NORTHEASTERN UNIVERSITY

Design Documentation

DATAWAREHOUSING & BUSINESS INTELLIGENCE

By

Team 2 - Data Wizards

Darshan Durve	**001898887**
Manya Raman	**001497103**
Adithya Prasad	**001476343**
Vedant Pednekar	**001434737**

Under the guidance of

Professor Vincent Lattuada

Document Revision History:

Date	Who	What	Version
3/15/2020	All Members	Initial proposal draft of the project submitted	**0.0**
3/22/2020	All Members	Final project proposal draft submitted	**1.0**
3/23/2020	All Members	Received feedback on proposal and went through it	
3/25/2020	Every member created one table	Created dim provider and dim geography. Loaded the data into staging using SSIS.	
3/27/2020	Manya and Darshan	Created the final ER diagram try to identify and identified lookups that can be used.	
3/30/2020	Darshan and Adithya	Data Profiling using SSIS	
3/30/2020	Manya and Adithya	EDA	
4/2/2020	Darshan and Vedant	Created lookups and reference tables using online datasets	
4/3/2020 – 4/4/2020	Manya, Darshan and Adithya	Started working on SSIS package to implement SCD and loaded using lookups for multiple years data.	
4/5/2020	Adithya and Manya	Continued working on SCD implementation	
4/6/2020-4/7/2020	All Members	Worked on the documentation by creating all transformation that need to be implemented in excel sheet to be submitted.	
4/8/2020	All Members	Worked on the documentation to be submitted today.	**2.0**

4/10/2020	All members	Added date dimension as recommended and made changes in the design document as recommended by the professor.	
4/11/2020- 4/12/2020	Adithya and Manya	Loaded all the datasets into staging table.	
4/13/2020-4/15/2020	Manya and Darshan	Did transformations and loaded the data into fact tables.	
4/16/2020	Darshan and Manya And Adithya	Decided to remove opioid table and add the inpatient table. Did all the process of staging and moving to fact for this table.	
4/17/2020	Adithya and Darshan	Give foreign key references and renaming tables and doing all the cleanups in SQL Server and everything.	
4/18/2020-4/19/2020	Darshan and Manya	Made OLAP cubes and hierarchies in SSAS.	
4/19/2020-4/21/2020	Darshan and Manya	Loaded data into tableau for visualization and created various dashboards.	
4/22/2020	Adithya and Vedant	Added and made required changes to the design document	
4/22/2020-4/23/2020	All Members	Made power point presentation and cleaned up everything for final submission	**3.0**
4/22/2020-4/24/2020	All Members	Gathered all the files to be submitted today and zipped and went over presentation	

Objective:

The main purpose of this project is to create a dynamic data warehouse consisting all the information on health care provider which will enable the users to identify the most optimal medical center for their healthcare needs. We hope to create comprehensive reports which provide details of the number of beneficiaries in each medical center and condition for which they are receiving treatments from the health care center. We also aim to provide users with costs associated with different medical services and drug services provided by each medical center which will allow users to understand treatment costs in each medical center. The final deliverable of this project is to create a dynamic data warehouse using ETL processes in order to create business intelligence dashboards about medical costs for various procedures and Medicare center performance across various states by doing the following:

- To analyze the amount the provider pays and the connection between the calculation of costs for medical procedures in different medical centers based on beneficiary conditions.
- To analyze the relation between the age of beneficiaries and the number of beneficiaries with chronic conditions treated under a specific provider.
- To analyze state-wise suppliers and their average payment info depending on the Durable Medical Equipment, Prosthetics, Orthotics and Supplies provided.
- Comparing the cost of medical procedures in each center with respect to the average cost for each condition of the beneficiary.
- Comparing the cost of medical equipment in each medical center with respect to the national average.
- Other ad-hoc analysis reports based on requirements.

Data Source Description:

All the datasets we selected and are described below are for multiple years from **2015 to 2017**.

A through research on the Federal Governments website for Medicare centers and Medicaid Services data provided the team with 4 data sets which align with the goals of the project, they are:

1. The Medicare Physician and Other Supplier National Provider Identifier (NPI) Aggregate Report - This subset of this file will be the one of the main fact table for the project. It contains 1.09million rows and 70 columns which includes data of all the healthcare providers of the nation, which can be drilled down to the city level. This also provides us with the exact amount each provider charged for their benefactors and the amount that was covered by Medicare and the actual amount payed by the patients which is extremely crucial for our analysis. It also contains the percentage of beneficiaries with every chronic disease which each provider has giving us the information about each provider in extreme detail.

   https://data.cms.gov/Medicare-Physician-Supplier/Medicare-Physician-and-Other-Supplier-National-Pro/n5qc-ua94

2. Medicare Referring DMEPOS HCPCS State Aggregate table - DMEPOS is short for Durable Medical Equipment, Prosthetics, Orthotics and Supplies. This dataset provides a comprehensive description of the DEMPOS product and services provided to Medicare beneficiaries ordered by physicians and other healthcare professionals. It proves to be useful by providing us with the average supplier payment amount and the average allowed Medicare amount categorized by state and DEMPOS product or service. It also provides with the number of providers offering these services to the number of suppliers rendering these services. This data set contains 40k rows and 15 columns as shown below:

https://data.cms.gov/Medicare-Durable-Medical-Equipment-DME-/Medicare-Referring-DMEPOS-HCPCS-State-Aggregate-ta/kpkt-855d

