Data Swan Inc: Medicaid and Medicare evaluation of sponsorship and ratings

Data Procurement and storage plan

Team #13: Ashek Ag Mohamed, Esneyder Gonzalez, Isiasha Gordon, Bethel Ikejiofor, Brian Noble

I. Data Procurement

A. Data Requirements & Sources:
Our data consists of 8 flat file sets saved in CSV format as well as one obtained through API connection.

  a. The following data files are needed:
    i. General Payment Data from 2020 and 2021
    ii. Research Payment Data from 2020 and 2021
    iii. Ownership Payment Data from 2020 and 2021
    iv. Physician Ratings
    v. General Payment Resource (supplemental)
    vi. Physician Profile Supplemental data (supplemental)

  b. Purpose: To build an ETL pipeline that will be used by the data analysts responsible for creating a comprehensive dashboard for decision-making

  c. Objective: The above data requirements provide payment data for the analysis and determination of a possible correlation between the payments and the ratings generated by patients

B. Data Collection:
  a. The data was manually downloaded from the CMS.gov website and will be uploaded to AWS S3 buckets.
  b. To assess the quality of the data and remove unusable columns, we generated a random sample of 10,000 entries for each data set and excluded columns with at least 30% missing values as long as they are not needed for data analysis identified by our client.
  c. Jupyter Notebook can be found here
  d. The physician and hospital ratings will need to be assessed using APIs(?)
  e. The tables listed below contian information regarding each data sets. A more concise layout can be found here

- All files are in csv file formats. For each remaining column there is an example of the content in that column, how many missing values remain, and the data type of that variable.
- 2020 General Payment Data:

- The original dataset consists of 5,823,162 rows and is 3.7 GB

2020 General Payment Data				
Column #	**Columns:**	**Example/Contents**	**Missing Value Count (10,000)**	**Dtype**
0	Change_Type	UNCHANGED	0	object
1	Covered_Recipient_Type	Covered Recipient Physician	0	object
2	Covered_Recipient_Profile_ID	557946	0	int64
3	Covered_Recipient_NPI	1841302403	0	int64
4	Covered_Recipient_First_Name	Vikas	40	object
5	Covered_Recipient_Last_Name	Pilly	41	object
6	Recipient_Primary_Business_Street_Address_Line1	400 International Drive	0	object
7	Recipient_City	Williamsville	0	object
8	Recipient_State	NY	0	object
9	Recipient_Zip_Code	14221-5771	0	object
10	Recipient_Country	United States	0	object
11	Covered_Recipient_Primary_Type_1	Medical Doctor	40	object
12	Covered_Recipient_Specialty_1	Allopathic & Osteopathic Physicians|Pain Medic...	44	object
13	Covered_Recipient_License_State_code1	OH	40	object
14	Submitting_Applicable_Manufacturer_or_Applicable_GPO_Name	Caerus Corp.	0	object
15	Applicable_Manufacturer_or_Applicable_GPO_Making_Payment_ID	1E+11	0	int64
16	Applicable_Manufacturer_or_Applicable_GPO_Making_Payment_Name	Caerus Corp.	0	object
17	Applicable_Manufacturer_or_Applicable_GPO_Making_Payment_State	MN	33	object
18	Applicable_Manufacturer_or_Applicable_GPO_Making_Payment_Country	United States	0	object
19	Total_Amount_of_Payment_USDollars	14875	0	float64
20	Date_of_Payment	8/19/20	0	datetime64[ns]
21	Number_of_Payments_Included_in_Total_Amount	1	0	int64
22	Form_of_Payment_or_Transfer_of_Value	Cash or cash equivalent	0	object
23	Nature_of_Payment_or_Transfer_of_Value	Consulting Fee	0	object
24	Physician_Ownership_Indicator	No	40	object
25	Third_Party_Payment_Recipient_Indicator	No Third Party Payment	0	object
26	Delay_in_Publication_Indicator	No	0	object
27	Record_ID	709231379	0	int64
28	Dispute_Status_for_Publication	No	0	object
29	Related_Product_Indicator	Yes	0	object
30	Covered_or_Noncovered_Indicator_1	Covered	2534	object
31	Indicate_Drug_or_Biological_or_Device_or_Medical_Supply_1	Drug	2534	object
32	Product_Category_or_Therapeutic_Area_1	Prenatal Vitamin & Mineral	2646	object

- The following columns were removed from the 2020 General Payment data set:

2020 General Dropped Columns
Charity_Indicator',
'Covered_Recipient_Middle_Name',
'Associated_Device_or_Medical_Supply_PDI_5',
'Covered_Recipient_Primary_Type_6',
'Associated_Drug_or_Biological_NDC_2',
'Associated_Device_or_Medical_Supply_PDI_2',
'Covered_Recipient_License_State_code5',
'Associated_Drug_or_Biological_NDC_3',
'Associated_Device_or_Medical_Supply_PDI_3',

'Covered_Recipient_Specialty_6',
'Covered_Recipient_Specialty_5',
'Covered_Recipient_Specialty_4',
'Covered_Recipient_Specialty_3',
'Covered_Recipient_Specialty_2',
'Covered_Recipient_Primary_Type_5',
'Covered_Recipient_Primary_Type_4',
'Covered_Recipient_Primary_Type_3',
'Covered_Recipient_Primary_Type_2',
'Recipient_Postal_Code',
'Recipient_Province',
'Associated_Drug_or_Biological_NDC_4',
'Associated_Device_or_Medical_Supply_PDI_4',
'Associated_Drug_or_Biological_NDC_5',
'Associated_Device_or_Medical_Supply_PDI_1',
'Covered_Recipient_License_State_code4',
'Teaching_Hospital_ID',
'Teaching_Hospital_CCN',
'Teaching_Hospital_Name',
'Covered_Recipient_License_State_code3',
'Contextual_Information',
'Name_of_Third_Party_Entity_Receiving_Payment_or_Transfer_of_Value',
'Third_Party_Equals_Covered_Recipient_Indicator',
'Covered_Recipient_Name_Suffix',
'Name_of_Drug_or_Biological_or_Device_or_Medical_Supply_5',
'Product_Category_or_Therapeutic_Area_5',
'Indicate_Drug_or_Biological_or_Device_or_Medical_Supply_5',
'Covered_or_Noncovered_Indicator_5',
'Product_Category_or_Therapeutic_Area_4',
'Name_of_Drug_or_Biological_or_Device_or_Medical_Supply_4',
'Covered_or_Noncovered_Indicator_4',
'Indicate_Drug_or_Biological_or_Device_or_Medical_Supply_4',
'Covered_or_Noncovered_Indicator_3',
'Indicate_Drug_or_Biological_or_Device_or_Medical_Supply_3',
'Product_Category_or_Therapeutic_Area_3',
'Name_of_Drug_or_Biological_or_Device_or_Medical_Supply_3',
'Covered_Recipient_License_State_code2',
'Name_of_Drug_or_Biological_or_Device_or_Medical_Supply_2',
'Covered_or_Noncovered_Indicator_2',
'Product_Category_or_Therapeutic_Area_2',
'Indicate_Drug_or_Biological_or_Device_or_Medical_Supply_2',
'State_of_Travel',
'City_of_Travel',
'Country_of_Travel',
'Associated_Drug_or_Biological_NDC_1',
'Recipient_Primary_Business_Street_Address_Line2'

