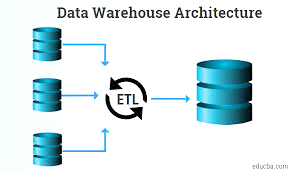
**Technical Report for Building Data Warehouse for**

**Drug Enforcement Administration**

****

**Rishav**: 10536571

**Nayil**: 10541185

**Vishal Tyagi**: 10535541

**Vivek Sanu**: 10534553

Contents

[1. Introduction 2](#_Toc27418083)

[1.1. Reasons for selecting the subject area AND DATA 2](#_Toc27418084)

[1.2. Key StakeHolders 3](#_Toc27418085)

[1.3. Vision and Goals 3](#_Toc27418086)

[1.4. Business requirements 3](#_Toc27418087)

[2. SCHEMA 3](#_Toc27418088)

[3. ETL 7](#_Toc27418089)

[4. VISUALIZATIONS AND REPORTS 12](#_Toc27418090)

[4.1. Visualizations 12](#_Toc27418091)

[4.2. Reports 16](#_Toc27418092)

[5. Include XML and Schema 20](#_Toc27418093)

[6. Graph Databases 23](#_Toc27418094)

[6.1. COMAPRISON to realtional databases 27](#_Toc27418095)

[7. Conclusions 27](#_Toc27418096)

[8. Bibliography 28](#_Toc27418097)

[Appendix A – Visualization Code 29](#_Toc27418098)

[Appendix B – Neo 4J code 30](#_Toc27418099)

[Appendix C – SQL Task (Drug Incident Mapping) 31](#_Toc27418100)

[Appendix D – Drugs categories 33](#_Toc27418101)

[Appendix E – R Code for data extraction and cleaning 34](#_Toc27418102)

[Appendix F – SQL TASK(Transforming null Values) 35](#_Toc27418103)

# 1. Introduction

In this project we have built a data warehouse for the Drug Enforcement Agency of USA, which would help them in understanding different drug consumption pattern across the state of Connecticut. The main source have been given by the state agency which contains the toxicology details and other relevant information of the individuals who have died due to drug overdose from the year 2012 to 2018 in the state of Connecticut.

We have followed the Star Schema architecture in developing the warehouse and is implemented in MS SQL Server. Various tools including SSIS, SSRS have been used for ETL and report generation process. A part of the project has also been replicated using XML and Neo4J technologies.

A total of 3 different datasets have been used in the project:

1. Dataset containing information of deceased who have consumed drugs and were the citizens of United States from different counties of Connecticut, and in total it contains around 5000 records.
2. Dataset containing the information about the average income of population in different counties of United States.
3. Dataset containing the details of drugs like price, category etc. which are popular in United States.

* By this Data Warehouse project, we have tried to draw important insights, which can be helpful if presented to authorities.
* Data which was collected from the different websites are mentioned under bibliography section.
* We have categorized the population based on age and for that we have referred WHO report specifically talking about the deaths in US. Categories are mentioned below:

1-14 – Children, 15-24 – Youth, 25-44 - Young Adulthood, 45-64 – Middle Adulthood, 65+- Senior Adulthood

## 1.1. Reasons for selecting the subject area AND DATA

Connecticut takes third place in terms of real GDP per capita of $64,511 for 2018 and a per-capita, current-dollar GDP of $73,643 in United States. According to the National Survey on Drug Use and Health (NSDUH), 9.10 % of Connecticut residents reported using drugs in recent past and this is more than national average of 8.82 %. Though Marijuana is common drug but here in Connecticut 3.48 % residents reported using illicit drug other than Marijuana which is more than average of 3.33 % national average.

These figures worried Drug Enforcement Administration and they would like to know the exact situation of the state by knowing the facts like:

* Consumption of which drug is significant in Connecticut, so by knowing this they can track the drug source.
* In which county drugs are more popular, this will help them to increase the surveillance in those counties.
* Finally, can alert people about the consequence of doing drugs by using facts and figures in their campaign.

So, by having this aim in mind we have done this project from the perspective US Drug Enforcement Administration analysis wing.

## 1.2. Key StakeHolders

Below are the key stakeholders:

* Drug Enforcement Administration (DEA)
* Inter-American Drug Abuse Control Commission (CICAD)
* Secretariat for Multidimensional Security (SMS)
* State and Local Police of Connecticut

## 1.3. Vision and Goals

The vision and goal of this project is to alert Drug Enforcement Administration about drugs which are popular in the Connecticut state and along with the name of counties which are highly vulnerable to drugs with other contributing factors.

## 1.4. Business requirements

Below questions can be answered after this study:

* Which drugs have caused highest deaths in Connecticut?
* Which county is highly vulnerable to drugs?
* What are the figures of death because of drugs overdose in last 7 years?
* Is there any relation between average income of people and drug user’s county wise?

# 2. SCHEMA

Datasets are huge and individually they will not solve the purpose, so in order to combine the data from multiple sources and store it in one place, we need to design a data warehouse. So, a star schema is developed in order to design the data warehouse because of following benefits:

It will help end-user to understand the application and easily navigate through it. And a well-designed schema helps users to quickly analyze large, multidimensional data sets. Main advantages of star schemas are:

* **Improves query performance**

Execution of queries runs faster in comparison to OLTP system because of star schema database design and reason for that is it contains small number of tables and clear join paths. Like join queries which are large and fetch data from multiple tables takes minimal time for execution. The dimensions in this design are connected to fact table and when two-dimension tables are used in a query, only one join path, intersecting the fact table, exists between those two tables. This design feature enforces accurate and consistent query results.

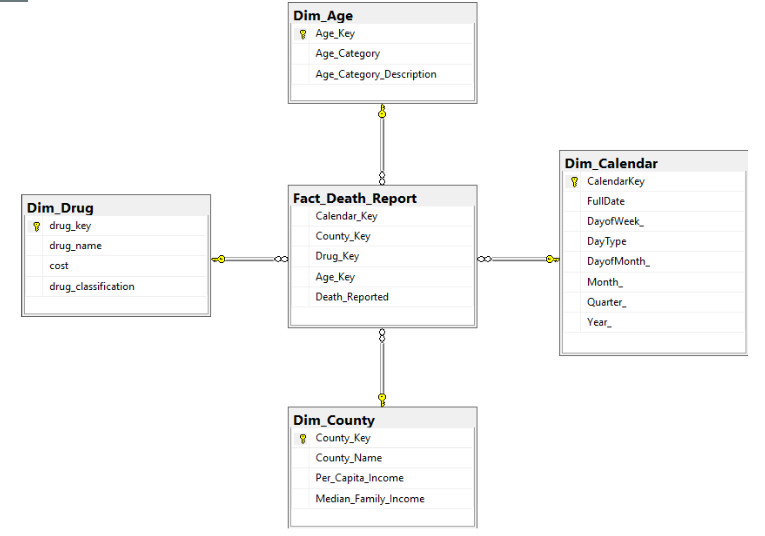
* **Administration and performance of Load**

Time required to load large batches of data significantly reduced in this design because structure simplicity. Also, the impact of load operation is reduced as we define the dimensions and facts separately. Altering the fact table is required only whenever new record is required to append and dimensions tables occasionally refreshed. So, it increases the efficiency of the system.

* **Easy to comprehend**

This design is easy to understand and optimized navigation through the database because of its uncomplex structure. End users can use the joins and fetch the information required without bothering about the underlying business. Users can also browse dimension table attributes before constructing a query.

**Star schema design**



**Details of Dimensions used in the project:**

* **Dim\_Drug**

In this dimension we have data of various drugs along with price, category according to US government, etc.

