

**Chronic Disease Analysis**

*In the*

*United States of America*

*by*

PRAVEEN TATA

SAMREEN NAYEEM

NIKHIL NANDA KUMAR

**Mohammad A. Rob**

School of Business

University of Houston – Clear Lake

Abstract

Data Warehouse’s have become a key factor in many companies in recent times. Companies generate large volumes of data on a daily basis, and this data needs to be handled in a way to make it useful in the long run. A Data Warehouse allows business professionals to access this data in a simple query centric environment, thus allowing them to review data trends in the form of reports. The benefit this provides is that large volumes of data can be accessed and compressed into reports which displays data trends in the way the user wants to view it.

Chronic Diseases are wide spread across the United States of America, and focusing resources to help prevent every chronic diseases across America makes the cost to the country go up, which inevitably leads to the patient paying more money to get medication and treatment. In such a situation a data warehouse can play a big role in enabling medical institutes to analyze chronic disease patient data across America, allowing them to pin point at the states in which a specific chronic disease is prevalent. This allows better resource management and in creating specific awareness for those diseases, allowing the cost to the state go down which leads to better overall health of the patients, a reduction in chronic diseases, and a lower payments.

In this project we have focused on creating a Chronic Disease Data Warehouse. We have explained in detail all the steps taken in creating a data warehouse and how the reports that are generated show a clear understanding of how these diseases can be controlled.

Table of Contents

[Introduction 1](#_Introduction)

[Problem Statement 1](#_Toc468717049)

[Why Data Warehouse 1](#_Toc468717049)

[Data Origins & Expected Value 1](#_Toc468717049)

[Methodology 2](#_Methodology)

Raw Data Format 2

Dimension Modelling 3

[Data Cleansing 8](#_Toc468717051)

[Data Transfer 8](#_Toc468717051)

[Star Schema 9](#_Toc468717051)

[Cube implementation 10](#_Cube_Implementation)

[Overview 10](#_Toc468717063)

[Creating Hierarchies 1](#_Toc468717064)2

[Reports 1](#_Reports)3

[Cardiovascular vs. Diabetes 1](#_Toc468717063)3

[Chronic Diseases based Male Mortality Rate 1](#_Toc468717064)4

[Diseases based on Gender 1](#_Toc468717063)5

[Mortality rate of various diseases 1](#_Toc468717064)6

[Mortality rate based on Region 17](#_Toc468717063)

[Pie Chart of Male Mortality 18](#_Toc468717064)

[Year based chronic disease mortality rate 19](#_Toc468717063)

[Conclusion 2](#_Conclusion)0

[References 2](#_References)1

# 

# Introduction

1. **Problem Statement**

A Chronic Disease is disease that lasts for a period longer than 3 months, these can include Heart Diseases, Cancer, Diabetes, Obesity, and arthritis to name a few. In the United States of America as of 2012 nearly 117 million people had one more chronic health conditions. Similarly, in 2010 nearly 48% of all deaths in the United States were due to chronic diseases. Looking at the data generated on a financial basis, in 2003 nearly $128 billion was invested by patients and medical institutes towards arthritis and its related conditions.

1. **Why Data Warehouse**

When looking at such numbers, medical institutes need a better alternative to track chronic diseases and help single out areas where a specific diseases or prevalent so as to focus specific resources to such areas to help curb those diseases, thus limiting the medical expenses the state incurs. As the resources are focused on curbing a specific disease, more people with such conditions will get quick attention at a cheaper rate, which will lead to a reduction in that disease and reduce costs over the long run. To be able to achieve such goals, a system where an institute can look at graphs and reports of all patient data across America and be able to pinpoint which locations need specific help is required. This is where a Data Warehouse becomes extremely useful.