- 2021 General Payment Data:
    - The original dataset consists of 11,408,130 rows and is 7.11 GB

2021 General Payment Data				
Column #	**Columns:**	**Example/Contents**	**Missing Value Count (10,000)**	**Dtype**
0	Change_Type	UNCHANGED	0	object
1	Covered_Recipient_Type	Covered Recipient Physician	0	object
2	Covered_Recipient_Profile_ID	92058	0	int64
3	Covered_Recipient_NPI	1043218118	0	int64
4	Covered_Recipient_First_Name	Ahad	37	object
5	Covered_Recipient_Last_Name	Mahootchi	37	object
6	Recipient_Primary_Business_Street_Address_Line1	6739 Gall Blvd	0	object
7	Recipient_City	Zephrhills	0	object
8	Recipient_State	FL	28	object
9	Recipient_Zip_Code	33542	28	object
10	Recipient_Country	United States	0	object
11	Covered_Recipient_Primary_Type_1	Medical Doctor	37	object
12	Covered_Recipient_Specialty_1	Allopathic & Osteopathic Physicians|Ophthalmology	39	object
13	Covered_Recipient_License_State_code1	FL	37	object
14	Submitting_Applicable_Manufacturer_or_Applicable_GPO_Name	Mobius Therapeutics, LLC	0	object
15	Applicable_Manufacturer_or_Applicable_GPO_Making_Payment_ID	1E+11	0	int64
16	Applicable_Manufacturer_or_Applicable_GPO_Making_Payment_Name	Mobius Therapeutics, LLC	0	object
17	Applicable_Manufacturer_or_Applicable_GPO_Making_Payment_State	MO	178	object
18	Applicable_Manufacturer_or_Applicable_GPO_Making_Payment_Country	United States	0	object
19	Total_Amount_of_Payment_USDollars	2500	0	float64
20	Date_of_Payment	5/26/21	0	datetime64[ns]
21	Number_of_Payments_Included_in_Total_Amount	1	0	int64
22	Form_of_Payment_or_Transfer_of_Value	Cash or cash equivalent	0	object
23	Nature_of_Payment_or_Transfer_of_Value	Compensation for services other than consultin...	0	object
24	Physician_Ownership_Indicator	No	606	object
25	Third_Party_Payment_Recipient_Indicator	No Third Party Payment	0	object
26	Delay_in_Publication_Indicator	No	0	object
27	Record_ID	754966348	0	int64
28	Dispute_Status_for_Publication	No	0	object
29	Related_Product_Indicator	Yes	0	object
30	Covered_or_Noncovered_Indicator_1	Covered	2869	object
31	Indicate_Drug_or_Biological_or_Device_or_Medical_Supply_1	Drug	2870	object
32	Product_Category_or_Therapeutic_Area_1	Ophthalmology	2960	object
33	Name_of_Drug_or_Biological_or_Device_or_Medical_Supply_1	Mitosol	2960	object

- The following columns were removed from the 2021 General Payment data set:

2021 General Dropped Columns
'Covered_Recipient_Middle_Name',
'Charity_Indicator',
'Associated_Drug_or_Biological_NDC_2',
'Covered_Recipient_Specialty_6',
'Associated_Drug_or_Biological_NDC_5',
'Associated_Drug_or_Biological_NDC_3',
'Covered_Recipient_License_State_code5',
'Recipient_Province',
'Covered_Recipient_Primary_Type_2',

'Covered_Recipient_Primary_Type_3',
'Covered_Recipient_Primary_Type_4',
'Covered_Recipient_Primary_Type_5',
'Covered_Recipient_Primary_Type_6',
'Covered_Recipient_Specialty_3',
'Covered_Recipient_Specialty_4',
'Covered_Recipient_Specialty_5',
'Covered_or_Noncovered_Indicator_5',
'Indicate_Drug_or_Biological_or_Device_or_Medical_Supply_5',
'Associated_Device_or_Medical_Supply_PDI_5',
'Product_Category_or_Therapeutic_Area_5',
'Name_of_Drug_or_Biological_or_Device_or_Medical_Supply_5',
'Covered_Recipient_Specialty_2',
'Covered_Recipient_License_State_code4',
'Associated_Device_or_Medical_Supply_PDI_4',
'Recipient_Postal_Code',
'Teaching_Hospital_ID',
'Teaching_Hospital_CCN',
'Teaching_Hospital_Name',
'Covered_Recipient_License_State_code3',
'Covered_Recipient_Name_Suffix',
'State_of_Travel',
'Name_of_Third_Party_Entity_Receiving_Payment_or_Transfer_of_Value',
'Country_of_Travel',
'City_of_Travel',
'Covered_Recipient_License_State_code2',
'Third_Party_Equals_Covered_Recipient_Indicator',
'Associated_Drug_or_Biological_NDC_4',
'Product_Category_or_Therapeutic_Area_4',
'Name_of_Drug_or_Biological_or_Device_or_Medical_Supply_4',
'Covered_or_Noncovered_Indicator_4',
'Indicate_Drug_or_Biological_or_Device_or_Medical_Supply_4',
'Associated_Device_or_Medical_Supply_PDI_3',
'Indicate_Drug_or_Biological_or_Device_or_Medical_Supply_3',
'Product_Category_or_Therapeutic_Area_3',
'Name_of_Drug_or_Biological_or_Device_or_Medical_Supply_3',
'Covered_or_Noncovered_Indicator_3',
'Associated_Device_or_Medical_Supply_PDI_2',
'Name_of_Drug_or_Biological_or_Device_or_Medical_Supply_2',
'Product_Category_or_Therapeutic_Area_2',
'Indicate_Drug_or_Biological_or_Device_or_Medical_Supply_2',
'Covered_or_Noncovered_Indicator_2',
'Contextual_Information',
'Associated_Device_or_Medical_Supply_PDI_1',
'Associated_Drug_or_Biological_NDC_1',
'Recipient_Primary_Business_Street_Address_Line2']

- 2020 Ownership Payment Data:
    - The original dataset consists of 3,377 rows and is 1.6 MB

2020 Ownership Payment Data				
Column #	Columns:	Example/Contents	Missing Value Count (10,000)	Dtype
0	Change_Type	UNCHANGED	0	object
1	Physician_Profile_ID	1187636	0	int64
2	Physician_NPI	1962422717	0	int64
3	Physician_First_Name	Dhananjaya	0	object
4	Physician_Last_Name	Kamisetti	0	object
5	Recipient_Primary_Business_Street_Address_Line1	30 S Cayuga Rd	0	object
6	Recipient_City	Williamsville	0	object
7	Recipient_State	NY	0	object
8	Recipient_Zip_Code	14221	0	object
9	Recipient_Country	United States	0	object
10	Physician_Primary_Type	Medical Doctor	77	object
11	Physician_Specialty	Allopathic & Osteopathic Physicians|Anesthesio...	0	object
12	Record_ID	750018311	0	int64
13	Total_Amount_Invested_USDollars	115295	0	float64
14	Value_of_Interest	656068	0	float64
15	Terms_of_Interest	Common and Preferred Shares	31	object
16	Submitting_Applicable_Manufacturer_or_Applicable_GPO_Name	Romark Laboratories, LC	0	object
17	Applicable_Manufacturer_or_Applicable_GPO_Making_Payment_ID	1E+11	0	int64
18	Applicable_Manufacturer_or_Applicable_GPO_Making_Payment_Name	Romark Laboratories, LC	0	object
19	Applicable_Manufacturer_or_Applicable_GPO_Making_Payment_State	FL	18	object
20	Applicable_Manufacturer_or_Applicable_GPO_Making_Payment_Country	United States	0	object
21	Dispute_Status_for_Publication	No	0	object
22	Interest_Held_by_Physician_or_an_Immediate_Family_Member	Physician Covered Recipient	0	object

