DWBI

Project Report

Factors affecting Mortality Rate

Authors:

Sunil Elangovan

Vasuki Manoharan

Nikheel Navanale

Praudeep Namakkal Balasubramani

TABLE OF CONTENTS

Revision History

Date	Version	Changes	Author
11/05/20	1	Initial Document	Team
11/15/10	1.1	Added revision history and table of contents	Team
11/15/20	1.2	Updated missing information on sources and explained metrics used	Team
11/15/20	2	Submitted for review	Team
12/1/20	2.1	Added new income data source by Zipcode and County	Team
12/2/20	2.2	Created ER diagram	Team
12/3/20	2.3	Created dimensional table	Team
12/4/20	2.4	Loaded the fact tables and dimensional tables in SSIS	Team
12/4/20	2.5	Cube designs	Team
12/4/20	2.6	Tableau Analysis Test	Team
12/4/20	3	Submitted for review	Team
12/16/20	4	Final Report submission	Team

A. Description of the data to be stored and moved into the analytics system

1. Data Sources and Description

Source 1: United States Substance Use Disorders and Intentional Injuries Mortality Rates by County 1980-2014

Description: The dataset contains Mortality rates by geographic variation, that is caused due to self-harm, alcohol use, drug use, and interpersonal violence for the years 1980, 1990, 2000, 2010, and 2014. Additionally, we have the mortality rate – age-standardized, for both sexes. Contributing organizations: Institute for Health Metrics and Evaluation (IHME)

Source 2: Socioeconomic Indicators – Education, Population, Unemployment

Description: Dataset contains socioeconomic indicators by county in the US, which is further sized down to rural-urban continuum code. It includes information about educational attainment for adults age 25 and older, population estimate and unemployment, and median household income for the years 1980, 1990, 2000, 2010, and 2014. Contributing organizations: Economic Research Service, USDA

Source 3: Median Household Income measures by County

Description: This dataset informs the income by County in the US. It includes the Median household income for the year 2014 matched with FIPS of the counties

Source 4: Accidents by County - 2010 and 2014

The Bureau of transportation provided why and when the accidents occurred along with the detail if alcohol was involved for the years 2010 and 2014. It included it's unique accident number also.

2. Description of Data Joins

The datasets were combined to check for correlation between the socio-economic factors in each county. The mortality table is combined with the socio economic factors and then accident tables are combined to it using year and location keys.

3. Data Transformation

- Transform all the FIPS (Federal Information Processing Standards) columns to character format.
- All the other columns should be in the float data type.
- Remove State extension name from Area name.
- Remove all values in the character data type under the Mortality rate column.

Note: FIPS code is a number that uniquely identifies a geographical location. County-level FIPS codes have five digits of which the first two are the FIPS code of the state to which the county belongs.

