Medicaid and Medicare evaluation of sponsorship

and ratings

Project Scoping Document  
  
Team #11:

* Lucius Anderson
* Chandrika Jones
* Anita Pinto
* Henry Rivera
* Anthony Robnett

# Project Topic

Topic Chosen:

* Topic 2: Data Swan Inc: Medicaid and Medicare evaluation of sponsorship and ratings

Why team members chose this topic?:

* Anderson:
  + Interested to see how individuals apart of medicaid and medicare experience healthcare
* Jones:
  + I work in finance, but I have an interest in moving into healthcare. This seems to be a good launching point into becoming more comfortable exploring big health data.
* Pinto:
  + This was a good opportunity for me to explore data sets related to Healthcare and an example of ever growing large data.
* Rivera:
  + My organization is in the pharmaceutical industry. Since I’m already exposed to it on a daily basis, I think I could help with the business context while continuing to develop my technical skills
* Robnett:
  + Having a background in public health, I always had an interest in learning about health informatics, so I found this topic relevant.

# Client Analysis

* a. What exactly needs to be measured?:
  + To determine if there is a correlation between the payments made by CMS to physicians and the ratings generated by patients.
    - Figure 1, Payment data:
      * Total count of payments.
      * Average payment amount (USD).
      * Sum of total amount of payment by recipient state.
      * Count of covered recipients by covered recipient types.
      * Sum of total amount of payment (USD) by recipient city.
    - Figure 2, Rating data:
      * Total count of ratings.
      * Count by Specialty.
      * Average of rating by year, month, day, and doctor name.
      * Sum of total amount of payment (USD) by doctor vs Average of rating.
      * Year, month, day, and rating platform vs Average of Rating.
* b. What are your variables?:
  + Figure 1, Payment data:
    - Date of Payment.
    - Product Category or Therapeutic Area.
    - Covered Recipient Type
    - Total Amount of Payment US Dollars
    - Recipient State.
    - Recipient City.
  + Figure 2, Rating data:
    - Date.
    - Rating platform.
    - Rating
    - Specialty
    - Hospital name.
    - Doctor name.
* c. Who will measure it?:
  + Payment data:
    - [Open Payments data](https://openpaymentsdata.cms.gov/summary) by the U.S. Centers for Medicare & Medicaid Services.
      * General Payments:
        + Payments that are not associated with a research study.
      * Ownership or Investment Interest:
        + Ownership and investment interest in companies, which describes both the actual dollar amount invested and the value of the ownership or investment interest. Records may have one or both of these values associated with them.
      * Research Payment:
        + Funding for a research project of study where the physician is named as a principal investigator.
  + Review data:
    - [Review Scraper API](https://help.datashake.com/article/34-supported-review-sites-scraper) data must come from Healthgrades, ZocDoc, WebMD, RateMDs, RealSelf, or Vitals.
* d. What is the time interval to be measured?:
  + The COVID-19 period from January 2020 to December 2021.

# Data Sources

[Open Payments](https://www.cms.gov/openpayments/resources)

* Overview: Open Payments data is publicly accessible information about payments and transfers of value that reporting entities make to covered recipients. Reporting entities are responsible for tracking these financial transactions annually and must submit this data to the Centers for Medicare & Medicaid Services (CMS).
* Key information:
  + Open Payments Data is recorded across three categories (datasets) for each Program Year:
    - **General Payments** - Payments or transfers of value that are not in connection with a research agreement or research protocol.
    - **Ownership and Investment Interests** - Information about the ownership or investment interests that physicians or their immediate family members have with the reporting entities.
    - **Research -** Information about the ownership or investment interests that physicians or their immediate family members have with the reporting entities.
  + Program Year data is available in ZIP files from 2015-2021. (Archives for 2013-2014 exist as well).
  + Physician Data is available for all years in a ZIP file.
  + API available to extract data.
* Data Format Availability:
  + ZIP Files
  + [API](https://openpaymentsdata.cms.gov/about/api)
* Strengths:
  + Robust dataset
  + Data dictionary/methodology documentation is available
  + Data can be extracted via API (automated means)
* Weakness:
  + Will need to decompress data in ETL pipelines if we use the ZIP files
  + [API documentation](https://openpaymentsdata.cms.gov/about/api) is a little confusing. (Will need to use JSON formatter to make page readable)
  + Will have to make sure we understand the data and methodology well when the data model and transformation process occurs

