ASSIGNMENT 3

DATA MANAGEMENT &WAREHOUSING ANALYTICS

SHRUTHI KALSAPURA RAMESH
B00822766

A. Sentiment Analysis

Sentiment Analysis is performed on the search tweets extracted with the help of tweepy library during Assignment 2. The format of the tweets is changed to fit the current assignment by considering only tweet text and eliminating other attributes like metadata.

Steps:

- Tweet texts are cleaned again to ensure accurate results.
- All the stop words were removed with the help of a text file that contains a set of stop words [3].
- Bag of words with their count is create from each tweet.
- Count of positive and negative words in tweets are calculated by comparing each word with a list of predefined positive and negative words [1].
- The polarity of tweets is inferred base on the count of negative and positive words.
- All the information with tweet text, bag of words and their count, matched positive and negative words, count of negative and positive tweets and final sentiment [5]. (Polarity) is stored in a Json file called *sentimentOutput.json* (figure 1).

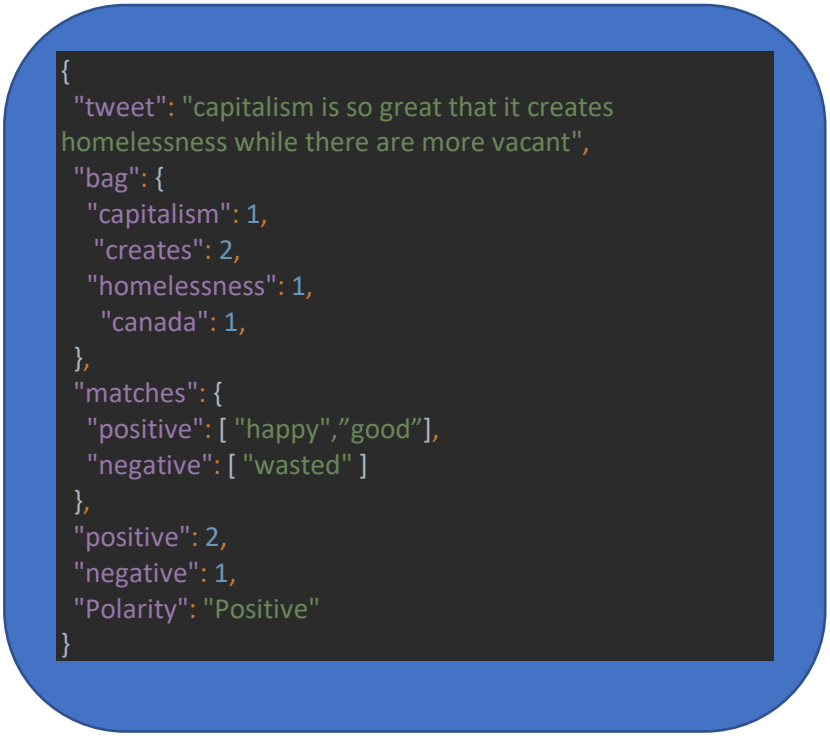


Figure 1: Output of sentiment analysis

B. Semantic Analysis

Semantic Analysis[6] is performed on the articles extracted from Reuter files during Assignment 2.

Steps for semantic analysis:

- Articles are cleaned to extract only text.
- A list of words in each article is created.
- Count of words in each article, and relative frequency of each word over total number of words is calculated.
- Term frequency (number of file a word appeared in- **df** and the total number of files- **N**) is calculated using the formula **Log10(N/df)** [2].
- Maximum relative frequency among all the documents is determined [2].
- Output of the sematic analysis is stored in a Json file with name *semanticOutput.json. (*figure 2*)*


```
{
 "canada": [
  {
   "count": 2,
   "totalwords": 133,
   "filename": "article1001.txt",
   "relFreq": 0.015037593984962405
  },
  {
   "count": 1,
   "totalwords": 77,
   "filename": "article1040.txt",
   "relFreq": 0.012987012987012988
  },
  {
   "count": 1,
   "totalwords": 220,
   "filename": "article1058.txt",
   "relFreq": 0.004545454545454545
  }]
}
```


Figure 2: Output of semantic analysis

C. Business Intelligence

I chose the following datasets for the assignment to work with IBM Cognos, which is a business intelligence tool, to analyse and represent data in a visual format [4].

Economics & Industry	Name: *Cannabis Income account* URL: *https://open.canada.ca/data/en/dataset/86a5c29c-0871-47ad-8da6-8a6b3992aea1* The dataset gives an insight of income by cannabis industry.
Education	Name: *Second Language Immersion Schools in Canada* URL: *https://open.canada.ca/data/en/dataset/2bfebd29-1a98-4c57-9134-93f1b18190ea* The dataset provides information about Second Language Immersion Schools in Canada based on province.
Vehicles	Name: *New motor vehicle sales, by type of vehicles* URL: https://open.canada.ca/data/en/dataset/f6e7e871-79b7-49e1-90a2-e3c913f1951d This dataset provides details of sales of vehicles in different provinces of Canada over the years.