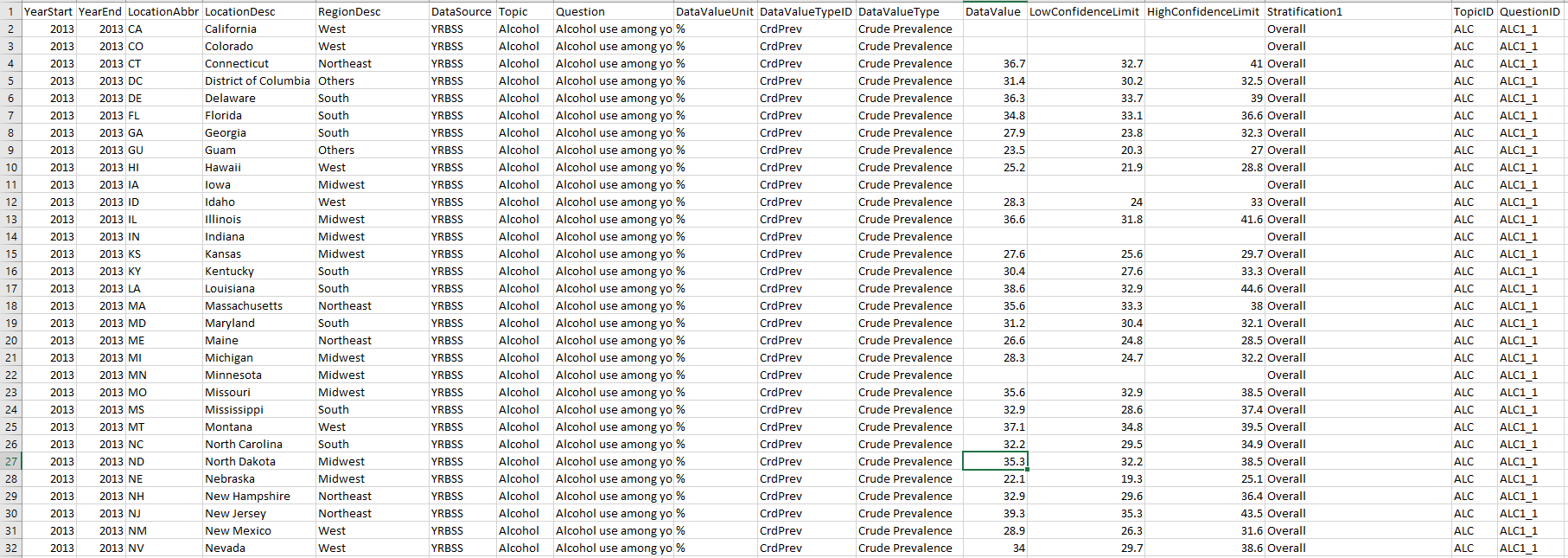
1. **Data Origins & Expected Value**

In our project, we have taken a large Chronic Disease Data Set from [www.data.gov](http://www.data.gov) which is a government website containing large dumps of data. This data set was then cleaned so as to remove gaps and information which is not valuable for analysis. We then standardized the columns of data and created dimensions and facts to help make the data query centric. The end result of doing all of this generates the output a data warehouse would, with simple query inputs resulting in detailed reports.

# Methodology

1. **Raw Data Format**

The raw data was taken from data.gov, a website which has an extensive collection of data sets in condensed into lines of records in an excel document. This external data is the initial stage of creating a dimension model. As can be seen below, all the data is clubbed into a long list of rows and columns. Querying such a Database is a not possible as there are no Dimensions or Facts connecting the data.



*Figure 1. Raw Data Set*

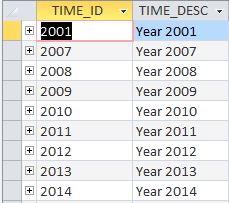
1. **Dimension Modelling**

Data warehouse and OLAP tools are based on a dimensional data model. A dimensional model is based on dimensions, facts, cubes, and schemas such as star and snowflake. Dimension models are utilized in a data warehouse, this immensely helps make the data query centric.

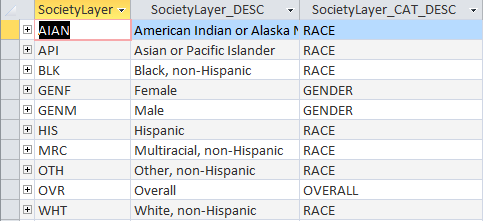
**Dimensions -** A dimension consists of attributes, which define the dimension. For example, a Time dimension could have year as an attribute. Furthermore dimensions have hierarchies, these hierarchies allow a user to be able to drill down or roll up data. An example of a hierarchy would be Year 🡪 Month 🡪 Day. A dimension table provides textual descriptions of a business dimension through attributes.