- The following columns were removed from the 2020 Ownership Payment data set:

2020 Ownership Dropped Columns
Physician_Middle_Name',
'Recipient_Primary_Business_Street_Address_Line2',
'Recipient_Province',
'Recipient_Postal_Code',
'Physician_Name_Suffix'

- 2021 Ownership Payment Data:
    - The original dataset consists of 3,129 rows and is 1.3 MB

2021 Ownership Payment Data				
Column #	Columns:	Example/Contents	Missing Value Count (10,000)	Dtype
0	Change_Type	UNCHANGED	0	object
1	Physician_Profile_ID	1187636	0	int64
2	Physician_NPI	1962422717	0	int64

2021 Ownership Payment Data				
3	Physician_First_Name	Dhananjaya	0	object
4	Physician_Last_Name	Kamisetti	0	object
5	Recipient_Primary_Business_Street_Address_Line1	30 S Cayuga Rd	0	object
6	Recipient_City	Williamsville	0	object
7	Recipient_State	NY	0	object
8	Recipient_Zip_Code	14221	0	object
9	Recipient_Country	United States	0	object
10	Physician_Primary_Type	Medical Doctor	0	object
11	Physician_Specialty	Allopathic & Osteopathic Physicians|Anesthesio...	92	object
12	Record_ID	750018311	0	int64
13	Total_Amount_Invested_USDollars	115295	0	float64
14	Value_of_Interest	656068	0	float64
15	Terms_of_Interest	Common and Preferred Shares	0	object
16	Submitting_Applicable_Manufacturer_or_Applicable_GPO_Name	Romark Laboratories, LC	0	object
17	Applicable_Manufacturer_or_Applicable_GPO_Making_Payment_ID	1E+11	0	int64
18	Applicable_Manufacturer_or_Applicable_GPO_Making_Payment_Name	Romark Laboratories, LC	0	object
19	Applicable_Manufacturer_or_Applicable_GPO_Making_Payment_State	FL	29	object
20	Applicable_Manufacturer_or_Applicable_GPO_Making_Payment_Country	United States	0	object
21	Dispute_Status_for_Publication	No	0	object
22	Interest_Held_by_Physician_or_an_Immediate_Family_Member	Physician Covered Recipient	0	object

- The following columns were removed from the 2021 Ownership Payment data set:

2021 Ownership Dropped Columns
'Recipient_Primary_Business_Street_Address_Line2',
'Physician_Middle_Name',
'Recipient_Province',
'Recipient_Postal_Code',
'Physician_Name_Suffix'

- 2020 Research Payment Data:
    - The original dataset consists of 654,771 rows and is 580.4 MB

2021 Research Payment Data				
Column #	Columns:	Example/Contents	Missing Value Count (10,000)	Dtype
0	Change_Type	UNCHANGED	0	object
1	Covered_Recipient_Type	Covered Recipient Physician	0	object
2	Teaching_Hospital_CCN	240010	2013	float64
3	Teaching_Hospital_ID	0	0	int64

2021 Research Payment Data				
4	Teaching_Hospital_Name	MAYO CLINIC HOSPITAL ROCHESTER	2013	object
5	Recipient_Primary_Business_Street_Address_Line1	10796 Pines Blvd Ste 205	0	object
6	Recipient_City	Pembroke Pines	0	object
7	Recipient_State	FL	0	object
8	Recipient_Zip_Code	33026	0	object
9	Recipient_Country	United States	0	object
10	Principal_Investigator_1_Covered_Recipient_Type	Covered Recipient Teaching Hospital	2427	object
11	Principal_Investigator_1_Profile_ID	768084	0	int64
12	Principal_Investigator_1_NPI	1215062989	0	int64
13	Principal_Investigator_1_First_Name	TUFIA	2427	object
14	Principal_Investigator_1_Last_Name	HADDAD	2427	object
15	Principal_Investigator_1_Business_Street_Address_Line1	200 1ST ST SW	2427	object
16	Principal_Investigator_1_City	ROCHESTER	2427	object
17	Principal_Investigator_1_State	MN	2427	object
18	Principal_Investigator_1_Zip_Code	55905-0001	2427	object
19	Principal_Investigator_1_Country	United States	2427	object
20	Principal_Investigator_1_Primary_Type_1	Medical Doctor	2427	object
21	Principal_Investigator_1_Specialty_1	Allopathic & Osteopathic Physicians|Internal M...	2427	object
22	Principal_Investigator_1_License_State_code1	MN	2427	object
23	Submitting_Applicable_Manufacturer_or_Applicable_GPO_Name	Sanofi and Genzyme US Companies	0	object
24	Applicable_Manufacturer_or_Applicable_GPO_Making_Payment_ID	1.00E+11	0	float64
25	Applicable_Manufacturer_or_Applicable_GPO_Making_Payment_Name	SANOFI US SERVICES INC.	0	object
26	Applicable_Manufacturer_or_Applicable_GPO_Making_Payment_State	NJ	1318	object
27	Applicable_Manufacturer_or_Applicable_GPO_Making_Payment_Country	United States	0	object
28	Related_Product_Indicator	Yes	0	object
29	Covered_or_Noncovered_Indicator_1	Non-Covered	1488	object
30	Total_Amount_of_Payment_USDollars	22499.91	0	float64
31	Date_of_Payment	8/3/20	0	datetime64[ns]
32	Form_of_Payment_or_Transfer_of_Value	Cash or cash equivalent	0	object
33	Preclinical_Research_Indicator	No	0	object
34	Delay_in_Publication_Indicator	No	0	object
35	Name_of_Study	A 56-week, Multicenter, Open-label, Active-con...	171	object
36	Dispute_Status_for_Publication	No	0	object
37	Record_ID	744032425	0	int64

- The following columns were removed from the 2020 Research Payment data set:

2020 Research Dropped Columns
Principal_Investigator_1_Business_Street_Address_Line2',
'Associated_Drug_or_Biological_NDC_1',
'Principal_Investigator_1_Middle_Name',
'Name_of_Drug_or_Biological_or_Device_or_Medical_Supply_1',
'Product_Category_or_Therapeutic_Area_1',