4. Data Warehouse Architecture

Conceptual Model

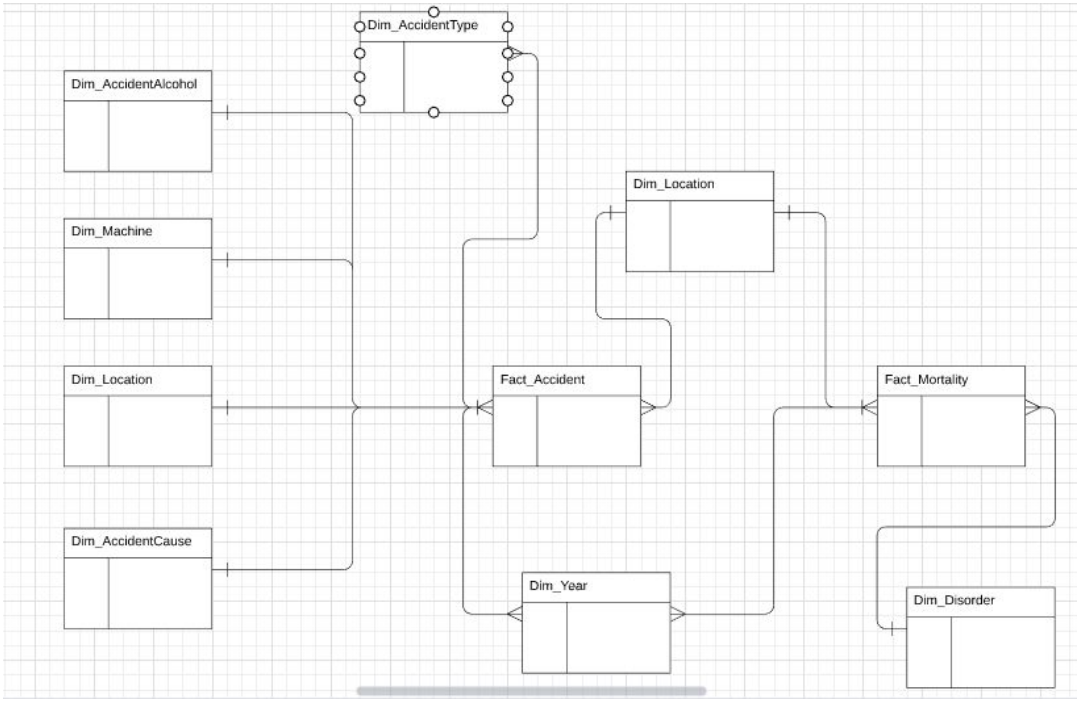


Fig 1

5. Dimensional Model

The dimensional model used follows Inmon's Top Down approach. We use two fact tables and with two linking dimension tables. There are no relationships between any dimensional tables and hence we have a star schema.

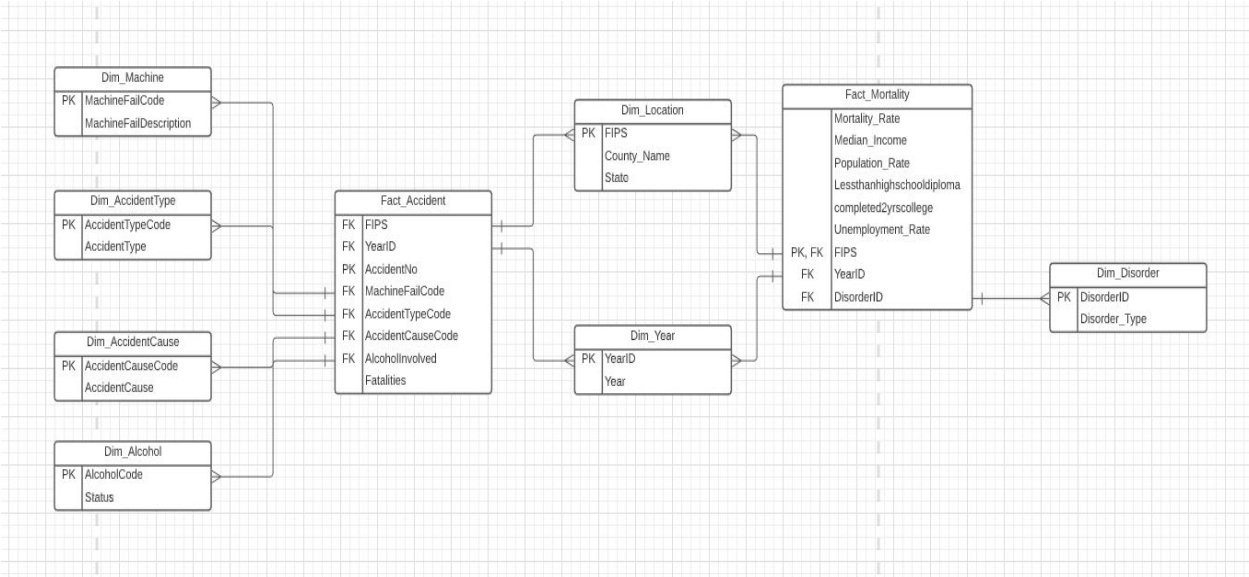


Fig 2

B. High-level data flow

1. Data Flow Diagram

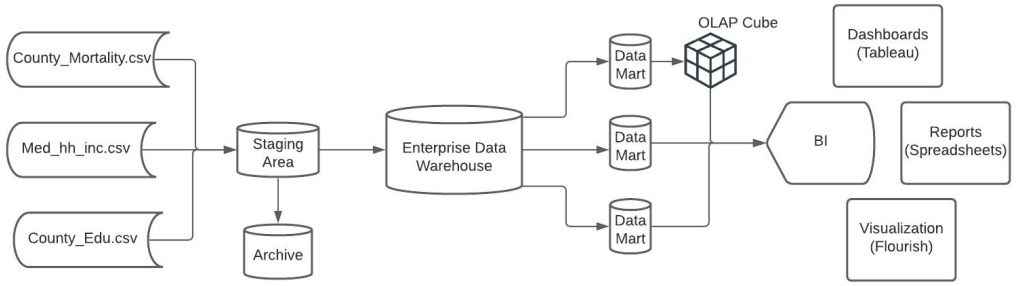


Fig 3

In this step we create destination tables for dimensions and fact tables to load in the data warehouse, from the CSV files. We use a lookup table with a correct county FIPS to match the values and clean the data while loading on to the staging stage.

3. Error Handling

Types of error handling performed:

- Checking if the (mortality,unemployment) rate values are within range
- Checking for character values in integer columns

4. Logging Process

Logs of any error handling mechanism should be saved in a separate table or file. It will contain information about error description, success/failure message, and time of failure. If the error has any major business impact, the ELT process will be stopped immediately.

