****

**Data Warehouse and Business Intelligence**

**Project Report**

**Group 10: Guardians**

**Group members:**

**Rakesh Reddy Birika**

**Jiachen Xu**

**Nishit Samir Ajmera**

**Introduction**

In this project we will deal with two of the datasets from a number of datasets provided by the professor. The datasets we are choosing are the following:

* CMS Medicare Part D Drug Spending
* Medicare Provider Utilization and Payment Data: Part D Prescriber

Specifically, CMS Medicare Part D Drug Spending data comprises the data from 2013-2017; and

Medicare Provider Utilization and Payment Data: Part D Prescriber comprises the data from 2013 to 2016. The datasets we chose are about prescription drugs prescribed by individual physicians and other health care providers and paid for under the Medicare Part D Prescription Drug Program. Medicare is the health insurance plan provided by the government for people above 65 years of age or for people with disabilities. We have chosen to use the data of Medicare Part D which deals with coverage related to prescription drugs. As a Business Intelligence

project, the process is as follows: Data Integration -> Data Warehousing -> Business Intelligence.

**Objectives**

We have selected two datasets to work on our project they are the **CMS Medicare Provider Utilization and Payment Data** and **CMS Medicare Part D Drug Spending Dashboard and Data.** We identified the primary key and foreign keys of the tables. After identification of the primary and foreign keys we will clean the source datasets to carry out further processes. We created a schema in which we created the dimensions and fact tables. Once the dimensions and facts have been created we populate these tables using ETL processes to create a data warehouse and build OLAP cubes. Our main objective to perform these steps is to get insight to make meaningful analysis and visualization to find the correlation between these two datasets.

**Dataset Description**

**CMS Medicare Provider Utilization and Payment Data**

This dataset provides the data on drugs prescribed by health care providers and covered under the Medicare part D program. The healthcare care providers have been provided with a unique 10 digit identification number called the National Provider Identifier. Since the NPI has unique values we will use it as a primary key.Also we will use Drug Name column as foreign key to refer to the other dataset “CMS Medicare Part D Drug Spending Data”. There are 21 columns in the dataset and around 25 million rows in each table from different years. The column names are as follows: npi, nppes\_provider\_last\_org\_name, nppes\_provider\_first\_name, nppes\_provider\_city, nppes\_provider\_state, specialty\_desc, description\_flag, drug\_name, generic\_name, bene\_count, total\_claim\_count, total\_30\_day\_fill\_count, total\_day\_supply, total\_drug\_cost, bene\_count\_ge65, bene\_count\_ge65\_redact\_flag, total\_claim\_count\_ge65, ge65\_redact\_flag, total\_30\_day\_fill\_count\_ge65, total\_day\_supply\_ge65, and total\_drug\_cost\_ge65.

**CMS Medicare Part D Drug Spending Dashboard & Data**

This dataset presents spending information for Medicare Part D drugs - drugs patients generally administer themselves and that are paid through the Medicare Part D program. Part D drug information is available from the Part D Prescription Drug Event (PDE) data, which are available for a subset of Medicare beneficiaries who choose to enroll in Part D (which represents approximately 70% of Medicare beneficiaries). The dataset is about Medicare Part D Drug Spending and Utilization from Calendar Years 2013 to 2017. There are 13 columns in the dataset, they are showing as follows: Brand Name, Generic Name, Manufacturer, Number of Manufacturers, Total Spending, Total Dosage Units, Total Claims, Total Beneficiaries, Average Spending Per Dosage Unit (Weighted), Average Spending Per Claim, Average Spending Per Beneficiary, Change in Average Spending Per Dosage Unit (2016-2017), and Annual Growth Rate in Average Spending Per Dosage Unit (2013-2017). The brand name is the primary key.

**Softwares and tools**

For most part of our project we have used SSIS and SSMS to clean the source data and transform it into dimensions and facts. A We used SSAS to build multidimensional OLAP cubes to ready these tables for analysis based on various measures. Also toad data modeler has been used to create the schema. Finally we used Tableau to visualize the data.

**Process**

**1. Data preprocessing**

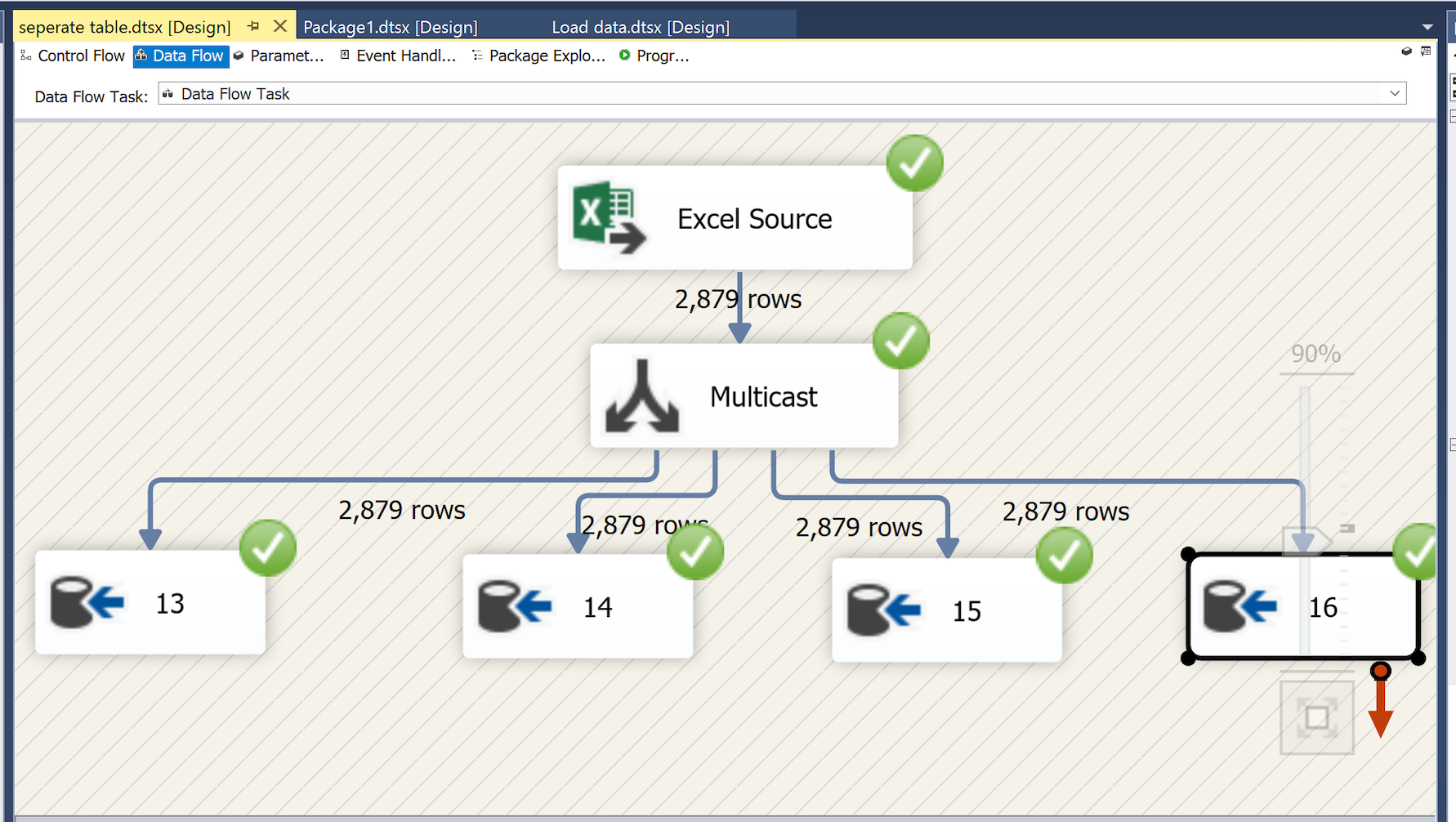
The datasets are from two sources, one is CMS drug spending, the other one is

PartD\_Prescriber. The CMS drug spending dataset contains the data from the year 2013 to 2016, they are stored in four tables respectively. The PartD\_Prescriber contains data from the year 2013 to 2016, and they are all stored in one table. In order to bring uniformity and load the data into our database to make dimensions and facts for further tasks, we splitted the PartD\_Prescriber database (using the multicast function in SSIS) and loaded the data from the year 2013 to 2016 into individual tables in our database. In addition, we have converted the data type in SSIS and load all the raw tables into relative staging tables.

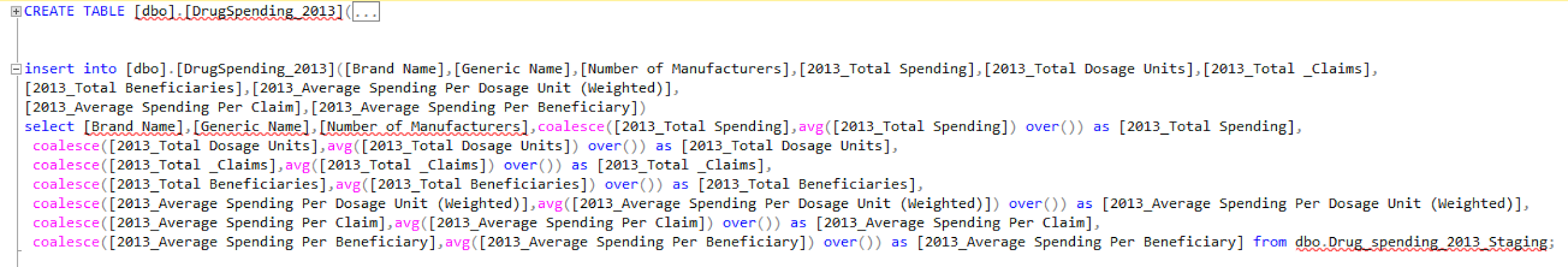
We noticed there were a lot of NULL values in the CMS drug spending datasets thus we have taken the mean values of the column by SQL query and replaced the NULL values in each column with the corresponding mean value of that column using the Coalesce function in SSMS. After seperating the tables into 4 years, the years were not specified in the tables. In order to prepare the data for proper analysis we added a year column to each of the corresponding tables using alter and update statement in SSMS. We then combined the 4 tables by years into a single staging table using the union all statement.

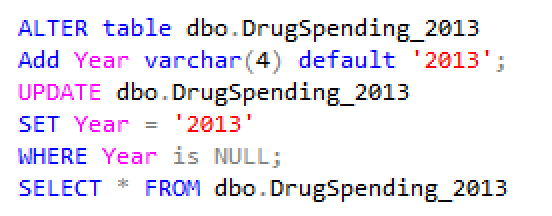
In the PartD\_Prescriber dataset, we checked the missing values of the columns with numeric values and replaced all the missing values with 0. There are also three columns with “flag” meaning which need to be ignored because they don’t have valuable meanings.