'Indicate_Drug_or_Biological_or_Device_or_Medical_Supply_1',
'Noncovered_Recipient_Entity_Name',
'Principal_Investigator_3_Province',
'Principal_Investigator_4_Name_Suffix',
'Principal_Investigator_4_Business_Street_Address_Line1',
'Principal_Investigator_4_Business_Street_Address_Line2',
'Principal_Investigator_4_City',
'Principal_Investigator_4_State',
'Principal_Investigator_4_Zip_Code',
'Principal_Investigator_4_Country',
'Principal_Investigator_4_Province',
'Principal_Investigator_4_Postal_Code',
'Principal_Investigator_4_Primary_Type_1',
'Principal_Investigator_4_Primary_Type_2',
'Principal_Investigator_4_Primary_Type_3',
'Principal_Investigator_4_Primary_Type_4',
'Principal_Investigator_4_Primary_Type_5',
'Principal_Investigator_4_Primary_Type_6',
'Principal_Investigator_4_Specialty_1',
'Principal_Investigator_4_Specialty_2',
'Principal_Investigator_4_Specialty_3',
'Principal_Investigator_4_Specialty_4',
'Principal_Investigator_4_Last_Name',
'Principal_Investigator_4_Middle_Name',
'Principal_Investigator_4_First_Name',
'Principal_Investigator_3_Specialty_3',
'Principal_Investigator_3_Primary_Type_1',
'Principal_Investigator_3_Primary_Type_2',
'Principal_Investigator_3_Primary_Type_3',
'Principal_Investigator_3_Primary_Type_4',
'Principal_Investigator_3_Primary_Type_5',
'Principal_Investigator_3_Primary_Type_6',
'Principal_Investigator_3_Specialty_1',
'Principal_Investigator_3_Specialty_2',
'Principal_Investigator_3_Specialty_4',
'Principal_Investigator_4_NPI',
'Principal_Investigator_3_Specialty_5',
'Principal_Investigator_3_Specialty_6',
'Principal_Investigator_3_License_State_code2',
'Principal_Investigator_3_License_State_code3',
'Principal_Investigator_3_License_State_code4',
'Principal_Investigator_3_License_State_code5',
'Principal_Investigator_4_Covered_Recipient_Type',
'Principal_Investigator_4_Profile_ID',
'Principal_Investigator_4_Specialty_5',
'Principal_Investigator_4_Specialty_6',
'Principal_Investigator_4_License_State_code1',
'Principal_Investigator_5_License_State_code1',
'Principal_Investigator_5_Primary_Type_5',
'Principal_Investigator_5_Primary_Type_6',
'Principal_Investigator_5_Specialty_1',
'Principal_Investigator_5_Specialty_2',
'Principal_Investigator_5_Specialty_3',

'Principal_Investigator_5_Specialty_4',
'Principal_Investigator_5_Specialty_5',
'Principal_Investigator_5_Specialty_6',
'Principal_Investigator_5_License_State_code2',
'Principal_Investigator_5_Primary_Type_3',
'Principal_Investigator_5_License_State_code3',
'Principal_Investigator_5_License_State_code4',
'Principal_Investigator_5_License_State_code5',
'Associated_Device_or_Medical_Supply_PDI_1',
'Associated_Device_or_Medical_Supply_PDI_2',
'Associated_Device_or_Medical_Supply_PDI_3',
'Associated_Device_or_Medical_Supply_PDI_4',
'Associated_Device_or_Medical_Supply_PDI_5',
'Principal_Investigator_5_Primary_Type_4',
'Principal_Investigator_5_Primary_Type_2',
'Principal_Investigator_4_License_State_code2',
'Principal_Investigator_5_Last_Name',
'Principal_Investigator_4_License_State_code3',
'Principal_Investigator_4_License_State_code4',
'Principal_Investigator_4_License_State_code5',
'Principal_Investigator_5_Covered_Recipient_Type',
'Principal_Investigator_5_Profile_ID',
'Principal_Investigator_5_NPI',
'Principal_Investigator_5_First_Name',
'Principal_Investigator_5_Middle_Name',
'Principal_Investigator_5_Name_Suffix',
'Principal_Investigator_5_Primary_Type_1',
'Principal_Investigator_5_Business_Street_Address_Line1',
'Principal_Investigator_5_Business_Street_Address_Line2',
'Principal_Investigator_5_City',
'Principal_Investigator_5_State',
'Principal_Investigator_5_Zip_Code',
'Principal_Investigator_5_Country',
'Principal_Investigator_5_Province',
'Principal_Investigator_5_Postal_Code',
'Principal_Investigator_3_Postal_Code',
'Principal_Investigator_3_License_State_code1',
'Principal_Investigator_3_Country',
'Principal_Investigator_1_Province',
'Principal_Investigator_2_Name_Suffix',
'Principal_Investigator_1_License_State_code5',
'Principal_Investigator_1_Specialty_6',
'Principal_Investigator_1_Specialty_5',
'Principal_Investigator_1_Specialty_4',
'Principal_Investigator_1_Specialty_3',
'Principal_Investigator_1_Specialty_2',
'Principal_Investigator_1_Primary_Type_6',
'Principal_Investigator_1_Primary_Type_5',
'Principal_Investigator_1_Primary_Type_4',
'Principal_Investigator_1_Primary_Type_3',
'Principal_Investigator_3_Zip_Code',
'Principal_Investigator_1_Postal_Code',
'Covered_Recipient_License_State_code5',

'Principal_Investigator_2_Postal_Code',
'Covered_Recipient_License_State_code4',
'Covered_Recipient_Specialty_6',
'Covered_Recipient_Specialty_5',
'Covered_Recipient_Specialty_4',
'Covered_Recipient_Specialty_3',
'Covered_Recipient_Specialty_2',
'Covered_Recipient_Primary_Type_6',
'Covered_Recipient_Primary_Type_5',
'Covered_Recipient_Primary_Type_4',
'Covered_Recipient_Primary_Type_3',
'Covered_Recipient_Primary_Type_2',
'Recipient_Postal_Code',
'Recipient_Province',
'Principal_Investigator_2_Province',
'Principal_Investigator_1_Primary_Type_2',
'Principal_Investigator_3_Last_Name',
'Principal_Investigator_2_Specialty_2',
'Principal_Investigator_2_Specialty_4',
'Principal_Investigator_2_Specialty_5',
'Principal_Investigator_2_Specialty_6',
'Principal_Investigator_2_License_State_code2',
'Principal_Investigator_2_License_State_code3',
'Principal_Investigator_2_License_State_code4',
'Principal_Investigator_2_License_State_code5',
'Principal_Investigator_3_Covered_Recipient_Type',
'Principal_Investigator_3_Profile_ID',
'Principal_Investigator_3_NPI',
'Principal_Investigator_3_First_Name',
'Principal_Investigator_3_Middle_Name',
'Principal_Investigator_3_Name_Suffix',
'Principal_Investigator_3_Business_Street_Address_Line1',
'Principal_Investigator_3_Business_Street_Address_Line2',
'Principal_Investigator_3_City',
'Principal_Investigator_3_State',
'Principal_Investigator_2_Specialty_3',
'Unnamed: 252',
'Principal_Investigator_2_Primary_Type_5',
'Principal_Investigator_2_Primary_Type_4',
'Principal_Investigator_2_Primary_Type_2',
'Principal_Investigator_2_Primary_Type_3',
'Principal_Investigator_2_Primary_Type_6',
'Associated_Drug_or_Biological_NDC_5',
'Expenditure_Category6',
'Expenditure_Category5',
'Principal_Investigator_1_License_State_code4',
'Principal_Investigator_2_Business_Street_Address_Line2',
'Expenditure_Category4',
'Expenditure_Category3',
'Expenditure_Category2',
'Covered_Recipient_License_State_code3',
'Associated_Drug_or_Biological_NDC_4',
'Principal_Investigator_2_Middle_Name',