5. First time vs monthly loads

Subsequent changes to the data will be tracked using Slowly Changing Dimensions (SCD). This tool gives us the option to maintain changes in the dimension table in the warehouse. It also gives us an option to keep a record of all the historical data along with the current data.

Type 1 : Accident type - slowly changing attribute

Type 2: Accident Cause - historical attribute

6. Data warehouse design along with the history

External Sources to staging area: Since the data is collected externally they don't follow a particular format. So there was a need to validate it before we load it into the warehouse. Hence we use the ETL tool. We extract the data, and place it in the staging table. Then after datatype transformation,error handling and logging we place it in their destination table and we finally load it into the warehouse.

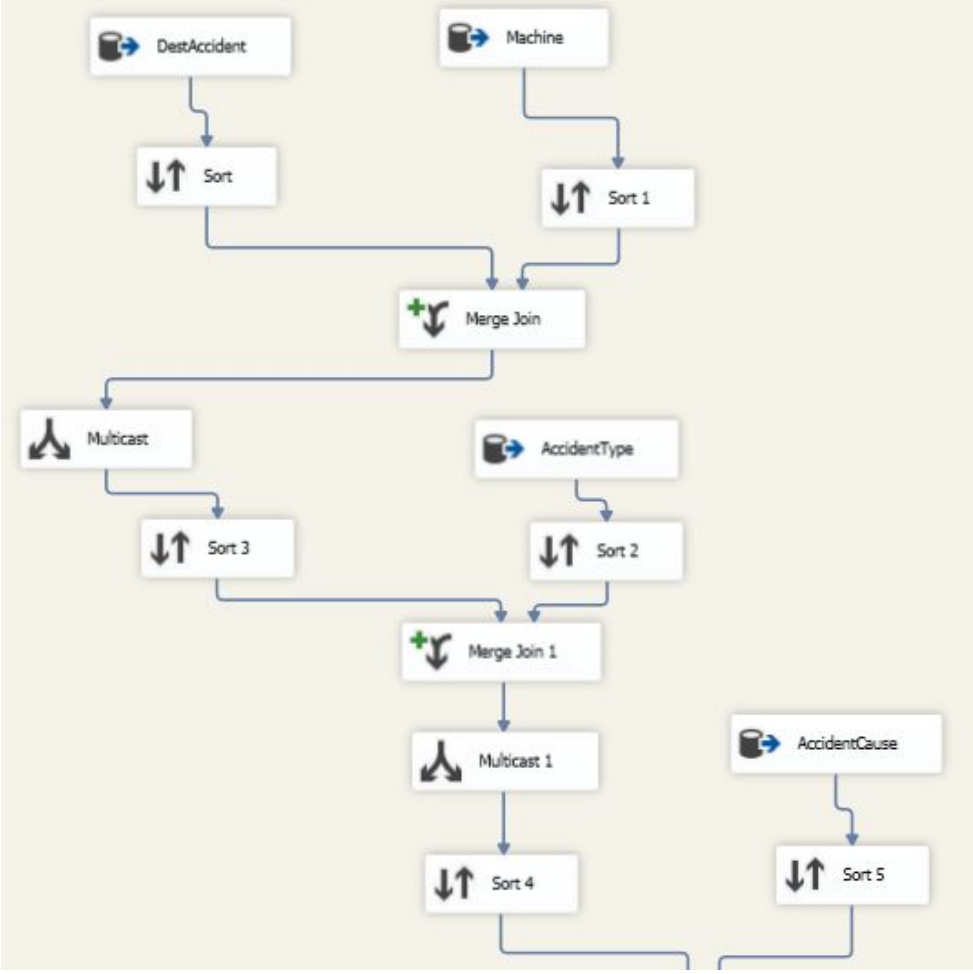


Fig 5

7. Data mart design with aggregates

Socio-economic factors in each county across the US shall be correlated/aggregated using FIPS code as the primary key. Datamart will have access to county data for a particular state for quick and easy access by their respective state healthcare providers. This enables business analysts to extract necessary information with respect to different county's within a particular state.

C. Data Validation

After analyzing and understanding the Unemployment, Education, and IHME data sets using SSIS.

- Remove ' , ', ' " ', ' - ', space, etc. in the column names.
- Same data types for each column.

- No length variation and unknown characters in the FIPS column.
- No null values in the Location, State, Area name column.
- No commas (,) in all the fields except Location, State, Area name.
- No duplicate column in Area name.

2. How will validations be logged/recorded

- Validations will be logged at regular intervals after each transformation process. Then all the logged files will be reported according to the updates to the data set.
- The following validation outcomes will be logged in a script file or an excel data sheet periodically.