BI Framework: Analysis of facts and dimensions

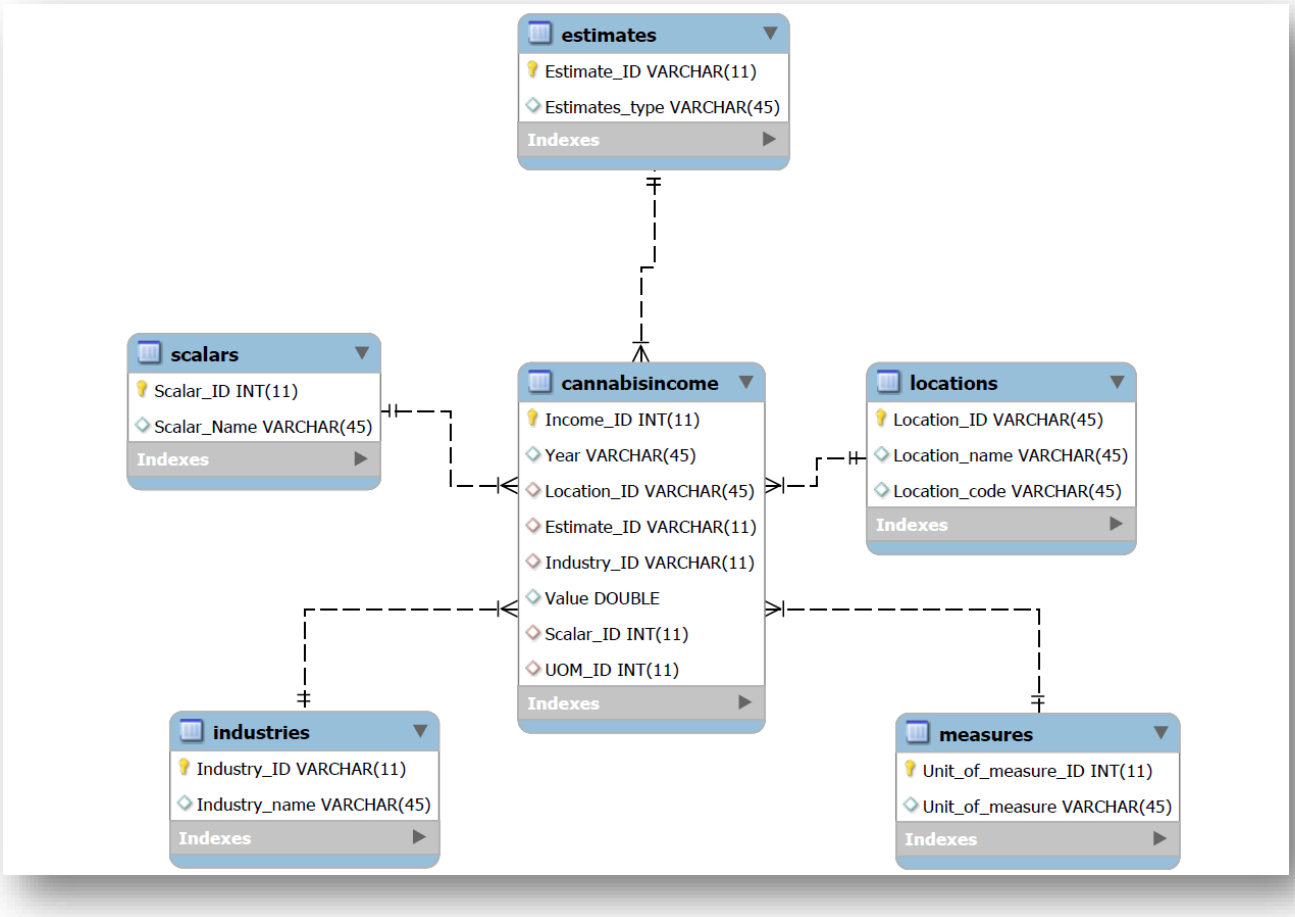
The following are the fact and dimension tables obtained, after cleaning the datasets, for domains Economics & Industry, Schools and Vehicles.

Economics & Industry

- Facts Table
  CannabisIncome(Income_ID, Year, Location_ID, Estimate_ID, Industry_ID, Value, Scalar_ID, UOM_ID)

- Dimension Tables & attributes
  Locations(Location_ID, Location_name, Location_code)
  Estimates(Estimate_ID, Estimate_type)
  Industries(Industry_ID, Industry_name)
  Scalars(Scalar_ID, Scalar_name)
  Measure(Unit_of_measure_ID, Unit_of_measure_name)