Below are the dimensions we have utilized to create the dimension model:

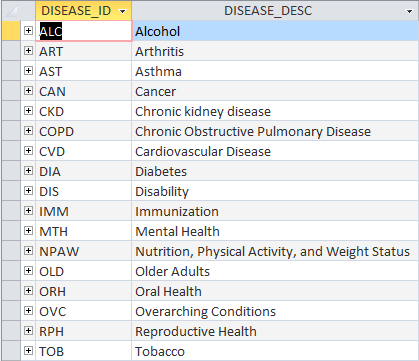
1. TimeDimension - This dimension is crucial as it allows the user to pick a specific year and see the effect chronic diseases has caused. It also allows the user to generate a report of the changes in chronic diseases over the years, which aids in seeing hospitalization and mortality trends. The Time Dimension consists of two attributes Time ID and Time Description.



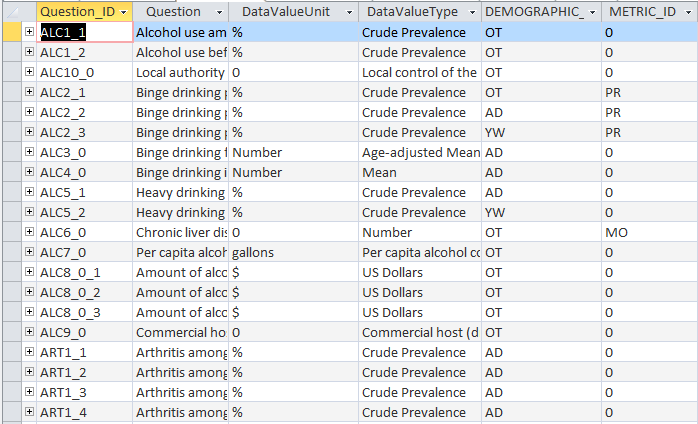
1. Society Layer Dimension – This dimension outlines the gender, race and ethnicity. America consists of people from different areas across the world, being able to select a certain type allows an institute to understand health trends. This dimension has hierarchies as gender can be further drill down to male and female. The attributes here are the abbreviation of the race/gender, the description and the detailed description of the same.



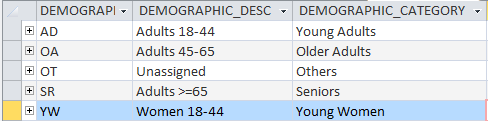
1. Disease Dimension – Chronic diseases are described in this dimension. So when a user queries the database, selecting a specific disease is easy. The attributes here are disease id and disease description.



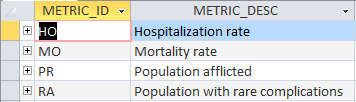
1. Question Dimension – Here all the chronic disease data is stored in the form of question tags, these tags allow the user to search for a specific condition of a specific chronic disease. The question tag is further drilled down to the Demographic and Metric Dimensions as we shall see.



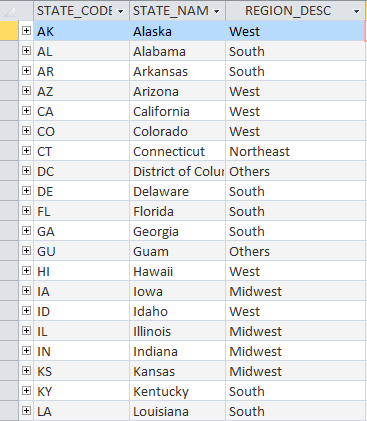
1. Demographic Dimension – This dimension outlines the various age groups of people affected by one or more conditions. It is linked to the question dimension. Even though dimension models are not relational in nature, some alterations are made to make understanding the data easier. The attributes here are id, description and category.



1. Metric Dimension – Metrics is the unit of measurement used. Primarily the unit of measurement utilized is mortality rate and hospitalization rate. The attributes are here the metric id and the description of the metric.