D. Analytics Design

1. OLAP cubes from star schema

An OLAP DB would be created to make a near-instantaneous analysis of data as it is easily searchable for specific as well as broad terms.

Below are the screenshots of steps that were taken to create, load and design the OLAP cube to analyze data.

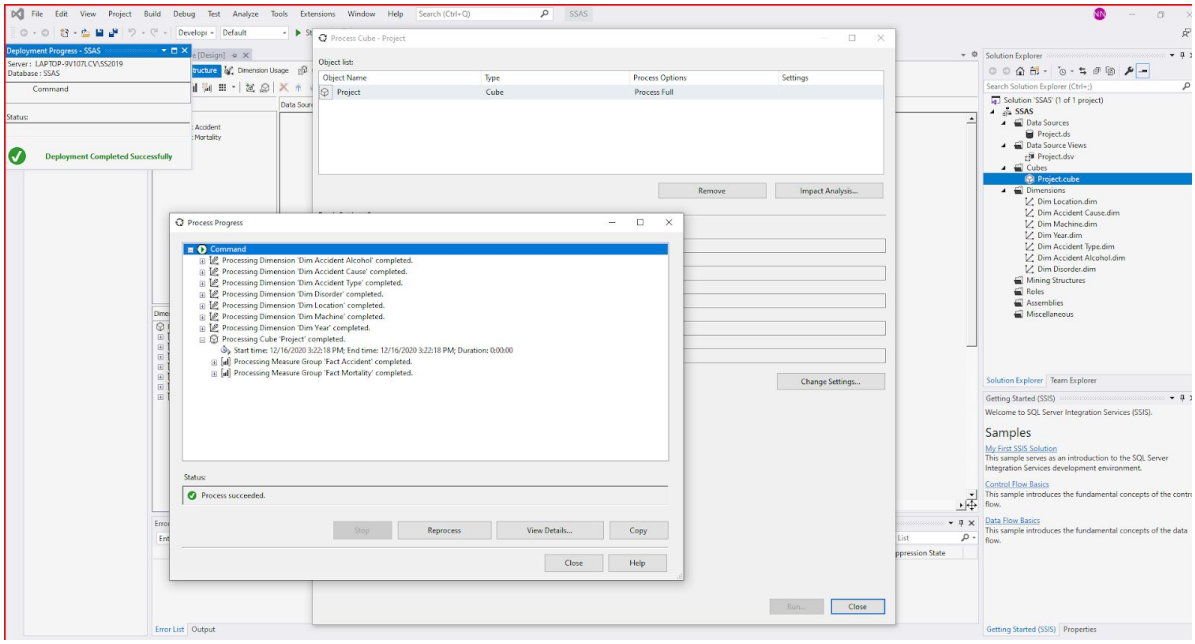


Fig 6 (Deployment)

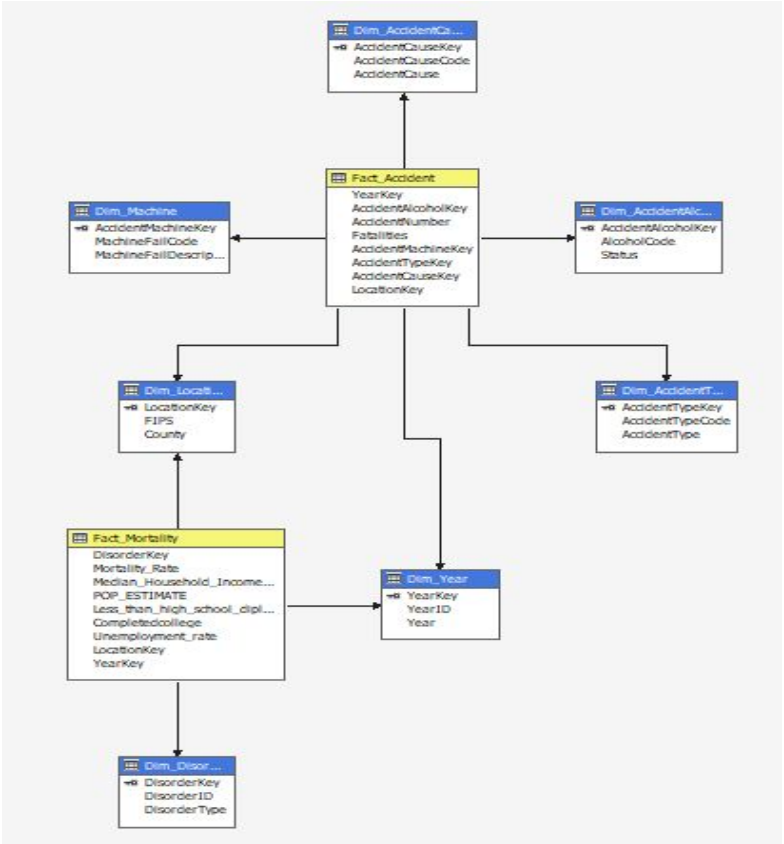


Fig 7 (Diagram of OLAP Cube)