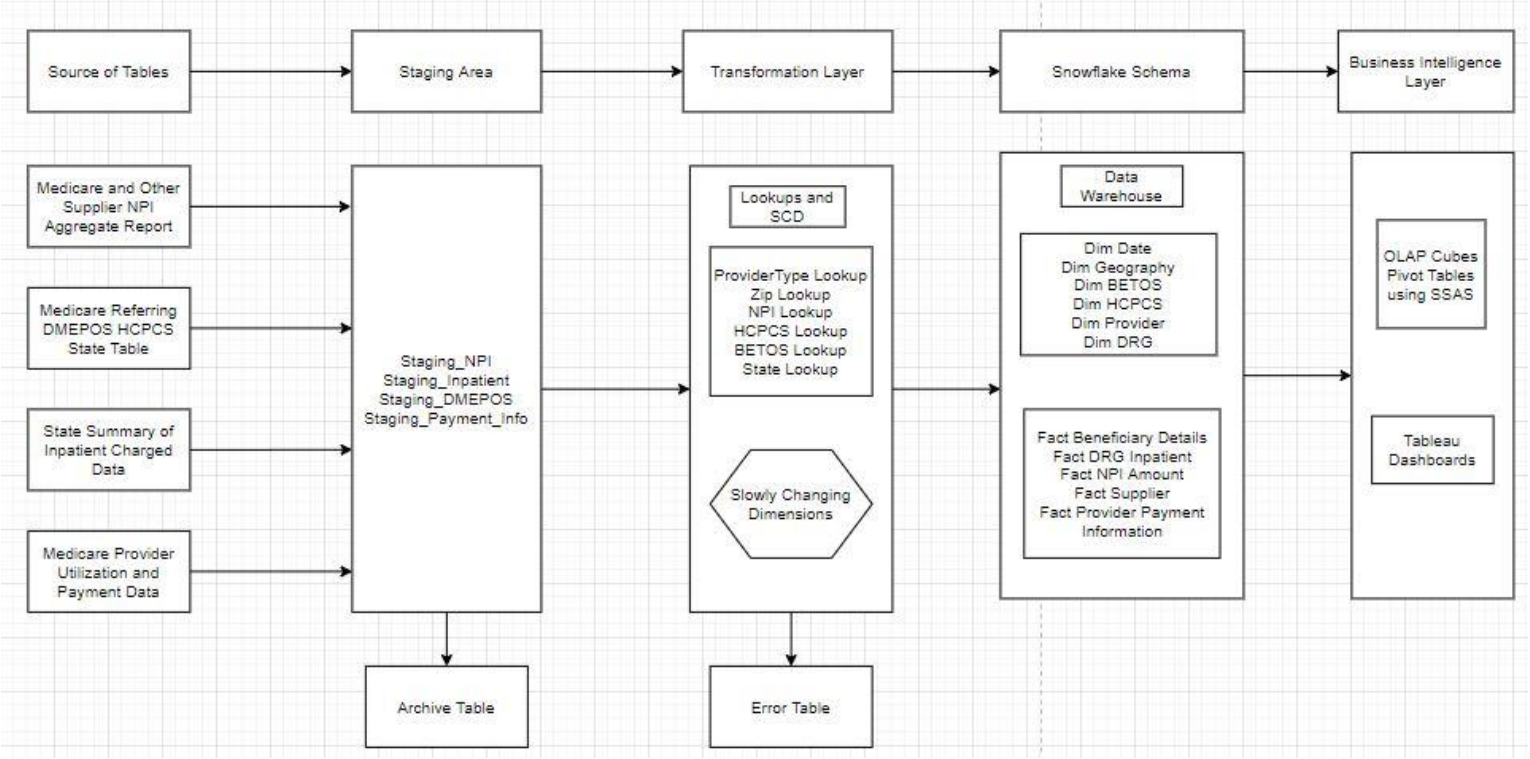
3. State Summary of Inpatient Charge Data- The Inpatient PUF includes information on utilization, payment (total payment and Medicare payment), and hospital-specific charges for the more than 3,000 U.S. hospitals that receive Medicare Inpatient Prospective Payment System (IPPS) payments. Each row is described as a drug per state. It consists of 27.5k rows and 6 columns. The entire dataset can be found below:

   https://data.cms.gov/Medicare-Inpatient/State-Summary-of-Inpatient-Charge-Data-by-Medicare/q5hc-zvkx

4. Medicare Provider Utilization and Payment Data - This dataset provides details on services and procedures provided to Medicare beneficiaries by physicians and other healthcare professionals. The Physician and Other Supplier PUF contains information on utilization, payment (allowed amount and Medicare payment), and submitted charges organized by National Provider Identifier (NPI), Healthcare Common Procedure Coding System (HCPCS) code, and place of service. This data set allows users to find out the average number of beneficiaries under each health care practitioner and the cost charged by each provide along with the amount covered by Medicare per HCPCS code. This data is extremely comprehensive consisting of 9.85 million rows and 25 columns. The entire data set can be seen below:

   https://data.cms.gov/Medicare-Physician-Supplier/Medicare-Provider-Utilization-and-Payment-Data-Phy/fs4p-t5eq

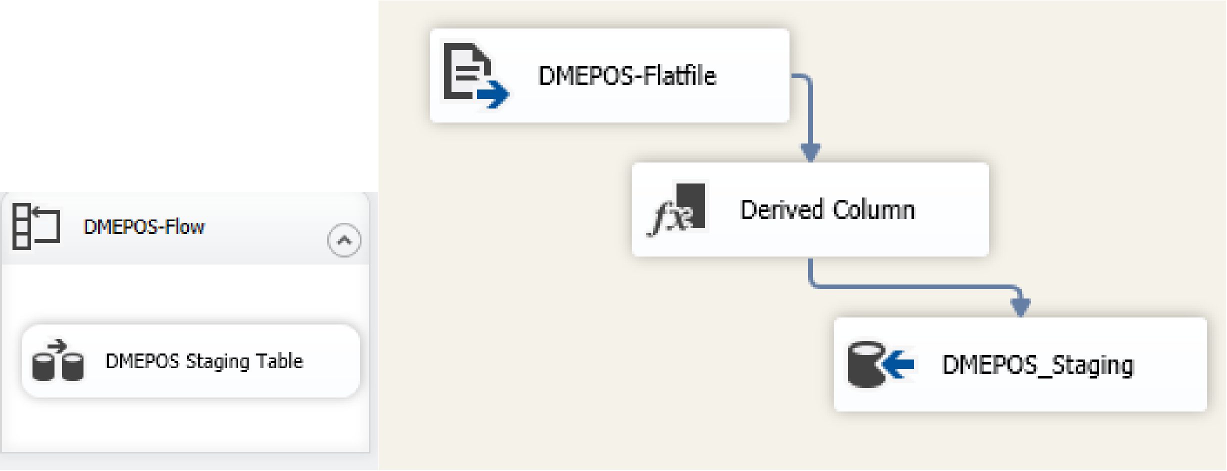
Source to Destination Mapping Flow:



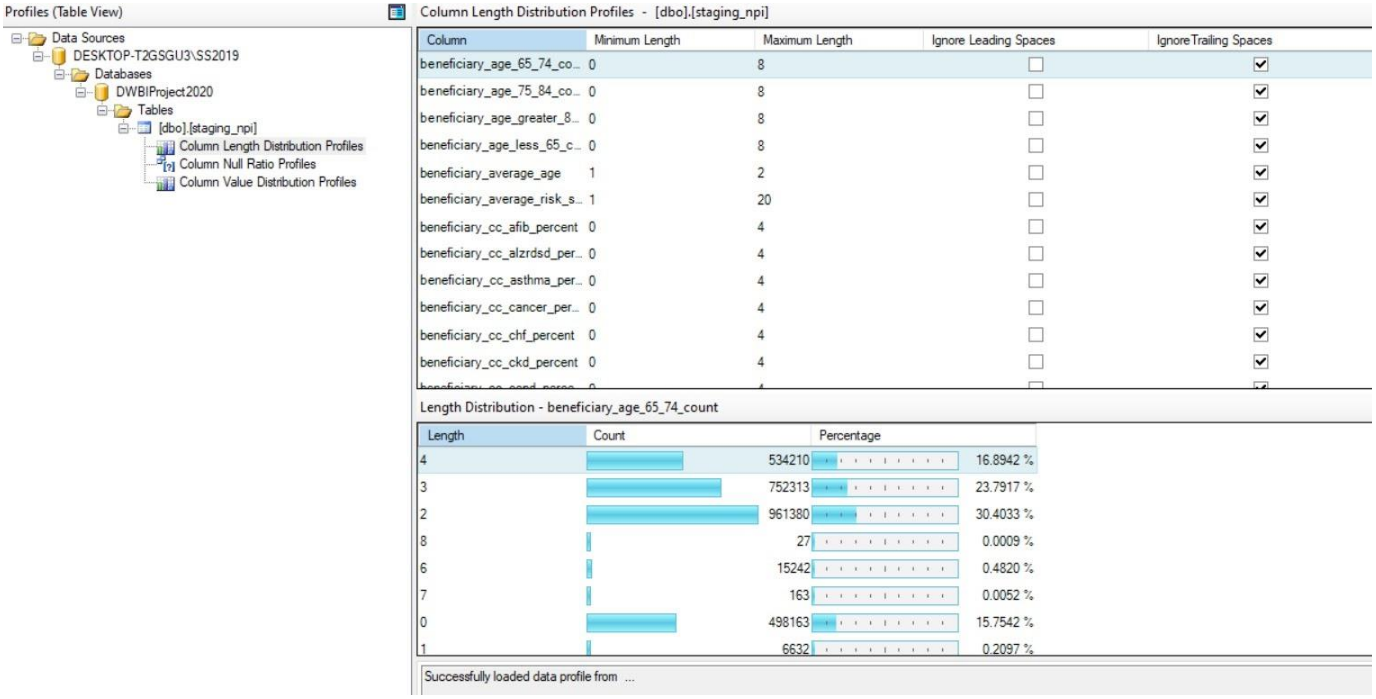
ETL Process:

Flow of the Data in the ELT process with potential tools:

1. Firstly, we created a for each loop container which has a Data Flow task to dump the sample dataset files from year 2015-2017 to the staging table. We added a Derived column component in between to create a 'Year' column for the indication of the year of the file being loaded. It took the FileName value of the dataset file containing the year.



2. Secondly, we profiled the above files using tools like Visual Studio which give us the in-depth analysis of columns as well as the whole dataset in form of column length and uniqueness, Null value analysis, accuracy and completeness etc. This helped us in understanding the transformations needed to be done on the data during the loading process.



Column Length Distribution Profiles - [dbo].[staging_npi]

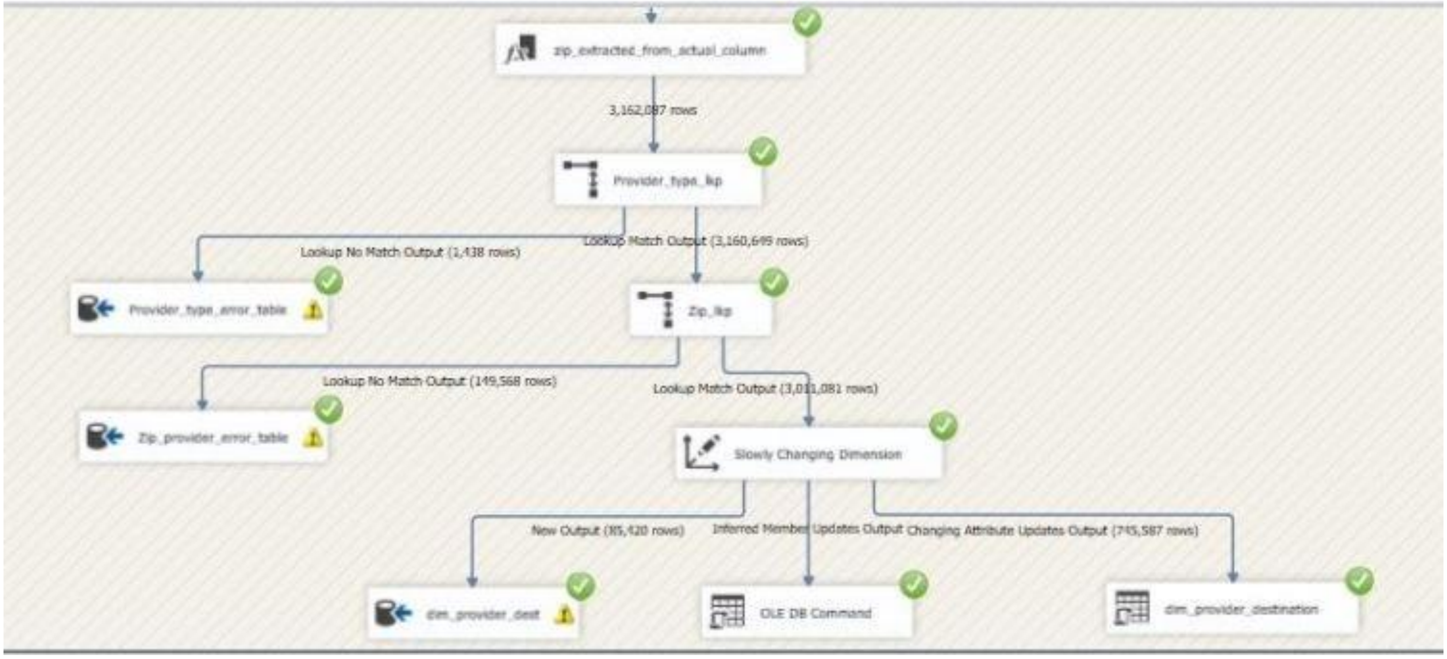
Column	Minimum Length	Maximum Length	Ignore Leading Spaces	IgnoreTrailing Spaces
beneficiary_age_65_74_co...	0	8	☐	☑
beneficiary_age_75_84_co...	0	8	☐	☑
beneficiary_age_greater_8...	0	8	☐	☑
beneficiary_age_less_65_c...	0	8	☐	☑
beneficiary_average_age	1	2	☐	☑
beneficiary_average_risk_s...	1	20	☐	☑
beneficiary_cc_afib_percent	0	4	☐	☑
beneficiary_cc_alzrdsd_per...	0	4	☐	☑
beneficiary_cc_asthma_per...	0	4	☐	☑
beneficiary_cc_cancer_per...	0	4	☐	☑
beneficiary_cc_chf_percent	0	4	☐	☑
beneficiary_cc_ckd_percent	0	4	☐	☑