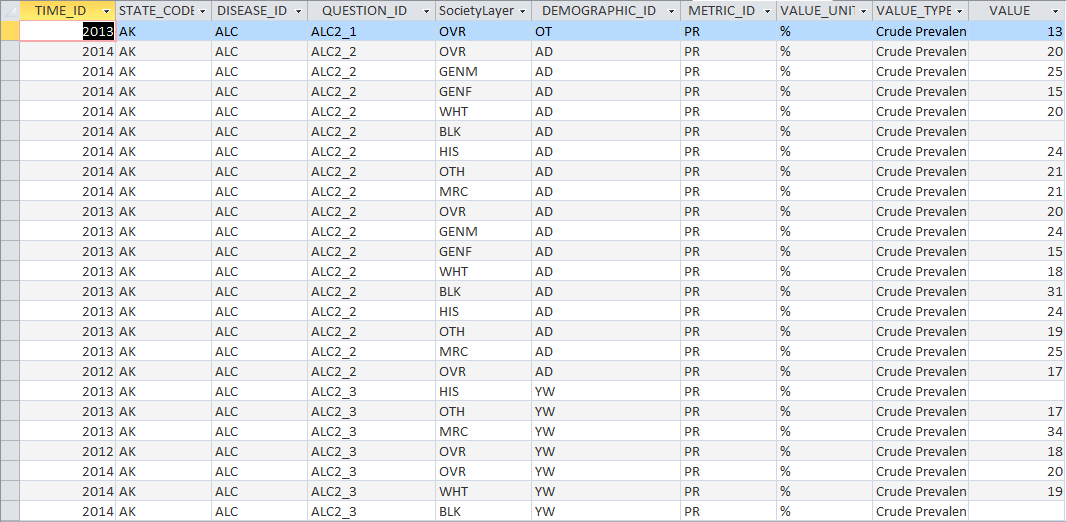


1. Location Dimension – This dimension defines the various states of America. It is divided via region and state, hence it has a hierarchy which can be drilled down. The attributes here are state code, state name and region name.



**Facts -** In data warehousing, a Fact table consists of the measurements, metrics or facts of a business process. It is located at the center of a star schema or a snowflake schema surrounded by dimension tables. This table is the primary table in a dimensional model, here all the measurements and metrics are stored.

The fact table in our project consists of 7 primary keys that connect to the various dimension tables respectively. The remaining are metrics of the business process. The primary key here is also called composite primary key as it is made up of a subset of foreign keys.



1. **Data Cleansing**

Not all the data at these sources is in the same format or utilize the same metrics, thus this data needs to be standardized and cleaned, to remove redundancy and improve performance. Data cleansing is also called data cleaning, here any inconsistencies and gaps are removed from the data set to reduce the overall size of the data and improve performance and overall quality. The data is also standardized, this removes anomalies like incorrect values and mismatched data. Data Sets consists of data grouped together from multiple sources.

In our project, the data set was cleaned, gaps were removed and the data was standardized prior to creating our database. The cleaning increased the responsiveness of the database, gaps removed redundancies and incorrect values, and standardizing the data meant when running a query the data returned would include results from all records of the database.

1. **Data Transfer**

Data.gov is the mother of our data warehouse. We used the dataset from the data.gov website which was quite intriguing, as we were about to study effects of diseases in health conscious and highly contradicting fast food world. The raw data in excel sheet had many columns but was not organized to form a data warehouse. Creating dimensions using the questions present in the dataset was our initial step. Questions were used to categorize diseases, race, and gender, hospitality and mortality rate.

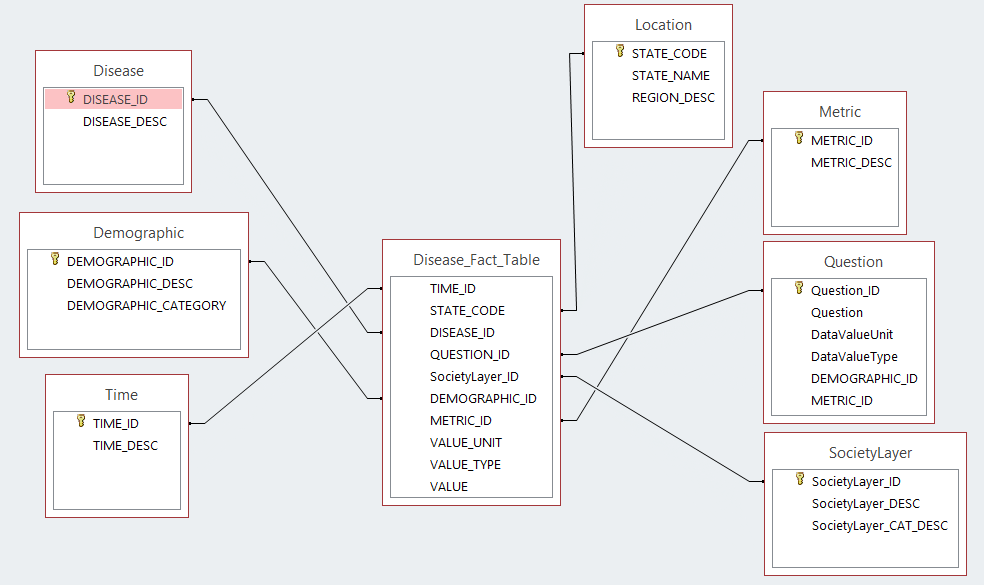
**Hierarchy** - Regions and states formed our main hierarchy which we named as ‘Location’. Also race and gender formed another hierarchy which allowed us to drill-down using slicing and dicing.