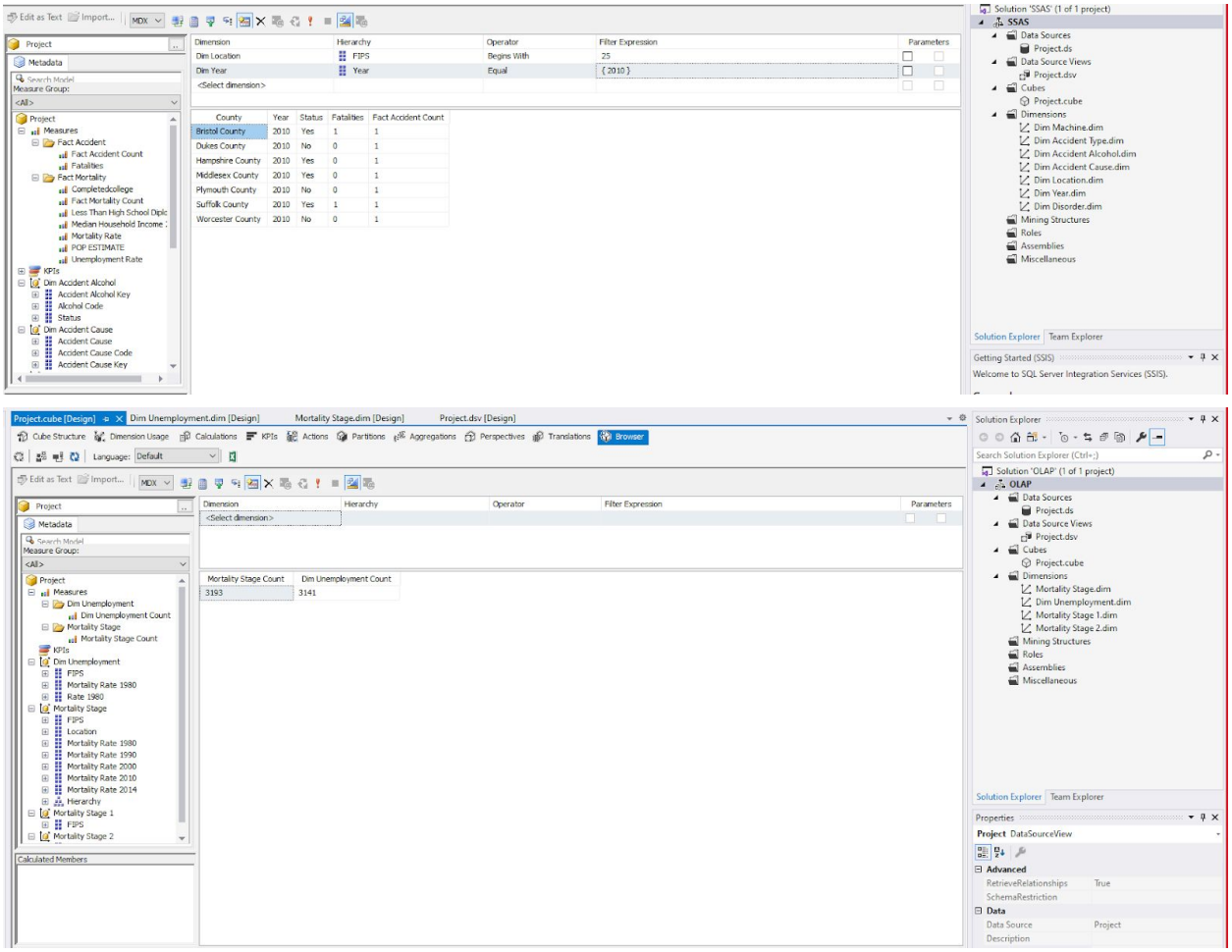


Fig 8 (Analysis)

2. Reporting tools and visualizations

We have planned to use visualization tools such as Tableau and Flourish to analyze data, create dashboard/operational dashboards, forecast results, and visualize meaningful information.

E. Analytics Outcomes

1. Why is this data interesting?

The data sets are collected from various sources on mortality rate, and accidents happening in the US due to various reasons like machine breakdown, accidents due to intoxication and how education affects mortality rate and accidents across US etc. We wanted to know the factors

that contribute to mortality. And to see, if other ways like accidents that are prominent nowadays, have an impact on the mortality rate.

2. Questions Answered

Case 1:

- Has there been an increase in the Mortality? If yes when?