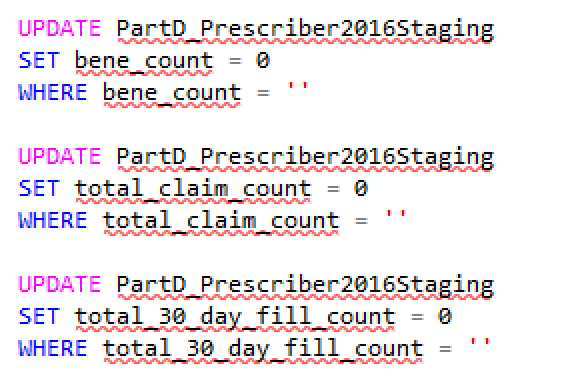
The picture below shows how we used multicast to load the data into different staging tables in our database.



The picture below shows how did we clean the drug\_spending dataset.

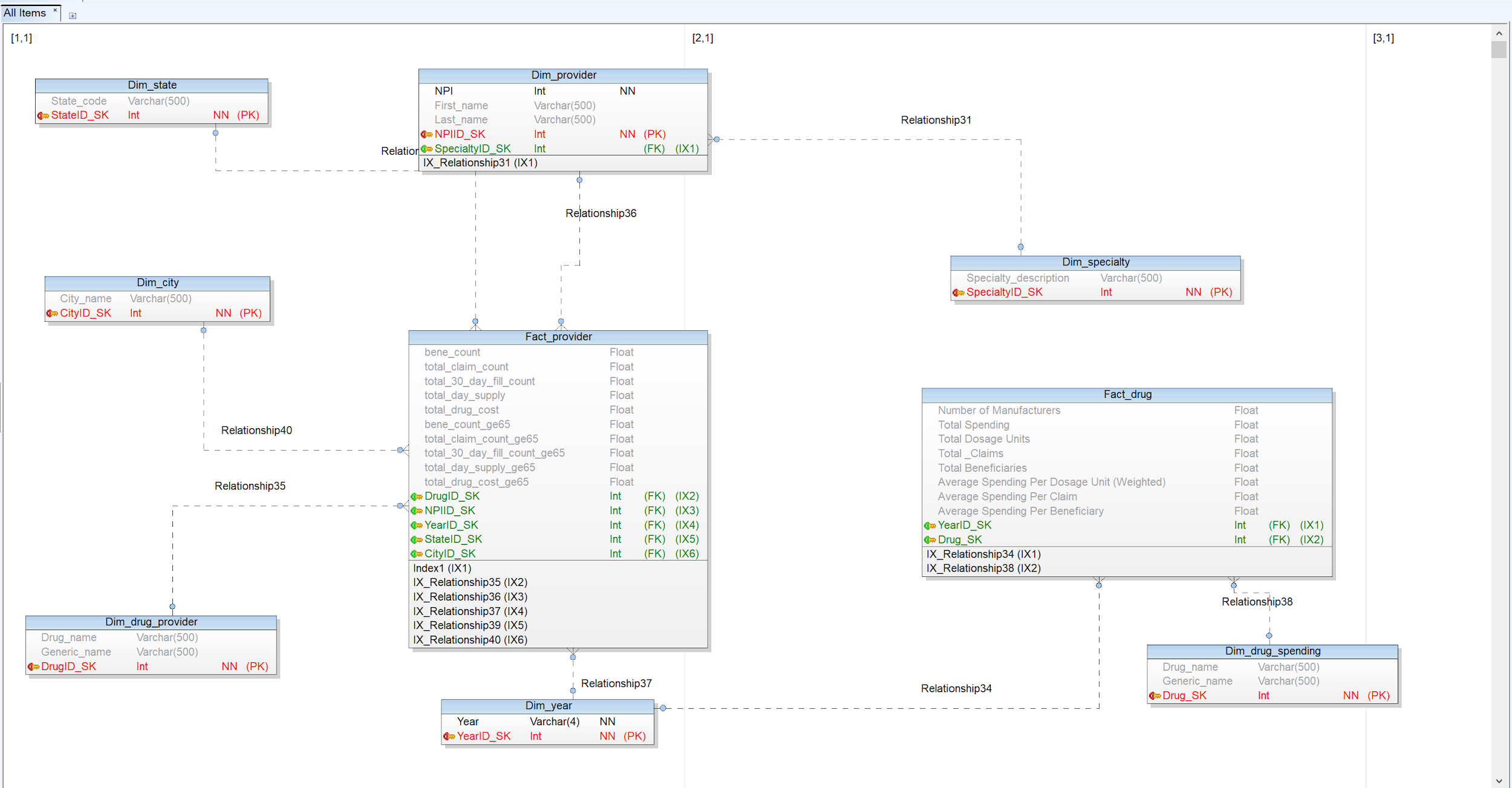






**2. Data Modeling**

Using Toad data modeler we have created surrogate keys for all the 7 dimensions which acted as references to two fact tables: Fact\_provider and Fact\_drug. The DDL script was then generated to create the dimensions and facts in the data warehouse. The data model for our database project shows below:

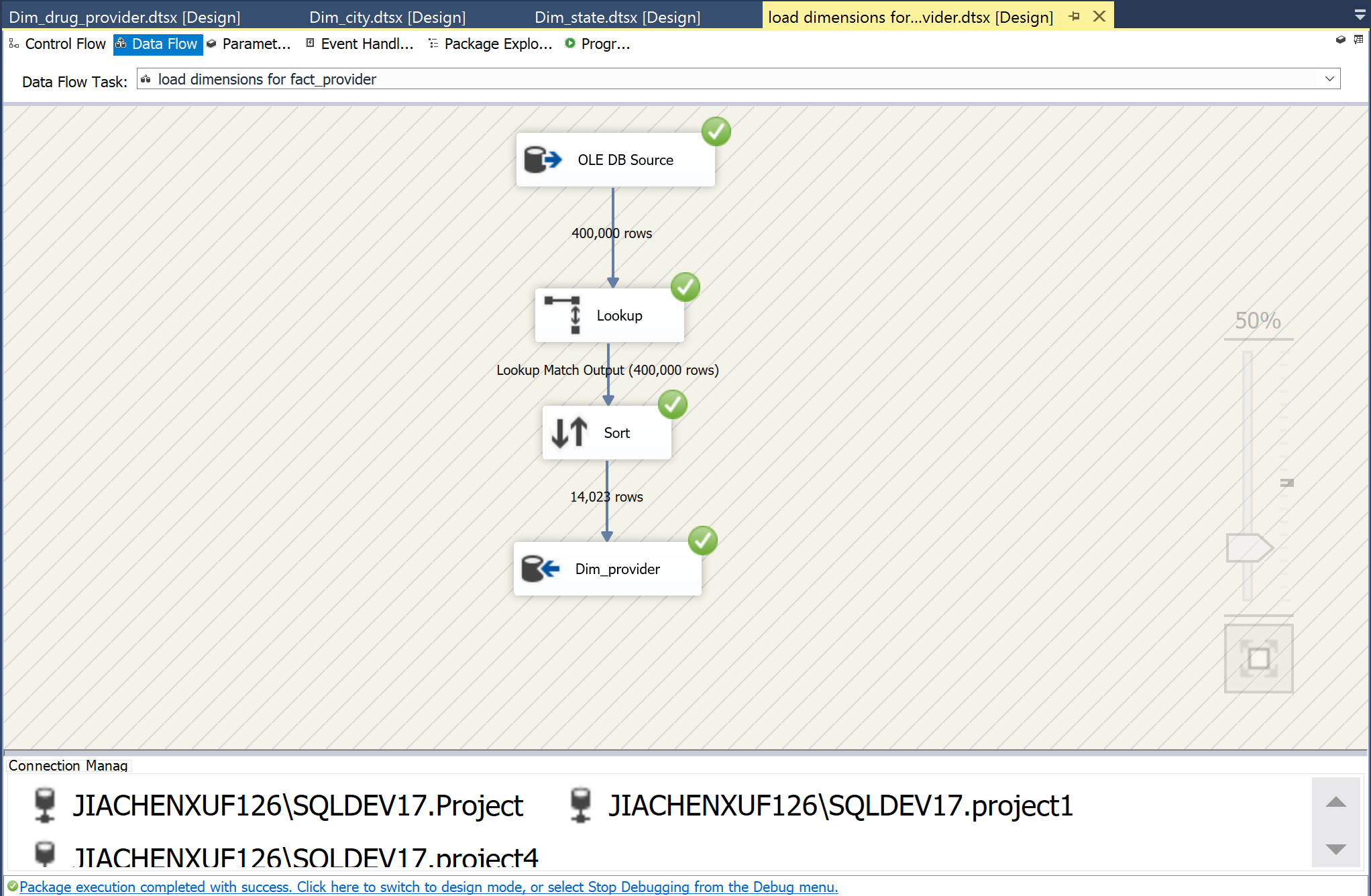


**3. Data populating**

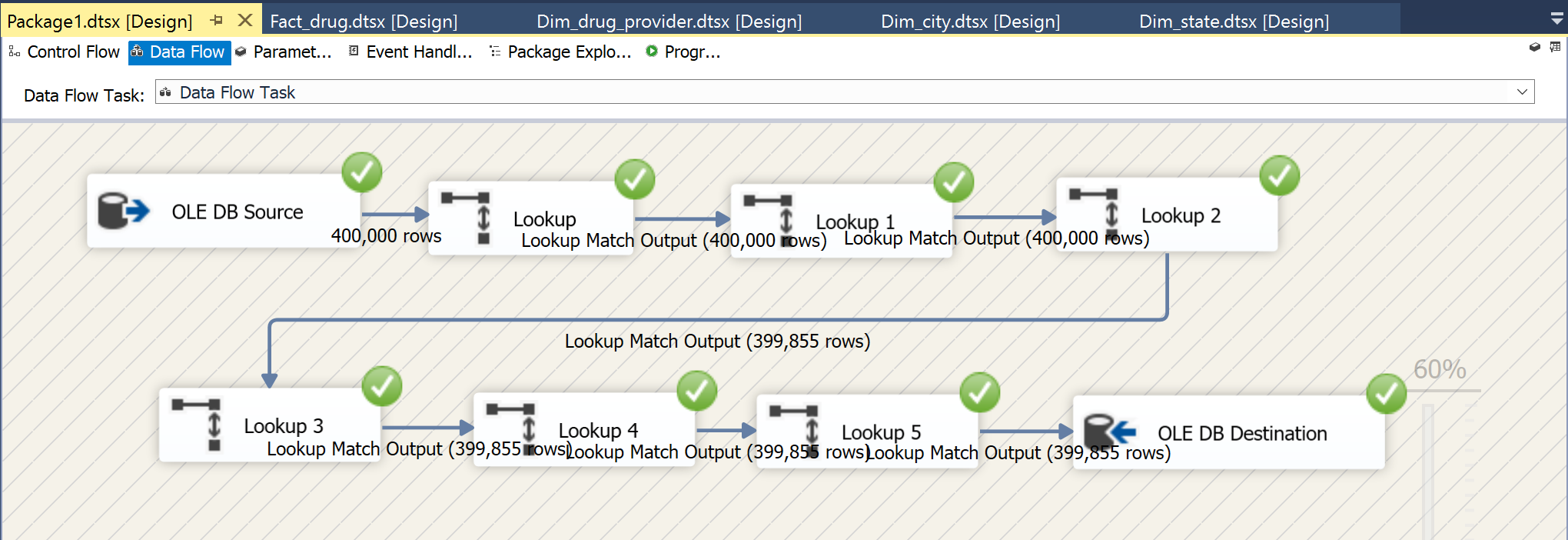
Using SSIS we have populated the dimensions and facts by looking up on surrogate keys and sorting the required dimensions.