**Dimensions/Facts** - We formed 7 dimension tables namely Time, Disease, Location, Questions, Metric, Society Layer, Demographic. Fact table used PK from all the dimension tables and also included its own ‘Value’ column which was used to sum and aggregate the results.

The data was then imported in MS Access and in turn imported in MS SQL Server to create cube and extract meaningful data into excel to generate Pivot tables.

**Star Schema**

We have taken the cleansed and standardized data, imported it into Microsoft Access and created the relationships as shown below. This is the initial stage, post which the database is created and then utilized to further create a cube which is outlined in the next steps.

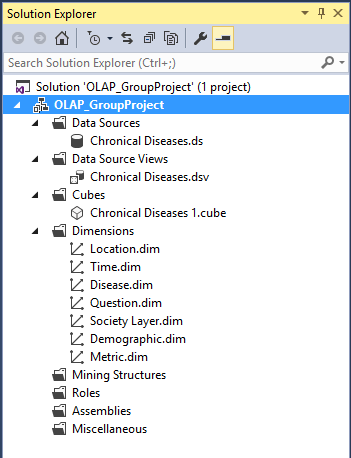


# Cube Implementation

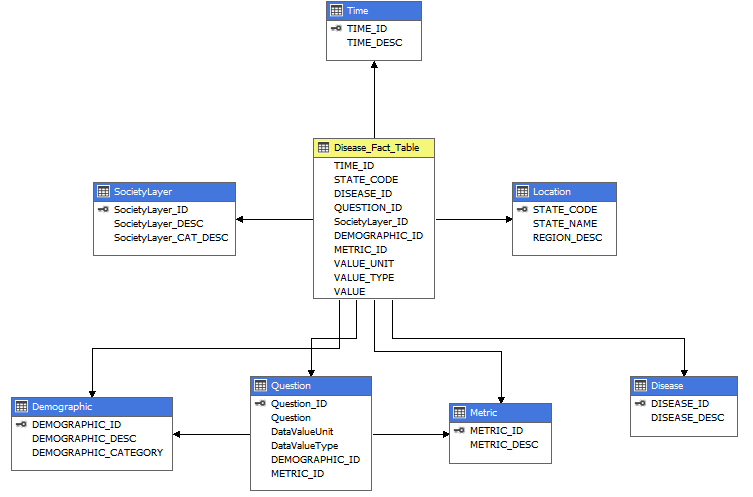
1. **Overview**

To create the cube we need to import the standardized database into Visual Studio. The database created in Microsoft Access is imported into Visual Studio and a Data Source is created. Once the data set is created we have to create a data source view which displays the tables in the database and their relationships, which results in a Star Schema diagram. The final step before creating the Cube is to create any dimension that has the primary hierarchy structure, in our case the Location Dimension is the main dimension over which our data is defined. Hence we create the location dimension and set the attribute hierarchy levels. When creating the cube, it auto detects the fact table along with the various possible dimensions present in the database.

Below is the Solution Explorer which shows the various Stages in creating the Chronic Disease OLAP Project.



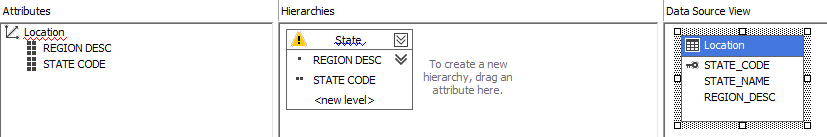
On completion of all the above steps, the Cube Structure is created. As can be seen below, all the titles in blue are the dimensions and the yellow title is the fact table. Just as in the star schema the fact table is at the center surrounded by the dimensions. Every dimension table must be linked to the fact table for optimal query analysis. As all textual data is meaningless without a metric to define the data. As the dimensions are auto generated on implementation of the cube in the following steps we make changes to suit our requirements for the Data Warehouse.