**Primary Key**: Drug\_Key

* **Dim\_Age**

This dimension has the data of various age categories according to the UN categorization of deaths and the description of the categories.

**Primary Key**: Age\_Key

* **Dim\_Calendar**

This dimension helps by allowing to perform all operations related to dates, months, years etc.

**Primary Key**: Calendar\_Key

* **Dim\_County**

This dimension contains the data of county name, per capita income and median family income.

**Primary Key**: County\_Key

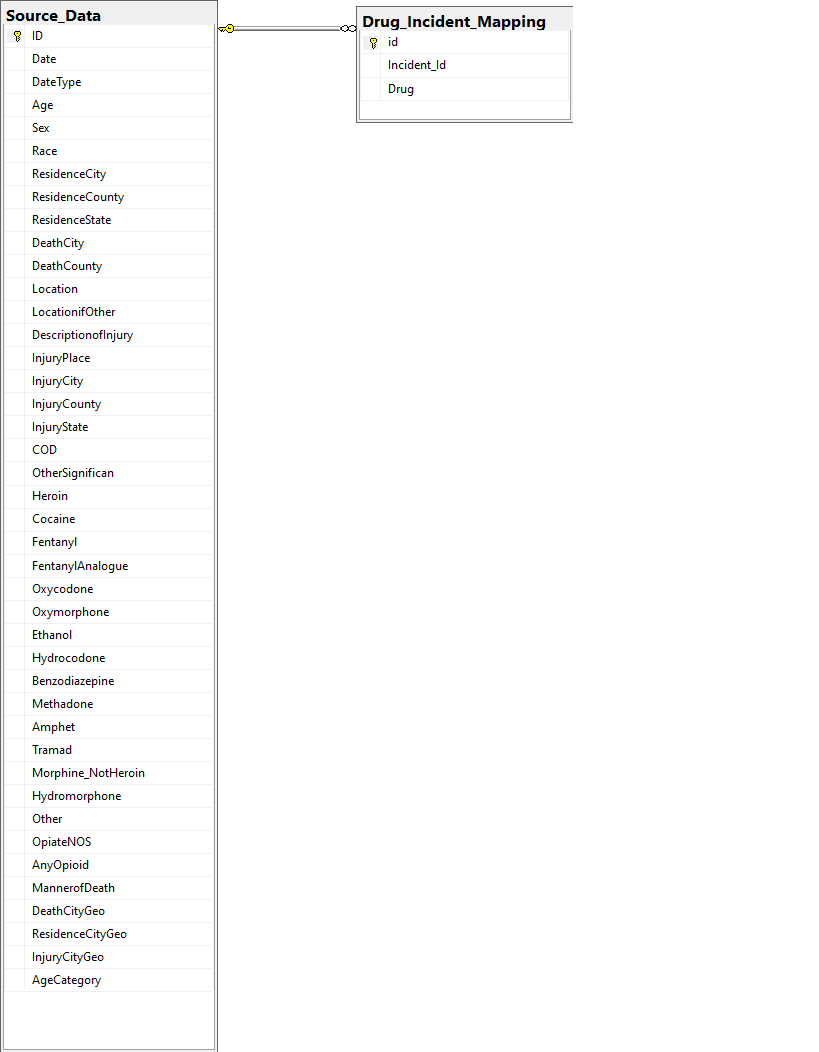
**Fact Table**

* **Fact\_Death\_Report**

This fact table contains the total number of death reported on a particular day at specific county caused due to a particular drug and falling under a specific age category. The fact is recorded in Death\_Reported.

**Primary Key**: (Calendar\_Key, County\_Key, Age\_Key, Drug\_Key), it is a composite key.

**Schema Used in Staging For ETL**



**Source\_Data**

The Source\_Data table used in staging process is a direct replica of the excel provided by the government with an addition of AgeCategory field, used to map AgeCategory to an incident.

**Drug\_Incident\_Mapping**

This table contains the list of all drugs consumed by the deceased in such a manner that it can be mapped to the dimension Dim\_Drug. The data in the source file contains the all the list in a single row making it unsuitable for a direct lookup in SSIS.

Id : Primary Key

Incident\_Id: Foreign key to id of Source\_Data

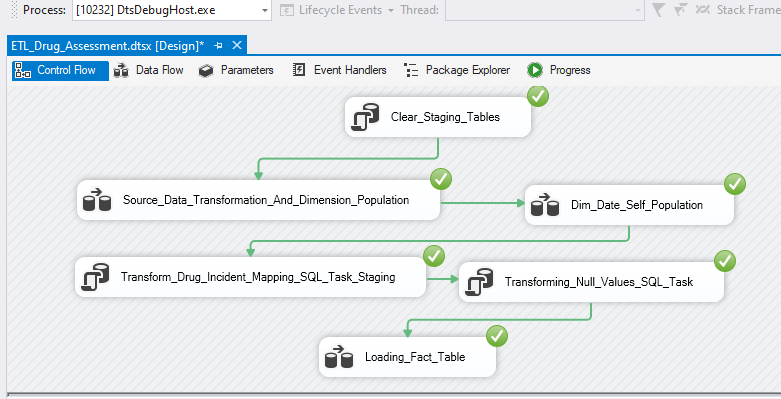
Drug: Name of drug used by the deceased with Incident\_Id in source data

# 3. ETL

In this project all stages of ETL (Extraction, Transformation and Loading) has been take care of only in one package.

The source data for Dim\_County was extracted using R and the script for the same is given in the appendix.

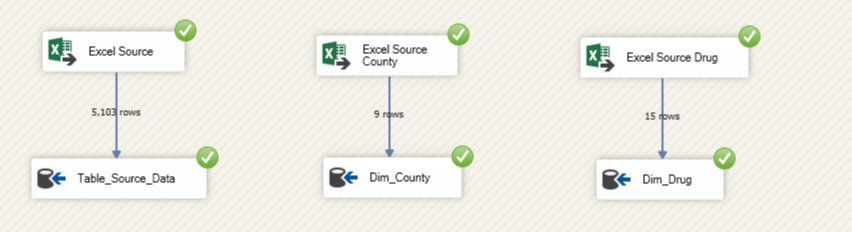
The following diagram represents the overall flow of the process right from extracting the data from various sources, staging and transforming the data and finally loading into the fact table.



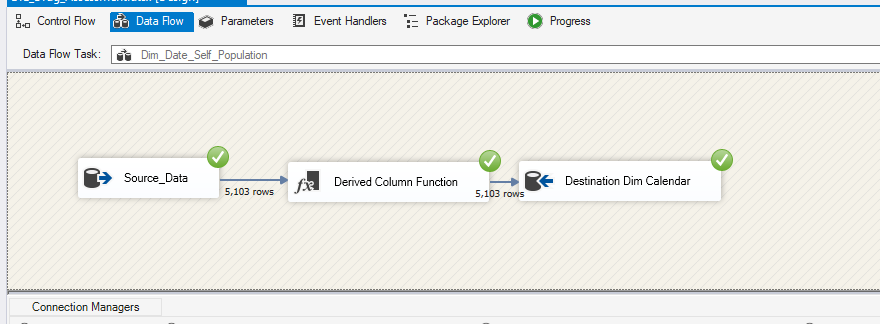
**Clear Staging Tables**

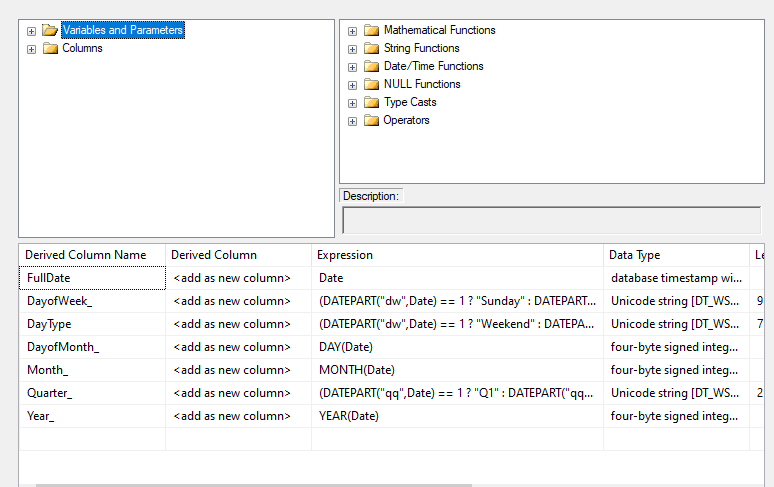
In this SQL Task, we empty the tables Data\_Source and Drug\_Incident\_Mapping.