Part of the process shows as below:

The following graph shows the loading process for dimension tables. We do look up based on the attributes from the populated dimension table and sort by certain attributes then delete duplicate columns, finally load the data into our dimension destinations



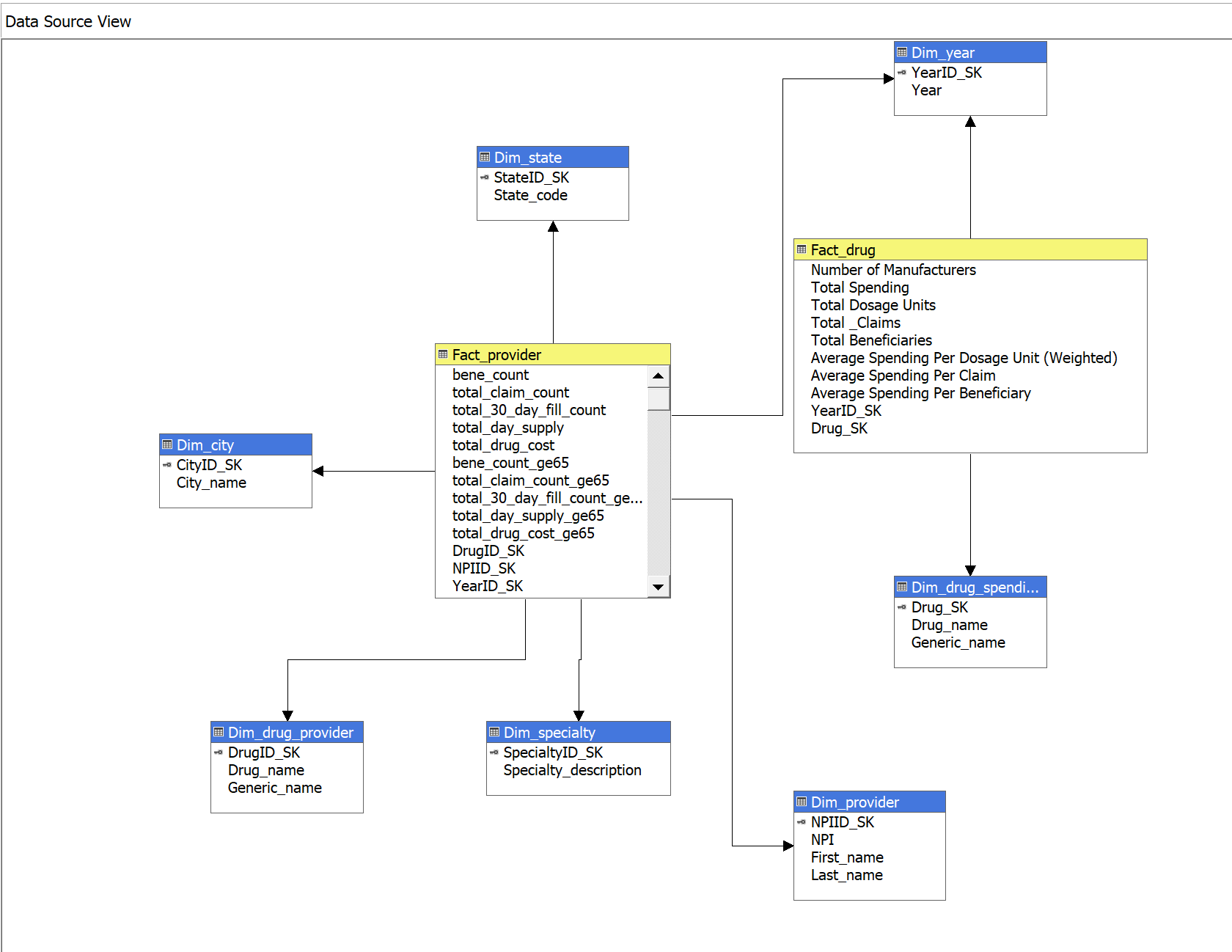
The graph below shows the populating process of fact tables, we do look up by the surrogate keys from all the dimension tables and then load the data into the fact in our database.



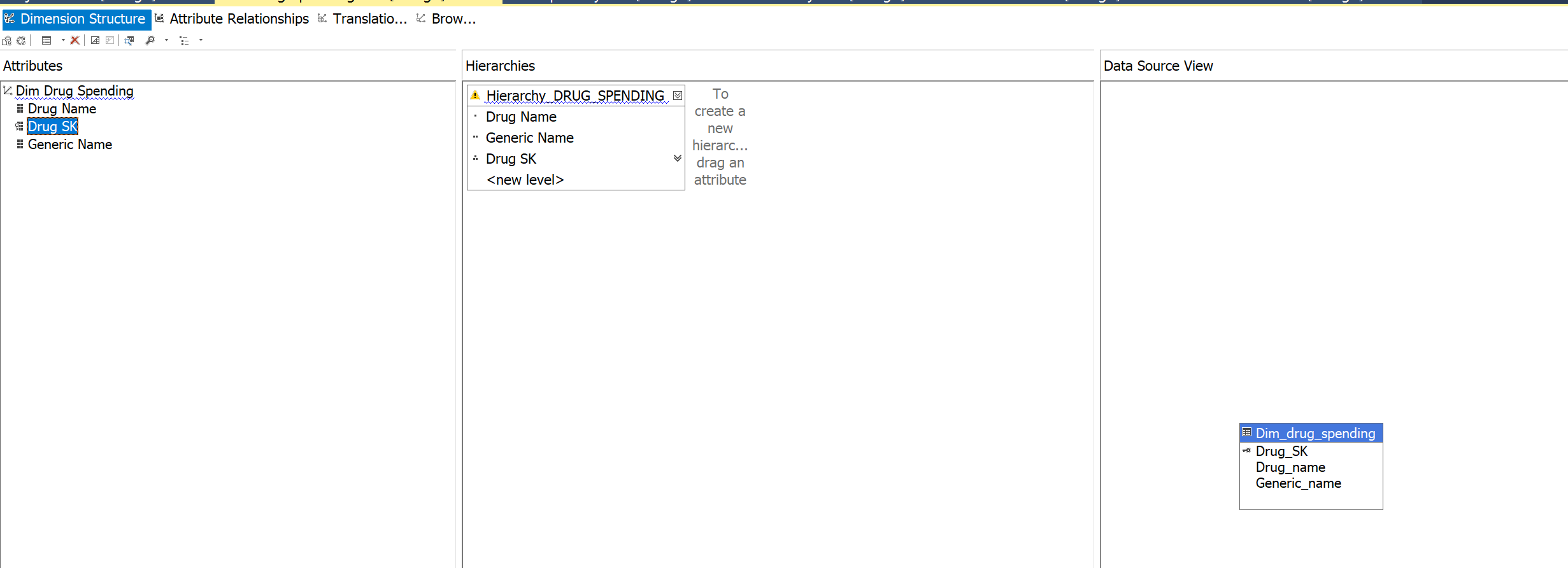
**4. OLAP cubes**

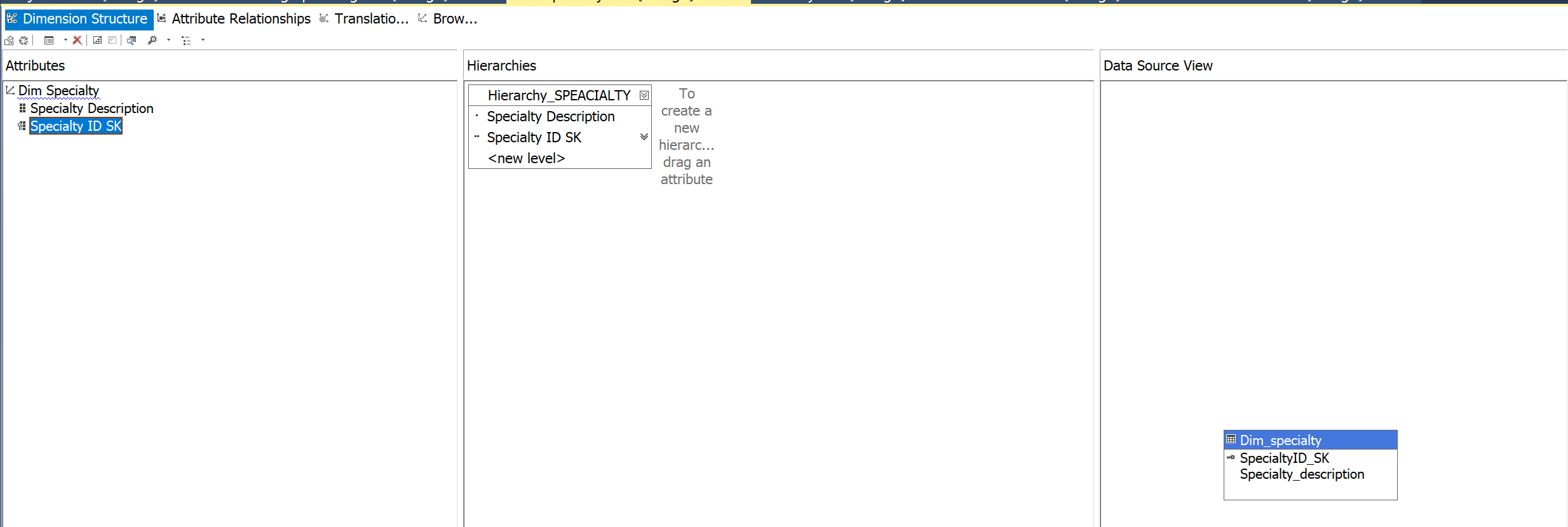
After we finished loading all the dimensions and facts, we have created the OLAP cubes using SSAS. We have created hierarchies for each of the dimensions to organise the fields into similar categories which makes it easy to keep track of the data.