Star Schema for Economics & Industry

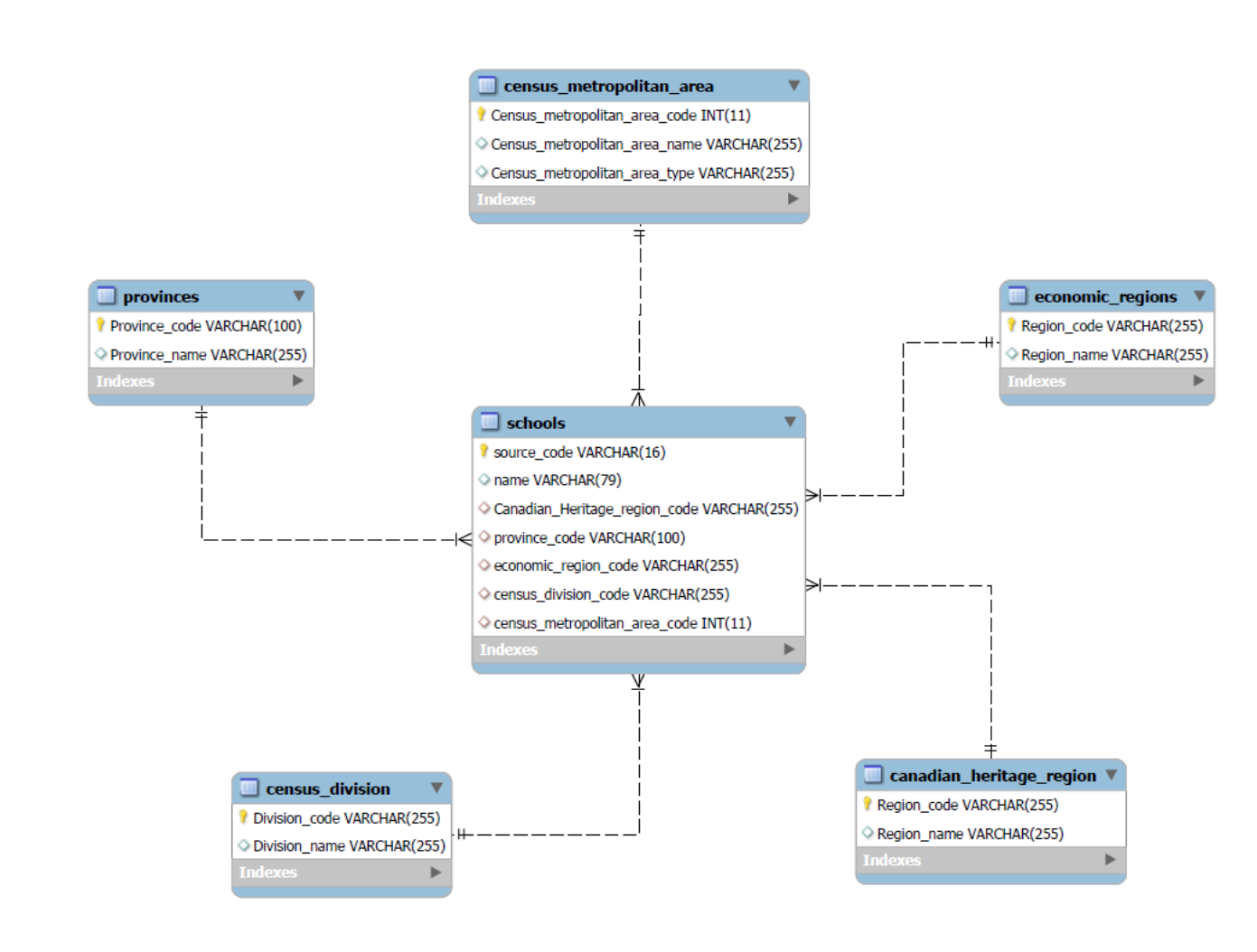


Schools

- Facts Table
  Schools(Source_code, Name, Canadian_Heritage_region_code, Province_code, economic_region_code, Census_division_code, Census_metropolitan_area_code)

- Dimension Tables & attributes
  Provinces(Province_code, Province_name)
  Census_metropolitan_area (census_metropolitan_area_code, census_metropolitan_area_name, census_metropolitan_aree_name)
  Economic_regions (Region_code, Region_name)
  Census_division (Division_code, Division_name)
  Canadian_heritage_region (Region_code, Region_name)

Star Schema for Schools

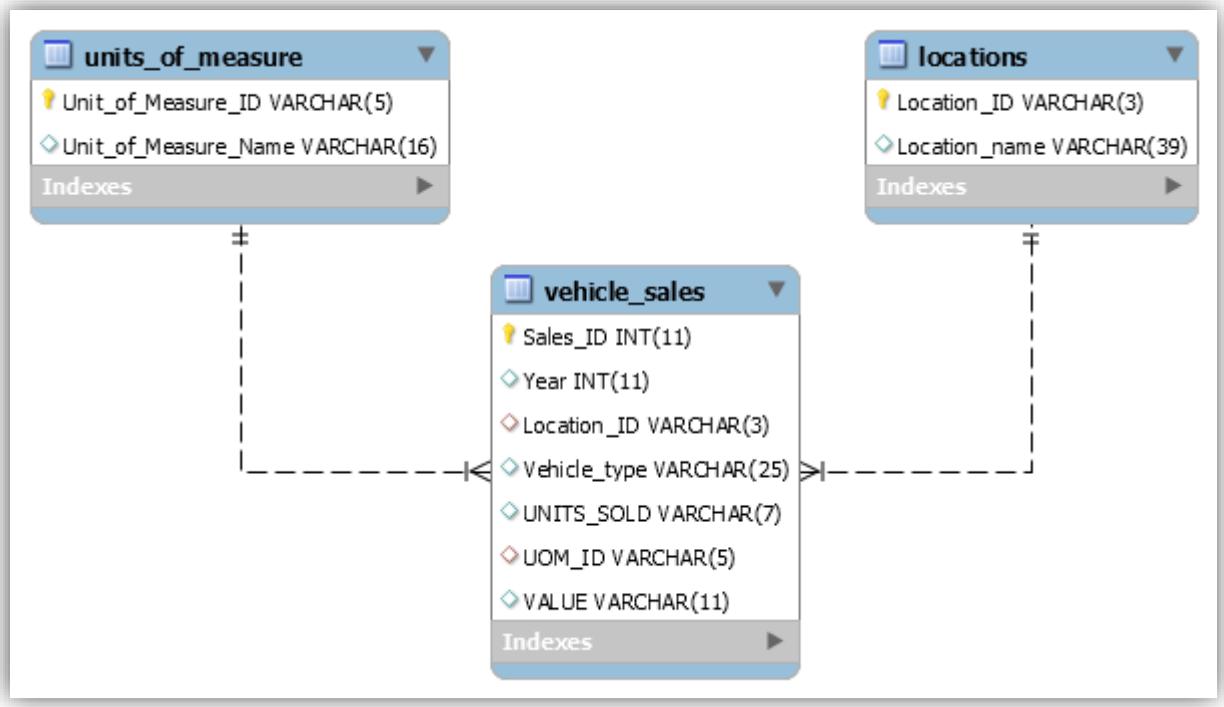


Vehicles

- Facts Table
  Vehicle_sales(Sales_ID, Year, Location_ID, Vehicle_type, Units_sold, UOM_ID,Value)

- Dimension Tables & attributes
  Units_of_measure(Unit_of_measure_ID, Unit_of_measure_name)
  Locations (Location_ID, Location_name)

Star Schema for Vehicles



IBM Congnos connection to MySQL

Based on the Lab tutorials provided, I connected to MySQL on EC2 instance of AWS from IBM Cognos.