  For example, here we see the Mortality rate of counties in Massachusetts in the year 1980

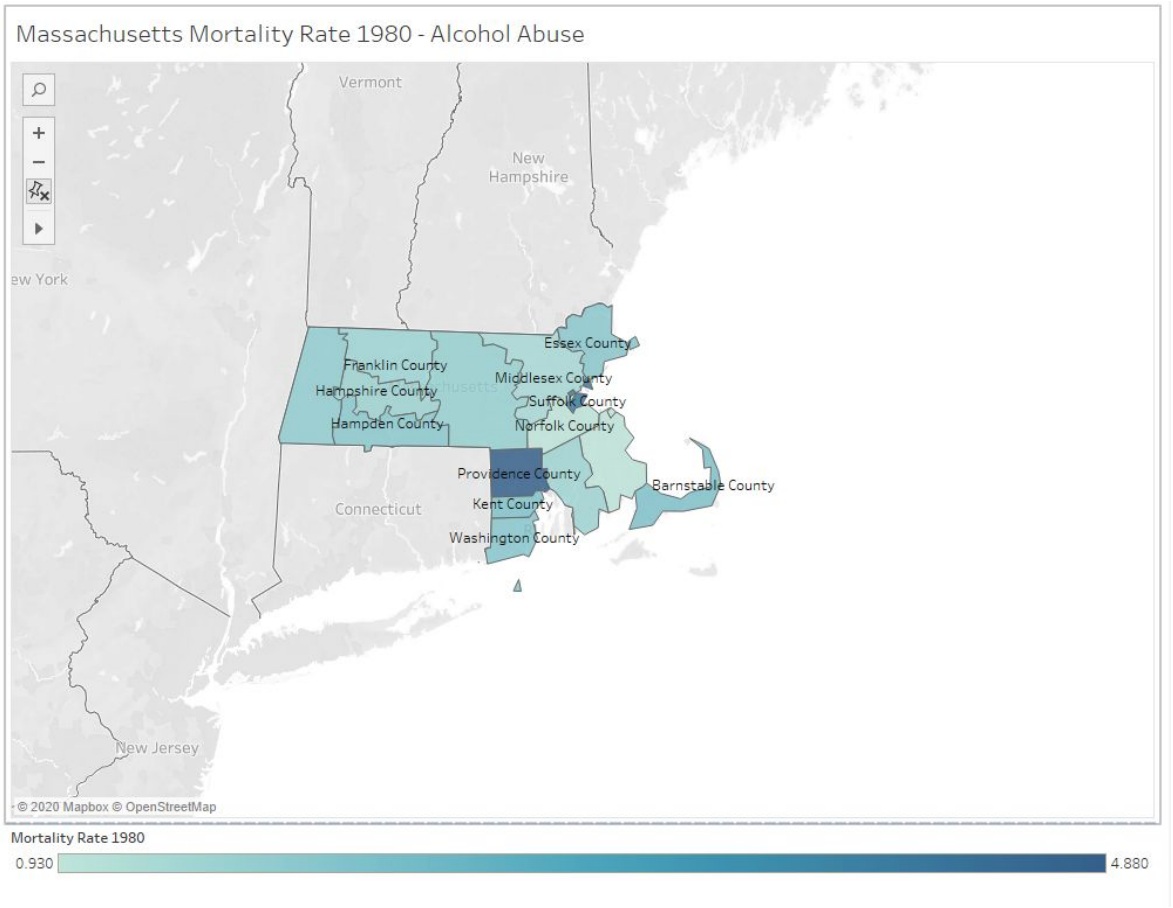
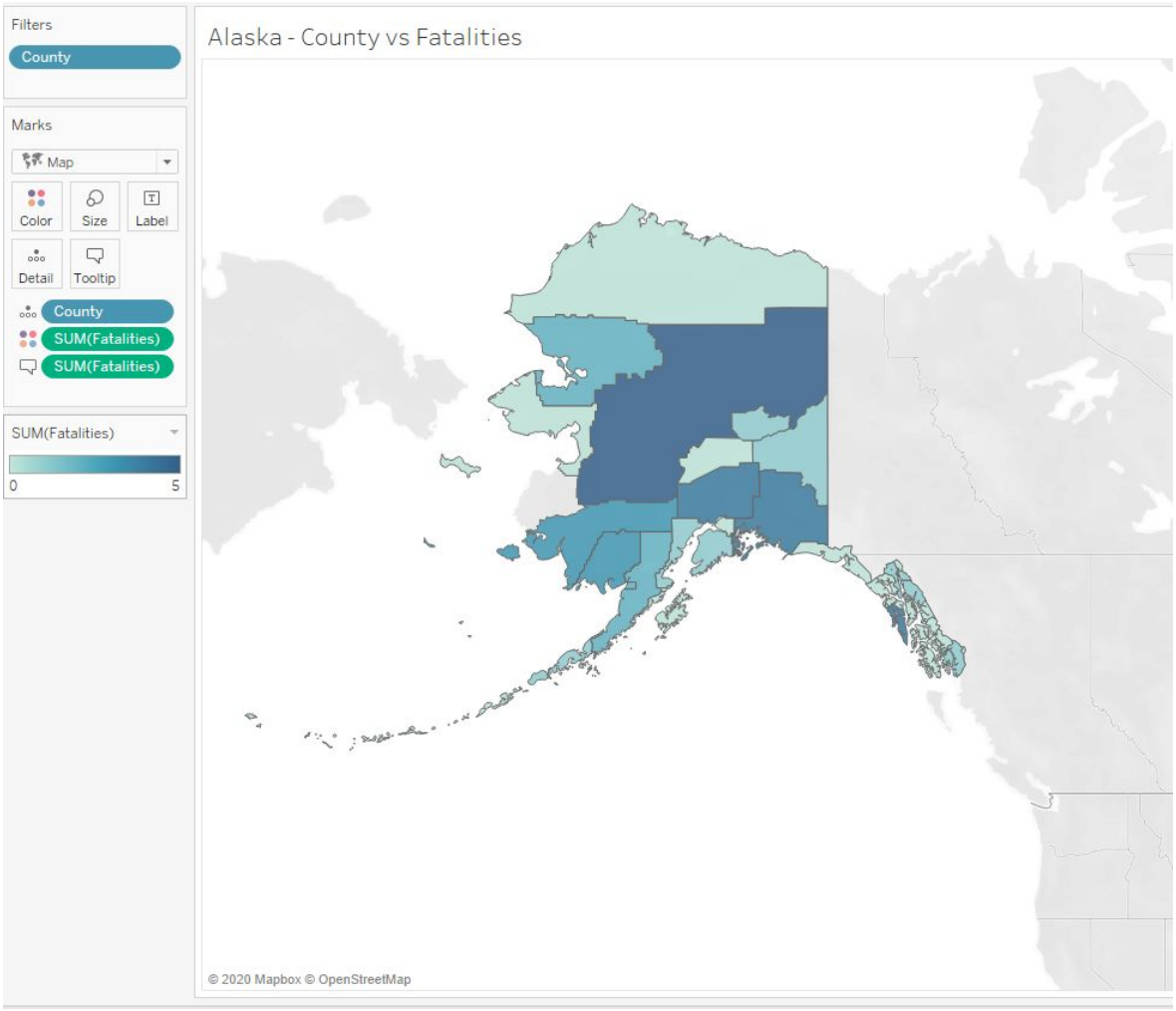