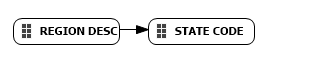
1. **Creating Hierarchies**

Our dimension model contains two dimensions with hierarchies, one is the Location Dimension and the other is the Society Layer.

**Location Dimension Hierarchy** – Once the dimension is created, we select the attributes we want to include in the hierarchy. Based on preference the attributes are placed in a way to allow the user to be able to drill down or roll up. As can be seen below, a user can drill down from region to state and roll up from state to region.



Once the hierarchies are set, the attribute relations need to be defined. As the relationship between regions and states does not change over time, region and state have a rigid relationship.



**Society Layer Hierarchy –** The society layer hierarchy is based on society category which is gender, race, ethnicity, which can be further drilled down to the description of such a category such as male and Hispanic.



Similar to the location attribute, as society layers do not frequently change, the relationship between the society category and its description is rigid.

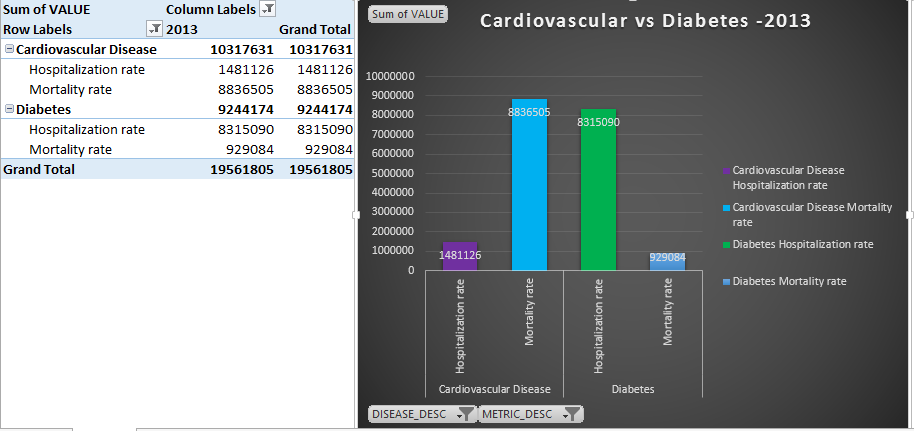


# Reports

1. Cardiovascular vs. Diabetes

The graph shown below displays the comparison between Cardiovascular Disease vs. Diabetes. Drilling down further, the graph compares hospitalization vs. mortality for both diseases for the year 2013.

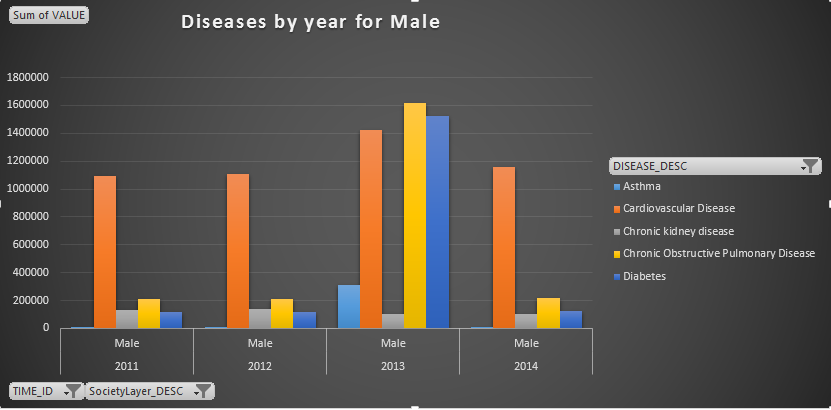
We can clearly conclude that more people have died because of cardiovascular disease and on the other hand more people were hospitalized due to diabetes during the year 2013.



1. Chronic Diseases based Male Mortality Rate

The graph shows the comparison of mortality ratedue to various diseases like Asthma, Cardiovascular, Chronic Kidney, Diabetes and other diseases. The comparison shows the mortality rate of Males in different years.