**Extracting Values into Dimensions and Staging Table**

Here we extract values form different sources in to the staging table “Source Data” and dimensions “Dim\_County” and “Dim\_Drug “.

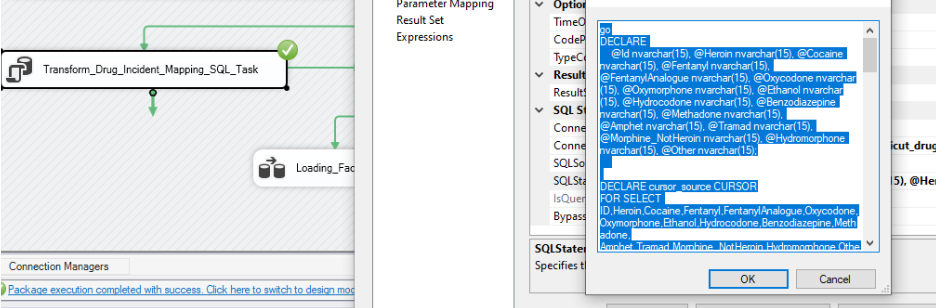
The dimension Dim\_Calendar is derived from the staging table “Source\_Data”





**Insertion in Drug\_Incident\_Mapping (Staging)**

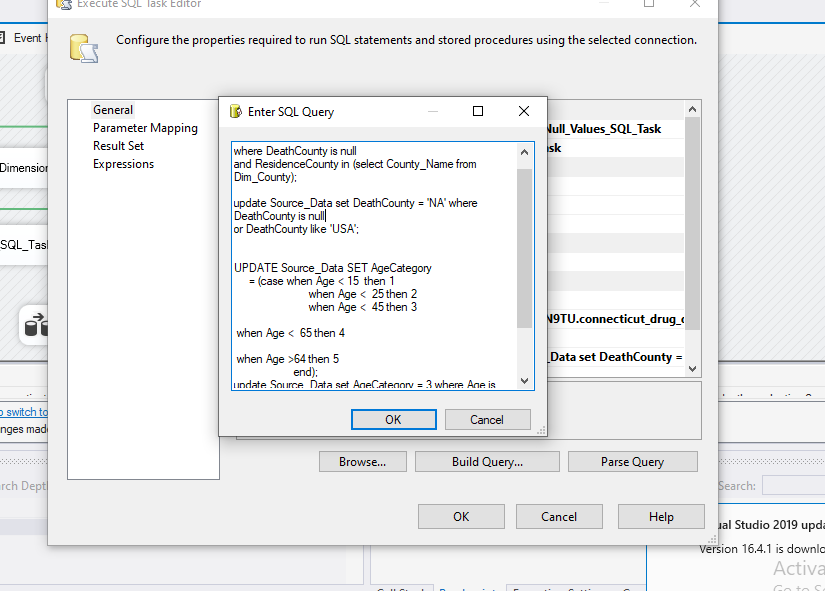
Insertion into this table is done using an “Sql Task Flow” component. The function used to the do the same can be found in Appendix C.



**Transformation of Data**

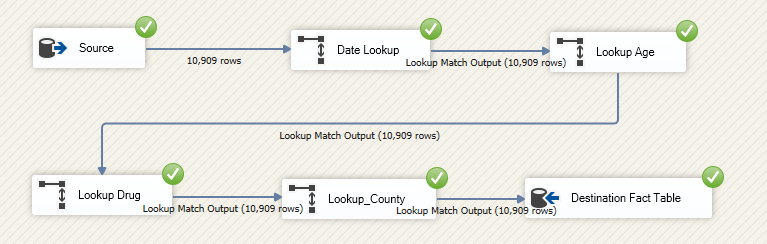
* **Replacing the NULL values** present in the field of death county with resident county, as here we have assumed if death county is null then incident might happen in the resident county of the deceased. And we build up this assumption based on trend in the data records as most of the time death counties and resident counties are same.
* **Categorizing the age** of the deceased so that data can make more sense and we can draw some useful insights from it. For this we have used the categorization provided under **Provisional Guidelines on Standard International Age Classification,** UN. They have provided statistical arguments to categorize the ages, so we built up on their findings.

The above transformations are done using an SQL Task component

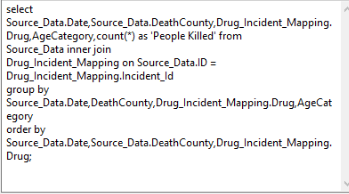


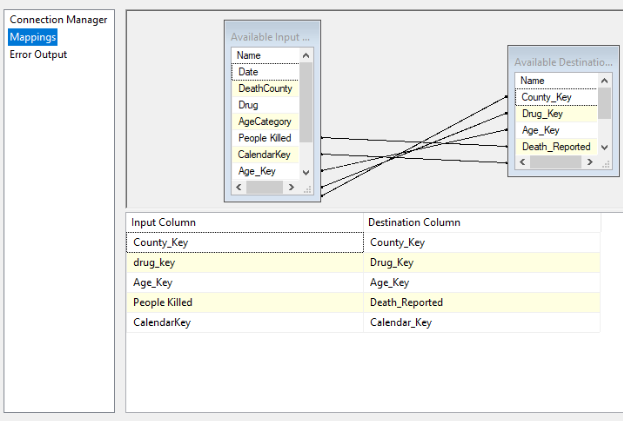
**Loading**

Using this data flow, we have used various lookups which loads the results into the fact table.



This is the SQL query we have used in the source to slect desired data for our project.



Final mapping of the source data and fact table, this helps to feed data into correct fields.