'Principal_Investigator_1_License_State_code3',
'Name_of_Drug_or_Biological_or_Device_or_Medical_Supply_5',
'Product_Category_or_Therapeutic_Area_5',
'Indicate_Drug_or_Biological_or_Device_or_Medical_Supply_5',
'Covered_or_Noncovered_Indicator_5',
'Principal_Investigator_2_City',
'Principal_Investigator_2_Zip_Code',
'Principal_Investigator_2_State',
'Principal_Investigator_2_Specialty_1',
'Principal_Investigator_2_Primary_Type_1',
'Principal_Investigator_2_NPI',
'Principal_Investigator_2_Country',
'Principal_Investigator_2_License_State_code1',
'Principal_Investigator_2_Covered_Recipient_Type',
'Principal_Investigator_2_Profile_ID',
'Principal_Investigator_2_Business_Street_Address_Line1',
'Principal_Investigator_2_Last_Name',
'Principal_Investigator_2_First_Name',
'Associated_Drug_or_Biological_NDC_3',
'Covered_Recipient_License_State_code2',
'Principal_Investigator_1_Name_Suffix',
'Covered_Recipient_Name_Suffix',
'Indicate_Drug_or_Biological_or_Device_or_Medical_Supply_4',
'Covered_or_Noncovered_Indicator_4',
'Product_Category_or_Therapeutic_Area_4',
'Name_of_Drug_or_Biological_or_Device_or_Medical_Supply_4',
'Name_of_Drug_or_Biological_or_Device_or_Medical_Supply_3',
'Product_Category_or_Therapeutic_Area_3',
'Indicate_Drug_or_Biological_or_Device_or_Medical_Supply_3',
'Covered_or_Noncovered_Indicator_3',
'Associated_Drug_or_Biological_NDC_2',
'Indicate_Drug_or_Biological_or_Device_or_Medical_Supply_2',
'Name_of_Drug_or_Biological_or_Device_or_Medical_Supply_2',
'Product_Category_or_Therapeutic_Area_2',
'Principal_Investigator_1_License_State_code2',
'Covered_or_Noncovered_Indicator_2',
'Expenditure_Category1',
'Research_Information_Link',
'Recipient_Primary_Business_Street_Address_Line2',
'Covered_Recipient_Middle_Name',
'Covered_Recipient_NPI',
'Covered_Recipient_First_Name',
'Covered_Recipient_Profile_ID',
'Covered_Recipient_Specialty_1',
'Covered_Recipient_License_State_code1',
'Covered_Recipient_Last_Name',
'Covered_Recipient_Primary_Type_1',
'ClinicalTrials_Gov_Identifier',
'Context_of_Research'

- 2021 Research Payment Data:
    - The original dataset consists of 672,765 rows and is 712 MB

2021 Ownership Payment Data

Column #	Columns:	Example/Contents	Missing Value Count (10,000)	Dtype
0	Change_Type	UNCHANGED	0	object
1	Covered_Recipient_Type	Covered Recipient Teaching Hospital	0	object
2	Teaching_Hospital_CCN	230297	1753	float64
3	Teaching_Hospital_ID	0	0	int64
4	Teaching_Hospital_Name	KARMANOS CANCER HOSPITAL	1753	object
5	Recipient_Primary_Business_Street_Address_Line1	4100 John R St	0	object
6	Recipient_City	Detroit	0	object
7	Recipient_State	MI	0	object
8	Recipient_Zip_Code	48201	0	object
9	Recipient_Country	United States	0	object
10	Principal_Investigator_1_Covered_Recipient_Type	Covered Recipient Physician	2086	object
11	Principal_Investigator_1_Profile_ID	0	0	int64
12	Principal_Investigator_1_NPI	0	0	int64
13	Principal_Investigator_1_First_Name	HADEEL	2088	object
14	Principal_Investigator_1_Last_Name	ASSAD	2086	object
15	Principal_Investigator_1_Business_Street_Address_Line1	4100 JOHN R ST	2086	object
16	Principal_Investigator_1_City	DETROIT	2086	object
17	Principal_Investigator_1_State	MI	2086	object
18	Principal_Investigator_1_Zip_Code	48201-2013	2086	object
19	Principal_Investigator_1_Country	United States	2086	object
20	Principal_Investigator_1_Primary_Type_1	Medical Doctor	2086	object
21	Principal_Investigator_1_Specialty_1	Allopathic & Osteopathic Physicians|Internal M...	2086	object
22	Principal_Investigator_1_License_State_code1	MI	2086	object
23	Submitting_Applicable_Manufacturer_or_Applicable_GPO_Name	Takeda Pharmaceuticals U.S.A., Inc.	0	object
24	Applicable_Manufacturer_or_Applicable_GPO_Making_Payment_ID	1E+11	0	int64
25	Applicable_Manufacturer_or_Applicable_GPO_Making_Payment_Name	Takeda Pharmaceuticals U.S.A., Inc.	0	object
26	Applicable_Manufacturer_or_Applicable_GPO_Making_Payment_State	IL	1454	object
27	Applicable_Manufacturer_or_Applicable_GPO_Making_Payment_Country	United States	0	object
28	Related_Product_Indicator	Yes	0	object
29	Covered_or_Noncovered_Indicator_1	Non-Covered	1344	object
30	Total_Amount_of_Payment_USDollars	31951.26	0	float64
31	Date_of_Payment	9/29/21	0	datetime64[ns]
32	Form_of_Payment_or_Transfer_of_Value	Cash or cash equivalent	0	object
33	Preclinical_Research_Indicator	No	0	object
34	Delay_in_Publication_Indicator	No	0	object
35	Name_of_Study	A STUDY TO EVALUATE THE SAFETY, TOLERABILITY A...	143	object
36	Dispute_Status_for_Publication	No	0	object
37	Record_ID	816658237	0	int64

- The following columns were removed from the 2021 Research Payment data set:

2021 Resarch Dropped Columns
'ClinicalTrials_Gov_Identifier',
'Associated_Drug_or_Biological_NDC_1',
'Principal_Investigator_1_Middle_Name',
'Name_of_Drug_or_Biological_or_Device_or_Medical_Supply_1',
'Product_Category_or_Therapeutic_Area_1',
'Indicate_Drug_or_Biological_or_Device_or_Medical_Supply_1',
'Noncovered_Recipient_Entity_Name',
'Principal_Investigator_4_Middle_Name',
'Principal_Investigator_4_Business_Street_Address_Line2',
'Principal_Investigator_4_Province',
'Principal_Investigator_4_Postal_Code',
'Principal_Investigator_4_Primary_Type_2',
'Principal_Investigator_4_Primary_Type_3',
'Principal_Investigator_4_Primary_Type_4',
'Principal_Investigator_4_Primary_Type_5',
'Principal_Investigator_4_Primary_Type_6',
'Principal_Investigator_4_Specialty_2',
'Principal_Investigator_4_Specialty_3',
'Principal_Investigator_4_Specialty_4',
'Principal_Investigator_4_Specialty_5',
'Principal_Investigator_4_Specialty_6',
'Principal_Investigator_4_License_State_code2',
'Principal_Investigator_4_License_State_code3',
'Principal_Investigator_4_Name_Suffix',
'Principal_Investigator_3_License_State_code5',
'Principal_Investigator_2_License_State_code4',
'Principal_Investigator_3_License_State_code4',
'Principal_Investigator_3_Name_Suffix',
'Principal_Investigator_3_Business_Street_Address_Line2',
'Principal_Investigator_3_Province',
'Principal_Investigator_3_Postal_Code',
'Principal_Investigator_3_Primary_Type_2',
'Principal_Investigator_3_Primary_Type_3',
'Principal_Investigator_3_Primary_Type_4',
'Principal_Investigator_3_Primary_Type_5',
'Principal_Investigator_3_Primary_Type_6',
'Principal_Investigator_3_Specialty_2',
'Principal_Investigator_3_Specialty_3',
'Principal_Investigator_3_Specialty_4',
'Principal_Investigator_3_Specialty_5',
'Principal_Investigator_3_License_State_code2',
'Principal_Investigator_3_License_State_code3',
'Principal_Investigator_4_License_State_code4',
'Principal_Investigator_4_License_State_code5',
'Principal_Investigator_5_Covered_Recipient_Type',
'Principal_Investigator_5_Profile_ID',
'Principal_Investigator_5_Primary_Type_6',
'Principal_Investigator_5_Specialty_1',
'Principal_Investigator_5_Specialty_2',
'Principal_Investigator_5_Specialty_3',