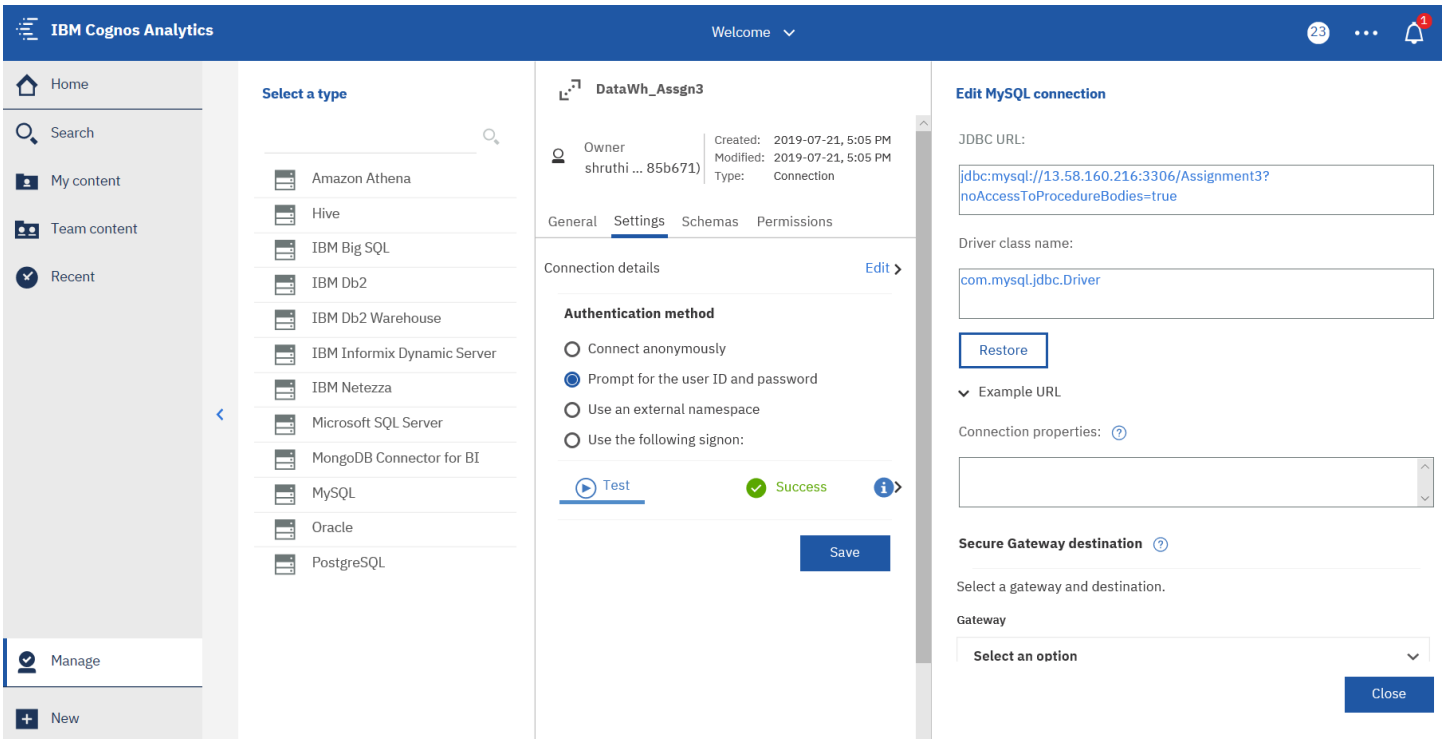


Figure: database server connection

BI Frameworks

Following are the relationships between fact and domain tables:

Economics & Industry

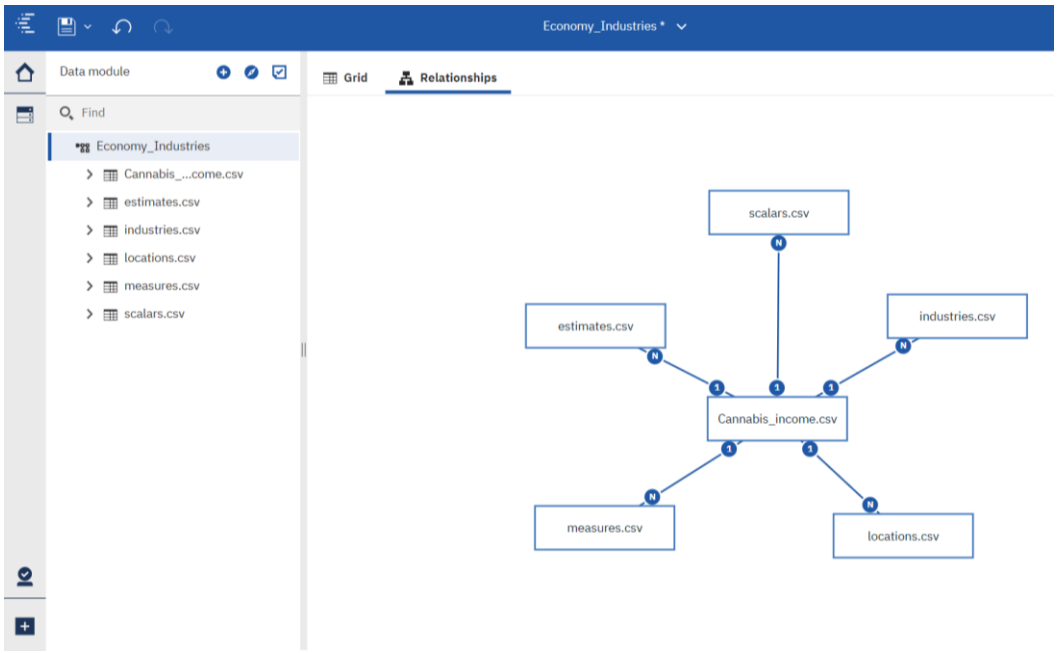


Figure: Relationship between fact table and domain tables for Economics & Industry