# 4. VISUALIZATIONS AND REPORTS

## 4.1. Visualizations

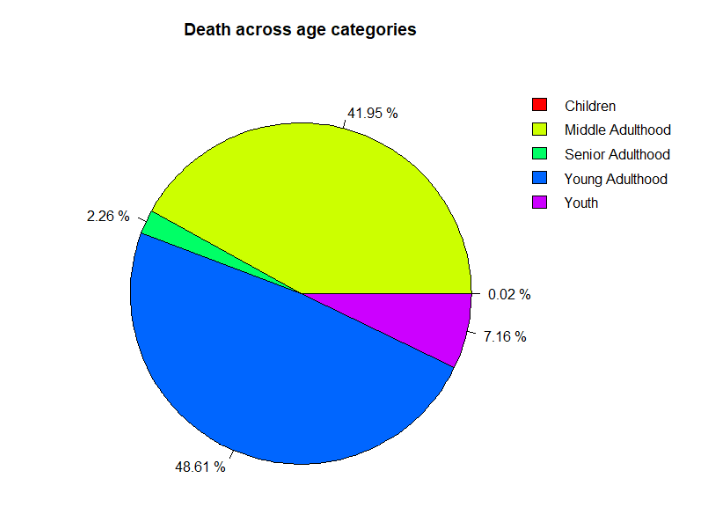
Data visualization is something which helps to read a big amount of data easily and helps to understand the trends in the data. This is something which is result of whole purpose of storing, manipulating and analyzing the data through different medias and finally presenting the outcomes using the data visualization methods to stakeholders. In our project it is crucial to study the trends and some important facts and figures of data. We have visualized our data by using tool: **R Programming**

**Visualization using R Programming**

Go to [**Appendix A**](#_Appendix_A_–) for visualization R code

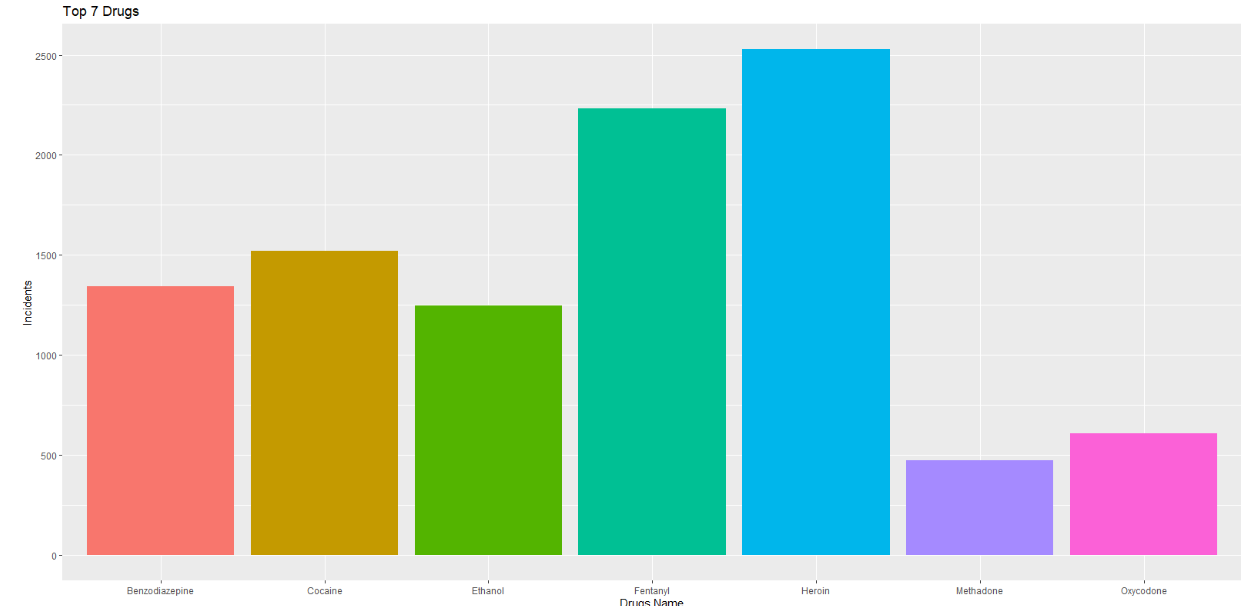
**Visualization 1:**

This pie chart shows the number of deaths according to the age category. We can say that most of the deaths because of drugs are from **young adult** category and this is followed by **middle adult.** So, sign of relief is children are not prone to this addiction and a lot can be done to protect them from this habit. But concerning categories Middle and Young adults needs to be targeted specifically through different campaigns and spreading awareness about drug usage.



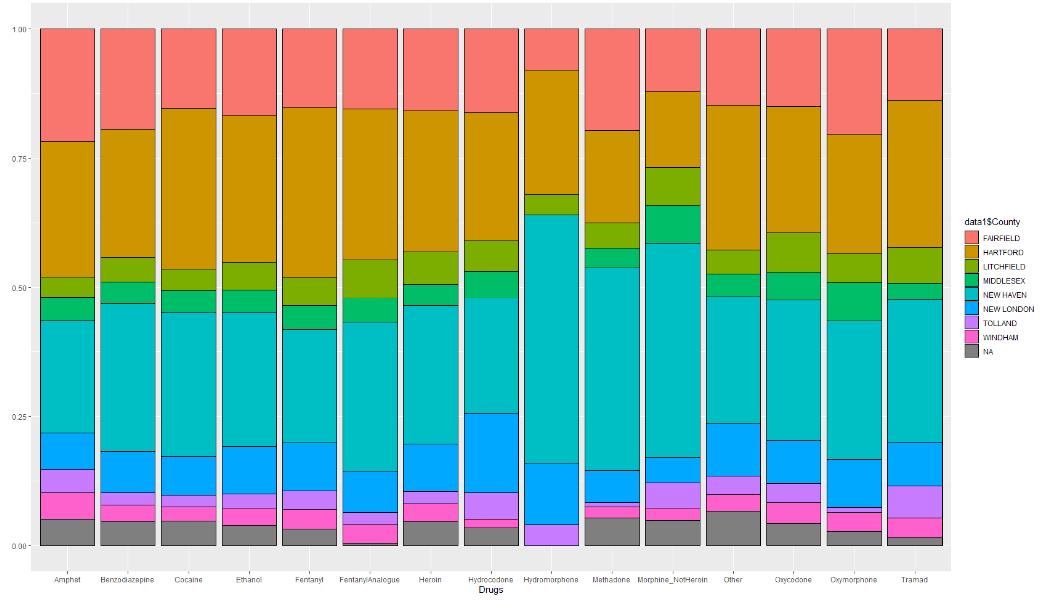
**Visualization 2: Top 7 Drugs in Connecticut**

This graph shows the top 7 drugs in terms of usage in the state and this number comes from the incidents reported. In this list Heroin is at the top and followed by Fentanyl. This shows that these drugs are really popular among the users and should be on list of departments to block the entry or manufacturing of specific list of chemicals in the country which on further processing can results into these drugs.



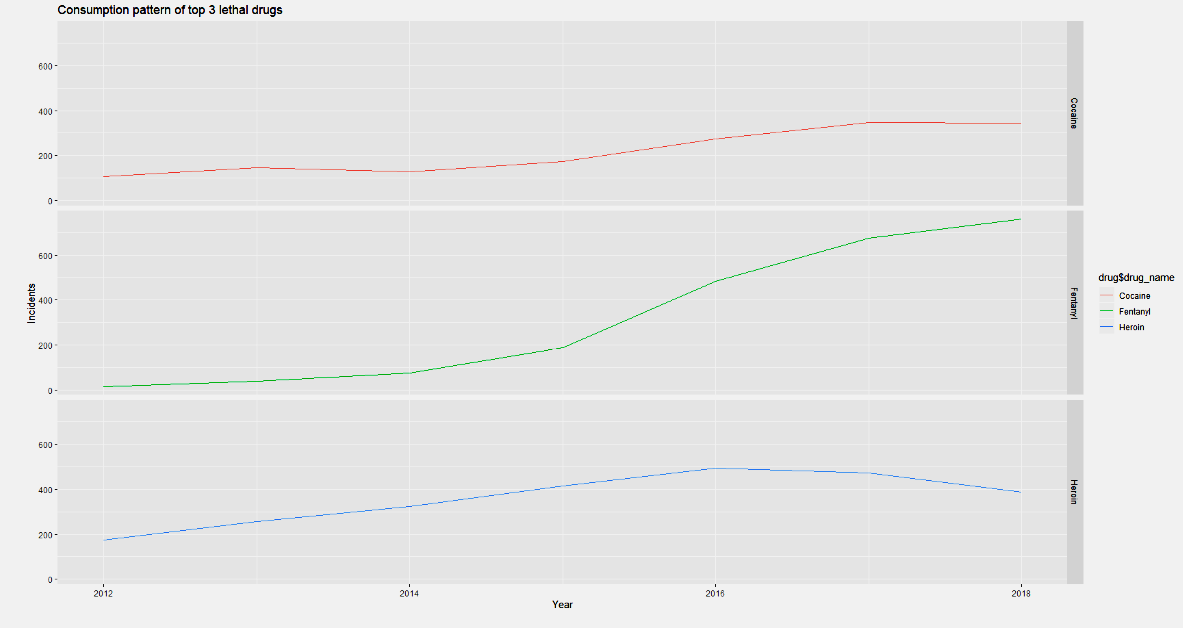
**Visualization 3: Drugs consumed by deceased in each county**