Length Distribution - beneficiary_age_65_74_count

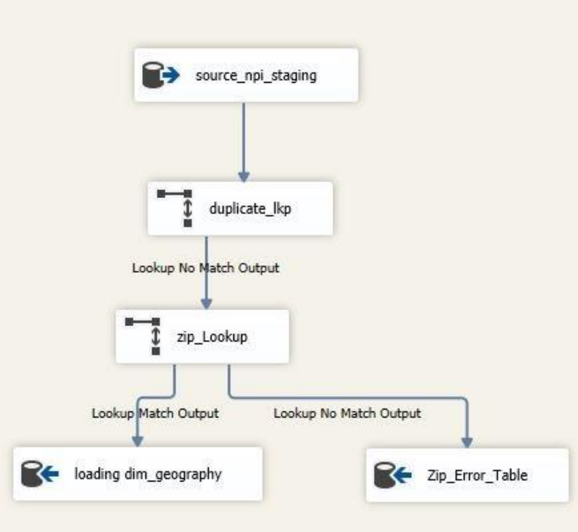
Length	Count	Percentage
4	534210	16.8942 %
3	752313	23.7917 %
2	961380	30.4033 %
8	27	0.0009 %
6	15242	0.4820 %
7	163	0.0052 %
0	498163	15.7542 %
1	6632	0.2097 %

Successfully loaded data profile from ...

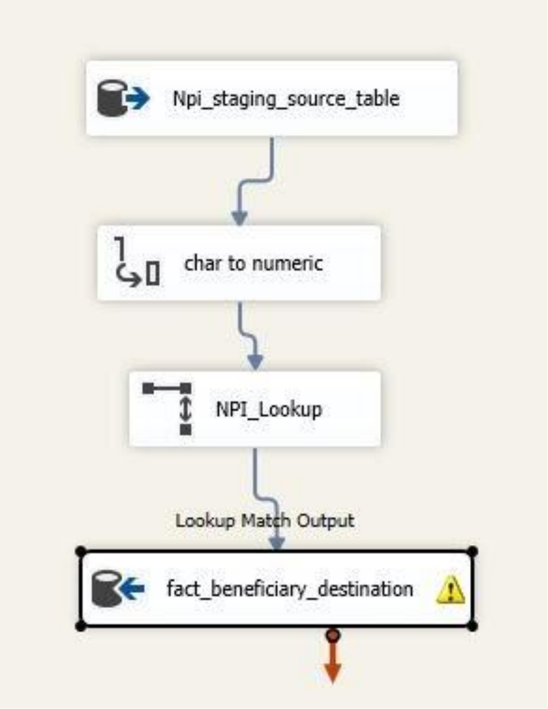
3. Going forward, the data from staging table is loaded into the Destination column while checking for changing fields via the Slowly Changing Dimension component. So that dimension like provide will not get added every year only if something changes then it will go into changing dimension output. For fact tables every year data gets loaded with year column as derived above. We also inserted the Lookup component in between to load only those records with fields which had reference in the Lookup tables. This way we get the accurate records in the Destination area


zip_extracted_from_actual_column
3,162,087 rows
Provider_type_lkp
Lookup No Match Output (1,438 rows)
Lookup Match Output (3,160,649 rows)
Provider_type_error_table
Zip_lkp
Lookup No Match Output (149,568 rows)
Lookup Match Output (3,011,081 rows)
Zip_provider_error_table
Slowly Changing Dimension
New Output (85,420 rows)
Inferred Member Updates Output
Changing Attribute Updates Output (745,587 rows)
dim_provider_dest
OLE DB Command
dim_provider_destination

4. We retrieved the Lookup table datasets from the internet and created Lookup tables for provider type, state, HCPCS type, DRG type BETOS id and zip code. Each was given and incremental reference ID which was then inserted to the destination tables via the Lookup up component while moving data from staging to destination



5. We also made some transformations to the data wherever necessary like converting the data format from varchar to its appropriate data type and also an extra column was added for HCPCS type where the codes were classified into two types.



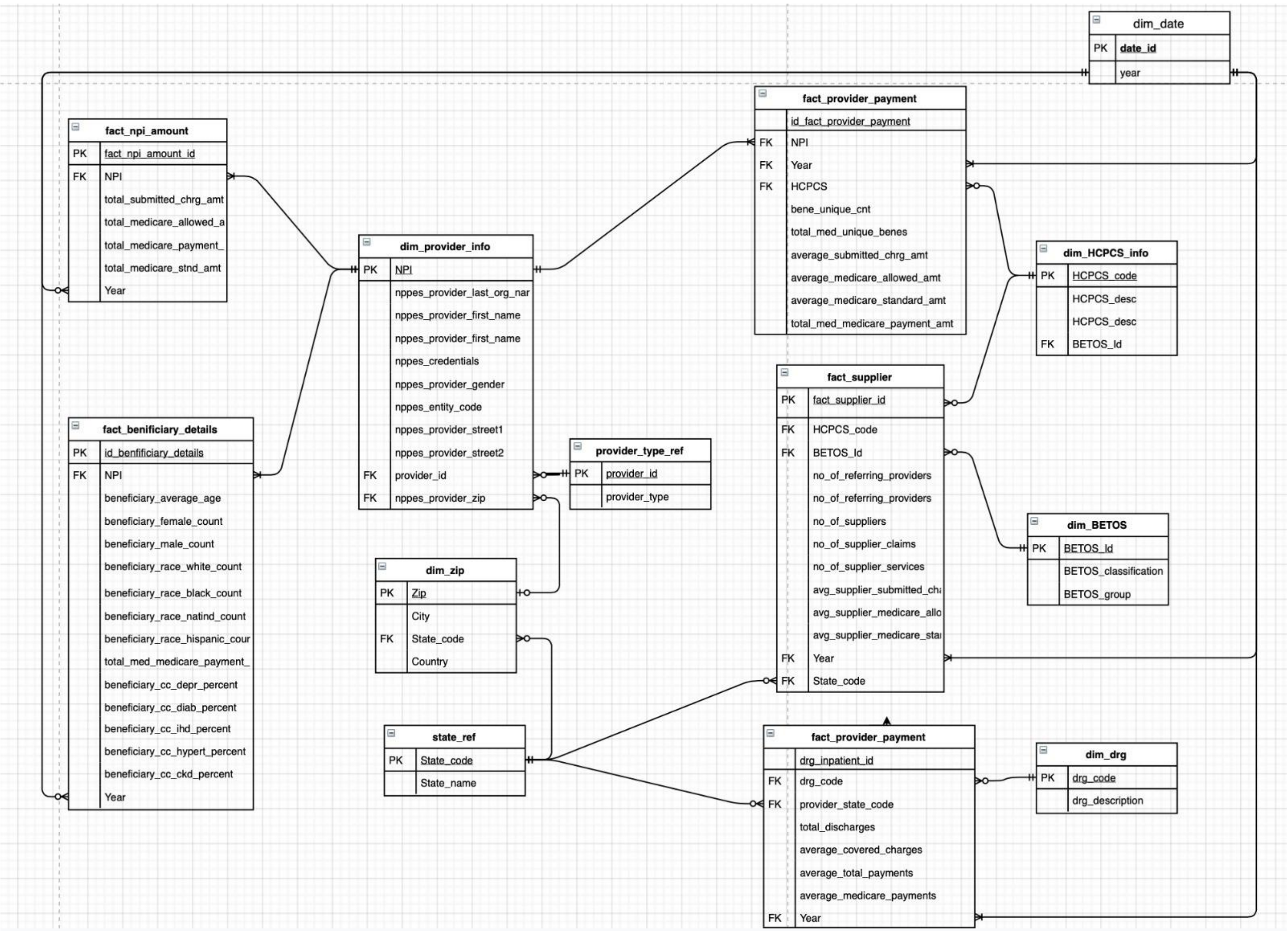
6. The unwanted records are then passed on to the Error tables for further analysis and inspection. We archived the data from the staging table so that initial unchanged load will be saved before the transformation to load it into our destination tables.