[Open Payments: Physician Profile Supplement](https://openpaymentsdata.cms.gov/dataset/23160558-6742-54ff-8b9f-cac7d514ff4e)

* Overview: The Physician Profile Supplement file contains information about physicians who have been indicated as recipients of payments, other transfers of value, or ownership and investment interest in payment records, as well as physicians who have been identified as principal investigators associated with research payment records published by Open Payments.
* Key Information:
  + This dataset contains only those physicians that have at least one published payment record in this cycle of the publication as of December 31, 2020
* Data Format Availability:
  + ZIP File
  + [API](https://openpaymentsdata.cms.gov/dataset/23160558-6742-54ff-8b9f-cac7d514ff4e#:~:text=Scroll%20to%20dataset-,API,-Get%20Email%20Updates)
* Strengths:
  + Physician profile data is a requirement
  + Robust dataset
  + Data dictionary/methodology documentation is available
  + Data can be extracted via API
* Weaknesses:
  + CSV download file is large
  + API documentation is a little confusing

[Data.CMS.GOV: Hospital All Owners](https://data.cms.gov/provider-characteristics/hospitals-and-other-facilities/hospital-all-owners)

* Overview: The Hospital All Owners Information dataset provides information on all owners of the hospitals. This data includes ownership information such as ownership name, ownership type, ownership address and ownership effective date.
* Key Information:
* Data Format Availability:
  + CSV
  + [API](https://data.cms.gov/provider-characteristics/hospitals-and-other-facilities/hospital-enrollments/api-docs)
* Strengths:
  + Dataset is key when joining on ratings data, to get ratings by hospital
  + Robust Dataset
  + Data dictionary/methodology documentation available
  + API documentation is well written
* Weaknesses:
  + \*Potential\* String classification of hospital names could be different from ratings data and make cleaning process more difficult

[Datashake API: Get Reviews](https://docs.datashake.com/reviewindex/get_reviews/)

* Overview: Datashake is API SaaS that fetches online reviews for your business. One API call to get reviews from 85+ websites without any technical overhead.
* Key information:
  + Datashake is an API SaaS that grabs review data from various platforms and allows businesses to grab that data via API.
  + Has data from 85+ platforms, but for our **Capstone project we are only interested in reviews from Healthgrades, ZocDoc, WebMD, RateMDs, RealSelf or Vitals**
* Data Format Availability:
  + [API](https://docs.datashake.com/reviewindex/get_reviews/)
* Strengths:
  + Data dictionary/methodology documentation available
  + API documentation is well written
  + Dataset is key to showcase ratings by hospital
* Weaknesses:
  + Depending on the engineering behind the process, data could be inconsistent if review data is built from web scraping
    - ZocDoc, WebMD, RateMDs, RealSelf, and Vitals are considered difficult web scraping websites by Datashake so we could see data inconsistencies or errors. [See here](https://help.datashake.com/article/34-supported-review-sites-scraper)
  + Pagination could cause pipeline to fail
  + Data aggregator SaaS APIs can sometimes fail so monitoring is key for this datasource

# Milestones

The team plans to identify each member's strengths and split the tasks accordingly.

| Date | Deliverable Category | Working Stage | Details | Team Members Working on |
| --- | --- | --- | --- | --- |
| Week 4, April 29th | Data Procurement & Storage | 1 | **Focus on procurement process, how we can make it efficient and scalable** | TBD |
| Week 5, May 6th | Data Procurement & Storage | 2 | Focus and prioritize high value datasets and their storage first. | TBD |
| **Week 6, May 13th** | Data Procurement & Storage | 3 | refine the deliverable and finalize the plan | TBD |
| **Week 7, May 20th** | Initial ETL | 1 | Focus on extraction stage of ETL | TBD |
| **Week 8, June 3rd** | Initial ETL | 2 | Focus on transformation stage of ETL pipeline | TBD |
| **Week 9, June 10th** | Initial ETL | 3 | Focus on loading stage of ETL pipeline. Refine and deliver | TBD |
| **Week 10, June 24th** | Data Modeling | 1 | Refine and finalize data model, deliver. | TBD |
| **Week 11, July 8th** | Initial Draft of Presentation | 1 | Make sure initial draft explains our processes and describes why we made certain decision | TBD |
| **Week 12, July 15th** | Final Presentation | 1 | Present final presentation. Graduate. Congrats to everyone | TBD |