Schools

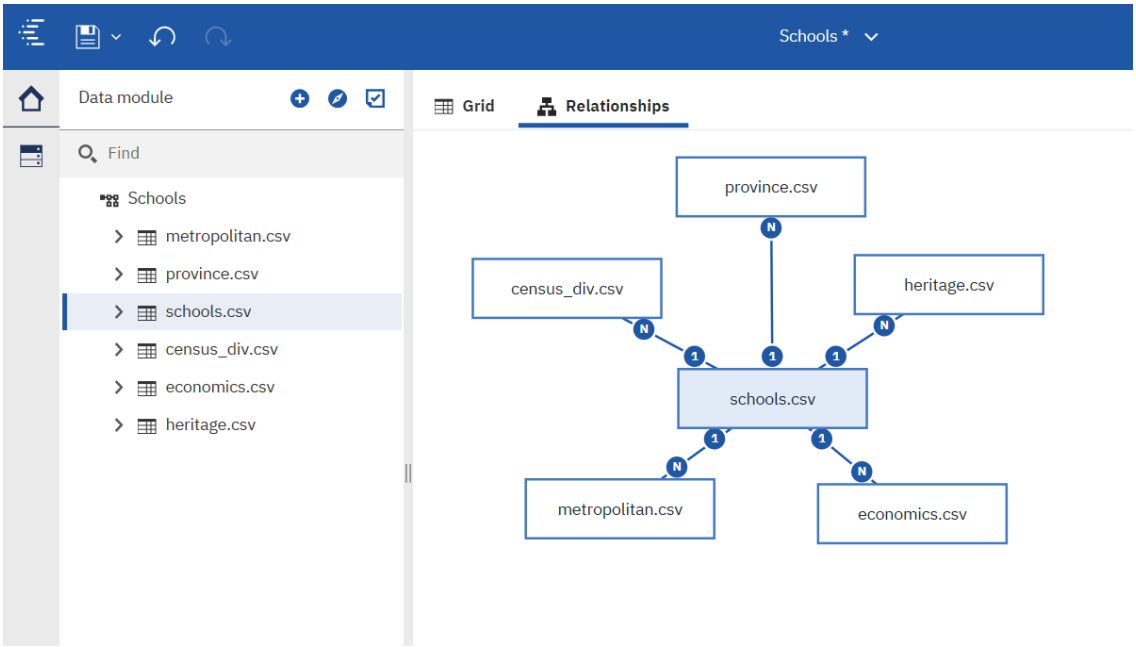


Figure: Relationship between fact table and domain tables for Schools

Vehicles

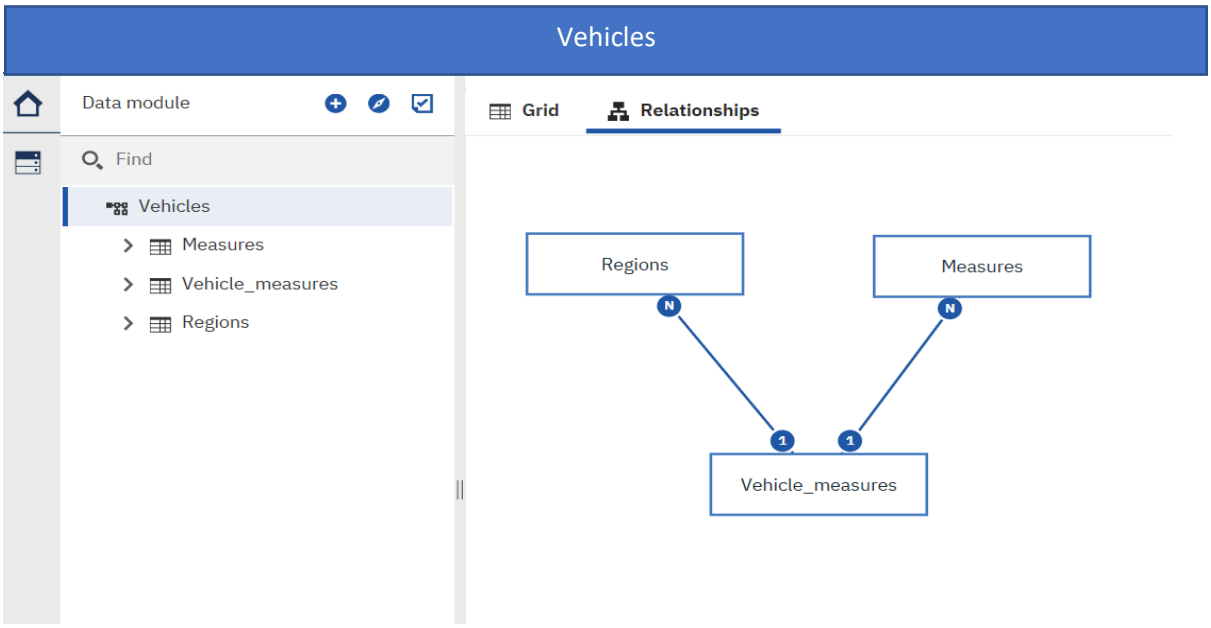
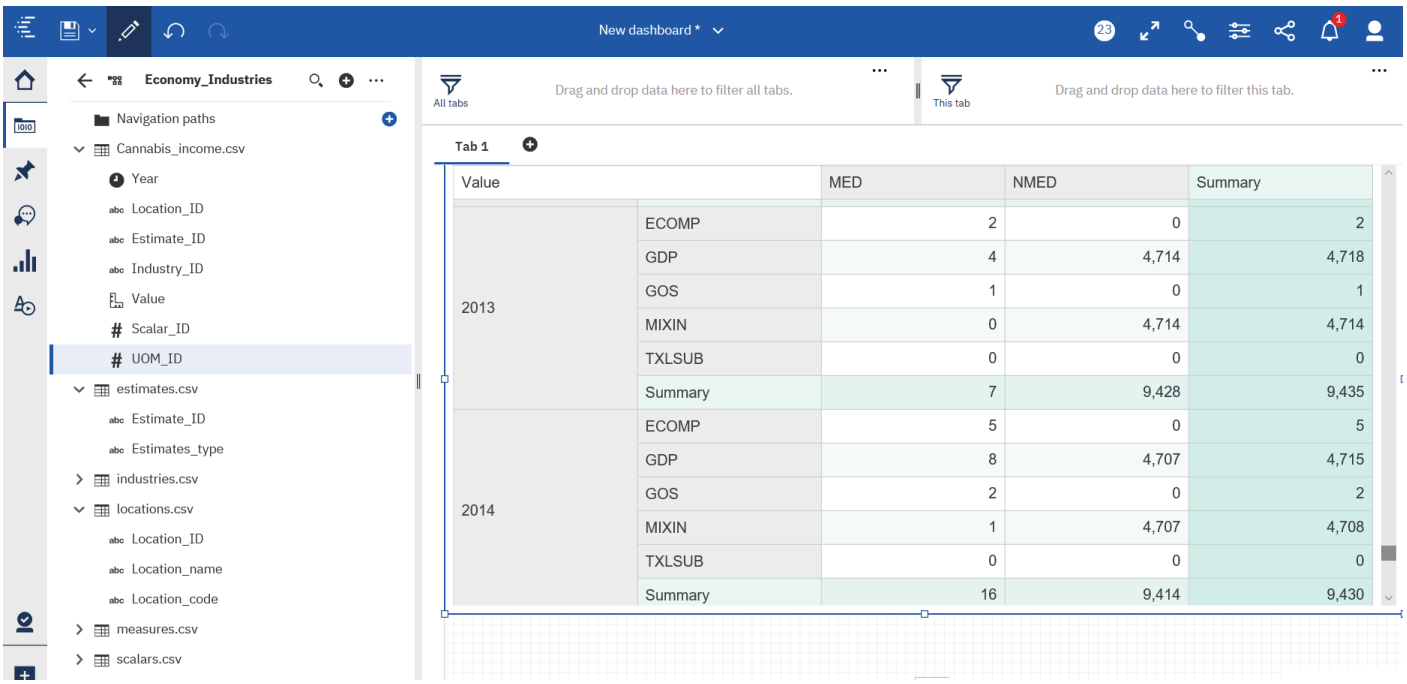


Figure: Relationship between fact table and domain tables for Vehicles

Multidimensional Data analysis

I have performed multidimensional data analysis on the data and generated the following graphs using IBM Congnos Tool.