This graph shows that in Hartford and in New Haven all drugs are being used by the people in large numbers, so we can say that in these two counties all drugs are under use and can’t blame a particular drug for most of the deaths. Tolland and Windham are the counties where there are minimal incidents happened because of drugs intake.

 **Drugs consumed by deceased in each county**

Visualization 4: **Consumption pattern of top 3 lethal drugs**

These line graphs are showing the trend of usage of three lethal drugs in the state, and we can see that use of cocaine is stagnant from late 2016 but for other two trend line is saying something interesting, like for Fentanyl usage has been increased exponentially from 2015 and it surpassed the other drugs usage significantly but for the Heroin usage is declining right from the beginning of 2016. So we can say that more people are dying from Fentanyl or more people are consuming it and this drug is in trend, means thing to worry about.

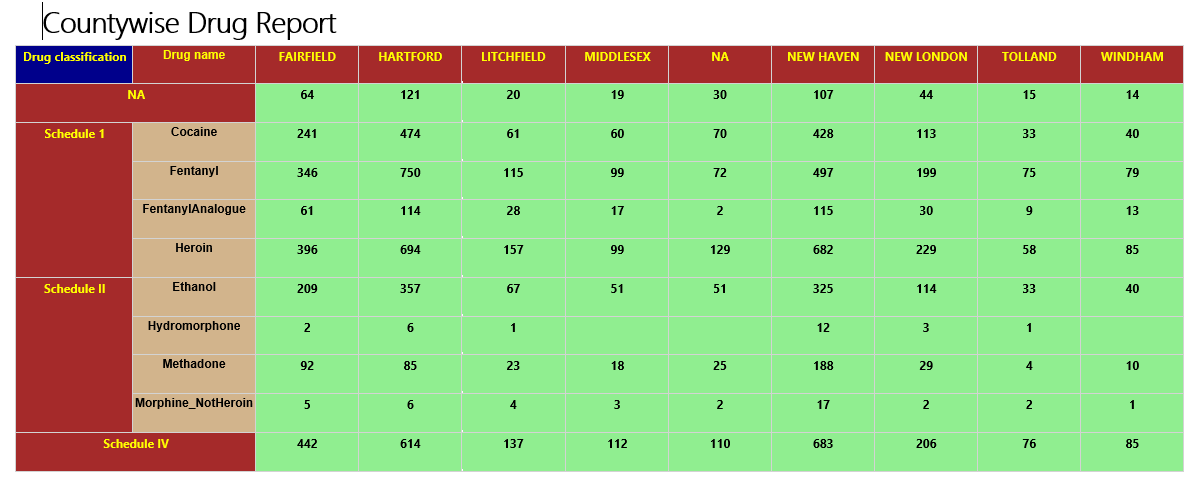


## 4.2. Reports

Report is the source of evidence and driver of making business metrics and estimations and strategy maker. Hence it is essential to produce report and study the data to make efficient and better decisions. Below are the reports produced in **SSRS**.

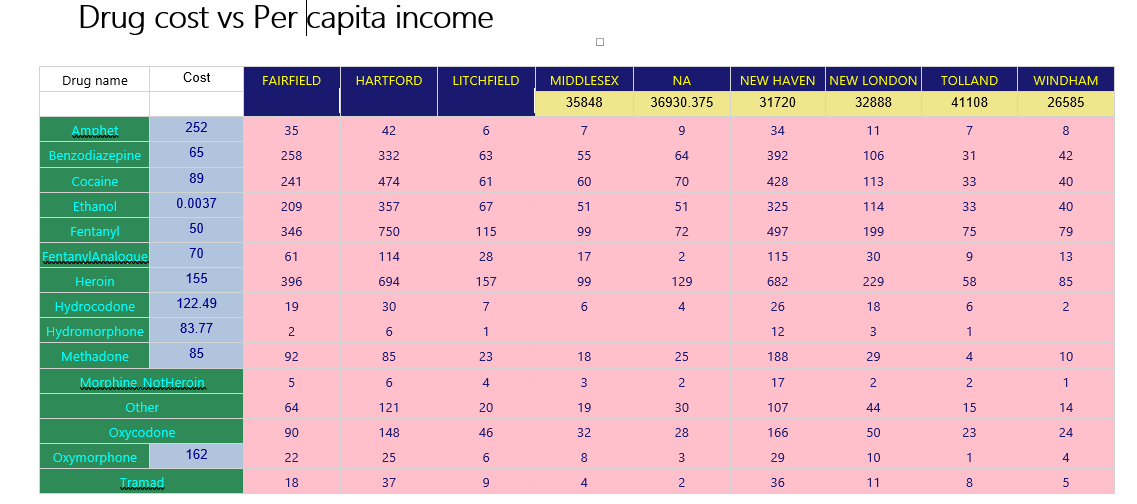
Report 1**: Drugs county wise**

This report is giving us the detail information of the drugs consumption in different counties according to the schedule of drugs defined by the federal government. And from this report we can see the most dangerous category of drug which is Schedule I containing drugs like Cocaine, Fentanyl and Heroin is widely used in Hartford, New Haven and followed by Fairfield. These figures are troublesome for the authorities and asking them to take some measures to stop this drug ASAP.



Report 2**: Drug Cost VS average Per capita Income**

With the available data here we have tried to find the relationship between the drug cost and its usage in relation to per capita income of the counties. And very interesting finding here is that county where per capita income is least or maximum having very less drug users. This could be because county with less avg. per capita income people over there can’t afford drugs. About county with high per capita income we can either people are educated and aware about the impact of drugs that’s why they are not doing it, or they can afford better medical facilities that’s why there are less incidents reported from those counties. Rest we can expensive drug i.e. Hydrocodone is less popular across the categories and Fentanyl which is avg priced is popular in the state. But Heroin despite of being most expensive is hitting top in the list of number of users, this could be because of its popularity or because of its availability.