'Principal_Investigator_5_Specialty_4',
'Principal_Investigator_5_Specialty_5',
'Principal_Investigator_5_Specialty_6',
'Principal_Investigator_5_License_State_code1',
'Principal_Investigator_5_License_State_code2',
'Principal_Investigator_5_License_State_code3',
'Principal_Investigator_5_License_State_code4',
'Principal_Investigator_5_License_State_code5',
'Expenditure_Category4',
'Expenditure_Category5',
'Expenditure_Category6',
'Principal_Investigator_5_Primary_Type_5',
'Principal_Investigator_5_Primary_Type_4',
'Principal_Investigator_5_Primary_Type_3',
'Principal_Investigator_5_Business_Street_Address_Line2',
'Principal_Investigator_5_NPI',
'Principal_Investigator_5_First_Name',
'Principal_Investigator_5_Middle_Name',
'Principal_Investigator_5_Last_Name',
'Principal_Investigator_5_Name_Suffix',
'Principal_Investigator_5_Business_Street_Address_Line1',
'Principal_Investigator_5_City',
'Principal_Investigator_5_Primary_Type_2',
'Principal_Investigator_5_State',
'Principal_Investigator_5_Zip_Code',
'Principal_Investigator_5_Country',
'Principal_Investigator_5_Province',
'Principal_Investigator_5_Postal_Code',
'Principal_Investigator_5_Primary_Type_1',
'Principal_Investigator_2_License_State_code5',
'Principal_Investigator_3_Specialty_6',
'Principal_Investigator_2_License_State_code3',
'Principal_Investigator_2_Postal_Code',
'Principal_Investigator_2_Name_Suffix',
'Principal_Investigator_1_Specialty_6',
'Principal_Investigator_1_Specialty_5',
'Principal_Investigator_1_Specialty_4',
'Principal_Investigator_1_Specialty_2',
'Principal_Investigator_1_Primary_Type_6',
'Principal_Investigator_1_Primary_Type_5',
'Principal_Investigator_1_Primary_Type_4',
'Principal_Investigator_1_Primary_Type_3',
'Principal_Investigator_1_Primary_Type_2',
'Principal_Investigator_1_Postal_Code',
'Principal_Investigator_1_Province',
'Covered_Recipient_License_State_code5',
'Covered_Recipient_Specialty_6',
'Covered_Recipient_Specialty_5',
'Covered_Recipient_Specialty_4',
'Covered_Recipient_Specialty_3',
'Covered_Recipient_Specialty_2',
'Covered_Recipient_Primary_Type_6',
'Covered_Recipient_Primary_Type_5',

'Covered_Recipient_Primary_Type_4',
'Covered_Recipient_Primary_Type_3',
'Covered_Recipient_Primary_Type_2',
'Recipient_Postal_Code',
'Recipient_Province',
'Principal_Investigator_2_Province',
'Principal_Investigator_1_Specialty_3',
'Principal_Investigator_2_Primary_Type_2',
'Principal_Investigator_2_Specialty_2',
'Principal_Investigator_2_Primary_Type_5',
'Principal_Investigator_2_Specialty_3',
'Principal_Investigator_2_Primary_Type_4',
'Principal_Investigator_2_Primary_Type_3',
'Principal_Investigator_2_Specialty_4',
'Principal_Investigator_2_Specialty_5',
'Principal_Investigator_2_Specialty_6',
'Principal_Investigator_2_Primary_Type_6',
'Principal_Investigator_4_Last_Name',
'Principal_Investigator_4_NPI',
'Principal_Investigator_4_Profile_ID',
'Principal_Investigator_4_Specialty_1',
'Principal_Investigator_4_License_State_code1',
'Principal_Investigator_4_Covered_Recipient_Type',
'Principal_Investigator_4_First_Name',
'Principal_Investigator_4_Primary_Type_1',
'Expenditure_Category2',
'Principal_Investigator_2_License_State_code2',
'Principal_Investigator_4_Country',
'Principal_Investigator_4_Zip_Code',
'Principal_Investigator_4_State',
'Principal_Investigator_4_City',
'Expenditure_Category3',
'Principal_Investigator_4_Business_Street_Address_Line1',
'Principal_Investigator_3_Profile_ID',
'Principal_Investigator_3_Primary_Type_1',
'Principal_Investigator_3_Covered_Recipient_Type',
'Principal_Investigator_3_State',
'Covered_Recipient_License_State_code4',
'Principal_Investigator_3_Zip_Code',
'Principal_Investigator_3_First_Name',
'Principal_Investigator_3_City',
'Principal_Investigator_3_Business_Street_Address_Line1',
'Principal_Investigator_3_NPI',
'Principal_Investigator_3_Last_Name',
'Principal_Investigator_3_Middle_Name',
'Principal_Investigator_3_License_State_code1',
'Principal_Investigator_3_Country',
'Principal_Investigator_3_Specialty_1',
'Associated_Drug_or_Biological_NDC_5',
'Principal_Investigator_1_License_State_code5',
'Principal_Investigator_2_Business_Street_Address_Line2',
'Covered_Recipient_License_State_code3',
'Principal_Investigator_1_License_State_code4',