7. We leveraged the Snowflake schema for building the EDW destination table. This is done in SQL Server Management Studio itself by altering the tables and adding foreign key constraints for each table and making sure that all the tables were in 3rd normal form which was then fed into the business intelligence layer.\

8. We performed analytics on the schema by forming an OLAP cube. We were able to create hierarchies for provider geography, provider and HCPCS which enabled us to create measure groups on SSAS such as number of beneficiaries by state over the 3 year time period, the average Medicare payments made by beneficiaries in each state over the same time period or the number suppliers associated with each medical center.

9. We then made visualizations to view the results and created reporting dashboards using calculated columns in tableau as shown below.

E-R Diagram: EDW Schema



	dim_date
PK	date_id
	year

	fact_npi_amount
PK	fact_npi_amount_id
FK	NPI
	total_submitted_chrg_amt
	total_medicare_allowed_a
	total_medicare_payment_
	total_medicare_stnd_amt
	Year

	fact_provider_payment
	id_fact_provider_payment
FK	NPI
FK	Year
FK	HCPCS
	bene_unique_cnt
	total_med_unique_benes
	average_submitted_chrg_amt
	average_medicare_allowed_amt
	average_medicare_standard_amt
	total_med_medicare_payment_amt

	dim_provider_info
PK	NPI
	nppes_provider_last_org_nar
	nppes_provider_first_name
	nppes_provider_first_name
	nppes_credentials
	nppes_provider_gender
	nppes_entity_code
	nppes_provider_street1
	nppes_provider_street2
FK	provider_id
FK	nppes_provider_zip

	dim_HCPCS_info
PK	HCPCS_code
	HCPCS_desc
	HCPCS_desc
FK	BETOS_Id

	fact_supplier
PK	fact_supplier_id
FK	HCPCS_code
FK	BETOS_Id
	no_of_referring_providers
	no_of_referring_providers
	no_of_suppliers
	no_of_supplier_claims
	no_of_supplier_services
	avg_supplier_submitted_cha
	avg_supplier_medicare_allo
	avg_supplier_medicare_stai
FK	Year
FK	State_code

	fact_benificiary_details
PK	id_benfificiary_details
FK	NPI
	beneficiary_average_age
	beneficiary_female_count
	beneficiary_male_count
	beneficiary_race_white_count
	beneficiary_race_black_count
	beneficiary_race_natind_count
	beneficiary_race_hispanic_cour
	total_med_medicare_payment_
	beneficiary_cc_depr_percent
	beneficiary_cc_diab_percent
	beneficiary_cc_ihd_percent
	beneficiary_cc_hypert_percent
	beneficiary_cc_ckd_percent
	Year

	provider_type_ref
PK	provider_id
	provider_type

	dim_BETOS
PK	BETOS_Id
	BETOS_classification
	BETOS_group

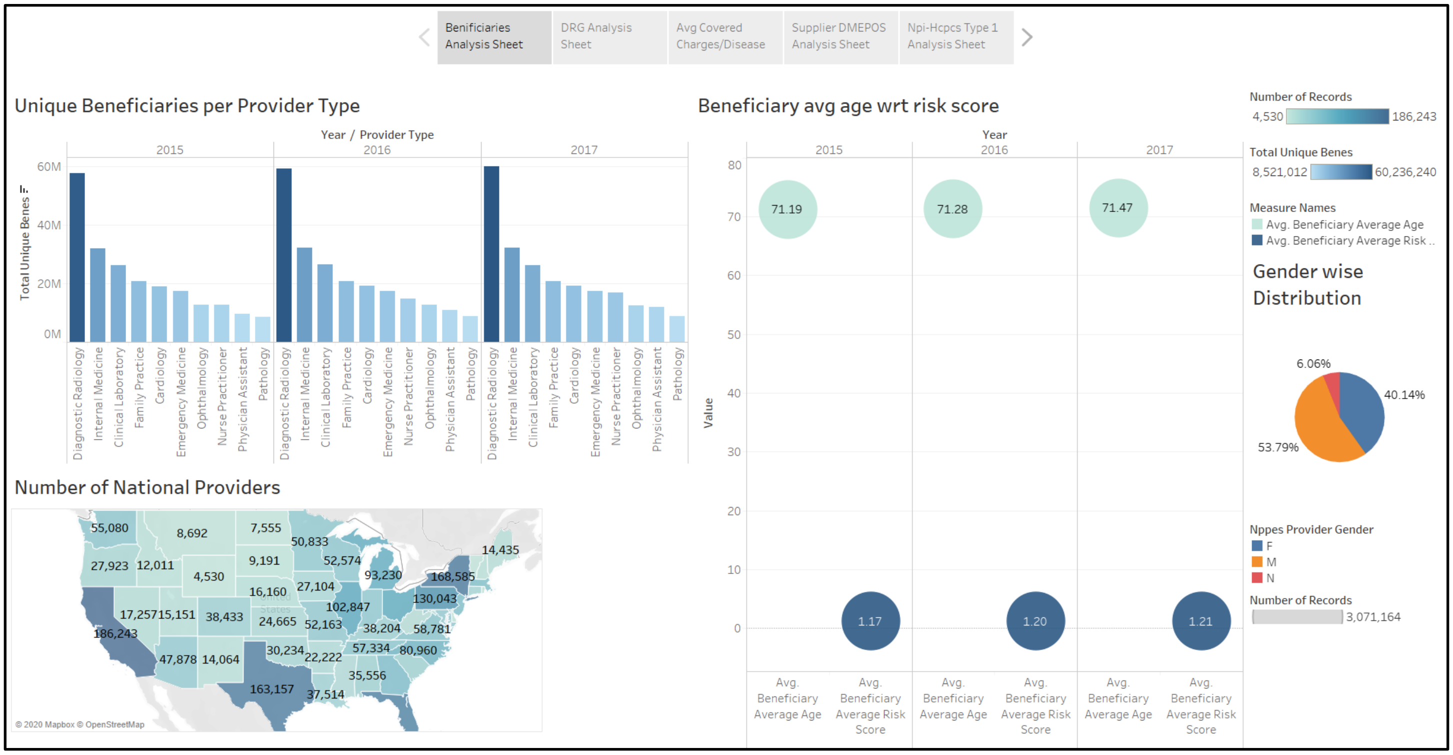
	dim_zip
PK	Zip
	City
FK	State_code
	Country