Fig 8 (Tableau Visualization)

Case 2:

- Does the income of the geographic region contribute to the mortality rate variation?

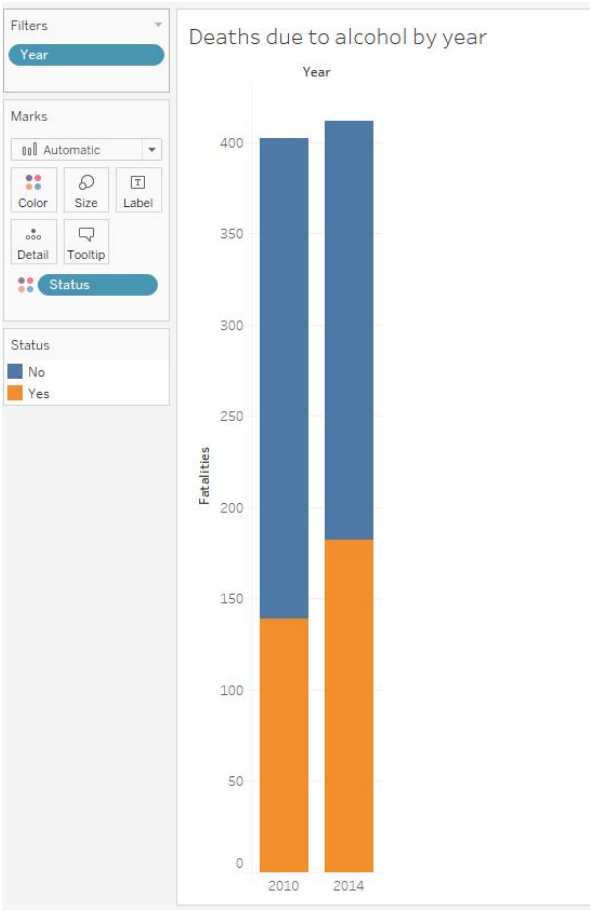
And these answers will lead to further questions that we will analyze as we move forward. We aim to inform target prevention and treatment to improve health that is currently affected by geographic inequalities.