**Report 3: Drugs in different age groups wise**

In this report we can find the most vulnerable age categories which are Middle adulthood and Young adults. We can also find the popular drugs in different age categories. This can help authorities to start targeted campaigns to aware the people about drugs and motivate them for healthy life.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Children** | **Middle Adulthood** | **Senior Adulthood** | **Young Adulthood** | | **Youth** |
| **Amphet** |  | 46 | 2 | | 92 | 19 |
| **Benzodiazepine** |  | 597 | 39 | | 615 | 92 |
| **Cocaine** |  | 677 | 30 | | 728 | 85 |
| **Ethanol** |  | 611 | 25 | | 550 | 61 |
| **Fentanyl** | 1 | 787 | 37 | | 1213 | 194 |
| **Fentanyl Analogue** |  | 143 | 7 | | 213 | 26 |
| **Heroin** | 1 |  | 45 | | 1313 | 243 |
| **Hydrocodone** |  | 66 | 10 | | 41 | 1 |
| **Hydromorphone** |  | 15 | 2 | | 7 | 1 |
| **Methadone** |  | 241 | 5 | | 208 | 20 |
| **Morphine\_NotHeroin** |  | 31 |  | | 11 |  |
| **Other** |  | 168 | 12 | | 220 | 34 |
| **Oxycodone** |  | 326 | 29 | | 230 | 22 |
| **Oxymorphone** |  | 59 | 2 | | 39 | 8 |
| **Tramad** |  | 70 | 12 | | 41 | 7 |

Report 4: **Drug Year wise**

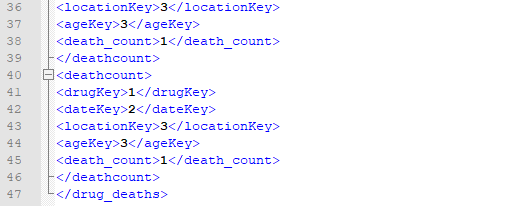
This report talks about the number of deaths by different drugs from 2012-2018. One shocking figure in this report is about the drug Fetanyl, in 2012 there were only 14 deceased because of this drug and this number shoots up to 760 in 2018 which almost 98% increase Y-o-Y in the death count.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **2012** | **2013** | **2014** | **2015** | **2016** | **2017** | **2018** |
| **Amphet** | 7 | 4 | 13 | 20 | 19 | 40 | 56 |
| **Benzodiazepine** | 49 | 78 | 157 | 221 | 242 | 330 | 266 |
| **Cocaine** | 105 | 147 | 127 | 174 | 275 | 347 | 345 |
| **Ethanol** | 61 | 83 | 126 | 176 | 255 | 290 | 256 |
| **Fentanyl** | 14 | 37 | 75 | 188 | 482 | 676 | 760 |
| **FentanylAnalogue** |  |  |  |  |  | 139 | 250 |
| **Heroin** | 174 | 257 | 325 | 416 | 494 | 474 | 389 |
| **Hydrocodone** | 15 | 19 | 15 | 20 | 20 | 15 | 14 |
| **Hydromorphone** |  |  |  |  |  | 16 | 9 |
| **Methadone** | 33 | 47 | 51 | 72 | 84 | 99 | 88 |
| **Morphine\_NotHeroin** |  |  |  | 12 | 16 | 12 | 2 |
| **Other** | 39 | 48 | 60 | 76 | 77 | 66 | 68 |
| **Oxycodone** | 70 | 74 | 101 | 95 | 110 | 95 | 62 |
| **Oxymorphone** | 30 | 16 | 29 | 6 | 7 | 8 | 12 |
| **Tramad** | 8 | 7 | 15 | 16 | 17 | 28 | 39 |

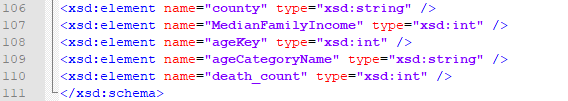
# 5. Include XML and Schema

XML: Extensible Markup Language (**XML**) is a markup language that defines a set of rules for encoding documents in a format that is both human-readable and machine-readable. The design goals of XML emphasize simplicity, generality, and usability across different systems.

**Above, we have shown a couple of instances of the dimension and the fact tables**.



XSD:



**XSD** (XML Schema Definition), specifies how to formally describe the elements in an Extensible Markup Language (XML) document. It can be used for validating an XML document to ensure it adheres to a specified structure that is defined in the XSD document.

In the above shown XSD Document we have specified the elements inside the dimensions including their data types and defined the structure of the fact.

# 6. Graph Databases

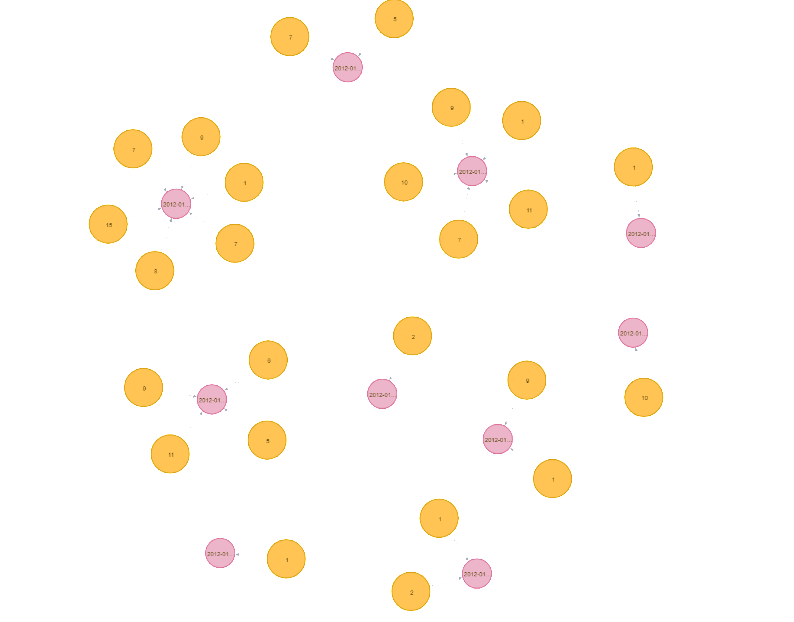
**Neo4j Graph Data Base:** Neo4j is one of the most popular graph database management system.

We have used the Cypher Query Language (See [Appendix B](#_Appendix_B_–) for Neo4j code) to import the data from the **Dimension and fact Tables.**

* We have created the dimensions and fact as nodes in the Neo4j Graph.
* We have further defined the relationship between the Dimension and the Fact nodes by relating dimensions: Age, Area, Drugs and Calendar to Drug Overdose Death (Fact)

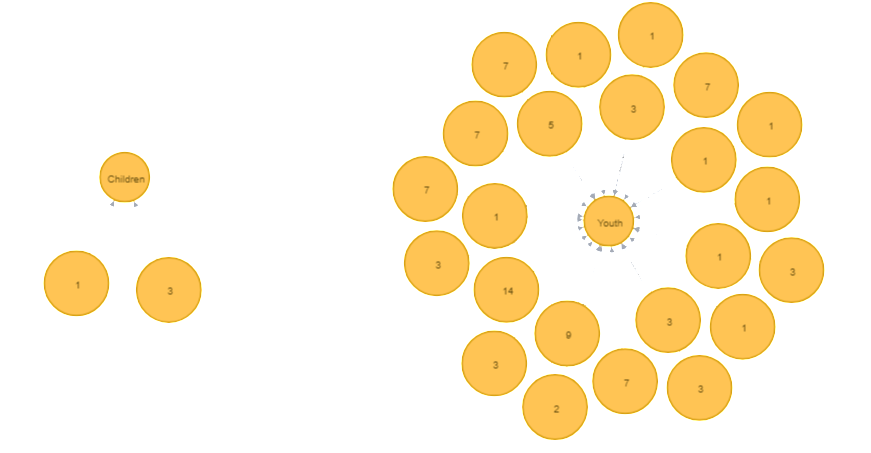
**Some Graphs:**

1) This graph shows the relationship between Calendar Node and Fact Node **[Fact-diedon->Calendar]**