# Timeline

| **Date** | **Deliverable** | **Details** |
| --- | --- | --- |
| **Week 1, April 8th** | Team Formation & Work on idea formation | *Introduction to Data Engineering* |
| **Week 2, April 15th** | Idea should be finalized & Start on Scoping | *Python Programming for Data Engineering* |
| **Week 3, April 22nd** | Project Scoping Completed | *Relational Databases and SQL* |
| **Week 4, April 29th** | Data clearing for Procurement | *Advanced SQL* |
| **Week 5, May 6th** | TBD | *Non Relational Databases* |
| **Week 6, May 13th** | TBD | *Data Procurement and Data Modeling* |
| **Week 7, May 20th** | [Data Procurement](https://docs.google.com/document/d/1U59G7AmjJA4huJorSbQI1a1kX6ImFT7jmzykVoPRmIo/edit#FiONEekCOxRDZbiajwR6gXfI4OMBjkM8ngfkOM/edit#gid=0) | *Working With Big Data* |
| **Week 8, June 3rd** | TBD | *Data Pipelines and Data Quality* |
| **Week 9, June 10th** | [Initial ETL Pipeline](https://docs.google.com/document/d/1sIuRONANs8lUEqoUmeSrTGim0f44HAYWiJP-A_KLLH0/edit) | *Data Infrastructure and Orchestration* |
| **]Week 10, June 24th** | [Data Modeling](https://docs.google.com/document/d/1DiAiCmqYiP0Ghk3GEL9A9KmtHgmOm6x_26SNY6s6jBM/edit?usp=sharing) | *Capstone Project Work* |
| **Week 11, July 8th** | Initial Draft of Presentation Complete | *Capstone Project Work* |
| **Week 12, July 15th** | Presentation rehearsed and ready | *Presentation of Capstone Projects and Graduation* |

# Team Organization

Rivera: I am a big proponent of communication, honesty, and having fun. I want everyone on the team to feel encouraged and comfortable to speak up and bring new ideas to the table. I will voice my opinion and thoughts, but I know when to step back for the good of the team. I do not want to be tied to one specific function. I like to consider myself a swiss army knife and can handle a wide range of technical and non technical tasks.

Pinto: I call myself a hybrid engineer with varied experience and roles from semiconductor industry to financial start-ups. I am flexible in my working style and interests and can modify my role based on the need of the project activity and timelines. Just as I am passionate about learning new skills and building code, I also love debugging errors and show stoppers. I am looking forward to working with my team members and learning from them.

Jones: I am very flexible in my roles, however, I am an excellent QA specialist. I have a keen eye for detail and no error/defect slips by me. I am all for open communication and my personal style leans toward a more direct method.

Lucius: I am flexible in my roles, for the project I believe I can specialize in architecting the data, coding , and strategy. I work great thinking high-level and overall strategizing a plan to achieve goals and deliver.

Robnett: As someone who is new to working with a team on a data project, I am eager to take on any role that can contribute to the success of the project. I am highly adaptable and willing to be flexible, whether it means coordinating and communicating with team members or performing any task required to meet project objectives. I believe that open communication and collaboration are essential to achieving our goals, and I am committed to working with the team to ensure that our project is completed successfully.

Concerns

* To create an effective timeline, we need to ensure a balanced approach to working on both the data modeling preparation and ETL pipeline portions concurrently.
* Regarding the deliverables for each week on the timeline, we are uncertain about the exact objectives we need to achieve to complete the submission assignments. We plan to leverage the skills learned in the lesson of that week to apply towards the assignment.
* We observed that there are three data sources within the payment data source, namely general payment, ownership payment, and research payment. As we move forward in the project, we will determine if it is optimal to incorporate all three payment data sources or prioritize the general payment data source.
* We still need to decide on the cloud repository tools for version management that we will use. We expect to receive guidance on this matter in the near future.