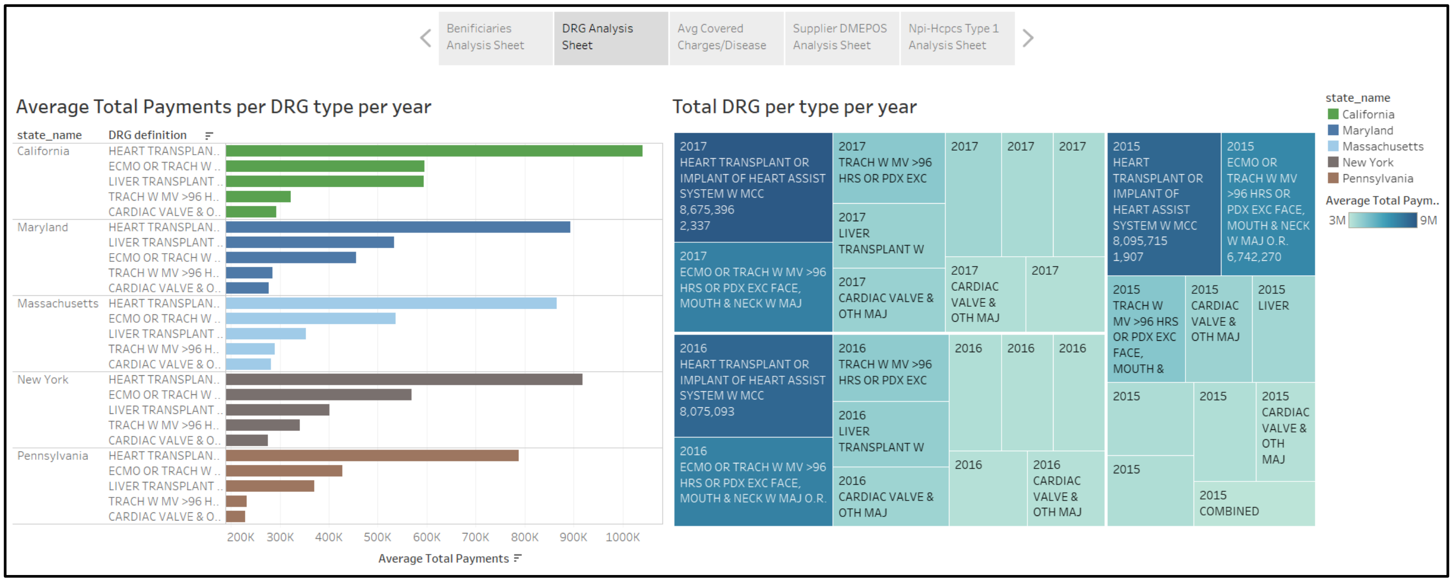
	state_ref
PK	State_code
	State_name

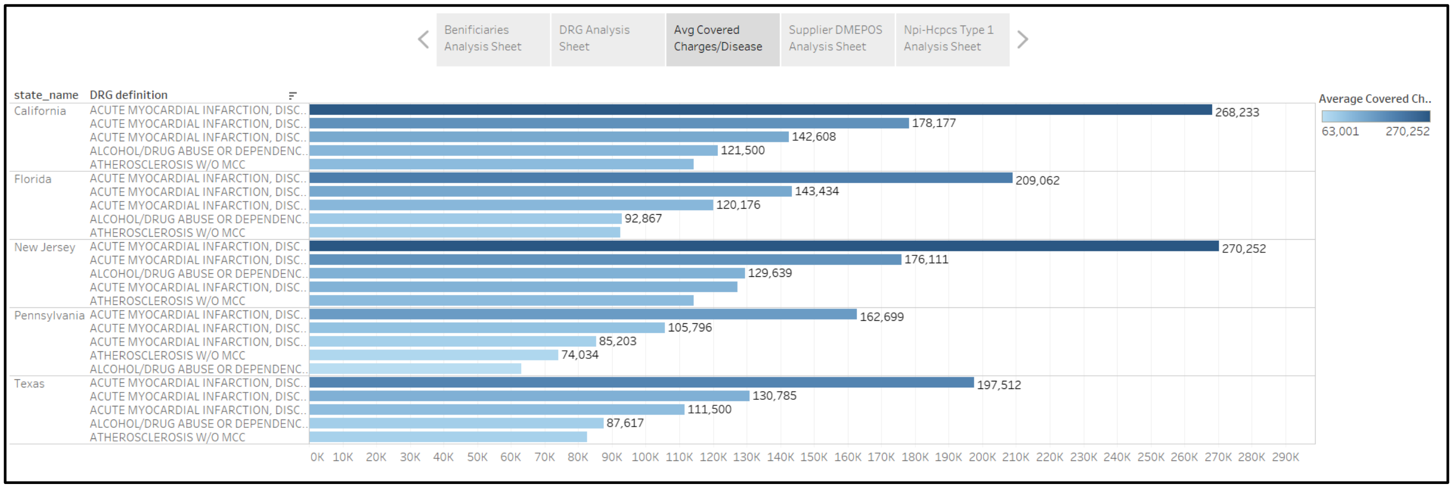
	fact_provider_payment
	drg_inpatient_id
FK	drg_code
FK	provider_state_code
	total_discharges
	average_covered_charges
	average_total_payments
	average_medicare_payments
FK	Year