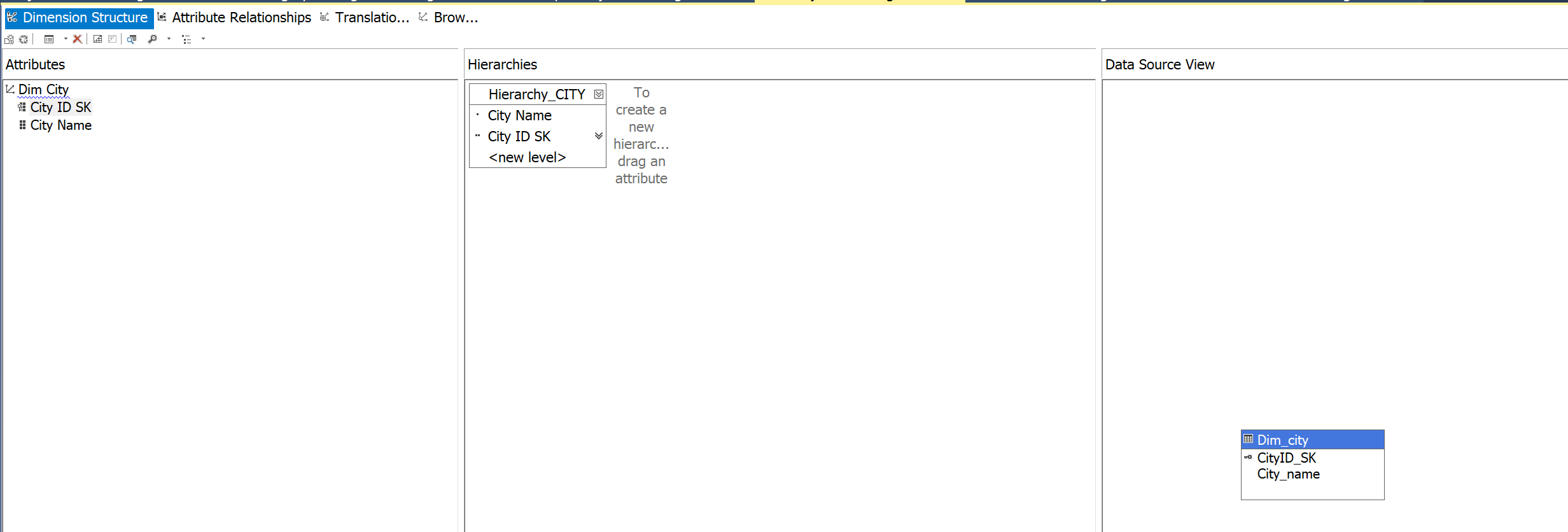
The data source view is shown below:

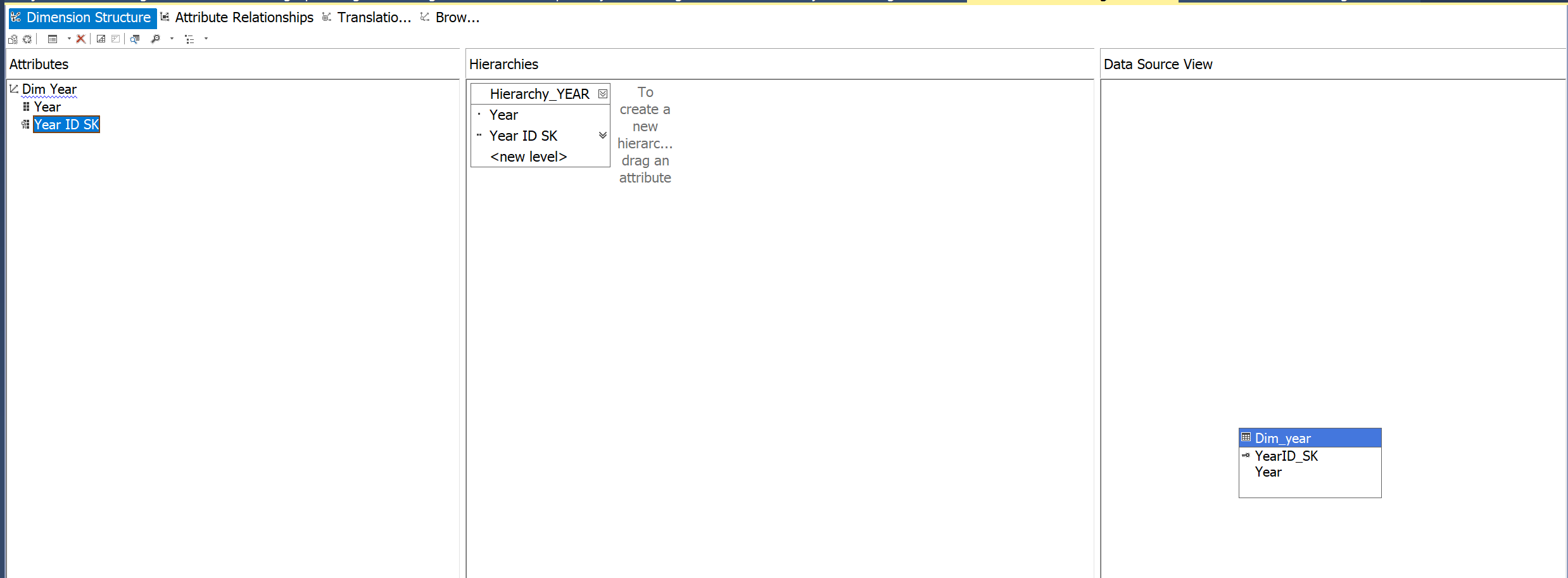


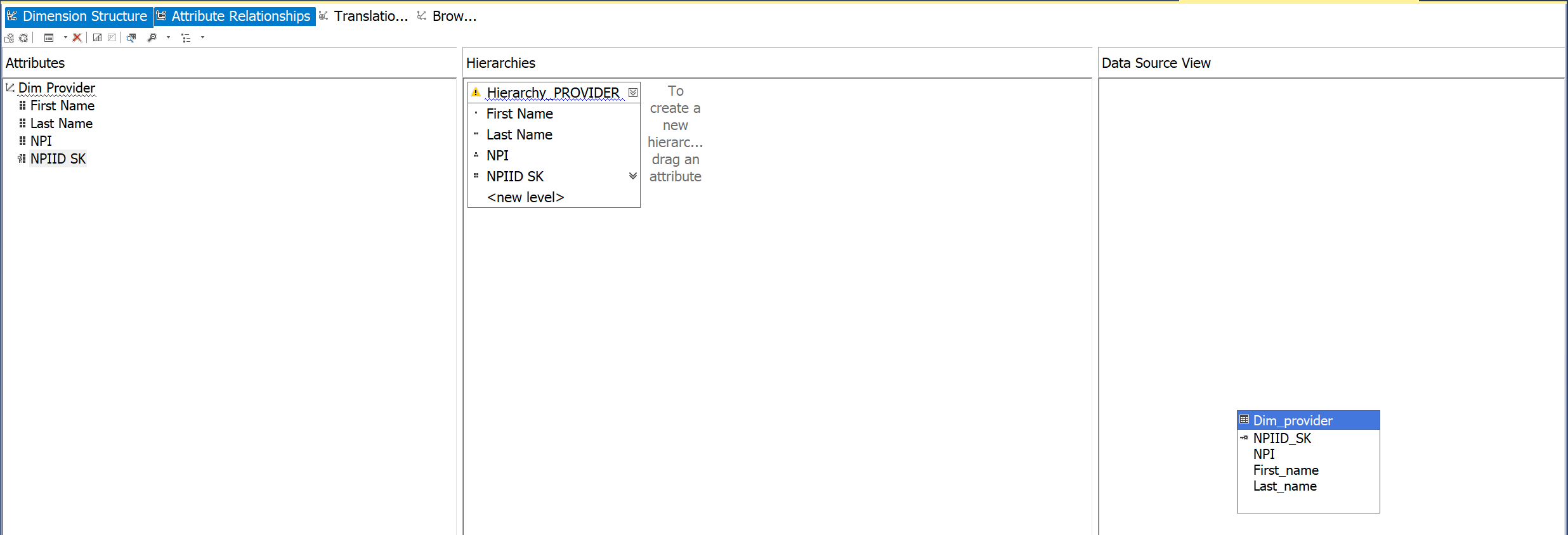
The hierarchies we created are shown as below:



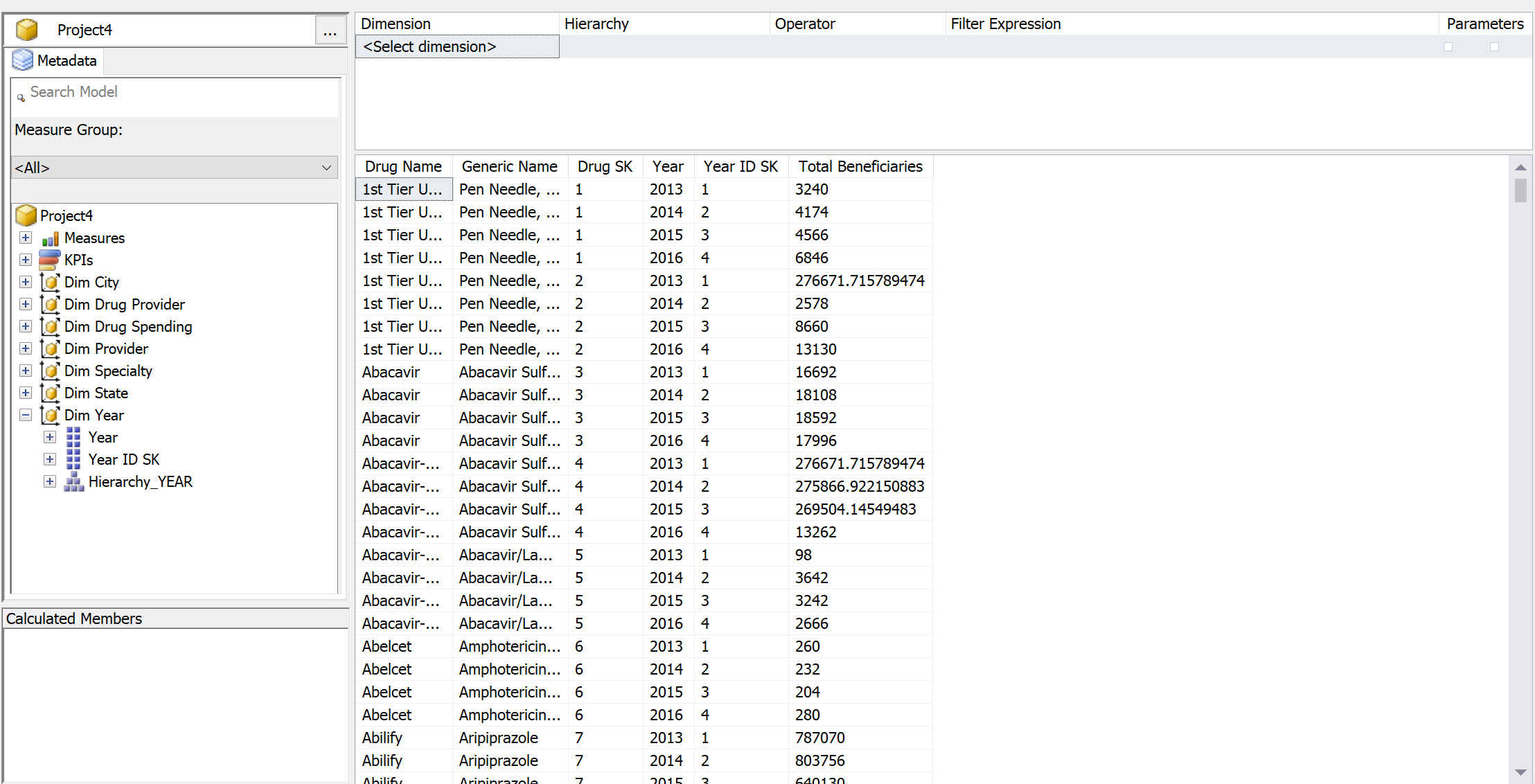








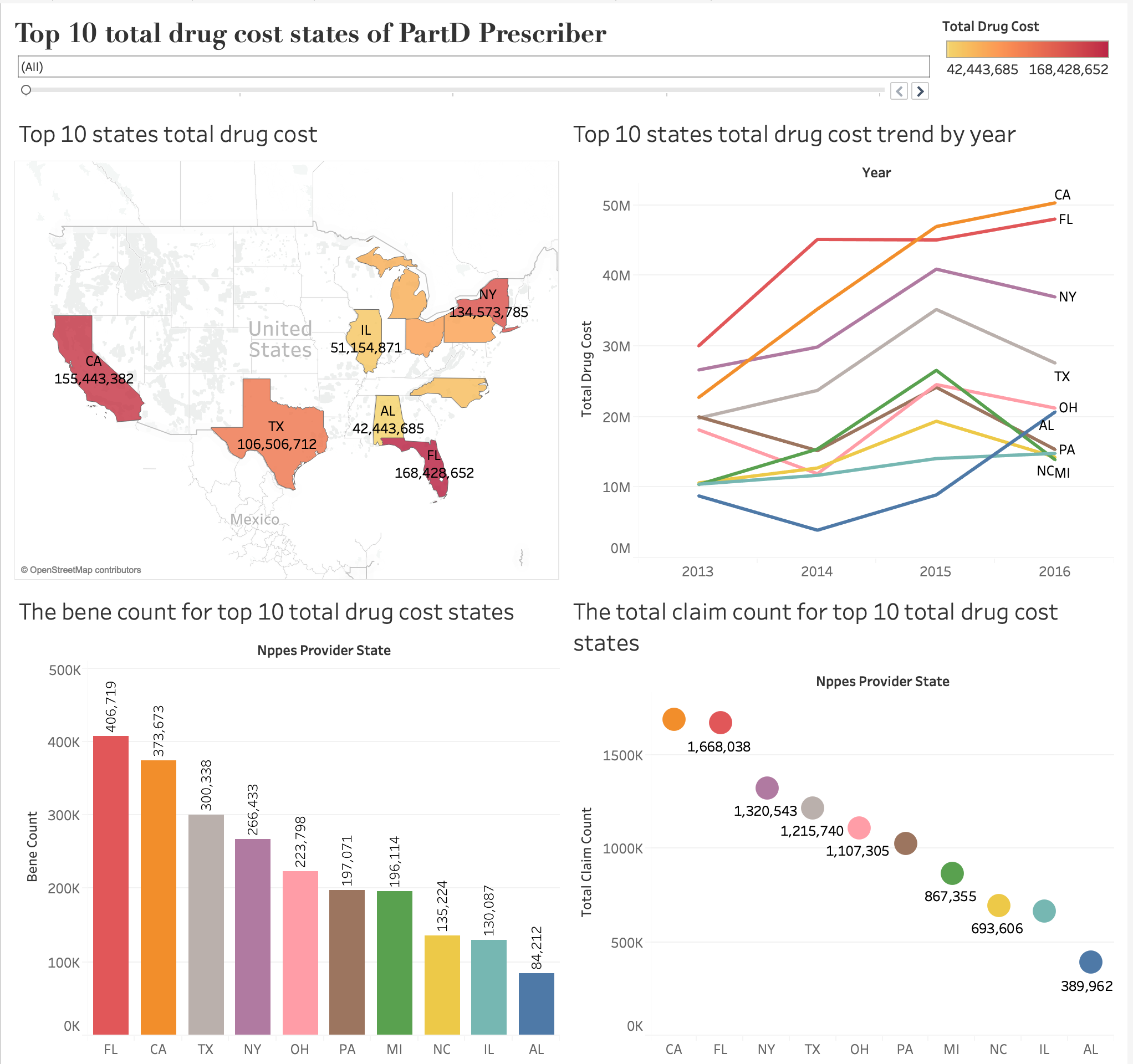
The cube browse results is shown below:



**5. Data Visualization**

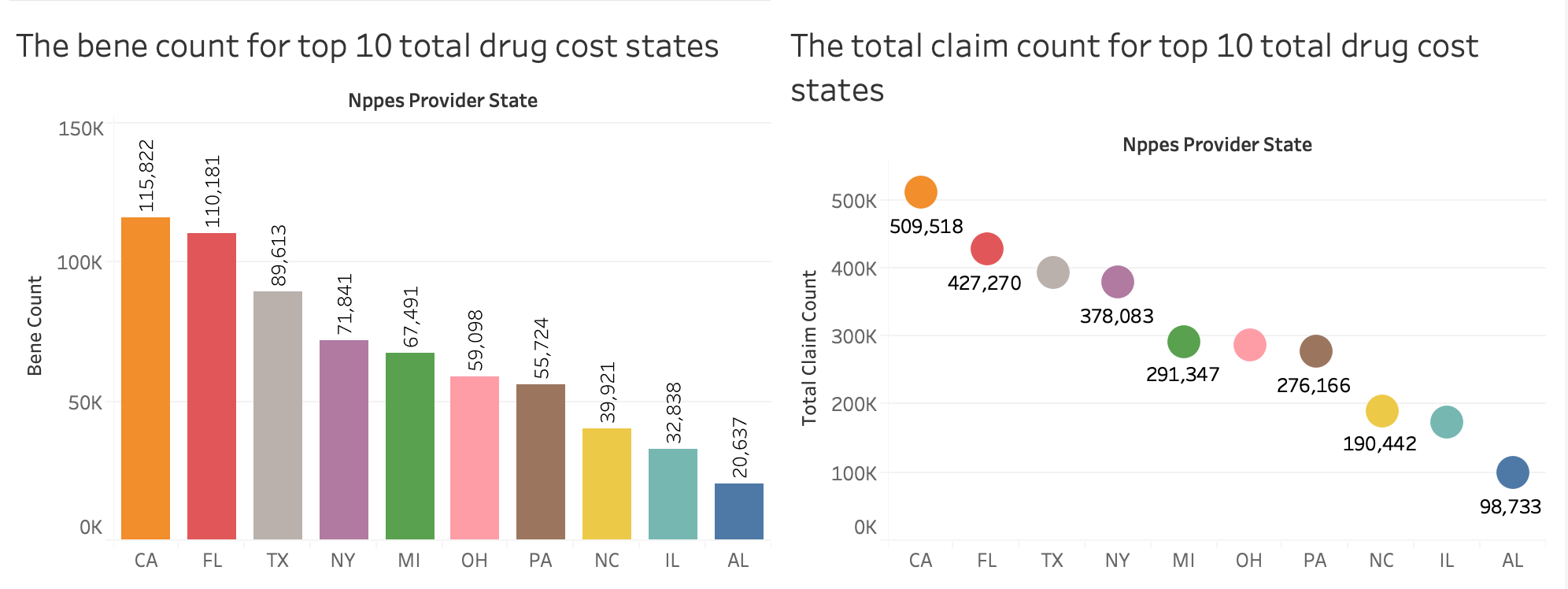
We have created our data visualization using the staging data to gain insights and trends of medicare part d prescriber and drug spending over 4 years from 2013 to 2016 in different states.

The interactive chart below shows the top 10 total drug cost states of PartD Prescriber and the corresponding beneficiaries count and total claim count for these 10 states. We can easily find from the chart that beneficiaries count and total claim count have positive correlations with total drug cost and the drug cost consuming differs from states to states as time went by. From the trend line chart, we can find only California, Florida, Illinois, and Alabama underwent an increasing trend in the total drug cost.