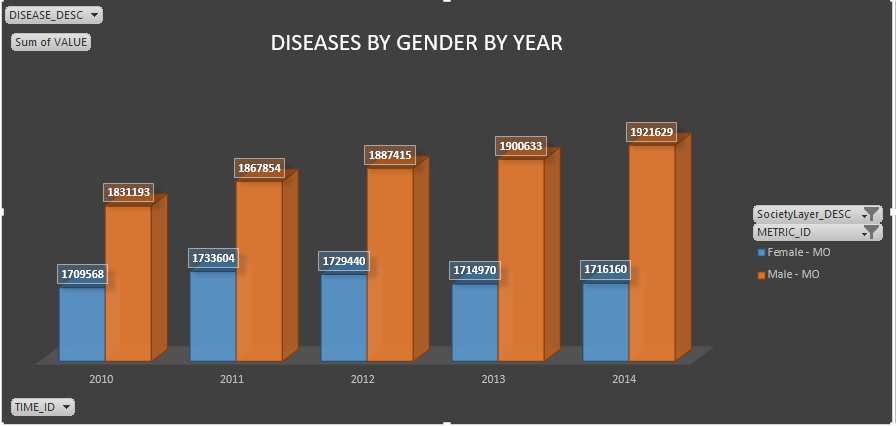
As we can see, the number of Males who died due to these diseases has a substantial increase per year, especially during 2013. The tremendous decline in the year 2014, shows Males are becoming health conscious.



1. Diseases based on Gender

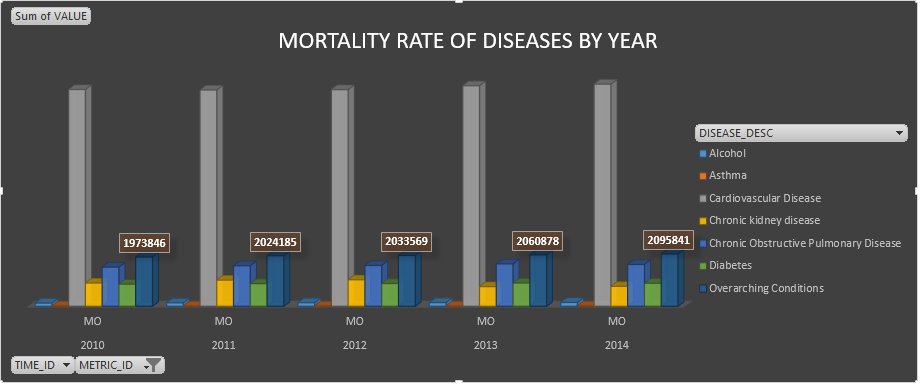
This graph shows the comparison of mortality rates of Males & Females based on year. It takes into account all chronic diseases.

This graph shows that males have a much higher mortality rate when compared to woman. Also per year the male mortality has a steady increase, whereas female mortality is nearly steady.



1. Mortality rate of various diseases

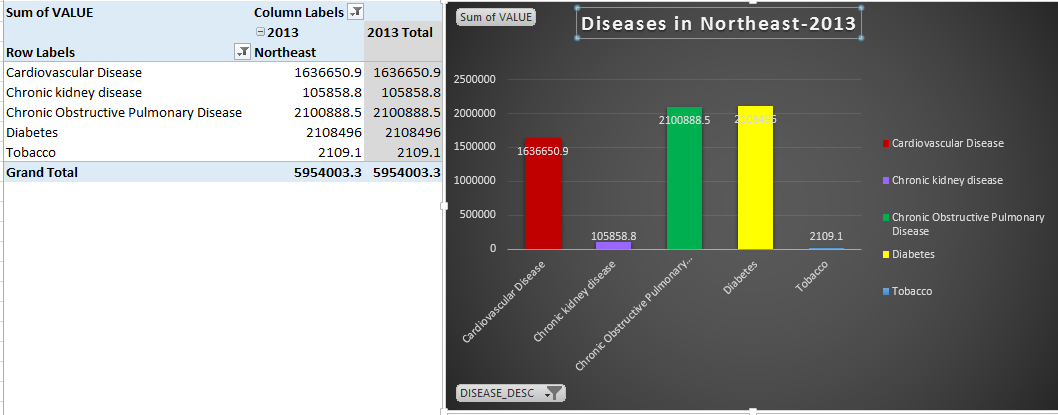
The graph shows the number of people who died due to different diseases on a yearly basis. This number has a small yet constant increase every year, except for Chronic Kidney Disease which seems to have had a reduction in deaths over 2013 and 2014. Fatalities caused due to cardiovascular diseases is one of the major causes of deaths.