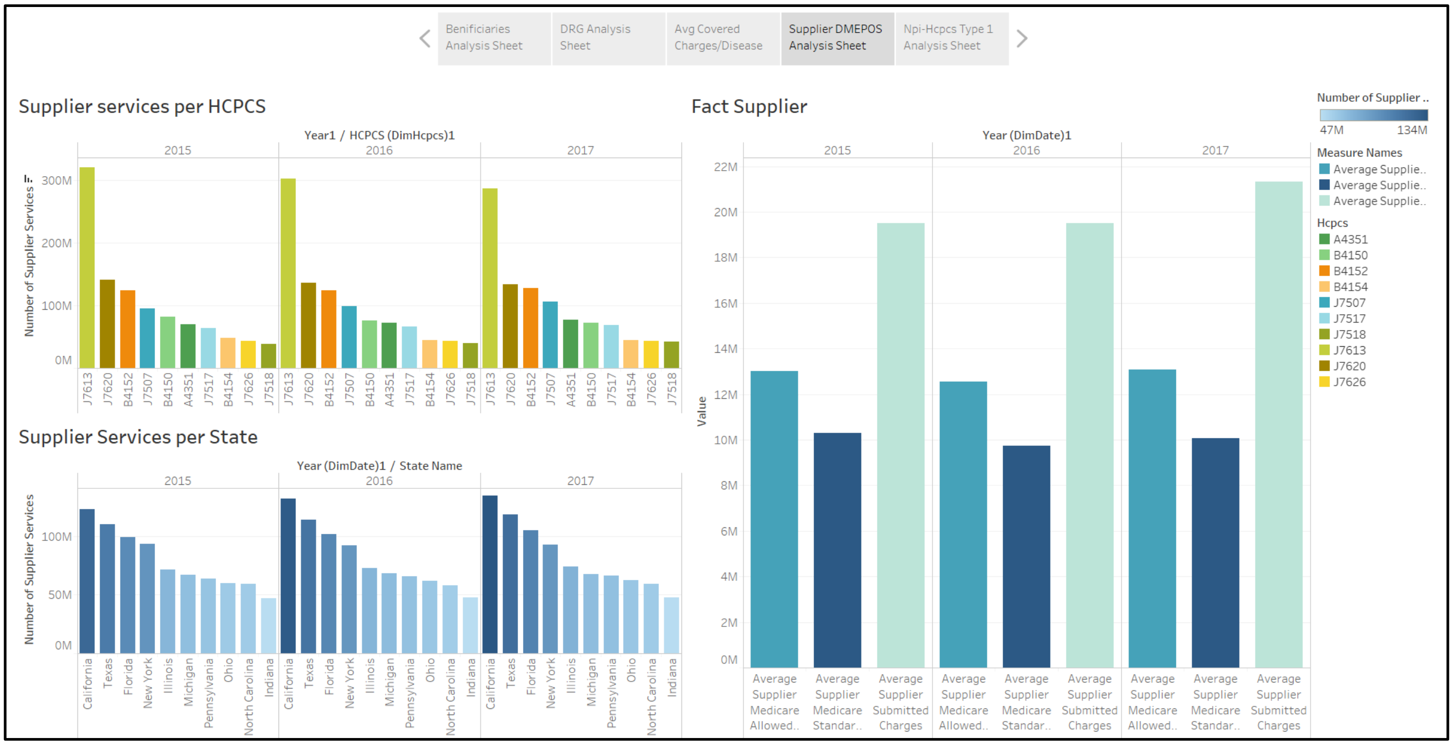
	dim_drg
PK	drg_code
	drg_description

Tableau Visualizations:




Beneficiaries Analysis Sheet
DRG Analysis Sheet
Avg Covered Charges/Disease
Supplier DMEPOS Analysis Sheet
Npi-Hcpcs Type 1 Analysis Sheet
Average Total Payments per DRG type per year
state_name
DRG definition
California
Maryland
Massachusetts
New York
Pennsylvania
HEART TRANSPLAN..
ECMO OR TRACH W ..
LIVER TRANSPLANT ..
TRACH W MV >96 H..
CARDIAC VALVE & O..
200K
300K
400K
500K
600K
700K
800K
900K
1000K
Average Total Payments
Total DRG per type per year
2017 HEART TRANSPLANT OR IMPLANT OF HEART ASSIST SYSTEM W MCC 8,675,396 2,337
2017 TRACH W MV >96 HRS OR PDX EXC
2017 LIVER TRANSPLANT W
2017 CARDIAC VALVE & OTH MAJ
2017 ECMO OR TRACH W MV >96 HRS OR PDX EXC FACE, MOUTH & NECK W MAJ
2016 HEART TRANSPLANT OR IMPLANT OF HEART ASSIST SYSTEM W MCC 8,075,093
2016 TRACH W MV >96 HRS OR PDX EXC
2016 LIVER TRANSPLANT W
2016 CARDIAC VALVE & OTH MAJ
2016 ECMO OR TRACH W MV >96 HRS OR PDX EXC FACE, MOUTH & NECK W MAJ O.R.
2015 HEART TRANSPLANT OR IMPLANT OF HEART ASSIST SYSTEM W MCC 8,095,715 1,907
2015 ECMO OR TRACH W MV >96 HRS OR PDX EXC FACE, MOUTH & NECK W MAJ O.R. 6,742,270
2015 TRACH W MV >96 HRS OR PDX EXC FACE, MOUTH &
2015 CARDIAC VALVE & OTH MAJ
2015 LIVER
2015 CARDIAC VALVE & OTH MAJ
2015 COMBINED
state_name
California
Maryland
Massachusetts
New York
Pennsylvania
Average Total Paym..
3M
9M