Value		MED	NMED	Summary
2013	ECOMP	2	0	2
	GDP	4	4,714	4,718
	GOS	1	0	1
	MIXIN	0	4,714	4,714
	TXLSUB	0	0	0
	Summary	7	9,428	9,435
2014	ECOMP	5	0	5
	GDP	8	4,707	4,715
	GOS	2	0	2
	MIXIN	1	4,707	4,708
	TXLSUB	0	0	0
	Summary	16	9,414	9,430

Figure: Crosstab - Cannabis Industry income (in thousand dollars) by medical and non-medical industries over the years.

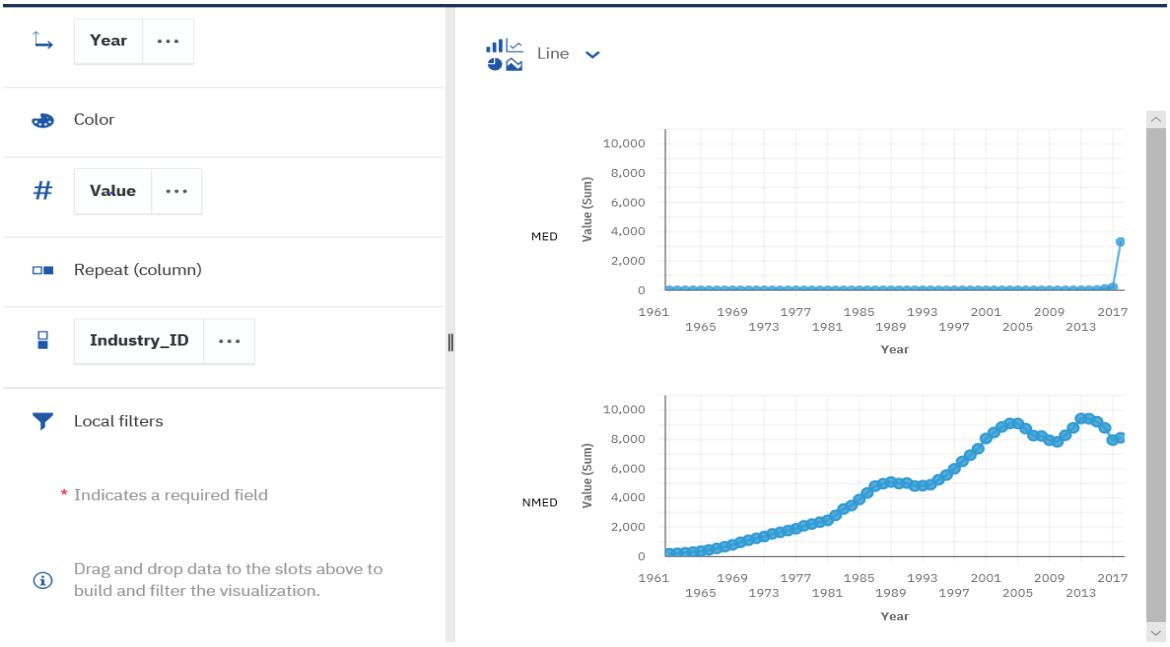
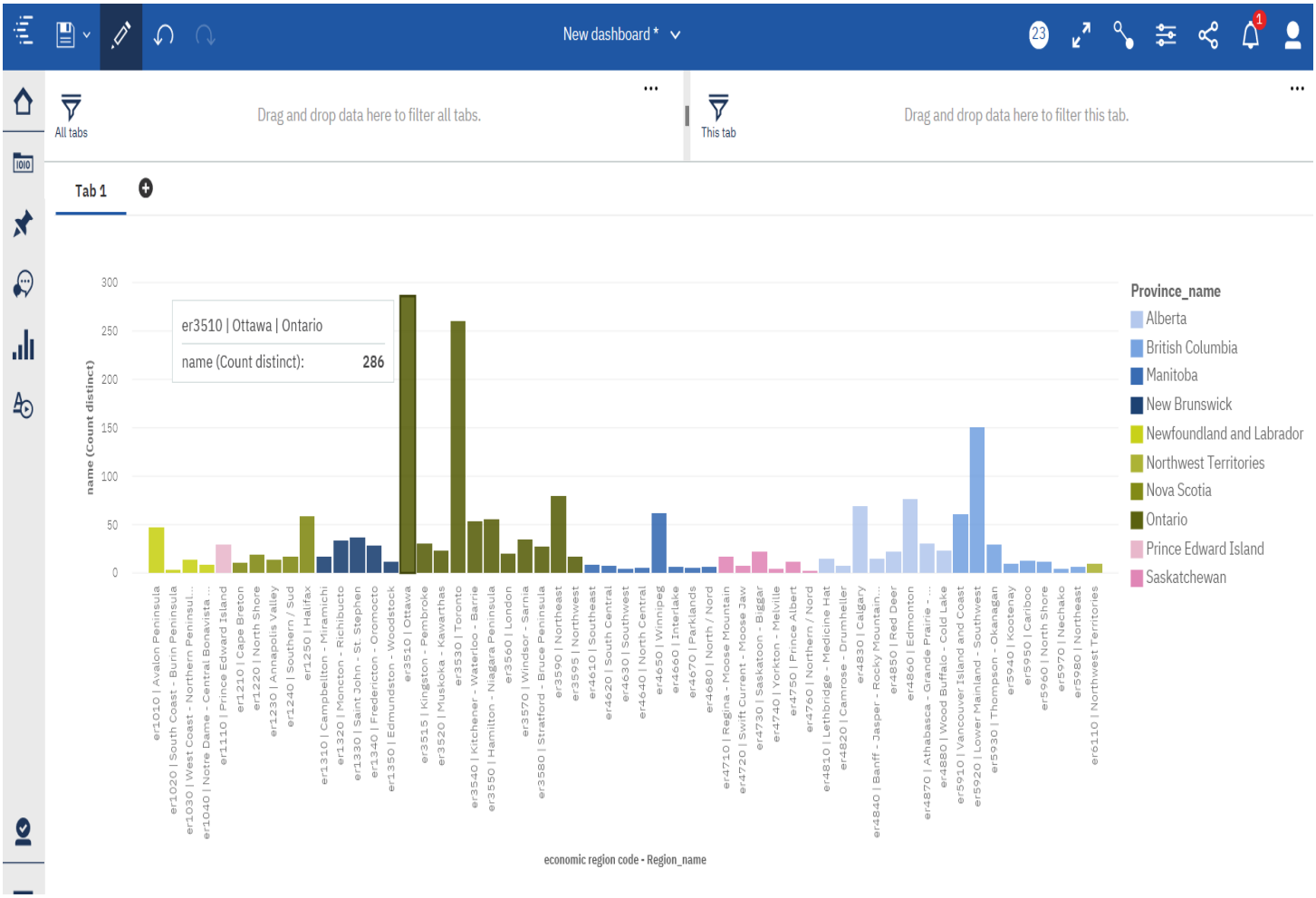