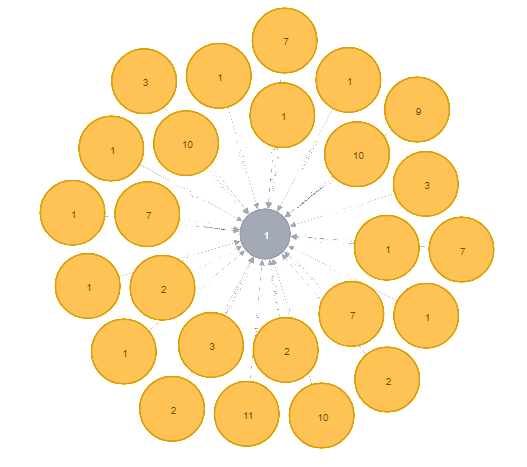
2) In this graph we have shown the relation between Age Node and Fact Node

**[Fact-diedat->Age]**



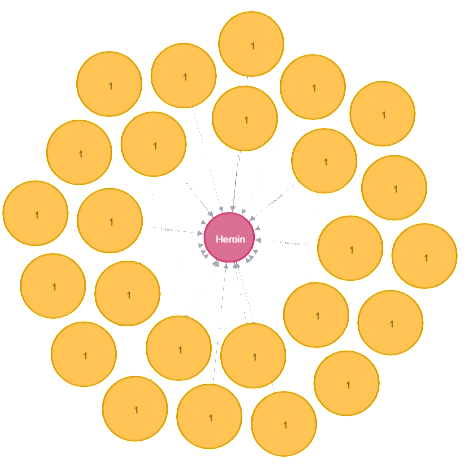
3) Above we have shown the relation between Area Node and Fact Node

**[Fact-diedatarea->Area]**



4) Above we have shown the relation between Drug Node and Fact Node

**[Fact-took->Drug]**



From these graphs, we can easily see the advantages of the Neo4j database over the relation database. It is visually appealing and very easily depicts the relationships between different nodes.

## 6.1. COMAPRISON to realtional databases

**Relational Database:**

A relational database is a set of formally described tables from which data can be accessed or reassembled in many ways without having to reorganize the database tables. The standard user and application programming interface (API) of a relational database is the Structured Query Language. SQL statements are used both for interactive queries for information from a relational database and for gathering data for reports.

**Graph Database:**

Graph Databases comes under category of NoSQL databases which use the graph data model comprised of vertices, which is an entity which gives relationship between two nodes. They are generally helpful because they highlight the links and relationships between relevant data similarly to how we use ourselves.

# 7. Conclusions

* Heroin is the drug which is widely used in the state of Connecticut and most prone are people of middle adulthood and youth.
* New Haven, Fairfield and Hartford are the counties where maximum incidents happened, so federal departments should take measures to tackle this issue ASAP by alerting local and state police.
* Fentanyl is the new drug which is getting popular nowadays, and usage of this has increased exponentially from last few years. So, this drug is new to market and authorities should take strict measures to stop this from getting viral to other sates too.
* Most of the victims are from age category Young and Medium Adulthood and this is the major point to concern about.
* Counties where per capita income is least or maximum having very less drug users comparison to other counties with average income.

# 8. Bibliography

**Dataset link:**

* **Kleykamp, T. (2019). *Accidental Drug Related Deaths 2012-2018*** *- Data.gov*. [online] Catalog.data.gov. Available at: [https://catalog.data.gov/dataset/accidental-drug-related-deaths-january-2012-sept-2015](https://catalog.data.gov/dataset/accidental-drug-related-deaths-january-2012-sept-2015%20) [Accessed 14 Dec. 2019].

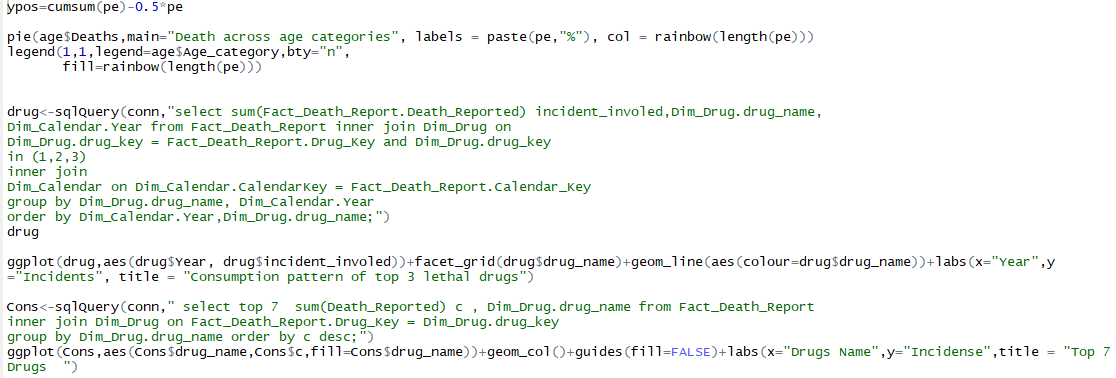
**Other references**:

* ***Connecticut Drug Control Update****.*[Online]. Available at: [*https://obamawhitehouse.archives.gov/sites/default/files/docs/state\_profile-connecticut.pdf*](https://obamawhitehouse.archives.gov/sites/default/files/docs/state_profile-connecticut.pdf )(Accessed: 14 December 2019).
* **Andriy Blokhin (2018) *5 States with the Highest Real GDP per Capita****.*[Online]. Available at: [*https://www.investopedia.com/articles/investing/112415/5-states-highest-gdp-capita.asp*](https://www.investopedia.com/articles/investing/112415/5-states-highest-gdp-capita.asp )(Accessed: 14 December 2019).
* **Sorcha Pollak (2019) *Price of heroin in Ireland second highest in Europe****.*[Online]. Available at: [*https://www.irishtimes.com/news/crime-and-law/price-of-heroin-in-ireland-second-highest-in-europe-1.3939263*](https://www.irishtimes.com/news/crime-and-law/price-of-heroin-in-ireland-second-highest-in-europe-1.3939263 )(Accessed: 14 December 2019).
* **Jeffrey Juergens (2019) *What Are Drug Classifications?***[Online]. Available at: [*https://www.addictioncenter.com/drugs/drug-classifications/*](https://www.addictioncenter.com/drugs/drug-classifications/ )(Accessed: 14 December 2019).
* **(2006) *National Shared Services Primary Care Reimbursement Service Information and Administrative Arrangements For Pharmacists****.*[Online]. Available at: [*https://www.hse.ie/eng/staff/pcrs/contractor-handbooks/pcrs-handbook-for-pharmacists.pdf* (](https://www.hse.ie/eng/staff/pcrs/contractor-handbooks/pcrs-handbook-for-pharmacists.pdf ()Accessed: 14 December 2019).
* **Department of International Economic and Social Affairs (1982) 'STATISTICAL PAPERS', Provisional Guidelines on Standard International Age Classification, (ST**/ESA/STAT/SER.M/74), pp. 11 [Online]. Available at: [https://unstats.un.org/unsd/publication/SeriesM/SeriesM\_74e.pdf](https://unstats.un.org/unsd/publication/SeriesM/SeriesM_74e.pdf%20) (Accessed: 14 December 2019).
* <https://www.drugs.com/price-guide/amphetamine-dextroamphetamine>
* **(18 October 2019) *List of Connecticut locations by per capita income,***Available at: [*https://en.wikipedia.org/wiki/List\_of\_Connecticut\_locations\_by\_per\_capita\_income*](https://en.wikipedia.org/wiki/List_of_Connecticut_locations_by_per_capita_income )(Accessed: 14 December 2019).