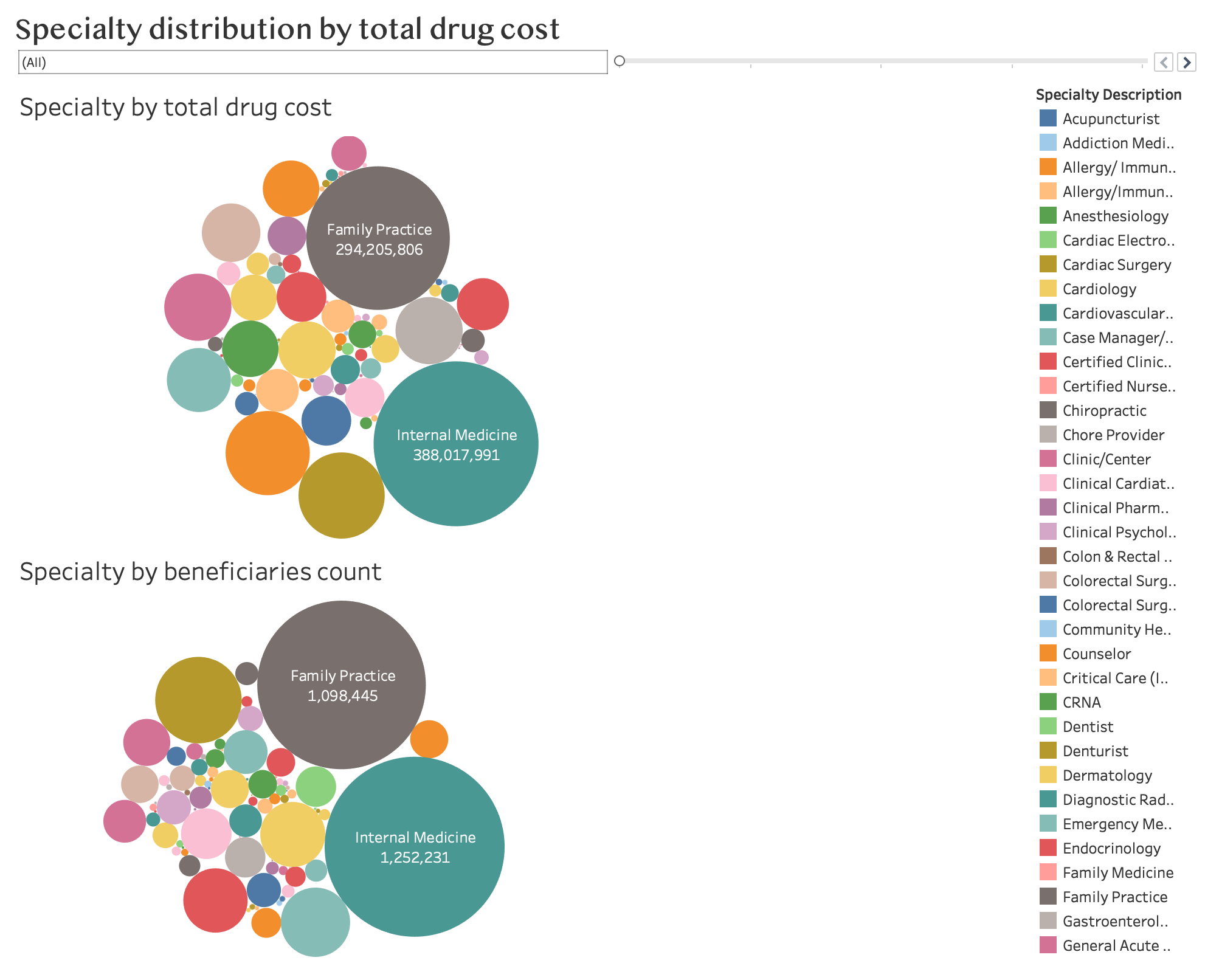


The graph below shows the beneficiaries count and total claim count of the top 10 total drug cost

states, it is clear to find out the orders of the 10 states in these two single graphs are exactly the same in the year 2015. Therefore, we can conclude that there is a positive correlation between beneficiaries count and total claim count.



The following bubble chart is about the specialty of the drug providers which gives us some clear and intuitive information that Internal Medicine and Family Practice are two most dominant specialties in every year.

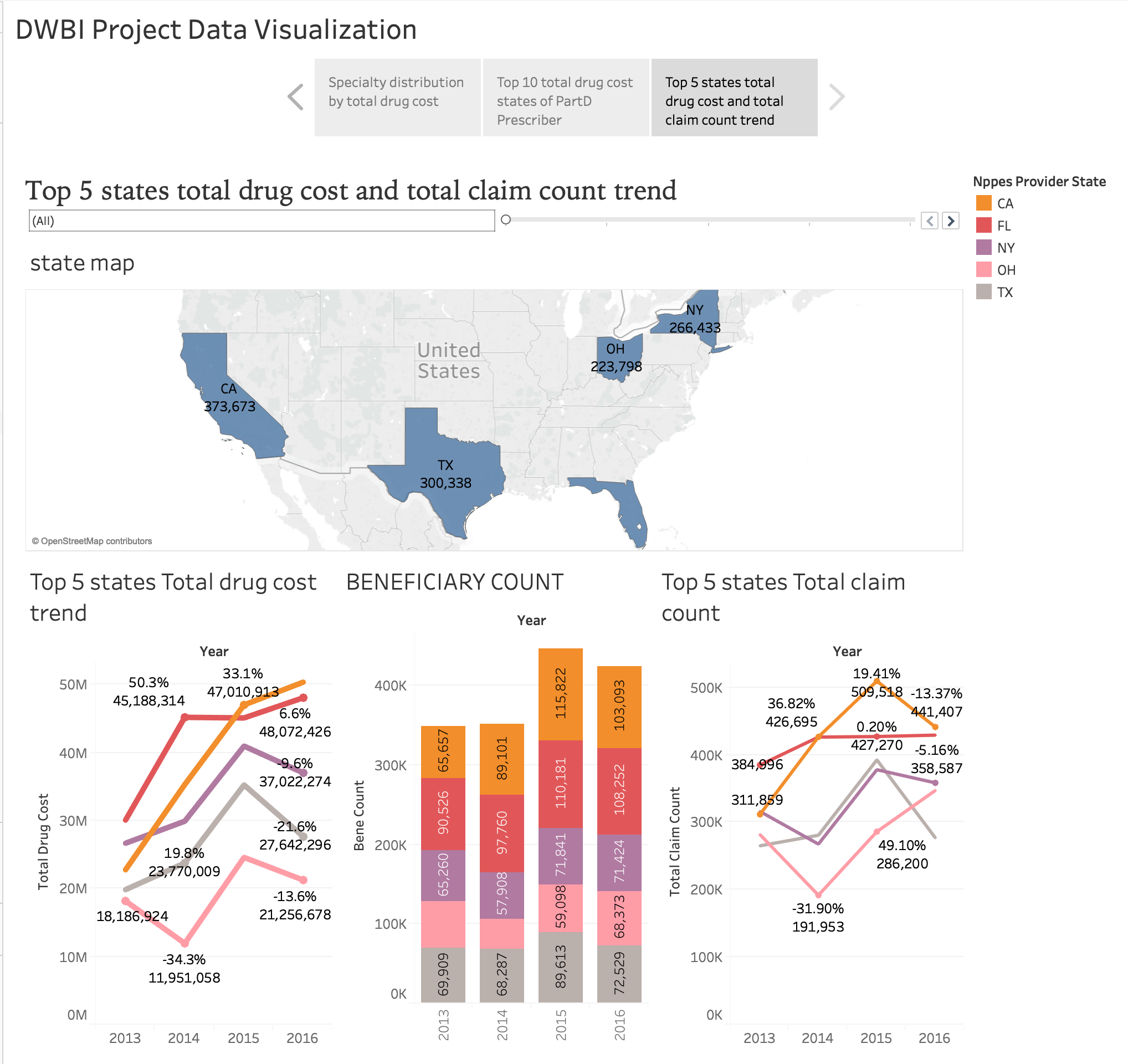


The below graph tells about the total drug cost and total claim count(Total Number of prescription fills for all the drugs) in 5 states during 2013-2016.

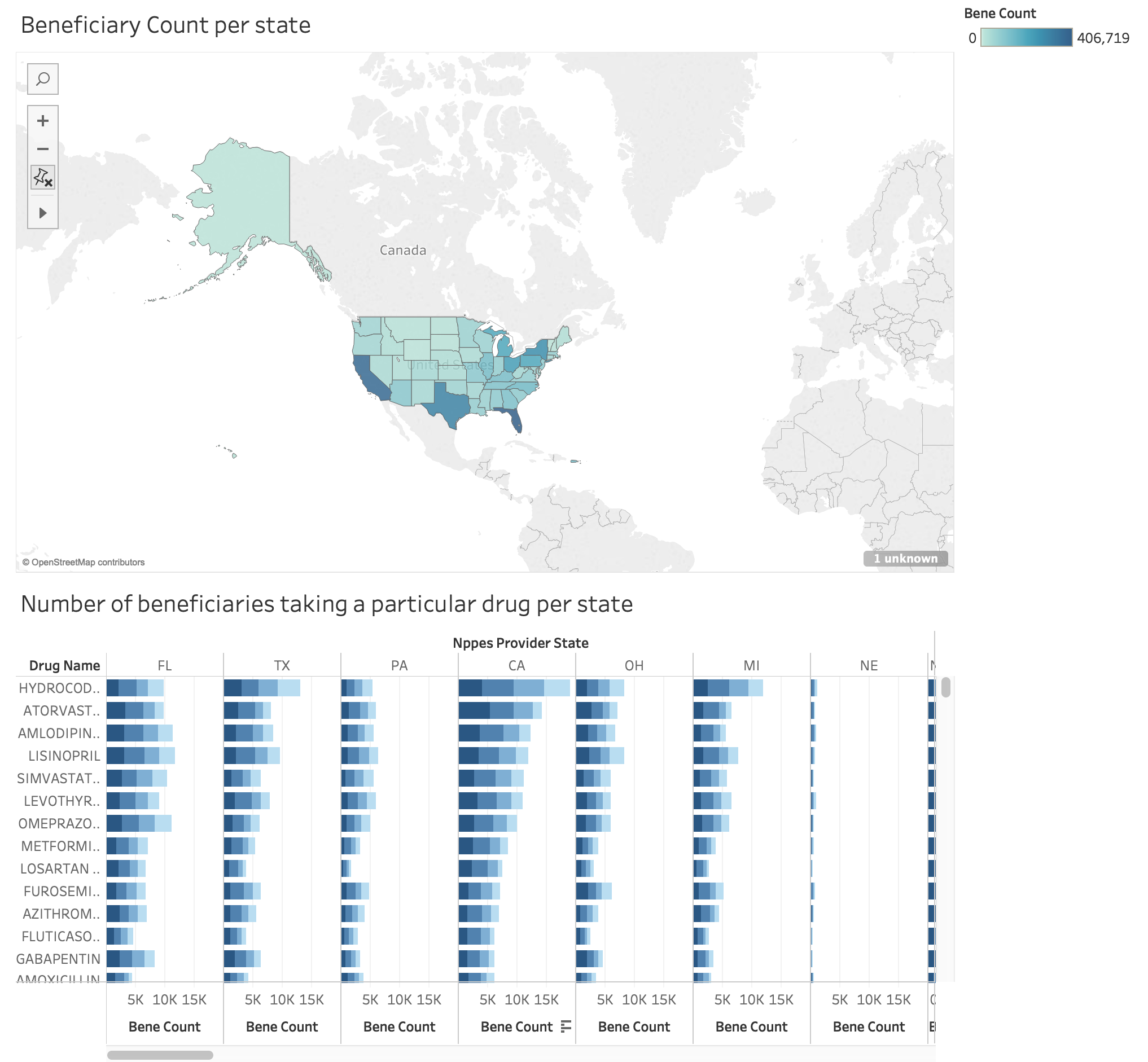
Trends in both the graphs were almost similar with california recording both highest drug cost and highest claim count and ohio recording the lowest counts

There is linear relationship between dug cost and claim count. As claim count increases drug cost also increases and vice-versa

In New york, Ohio , Texas highest percentage change was recorded during 2015 for both the measures whereas in California and Florida highest percent change was recorded during 2014.