Figure: Cannabis Industry income (in thousand dollars) by medical and non-medical industries over the years.



<u>Figure</u>: Number of Schools in Canada against Provisions and Economic area

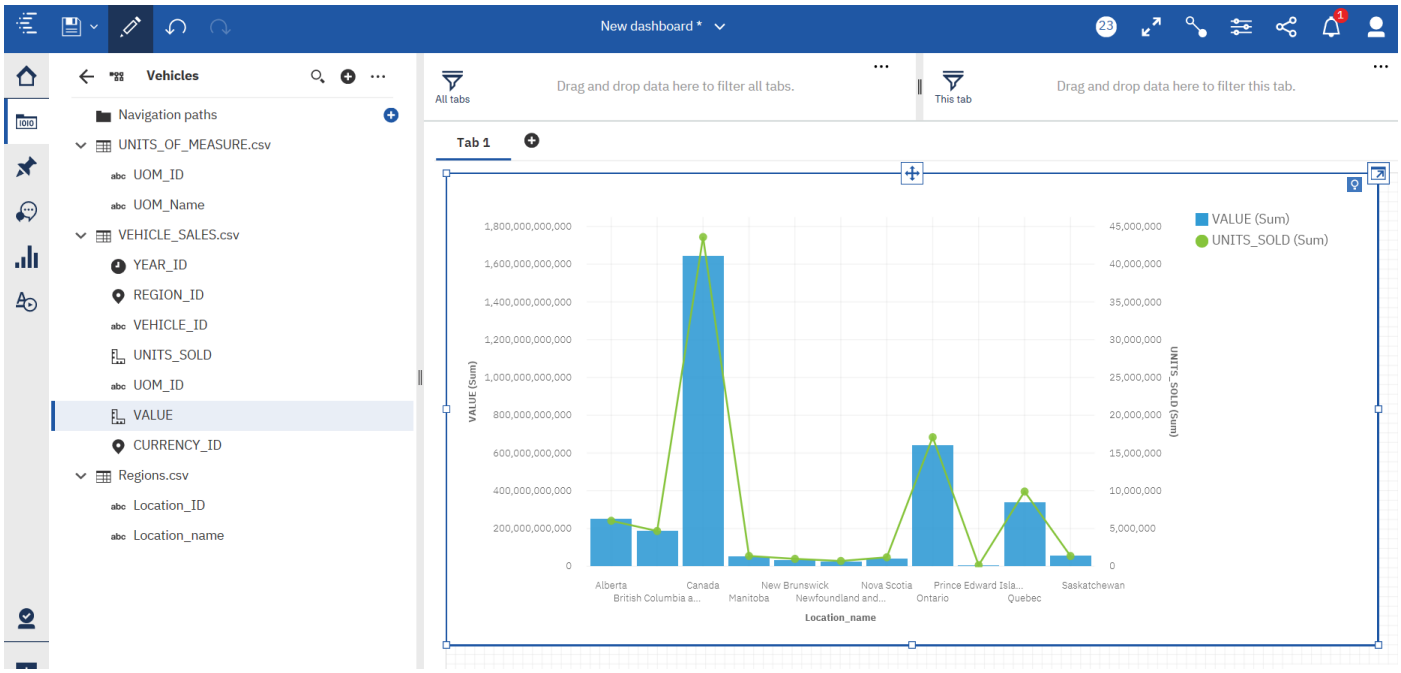


Figure: Number of vehicles sold and the sum of their value based on location.

