beneficiary-enrollment/medicare-and-medicaid-reports/medicare-monthly-enrollment/data****.*

*-* Health. Data. GOV New York State Hospital Inpatient Discharges (SPARCS De-Identified): 2022

https://health.data.ny.gov/Health/Hospital-Inpatient-Discharges-SPARCS-De-Identified/5dtw-tffi/about\_data

**Methodology: SEE BELLOW**

*.*

*Describe your preprocessing techniques. How was the data organized?*

*What was the workflow?*

*What tools were used to organize and better understand the data?*

*How was the data cleaned?*

*Were there any typos?*

*How was this corrected?*

*Any data type conversions, e.g. string to date values?*

*Any other formatting, e.g. renaming columns or values?*

*Were any missing values replaced? If so, describe why?*

*How will replacing the values impact the analysis?*

*Are there any exclusions (columns, rows, or values)?*

*Include any other methods in processing and organizing the data.*

**Findings:**

**Medicare Monthly Enrollment**

1. **Data Size:** *476,997 rows 26 columns*
2. **Missing Values:** *no missing values*
3. **Duplicates:** *no duplicates*
4. **Dates: *2013-2023***
5. **Additional Findings:** *see bellow.*

New York State Hospital Inpatient Discharges (SPARCS De-Identified): 2022

1. **Data Size: 2,061,634 rows 33 columns**
2. **Missing Values:** *total blank cells 5,913,717*
3. **Duplicates:** *no duplicates*
4. **Dates: *2022***
5. **Additional Findings:** *see bellow.*

**Attach any SQL or Python code in a separate file**

**Medicare Enrollment data set**

I began my analysis by downloading a CSV file containing Medicare enrollment data and loading it into Excel. My initial step involved understanding the organization of the dataset before performing any data cleaning. The dataset consists of 26 columns and 476,997 rows, covering the years 2013 to 2023. The first six columns include the year, month, geographical level (national, state, county), state abbreviation, state name (including US Territories), and county name.

I noticed the data was aggregated by year and geographical levels (national, state, county). I applied filters to separate the totals by these levels, creating individual tables for national by year, national by month/year, state by year, and state by month/year.

The dataset included 144 rows labeled as 'UNKNOWN,' which appeared to separate data between states and years. These rows were excluded from the analysis. Additionally, there were 3,300,941 cells marked with '\*', indicating that the plan is not offered in that state or territory. Of these, 2,717 belonged to the 'UNKNOWN' rows.

The dataset spans from 2013 to 2023, revealing trends such as a decrease in enrollment for original Medicare benefits (Part A and B) from 2019 to 2023, alongside an increase in Medicare Advantage enrollments.

**State Hospital Inpatient Discharges dataset**

I added the New York State Hospital Inpatient Discharges dataset for 2022. This dataset includes information on hospital service areas, which consist of counties with hospitals. Out of 62 counties in NY, 56 are listed as having hospitals. This discrepancy poses a challenge in comparing Medicare usage across all counties. Rows with missing service area or county information were labeled as 'Unk' to maintain the integrity of the analysis. Note: given the timeframe I will also see how covid 19 impacted hospital usage for this year.

Using SQL, I reviewed the New York State Hospital Inpatient Discharges data to identify demographic information, analyze county-level hospital visits, admission types and diagnosis, and evaluate charges and costs. The primary filter for the dataset was visits paid primarily with Medicare.

Once the SQL queries are finalized, the data will be imported into Tableau for visualization. I will build a dashboard to present my findings, aimed at healthcare providers, administrators, policymakers, insurance companies, and payers. The goal is to provide actionable insights and recommendations to improve healthcare delivery and patient outcomes based on my analysis.