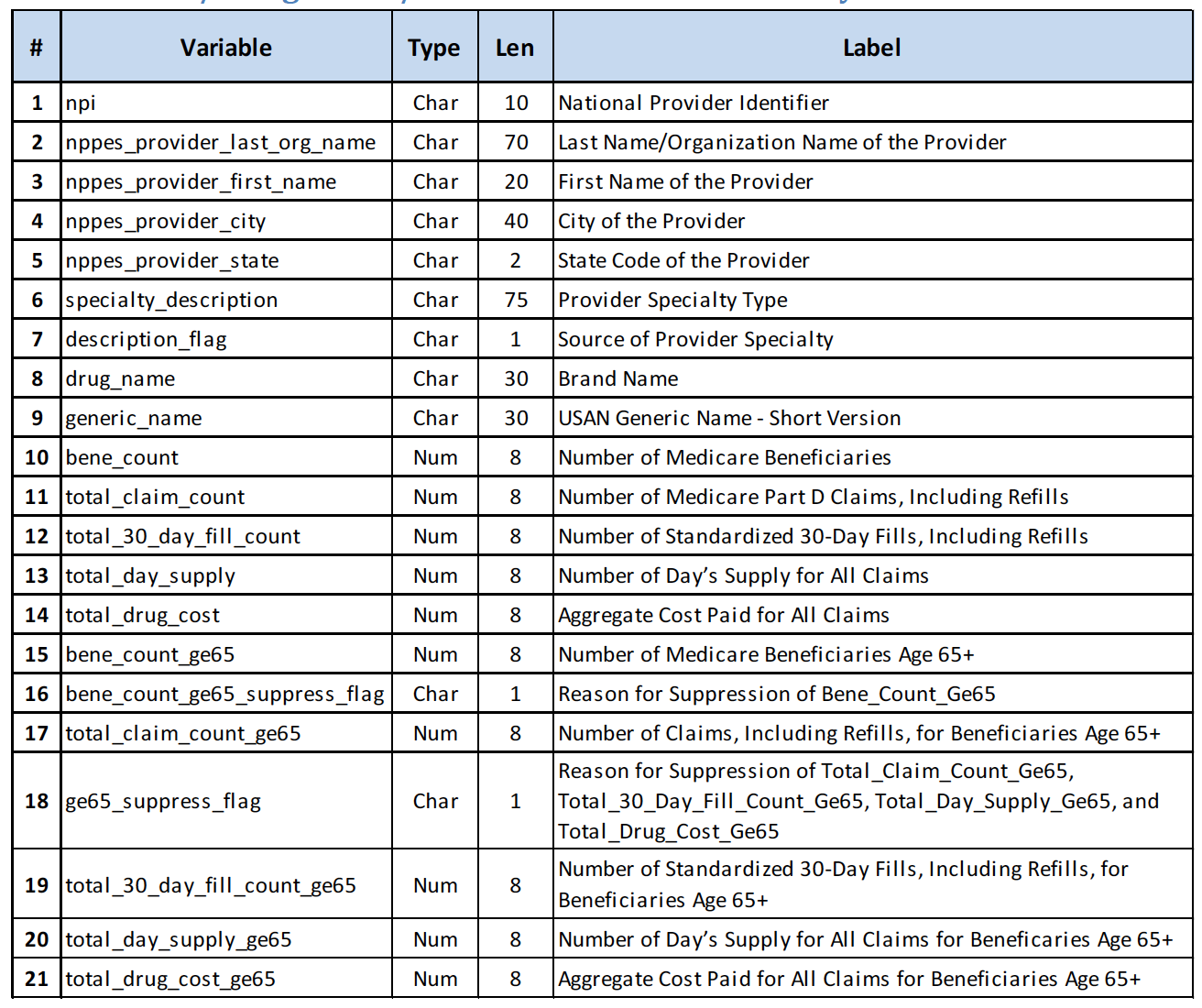


In the dashboard below the map shows the number of beneficiaries per state. The graph shows which drug is being consumed per state. Hydrocodone(narcotic painkiller) is being used the most. People are abusing this drug since it is an addictive and around 10-15% of beneficiaries taking it have died of overdose. The other top 5 heavily consumed drugs are all to control cholestrol and heart blockage indicting the rising obesity levels in these states. Also I found in this graph is the states physically active jobs these drugs were not consumed as much compared to other drugs while the states which mostly have sedentary jobs these drugs were consumed a lot.

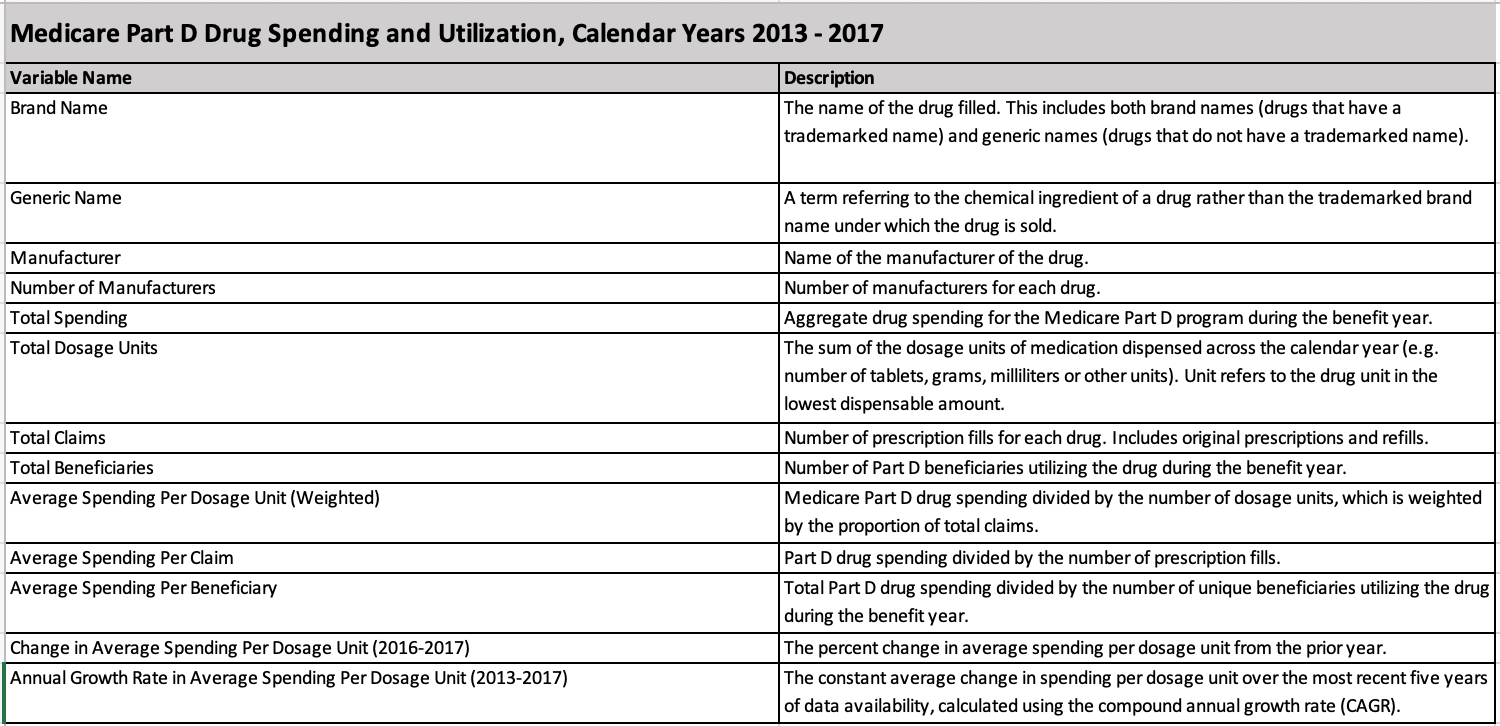


**Appendix**

**1. The Attributes meanings for PartD\_Provider dataset**



**2. The Attributes meanings for PartD\_Provider dataset**



**3. The SQL queries for this project**

**Cleaning drug spending data:**

CREATE TABLE [dbo].[DrugSpending\_2013](

[Brand Name] [nvarchar](255) NULL,

[Generic Name] [nvarchar](255) NULL,

[Number of Manufacturers] [float] NULL,

[2013\_Total Spending] [float] NULL,

[2013\_Total Dosage Units] [float] NULL,

[2013\_Total \_Claims] [float] NULL,

[2013\_Total Beneficiaries] [float] NULL,

[2013\_Average Spending Per Dosage Unit (Weighted)] [float] NULL,

[2013\_Average Spending Per Claim] [float] NULL,

[2013\_Average Spending Per Beneficiary] [float] NULL

) ON [PRIMARY]

insert into [dbo].[DrugSpending\_2013]([Brand Name],[Generic Name],[Number of Manufacturers],[2013\_Total Spending],[2013\_Total Dosage Units],[2013\_Total \_Claims],