1. **Mortality rate based on Region**

The graph shows the mortality rate in the Northeast region for the year 2013 for different diseases. Fatalities due to Diabetes and Chronic Pulmonary diseases are extremely high.

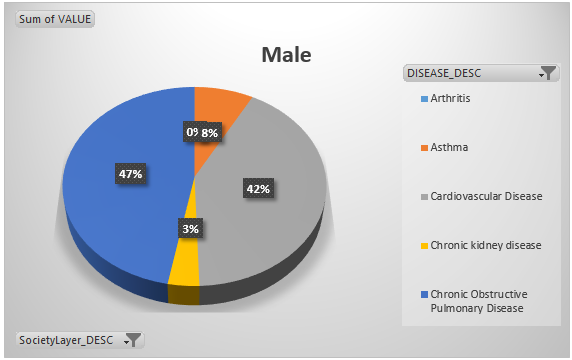
Such information can aid medical institutes to focus resources in that area to help reduce those specific cardiovascular diseases. As Tabaco isn’t a major cause of the high mortality rate, resources can be directed to raise awareness of Diabetes, thus helping patients and people. This will lead to a reduction of those diseases over time in those areas



1. Pie Chart of Male Mortality

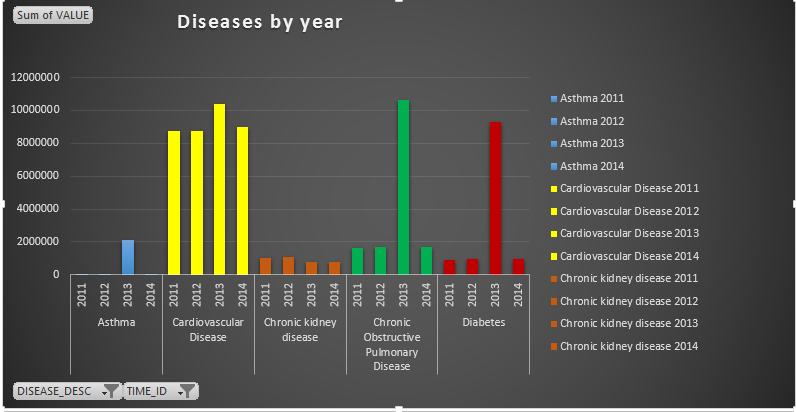
Overall male mortality rate by different diseases. Cardiovascular Disease and Obstructive Pulmonary disease have the most affect – 47% and 42% respectively.

Looking at a Pie Chart gives us a better understanding when comparing multiple objects, in this case chronic diseases. This simple yet detailed chart can aid hospitals understand patients affected and help raise awareness.



1. Year based chronic disease mortality rate

The graph shows the comparison of the number of people who died during the years 2011 – 2014. The Data Warehouse can be drilled down to a deep layer with multiple filters. Here we are referencing multiple diseases per year and using mortality rate as the metric. As can be seen cardiovascular disease has a very high mortality rate over the whole period. Whereas the remaining diseases only have a spike during the year 2013.



# Conclusion

As can be clearly seen with the results acquired from the various reports, a data warehouse can take a huge collection of raw data and convert this data into user-friendly analysis data, allowing business users to utilize such information to further improve their organization.

In our case, the data allows medical institutes to query the data warehouse and help understand the various chronic disease trends in America and take necessary action to help reduce the mortality rate and hospitalizations caused by these diseases.

This report has not only helped us understand the usage of a data warehouse but taught us how to execute and see the benefits in real-time. It has also taught us how to take a large dump of data and turn it into valuable data. Nearly every company today is moving to create data warehouses to handle terabytes of data, having this knowledge already is a massive benefit for anyone entering the professional world.

# References

* [*https://www.cdc.gov/chronicdisease/overview*](https://www.cdc.gov/chronicdisease/overview/)
* [*SQL Server Analysis Services 2012*](https://mis.uhcl.edu/rob/Course/DW/Resources/SQL%20Server%202012%20Analysis%20Services%20Multidimensional%20Modeling.pdf)
* *https://www.data.gov/*