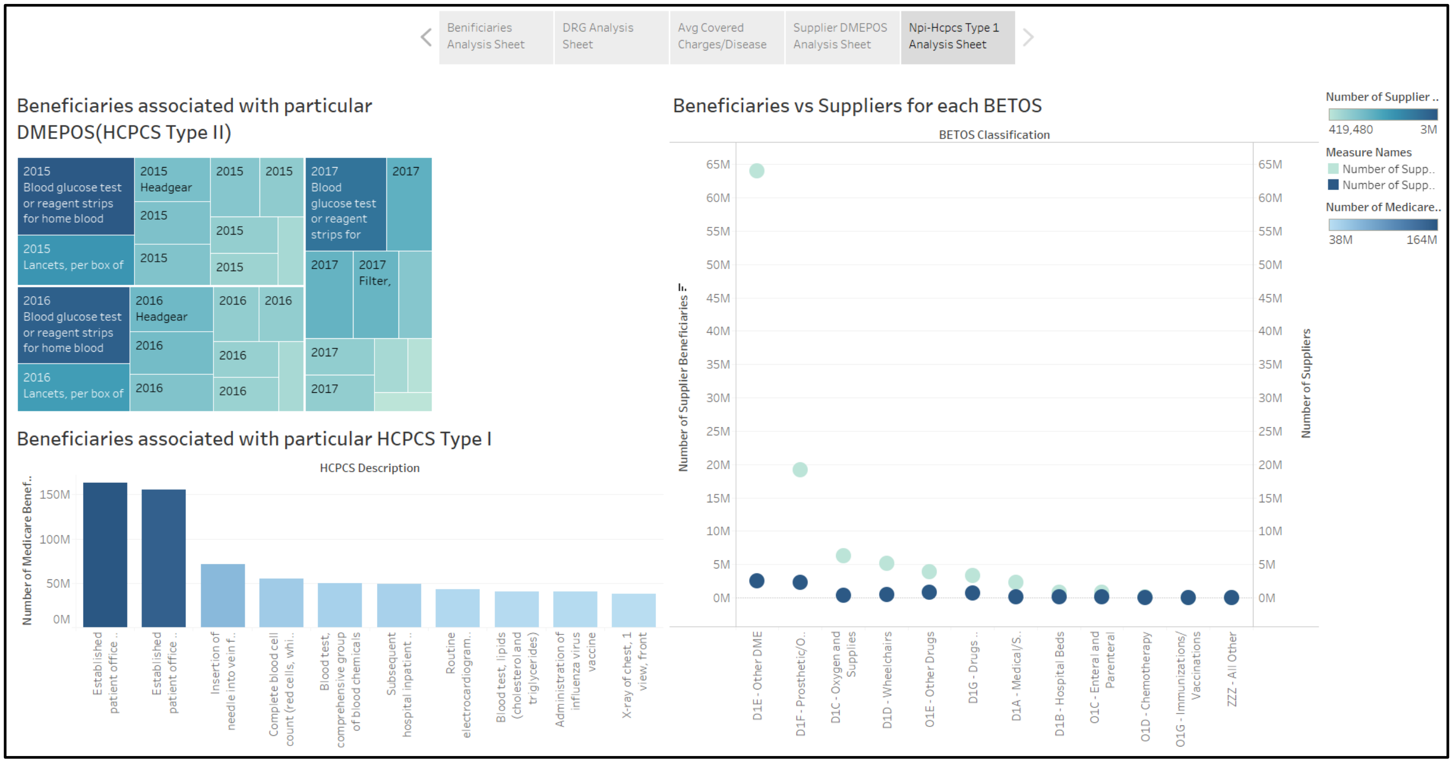




Supplier services per HCPCS

Supplier Services per State

Fact Supplier


Benificiaries Analysis Sheet
DRG Analysis Sheet
Avg Covered Charges/Disease
Supplier DMEPOS Analysis Sheet
Npi-Hcpcs Type 1 Analysis Sheet
2015 Blood glucose test or reagent strips for home blood
2015 Headgear
2015
2015
2017 Blood glucose test or reagent strips for
2017
2015 Lancets, per box of
2016 Blood glucose test or reagent strips for home blood
2016 Headgear
2016
2017 Filter,
2016 Lancets, per box of
HCPCS Description
Number of Medicare Benef..
150M
100M
50M
0M
Established patient office ..
Established patient office ..
Insertion of needle into vein f..
Complete blood cell count (red cells, whi..
Blood test, comprehensive group of blood chemicals
Subsequent hospital inpatient ..
Routine electrocardiogram..
Blood test, lipids (cholesterol and triglycerides)
Administration of influenza virus vaccine
X-ray of chest, 1 view, front
BETOS Classification
Number of Supplier Beneficiaries
Number of Suppliers
65M
60M
55M
50M
45M
40M
35M
30M
25M
20M
15M
10M
5M
0M
D1E - Other DME
D1F - Prosthetic/O..
D1C - Oxygen and Supplies
D1D - Wheelchairs
O1E - Other Drugs
D1G - Drugs ..
D1A - Medical/S..
D1B - Hospital Beds
O1C - Enteral and Parenteral
O1D - Chemotherapy
O1G - Immunizations/ Vaccinations
ZZZ - All Other
Number of Supplier ..
419,480
3M
Measure Names
Number of Supp..
Number of Supp..
Number of Medicare..
38M
164M

Beneficiaries associated with particular DMEPOS(HCPCS Type II)

Beneficiaries associated with particular HCPCS Type I

Beneficiaries vs Suppliers for each BETOS

Insights:

The analysis of the different datasets yields some interesting insights which help us in creating visualizations, these insights are as follows:

- Diagnostic Radiology has the most consistent unique consumers over the years for all US states
- The Average Risk score increases as Average Beneficiary Age increases proportionally
- California #1 in terms of total payments for diseases paid by the Insurance Companies
- Health transplant /Implant has the highest paid amount for all the years ~ $8,675,396, while average coverage charges is the best for Acute Myocardial Infarction in all the states
- Maximum number of Supplier Services involve supply of medical drug Albuterol and the highest number of supplier's service to California
- Average allowed amount & the standard amount allotted to suppliers remains constant while the Average charges submitted by supplier keeps on increasing YoY
- For the type 2 services - Blood glucose test is #1 supplied equipment by the suppliers prescribed by the maximum number of healthcare providers. It had the beneficiary count ~ 3,470,023 in 2015 which lowered down to 2,901,900 in 2017
- For the type 1 services - Outpatient visit (typically 15 mins) had the greatest number of beneficiaries all the years combined, followed by outpatient visit (25 mins) and Needle insertion for blood sample category

Results:

- The Medicare Data warehouse provided a granular insight on demographic & economic evaluation of Medicare companies, providers, equipment suppliers & beneficiaries
- Health care providers and the companies have invested the most in sectors of Diagnostic Radiology, Internal Medicine and Clinical Laboratory with regards to insurance plans
- Considering the volume of healthcare providers in the California, Insurance companies in that state repay the most dollar amount
- Heart Transplant has the highest repayment statistics for all the years for all the top repaid states
- The best coverage is received by beneficiaries residing in New Jersey state
- Albuterol is presumably the best in-demand product in market since it has the most number of suppliers in US
- Suppliers & medical manufacturers supplying glucose test equipment are making the most money
- Consumers are likely to invest in Healthcare plans involving Outpatient visits and blood test labs
- Overall, this warehouse digs deep and provides strategies as to how to grow the insurance business as well provide the consumers with dynamic view about the best plans to invest in & most optimal medical centers to visit for each chronic condition