The above figure visualises the Fatalities in each county located in Alaska. The higher the fatalities, darker is the colour. Through this pattern we can observe that Yukon-koyukuk is the county with highest fatality followed by Matanuska- Susitna and Chugach counties respectively.

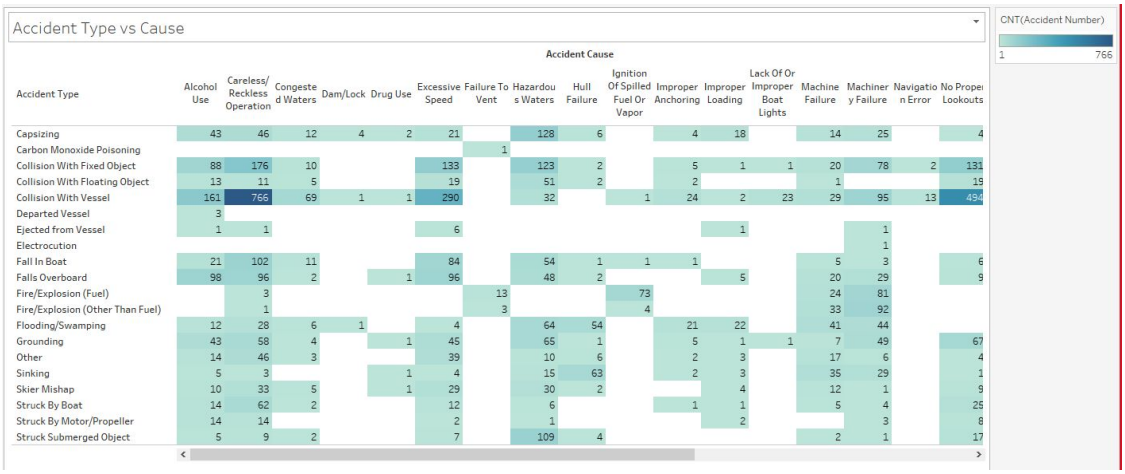
Case 3:

The above bar chart represents the total number of deaths that occured in 2010 and 2014 due to alcohol consumption. It's evident that the deaths were high during the year 2014 and less during 2010.



Case 4:

A heat map is typically used to display the data along with colors. The above image depicts a heat map of correlation between Accident types and the Causes of an accident. We can infer that careless/reckless operation is highly correlated with collision with vessels with a value of 766.



Accident Type	Alcohol Use	Careless/ Reckless Operation	Congested Waters	Dam/Lock	Drug Use	Excessive Speed	Failure To Vent	Hazardous Waters	Hull Failure	Ignition Of Spilled Fuel Or Vapor	Improper Anchoring	Improper Loading	Lack Of Or Improper Boat Lights	Machine Failure	Machinery Failure	Navigation Error	No Proper Lookouts
Capsizing	43	46	12	4	2	21		128	6		4	18		14	25		4
Carbon Monoxide Poisoning							1										
Collision With Fixed Object	88	176	10			133		123	2		5	1	1	20	78	2	131
Collision With Floating Object	13	11	5			19		51	2		2			1			19
Collision With Vessel	161	766	69	1	1	290		32		1	24	2	23	29	95	13	494
Departed Vessel	3																
Ejected from Vessel	1	1				6						1			1		
Electrocution															1		
Fall In Boat	21	102	11			84		54	1	1	1			5	3		6
Falls Overboard	98	96	2		1	96		48	2			5		20	29		9
Fire/Explosion (Fuel)		3					13			73				24	81		
Fire/Explosion (Other Than Fuel)		1					3			4				33	92		
Flooding/Swamping	12	28	6	1		4		64	54		21	22		41	44		
Grounding	43	58	4		1	45		65	1		5	1	1	7	49		67
Other	14	46	3			39		10	6		2	3		17	6		4
Sinking	5	3			1	4		15	63		2	3		35	29		1
Skier Mishap	10	33	5		1	29		30	2			4		12	1		9
Struck By Boat	14	62	2			12		6			1	1		5	4		25
Struck By Motor/Propeller	14	14				2		1				2			3		8
Struck Submerged Object	5	9	2			7		109	4					2	1		17

F. Appendix

Professors Comments:

11/15/2020

Add a revision history

What years is the second data set available for

What is a fips code – if you use a term you need to explain it

Why do you mention covid

11/27/2020

validation section missing info / incomplete

talk about the olap design

I still don't understand why you mention covid

You should bring in some zip code data from another source

Also look for income by zip code from another source

References:

https://www.transtats.bts.gov/TableInfo.asp

http://www.learnmsbitutorials.net/ssis-star-schema-designing.php

https://bennyaustin.com/2013/08/19/processcube/

https://datawarehouseinfo.com/advanced-dimensional-modeling-techniques/