'Associated_Drug_or_Biological_NDC_4',
'Principal_Investigator_2_Middle_Name',
'Associated_Drug_or_Biological_NDC_3',
'Principal_Investigator_2_City',
'Principal_Investigator_2_Business_Street_Address_Line1',
'Principal_Investigator_2_License_State_code1',
'Principal_Investigator_2_Country',
'Principal_Investigator_2_Last_Name',
'Principal_Investigator_2_Specialty_1',
'Principal_Investigator_2_First_Name',
'Principal_Investigator_2_NPI',
'Principal_Investigator_2_Profile_ID',
'Principal_Investigator_2_Zip_Code',
'Principal_Investigator_2_Primary_Type_1',
'Principal_Investigator_2_Covered_Recipient_Type',
'Principal_Investigator_2_State',
'Principal_Investigator_1_Name_Suffix',
'Covered_Recipient_Name_Suffix',
'Principal_Investigator_1_License_State_code3',
'Covered_Recipient_License_State_code2',
'Expenditure_Category1',
'Associated_Drug_or_Biological_NDC_2',
'Associated_Device_or_Medical_Supply_PDI_5',
'Associated_Device_or_Medical_Supply_PDI_4',
'Name_of_Drug_or_Biological_or_Device_or_Medical_Supply_5',
'Product_Category_or_Therapeutic_Area_5',
'Associated_Device_or_Medical_Supply_PDI_3',
'Indicate_Drug_or_Biological_or_Device_or_Medical_Supply_5',
'Covered_or_Noncovered_Indicator_5',
'Name_of_Drug_or_Biological_or_Device_or_Medical_Supply_4',
'Product_Category_or_Therapeutic_Area_4',
'Indicate_Drug_or_Biological_or_Device_or_Medical_Supply_4',
'Covered_or_Noncovered_Indicator_4',
'Associated_Device_or_Medical_Supply_PDI_2',
'Name_of_Drug_or_Biological_or_Device_or_Medical_Supply_3',
'Product_Category_or_Therapeutic_Area_3',
'Indicate_Drug_or_Biological_or_Device_or_Medical_Supply_3',
'Covered_or_Noncovered_Indicator_3',
'Name_of_Drug_or_Biological_or_Device_or_Medical_Supply_2',
'Product_Category_or_Therapeutic_Area_2',
'Principal_Investigator_1_License_State_code2',
'Indicate_Drug_or_Biological_or_Device_or_Medical_Supply_2',
'Covered_or_Noncovered_Indicator_2',
'Research_Information_Link',
'Recipient_Primary_Business_Street_Address_Line2',
'Covered_Recipient_Middle_Name',
'Associated_Device_or_Medical_Supply_PDI_1',
'Covered_Recipient_NPI',
'Covered_Recipient_First_Name',
'Covered_Recipient_License_State_code1',
'Covered_Recipient_Last_Name',
'Covered_Recipient_Primary_Type_1',
'Covered_Recipient_Specialty_1',

'Covered_Recipient_Profile_ID',
'Context_of_Research',
'Principal_Investigator_1_Business_Street_Address_Line2'

- Physician Scores/Rating
    - The original dataset consists of 698,884 data points and is 51.4 MB

Physician Ratings				
Column #	Columns:	Example/Contents	Missing Value Count (10000)	Dtype
0	NPI	1003028556	0	int64
1	Org_PAC_ID	143209809	1749	float64
2	lst_nm	HEIDARI	0	object
3	frst_nm	NEDA	1	object
4	source	individual	0	object
5	Quality_category_score	94.641	2773	float64
6	IA_category_score	20	2647	float64
7	final_MIPS_score_without_CPB	76.9526	0	float64
8	final_MIPS_score	82.1725	0	float64

- The following columns were removed from the Physician Ratings data set:

Physican Ratings Dropped Columns
'Covered_Recipient_Profile_Alternate_Middle_Name',
'Covered_Recipient_Profile_Address_Line_2',
'Covered_Recipient_Profile_License_State_Code_2',
'Covered_Recipient_Profile_Alternate_First_Name',
'Covered_Recipient_Profile_Alternate_Last_Name',
'Associated_Covered_Recipient_Profile_ID_2',
'Covered_Recipient_Profile_Province_Name',
'Associated_Covered_Recipient_Profile_ID_1',
'Covered_Recipient_Profile_OPS_Taxonomy_6',
'Covered_Recipient_Profile_OPS_Taxonomy_5',
'Covered_Recipient_Profile_OPS_Taxonomy_4',
'Covered_Recipient_Profile_Alternate_Suffix',
'Covered_Recipient_Profile_Suffix',
'Covered_Recipient_Profile_License_State_Code_5',
'Covered_Recipient_Profile_OPS_Taxonomy_3',
'Covered_Recipient_Profile_License_State_Code_4',
'Covered_Recipient_Profile_License_State_Code_3',
'Covered_Recipient_Profile_OPS_Taxonomy_2'

- Physician Profile Supplement
    - The original dataset consists of 1,048,575 data points and is 346.6 MB

2021 Ownership Payment Data				
Column #	Columns:	Example/Contents	Missing Value Count	Dtype
0	Covered_Recipient_Profile_Type	Covered Recipient Physician	0	object

1	Covered_Recipient_Profile_ID	87870	0	int64
2	Covered_Recipient_NPI	1.94E+09	2	float64
3	Covered_Recipient_Profile_First_Name	EDWARD	0	object
4	Covered_Recipient_Profile_Middle_Name	W	2740	object
5	Covered_Recipient_Profile_Last_Name	HAMILTON	1	object
6	Covered_Recipient_Profile_Address_Line_1	4215 WOODRUFF RD	0	object
7	Covered_Recipient_Profile_City	COLUMBUS	0	object
8	Covered_Recipient_Profile_State	GA	0	object
9	Covered_Recipient_Profile_Zipcode	31904-6889	0	object
10	Covered_Recipient_Profile_Country_Name	UNITED STATES	0	object
11	Covered_Recipient_Profile_Primary_Specialty	Allopathic & Osteopathic Physicians|Family Med...	285	object
12	Covered_Recipient_Profile_OPS_Taxonomy_1	207Q00000X	9	object
13	Covered_Recipient_Profile_License_State_Code_1	GA	0	object

- The following columns were removed from the Physcian Profile data set:

Physican Profile Dropped Columns
' PI_category_score',
' facility_ccn',
' facility_lbn',
' Cost_category_score'

C. Data transformation and integration:

- The following steps were taken to clean/process the data:
  - ➢ Identify the data quality issues:
    - ○ Began by assessing the quality of the data and identifying potential issues. These issues included missing values, duplicate entries, inconsistent formats, outliers, or incorrect data types.
  - ➢ Handle missing data:
    - ○ Complete Case Analysis. In this approach, columns that had a high percentage of missing values, greater than 30%, were entirely removed from the dataset. This method was used because the missing values were believed to be missing completely at random and the columns with missing values were not considered crucial for the analysis or did not contain substantial information
  - ➢ Remove duplicates:
    - ○ Duplicates can skew analysis and lead to incorrect conclusions. Identify and remove duplicate entries from the dataset.
    - ○ There were no duplicates or significant outliers present in the data sets
  - ➢ Address inconsistent formats:
    - ○ The following data types were accessed and changed:

            - ID numbers were converted to integers. Generally it is more efficient to use integer data type for columns containing ID numbers with no decimals. The integer data type requires less storage space compared to float64, as it does not store decimal places. Additionally, integer operations tend to be faster and more efficient than floating-point operations.
            - Dates of payments were converted to "datetime" from strings. String data types for dates can limit the ability to perform date-specific operations and calculations.
            - remove any other unnecessary columns

- Inspect for data errors:
    - Outlier is in the data were assessed/viewed utilizing z-scores.
    - There were no outliers needing to be removed in the data sets
    - There were no duplicate entries in the datasets
    - Standardization of zip codes across data sets

- Validate data integrity:
    - The following step were taken to validate the data:
        - Data profiling
        - Data validation checks
            - format and data types checks
            - Consistency checks
    - CMS.gov is the official website of the Centers for Medicare & Medicaid Services (CMS), a federal agency in the U.S. Department of Health and Human Services. CMS is responsible for administering healthcare programs such as Medicare and Medicaid. The data provided on CMS.gov is sourced from various programs and initiatives and is intended to provide information on healthcare services, costs, providers, and quality measures.