# Appendix A – Visualization Code

**ORD**

**require(RODBC)**



**Customers.CustomerID;**

**Go Back to** [**Visualization**](#_4.1._Visualizations)

# Appendix B – Neo 4J code

load csv with headers from "file:///Dim\_Age.csv" as row create (c:Age) set c = row {ageKey:row.Age\_Key, ageCat:row.Age\_Category, ageCatName:row.Age\_Category\_Name } return c

create constraint on (c:Age) assert c.ageKey is unique

load csv with headers from "file:///Dim\_Drug.csv" as row create (p:Drug) set p = row{drugkey:row.drug\_key,Name:row.drug\_name, Cost:toFloat(row.cost), drugClass:row.drug\_classification} return p

create constraint on (p:Drug) assert p.drugkey is unique

load csv with headers from "file:///Dim\_County.csv" as row create (a:Area) set a = row{areakey:row.Area\_Key,Name:row.Area\_name, Percapita\_income:toFloat(row. Per\_Capita\_Income), family\_income:row.Median\_Family\_Income} return a

create constraint on (a:Area) assert a.areakey is unique

load csv with headers from "file:///Dim\_Calendar.csv" as row create (d:Calendar) set d= row{calKey:row.CalendarKey,name:row.FullDate, day:row.dayOfWeek,dayType:row.dayType,dayOfMonth:row.DayOfMonth, month:row.Month\_, year:row.Year\_} return d

create constraint on (d:Calendar) assert d.calKey is unique

load csv with headers from "file:///Fact\_Death\_Report.csv" as row create(f:Drug\_Overdose\_Death) set f = row{calKey:row.Calendar\_Key, areaKey:row.Area\_Key, drugKey:row.Drug\_Key, agelKey:row.Age\_Key,deathreported:row.Death\_Reported} return f

MATCH (c:Age), (f:Drug\_Overdose\_Death) where f.agelKey=c.ageKey create (f)-[r:diedat]- >(c) return f,c, r

MATCH (p:Drug), (f:Drug\_Overdose\_Death) where f.drugKey=p.drugkey create (f)-[r:took]- >(p) return f,p, r

MATCH (a:Area), (f:Drug\_Overdose\_Death) where f.areaKey=a.areakey create (f)-[r:diedatarea]- >(a) return f,a, r

MATCH (d:Calendar), (f:Drug\_Overdose\_Death) where f.calKey=d.calKey create (f)-[r:diedon]- >(d) return f,d,r

# Appendix C – **SQL Task (Drug Incident Mapping**)

DECLARE

@Id nvarchar(15), @Heroin nvarchar(15), @Cocaine nvarchar(15), @Fentanyl nvarchar(15), @FentanylAnalogue nvarchar(15), @Oxycodone nvarchar(15), @Oxymorphone nvarchar(15), @Ethanol nvarchar(15), @Hydrocodone nvarchar(15), @Benzodiazepine nvarchar(15), @Methadone nvarchar(15),

@Amphet nvarchar(15), @Tramad nvarchar(15), @Morphine\_NotHeroin nvarchar(15), @Hydromorphone nvarchar(15), @Other nvarchar(15);

DECLARE cursor\_source CURSOR

FOR SELECT ID,Heroin,Cocaine,Fentanyl,FentanylAnalogue,Oxycodone,Oxymorphone,Ethanol,Hydrocodone,Benzodiazepine,Methadone,

Amphet,Tramad,Morphine\_NotHeroin,Hydromorphone,Other

from Source\_Data;

OPEN cursor\_source;

FETCH NEXT FROM cursor\_source INTO

@Id,@Heroin,@Cocaine,@Fentanyl,@FentanylAnalogue,@Oxycodone,@Oxymorphone,@Ethanol,@Hydrocodone,@Benzodiazepine,@Methadone,

@Amphet,@Tramad,@Morphine\_NotHeroin,@Hydromorphone,@Other

WHILE @@FETCH\_STATUS = 0

BEGIN

if (@Heroin is not null)

insert into Drug\_Incident\_Mapping values (@Id,'Heroin');

if (@Cocaine is not null)

insert into Drug\_Incident\_Mapping values (@Id,'Cocaine');

if (@Fentanyl is not null)

insert into Drug\_Incident\_Mapping values (@Id,'Fentanyl');

if (@FentanylAnalogue is not null)

insert into Drug\_Incident\_Mapping values (@Id,'FentanylAnalogue');

if (@Oxycodone is not null)

insert into Drug\_Incident\_Mapping values (@Id,'Oxycodone');

if (@Oxymorphone is not null)

insert into Drug\_Incident\_Mapping values (@Id,'Oxymorphone');

if (@Ethanol is not null)

insert into Drug\_Incident\_Mapping values (@Id,'Ethanol');

if (@Hydrocodone is not null)

insert into Drug\_Incident\_Mapping values (@Id,'Hydrocodone');

if (@Benzodiazepine is not null)

insert into Drug\_Incident\_Mapping values (@Id,'Benzodiazepine');

if (@Methadone is not null)

insert into Drug\_Incident\_Mapping values (@Id,'Methadone');

if (@Amphet is not null)

insert into Drug\_Incident\_Mapping values (@Id,'Amphet');

if (@Tramad is not null)

insert into Drug\_Incident\_Mapping values (@Id,'Tramad');

if (@Morphine\_NotHeroin is not null)

insert into Drug\_Incident\_Mapping values (@Id,'Morphine\_NotHeroin');

if (@Hydromorphone is not null)

insert into Drug\_Incident\_Mapping values (@Id,'Hydromorphone');

if (@Other is not null)

insert into Drug\_Incident\_Mapping values (@Id,'Other');

FETCH NEXT FROM cursor\_source INTO

@Id,@Heroin,@Cocaine,@Fentanyl,@FentanylAnalogue,@Oxycodone,@Oxymorphone,@Ethanol,@Hydrocodone,@Benzodiazepine,@Methadone,

@Amphet,@Tramad,@Morphine\_NotHeroin,@Hydromorphone,@Other

END;

CLOSE cursor\_source;

DEALLOCATE cursor\_source;

Go

# Appendix D – Drugs categories

* **Schedule I**

Schedule I drugs have the most regulations and harshest penalties of any drugs. Schedule I drugs have no legitimate accepted medical use and a high potential for abuse.

* **Schedule II**

Schedule II drugs have more regulations and harsher penalties than any drug classification other than Schedule I drugs. Schedule II drugs have a legitimate accepted medical use, a high potential for abuse, and a severe dependence risk.

* **Schedule III**

Schedule III drugs have more regulations and harsher penalties than Schedule IV drugs and fewer regulations and less severe penalties than Schedule II drugs. Schedule III drugs have a legitimate acceptable medical purpose, have a lower abuse potential than Schedule I and II drugs, and have a moderate or low potential for addiction.

* **Schedule IV**

Schedule IV drugs have regulations and penalties in between those of Schedule V and Schedule III drugs. Schedule IV drugs have a legitimate accepted medical purpose, have a low potential for abuse, and have a low potential for addiction.

* **Schedule V**

Schedule V drugs have the fewest regulations and lowest penalties of any federal drug classification.

# Appendix E – R Code for data extraction and cleaning

# Appendix F – SQL TASK(Transforming null Values)

update Source\_Data set DeathCounty = ResidenceCounty where DeathCounty is null

and ResidenceCounty in (select County\_Name from Dim\_County);

update Source\_Data set DeathCounty = 'NA' where DeathCounty is null

or DeathCounty like 'USA';

UPDATE Source\_Data SET AgeCategory

= (case when Age < 15 then 1

when Age < 25 then 2

when Age < 45 then 3

when Age < 65 then 4

when Age >64 then 5

end);

update Source\_Data set AgeCategory = 3 where Age is null;

update Dim\_County set County\_Name = UPPER(County\_Name);