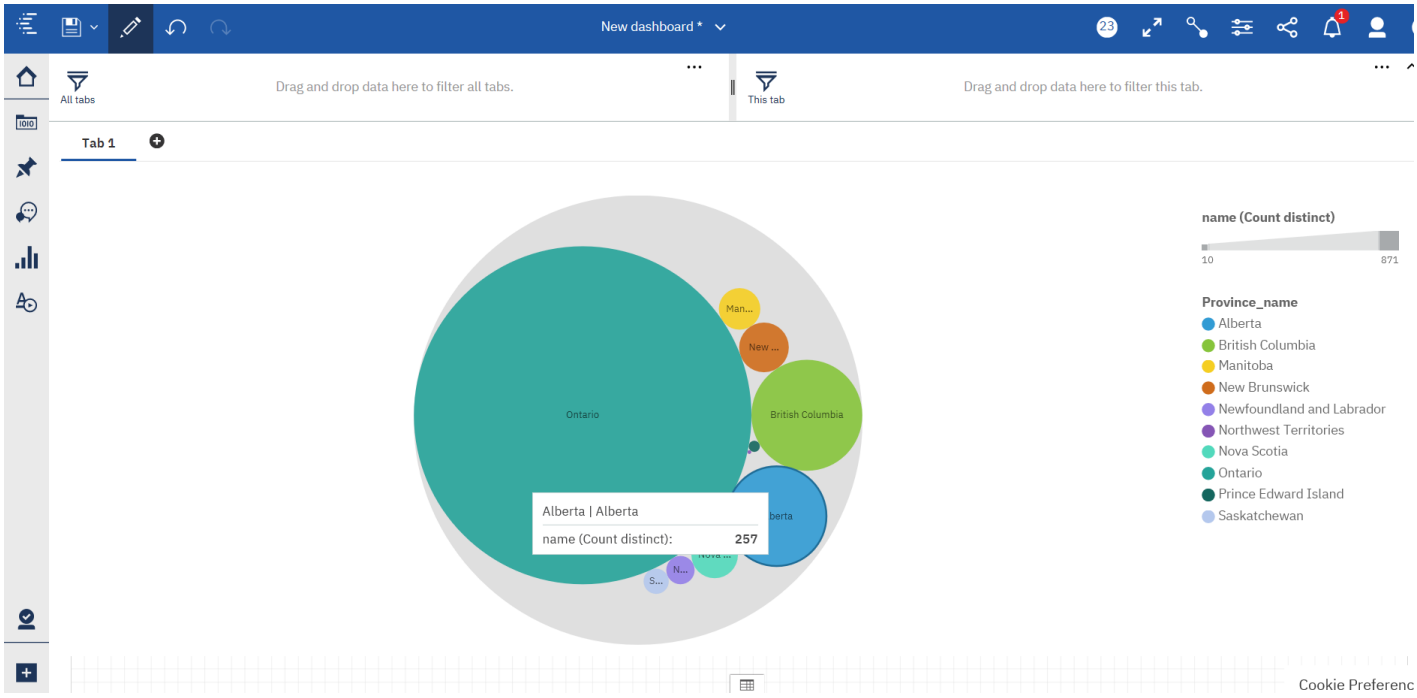


Figure: Number of schools in each province.

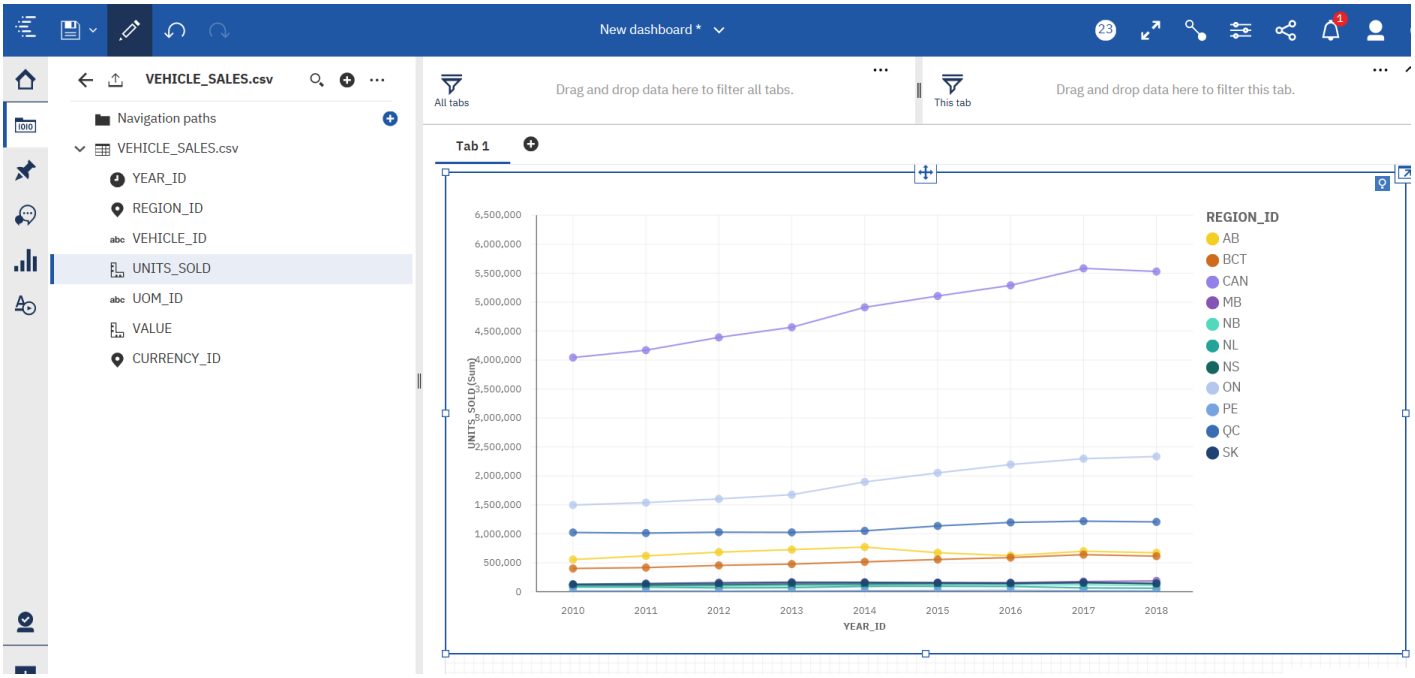


Figure: Growth in vehicle sales over the years.

References:

[1]"Sentiment Analysis Resources – Positive Words – Negative words", *POSITIVE WORDS RESEARCH*, 2019. [Online]. Available: https://positivewordsresearch.com/sentiment-analysis-resources/. [Accessed: 22- Jul- 2019].

[2]"math — Mathematical functions — Python 3.7.4 documentation", *Docs.python.org*, 2019. [Online]. Available: https://docs.python.org/3/library/math.html. [Accessed: 22- Jul- 2019].

[3]"Download Stop Word List - XPO6", *XPO6*, 2019. [Online]. Available: http://xpo6.com/download-stop-word-list/. [Accessed: 22- Jul- 2019].

[4]"Cognos Tutorial", *www.tutorialspoint.com*, 2019. [Online]. Available: https://www.tutorialspoint.com/cognos/. [Accessed: 22- Jul- 2019].

[5]"Sentiment Analysis: Nearly Everything You Need to Know | MonkeyLearn", *MonkeyLearn*, 2019. [Online]. Available: https://monkeylearn.com/sentiment-analysis/. [Accessed: 22- Jul- 2019].

[6]"Natural Language Process semantic analysis: definition -", *Expertsystem.com*, 2019. [Online]. Available: https://www.expertsystem.com/natural-language-process-semantic-analysis-definition/. [Accessed: 22- Jul- 2019].