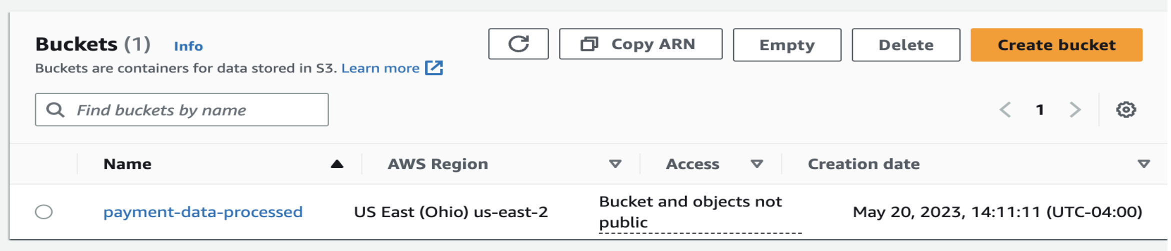
II. Data Storage Plan

A. **Source Data Storage**:

a. Process for data be storage and accessibility:

i. Initially stored on cloud drive & hard drive

1. The raw data will be uploaded on a Spark data cache for initial processing to remove unnecessary columns.
    a. The raw data will be uploaded using the CLI.
    b. We will apply our Python scripts saved in GitHub.
2. The refined data will be stored in S3 buckets.
    a. We clear the data cache if necessary.



Buckets (1) Info

Buckets are containers for data stored in S3. Learn more

	Name	AWS Region	Access	Creation date
○	payment-data-processed	US East (Ohio) us-east-2	Bucket and objects not public	May 20, 2023, 14:11:11 (UTC-04:00)

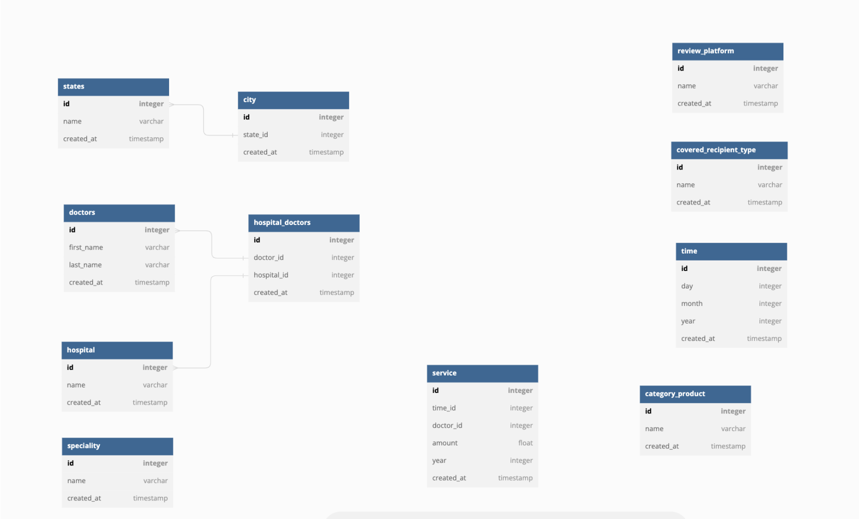
ii. Specifications:

1. Specific structure:
   a. CSV files
2. Naming conventions:
   a. 2020-general-date
   b. 2021-general-date
   c. 2020-ownership-date
   d. 2021-ownership-date
   e. 2020-research-date
   f. 2021-research-date
   g. physician-scores-date
   h. Physician-profiles-date

B. **Data Warehouse (drafted plan)**:

a. data storage plan for the warehouse : AWS S3

b. Draft data schema for data modeling: https://dbdiagram.io/d/646696f1dca9fb07c45aa900



states	
id	integer
name	varchar
created_at	timestamp

city	
id	integer
state_id	integer
created_at	timestamp

doctors	
id	integer
first_name	varchar
last_name	varchar
created_at	timestamp

hospital_doctors	
id	integer
doctor_id	integer
hospital_id	integer
created_at	timestamp

hospital	
id	integer
name	varchar
created_at	timestamp

speciality	
id	integer
name	varchar
created_at	timestamp

service	
id	integer
time_id	integer
doctor_id	integer
amount	float
year	integer
created_at	timestamp

review_platform	
id	integer
name	varchar
created_at	timestamp

covered_recipient_type	
id	integer
name	varchar
created_at	timestamp

time	
id	integer
day	integer
month	integer
year	integer
created_at	timestamp

category_product	
id	integer
name	varchar
created_at	timestamp

III. Resource table

- The table below organizes all data resources for easy access and reference
- Overall Table resource: click here

Resource Name	Type	Description	Format	Size	Source link	Storage place
2020 General Payment Data	File	This data is from CMS.gov and provides all general (non-research, non-ownership related) payments from the 2020 program year	CSV	3.7 GB	https://openpaymentsdata.cms.gov/dataset/a08c4b30-5cf3-4948-ad40-36f404619019	S3 Bucket
2021 General Payment Data	File	This data is from CMS.gov and provides all general (non-research, non-ownership related) payments from the 2021 program year	CSV	7.11 GB	https://openpaymentsdata.cms.gov/dataset/0380bbeb-aea1-58b6-b708-829f92a48202	S3 Bucket
2020 Ownership Payment Data	File	This data is from CMS.gov and provides all ownership and investment payments from the 2020 program year	CSV	1.6 MB	https://openpaymentsdata.cms.gov/dataset/a9a0bf48-6b96-4589-b4c2-3c5dcfbeaca2	S3 Bucket
2021 Ownership Payment Data	File	This data is from CMS.gov and provides all ownership and investment payments from the 2021 program year	CSV	1.3 MB	https://openpaymentsdata.cms.gov/dataset/b0c03b8d-06df-58f2-8ce2-4daeffee147e	S3 Bucket
2020 Research Payment Data	File	This data is from CMS.gov and provides research-related payments from the 2020 program year	CSV	580.4 MB	https://openpaymentsdata.cms.gov/dataset/9c248e7e-7c7f-478b-ab84-ce0919d72c1c	S3 Bucket
2021 Research Payment Data	File	This data is from CMS.gov and provides research-related payments from the 2021 program year	CSV	712 MB	https://openpaymentsdata.cms.gov/dataset/ce1d28dd-0094-5060-a036-580329439600	S3 Bucket
Physician Ratings	File	This file contains Merit-Based Incentive Payment System (MIPS) Final Scores and performance category scores for clinicians. For further details on 2021 MIPS scoring, see the 2021 Traditional MIPS Scoring Guide.	CSV	51.4 MB	https://data.cms.gov/provider-data/search?theme=Doctors%20and%20clinicians	S3 Bucket
Physician Profile Supplement	File	This data is from CMS.gov and provides information on physicians who have received payments or have ownership/investment interest	CSV	346.6 MB	https://openpaymentsdata.cms.gov/dataset/23160558-6742-54ff-8b9f-cac7d514ff4e	S3 Bucket
		Total Capacity		**12.90029 GB**		