[2013\_Total Beneficiaries],[2013\_Average Spending Per Dosage Unit (Weighted)],

[2013\_Average Spending Per Claim],[2013\_Average Spending Per Beneficiary])

select [Brand Name],[Generic Name],[Number of Manufacturers],coalesce([2013\_Total Spending],avg([2013\_Total Spending]) over()) as [2013\_Total Spending],

coalesce([2013\_Total Dosage Units],avg([2013\_Total Dosage Units]) over()) as [2013\_Total Dosage Units],

coalesce([2013\_Total \_Claims],avg([2013\_Total \_Claims]) over()) as [2013\_Total \_Claims],

coalesce([2013\_Total Beneficiaries],avg([2013\_Total Beneficiaries]) over()) as [2013\_Total Beneficiaries],

coalesce([2013\_Average Spending Per Dosage Unit (Weighted)],avg([2013\_Average Spending Per Dosage Unit (Weighted)]) over()) as [2013\_Average Spending Per Dosage Unit (Weighted)],

coalesce([2013\_Average Spending Per Claim],avg([2013\_Average Spending Per Claim]) over()) as [2013\_Average Spending Per Claim],

coalesce([2013\_Average Spending Per Beneficiary],avg([2013\_Average Spending Per Beneficiary]) over()) as [2013\_Average Spending Per Beneficiary] from dbo.Drug\_spending\_2013\_Staging;

ALTER table dbo.DrugSpending\_2013

Add Year varchar(4) default '2013';

UPDATE dbo.DrugSpending\_2013

SET Year = '2013'

WHERE Year is NULL;

SELECT \* FROM dbo.DrugSpending\_2013

**Cleaning PartD\_provider data:**

UPDATE PartD\_Prescriber2016Staging

SET bene\_count = 0

WHERE bene\_count = ''

UPDATE PartD\_Prescriber2016Staging

SET total\_claim\_count = 0

WHERE total\_claim\_count = ''

UPDATE PartD\_Prescriber2016Staging

SET total\_30\_day\_fill\_count = 0

WHERE total\_30\_day\_fill\_count = ''

UPDATE PartD\_Prescriber2016Staging

SET total\_day\_supply = 0

WHERE total\_day\_supply = ''

UPDATE PartD\_Prescriber2016Staging

SET total\_drug\_cost = 0

WHERE total\_drug\_cost = ''

UPDATE PartD\_Prescriber2016Staging

SET bene\_count\_ge65 = 0

WHERE bene\_count\_ge65 = ''

UPDATE PartD\_Prescriber2016Staging

SET total\_claim\_count\_ge65 = 0

WHERE total\_claim\_count\_ge65 = ''

UPDATE PartD\_Prescriber2016Staging

SET total\_30\_day\_fill\_count\_ge65 = 0

WHERE total\_30\_day\_fill\_count\_ge65 = ''

UPDATE PartD\_Prescriber2016Staging

SET total\_day\_supply\_ge65 = 0

WHERE total\_day\_supply\_ge65 = ''

UPDATE PartD\_Prescriber2016Staging

SET total\_drug\_cost\_ge65 = 0

WHERE total\_drug\_cost\_ge65 = ''

ALTER table dbo.PartD\_Prescriber2016Staging

Add Year varchar(4) default '2016';

UPDATE dbo.PartD\_Prescriber2016Staging

SET Year = '2016'

WHERE Year is NULL;

SELECT \* FROM dbo.PartD\_Prescriber2016Staging

**Generate dimensions and facts:**

-- Create tables section -------------------------------------------------

-- Table Dim\_state

create database project4;

use project4;

CREATE TABLE [Dim\_state]

(

[State\_code] Varchar(500) NULL,

[StateID\_SK] Int IDENTITY(1,1) NOT NULL

)

go

-- Add keys for table Dim\_state

ALTER TABLE [Dim\_state] ADD CONSTRAINT [PK\_Dim\_state] PRIMARY KEY ([StateID\_SK])

go

-- Table Dim\_city

CREATE TABLE [Dim\_city]

(

[City\_name] Varchar(500) NULL,

[CityID\_SK] Int IDENTITY(1,1) NOT NULL

)

go

-- Add keys for table Dim\_city

ALTER TABLE [Dim\_city] ADD CONSTRAINT [PK\_Dim\_city] PRIMARY KEY ([CityID\_SK])

go

-- Table Dim\_specialty

CREATE TABLE [Dim\_specialty]

(

[Specialty\_description] Varchar(500) NULL,

[SpecialtyID\_SK] Int IDENTITY(1,1) NOT NULL

)

go

-- Add keys for table Dim\_specialty

ALTER TABLE [Dim\_specialty] ADD CONSTRAINT [PK\_Dim\_specialty] PRIMARY KEY ([SpecialtyID\_SK])

go

-- Table Dim\_drug\_provider

CREATE TABLE [Dim\_drug\_provider]

(

[Drug\_name] Varchar(500) NULL,

[Generic\_name] Varchar(500) NULL,

[DrugID\_SK] Int IDENTITY(1,1) NOT NULL

)

go

-- Add keys for table Dim\_drug\_provider

ALTER TABLE [Dim\_drug\_provider] ADD CONSTRAINT [PK\_Dim\_drug\_provider] PRIMARY KEY ([DrugID\_SK])

go

-- Table Dim\_provider

CREATE TABLE [Dim\_provider]

(

[NPI] Int NOT NULL,

[First\_name] Varchar(500) NULL,

[Last\_name] Varchar(500) NULL,

[NPIID\_SK] Int IDENTITY(1,1) NOT NULL

)

go

-- Add keys for table Dim\_provider

ALTER TABLE [Dim\_provider] ADD CONSTRAINT [PK\_Dim\_provider] PRIMARY KEY ([NPIID\_SK])

go

-- Table Dim\_year

CREATE TABLE [Dim\_year]

(

[Year] Varchar(4) NOT NULL,

[YearID\_SK] Int IDENTITY(1,1) NOT NULL

)

go

-- Add keys for table Dim\_year

ALTER TABLE [Dim\_year] ADD CONSTRAINT [PK\_Dim\_year] PRIMARY KEY ([YearID\_SK])

go

-- Table Fact\_drug

CREATE TABLE [Fact\_drug]

(

[Number of Manufacturers] Float NULL,

[Total Spending] Float NULL,

[Total Dosage Units] Float NULL,

[Total \_Claims] Float NULL,

[Total Beneficiaries] Float NULL,

[Average Spending Per Dosage Unit (Weighted)] Float NULL,

[Average Spending Per Claim] Float NULL,

[Average Spending Per Beneficiary] Float NULL,

[YearID\_SK] Int NULL,

[Drug\_SK] Int NULL

)

go

-- Create indexes for table Fact\_drug

CREATE INDEX [IX\_Relationship34] ON [Fact\_drug] ([YearID\_SK])

go

CREATE INDEX [IX\_Relationship38] ON [Fact\_drug] ([Drug\_SK])

go

-- Table Fact\_provider

CREATE TABLE [Fact\_provider]

(

[bene\_count] Float NULL,

[total\_claim\_count] Float NULL,

[total\_30\_day\_fill\_count] Float NULL,

[total\_day\_supply] Float NULL,

[total\_drug\_cost] Float NULL,

[bene\_count\_ge65] Float NULL,

[total\_claim\_count\_ge65] Float NULL,

[total\_30\_day\_fill\_count\_ge65] Float NULL,

[total\_day\_supply\_ge65] Float NULL,

[total\_drug\_cost\_ge65] Float NULL,

[DrugID\_SK] Int NULL,

[NPIID\_SK] Int NULL,

[YearID\_SK] Int NULL,

[StateID\_SK] Int NULL,

[CityID\_SK] Int NULL,

[SpecialtyID\_SK] Int NULL

)

go

-- Create indexes for table Fact\_provider

go

CREATE INDEX [IX\_Relationship35] ON [Fact\_provider] ([DrugID\_SK])

go

CREATE INDEX [IX\_Relationship36] ON [Fact\_provider] ([NPIID\_SK])

go

CREATE INDEX [IX\_Relationship37] ON [Fact\_provider] ([YearID\_SK])

go

CREATE INDEX [IX\_Relationship39] ON [Fact\_provider] ([StateID\_SK])

go

CREATE INDEX [IX\_Relationship40] ON [Fact\_provider] ([CityID\_SK])

go

CREATE INDEX [IX\_Relationship41] ON [Fact\_provider] ([SpecialtyID\_SK])

go

-- Table Dim\_drug\_spending

CREATE TABLE [Dim\_drug\_spending]

(

[Drug\_name] Varchar(500) NULL,

[Generic\_name] Varchar(500) NULL,

[Drug\_SK] Int IDENTITY(1,1) NOT NULL

)

go

-- Add keys for table Dim\_drug\_spending

ALTER TABLE [Dim\_drug\_spending] ADD CONSTRAINT [PK\_Dim\_drug\_spending] PRIMARY KEY ([Drug\_SK])

go

-- Create foreign keys (relationships) section -------------------------------------------------

ALTER TABLE [Fact\_drug] ADD CONSTRAINT [Relationship34] FOREIGN KEY ([YearID\_SK]) REFERENCES [Dim\_year] ([YearID\_SK]) ON UPDATE NO ACTION ON DELETE NO ACTION

go

ALTER TABLE [Fact\_provider] ADD CONSTRAINT [Relationship35] FOREIGN KEY ([DrugID\_SK]) REFERENCES [Dim\_drug\_provider] ([DrugID\_SK]) ON UPDATE NO ACTION ON DELETE NO ACTION

go

ALTER TABLE [Fact\_provider] ADD CONSTRAINT [Relationship36] FOREIGN KEY ([NPIID\_SK]) REFERENCES [Dim\_provider] ([NPIID\_SK]) ON UPDATE NO ACTION ON DELETE NO ACTION

go

ALTER TABLE [Fact\_provider] ADD CONSTRAINT [Relationship37] FOREIGN KEY ([YearID\_SK]) REFERENCES [Dim\_year] ([YearID\_SK]) ON UPDATE NO ACTION ON DELETE NO ACTION

go

ALTER TABLE [Fact\_drug] ADD CONSTRAINT [Relationship38] FOREIGN KEY ([Drug\_SK]) REFERENCES [Dim\_drug\_spending] ([Drug\_SK]) ON UPDATE NO ACTION ON DELETE NO ACTION

go

ALTER TABLE [Fact\_provider] ADD CONSTRAINT [Relationship39] FOREIGN KEY ([StateID\_SK]) REFERENCES [Dim\_state] ([StateID\_SK]) ON UPDATE NO ACTION ON DELETE NO ACTION

go

ALTER TABLE [Fact\_provider] ADD CONSTRAINT [Relationship40] FOREIGN KEY ([CityID\_SK]) REFERENCES [Dim\_city] ([CityID\_SK]) ON UPDATE NO ACTION ON DELETE NO ACTION

go

ALTER TABLE [Fact\_provider] ADD CONSTRAINT [Relationship41] FOREIGN KEY ([SpecialtyID\_SK]) REFERENCES [Dim\_specialty] ([SpecialtyID\_SK]) ON UPDATE NO ACTION ON DELETE NO ACTION

go