8/5/2018

Data Management, Warehousing Analytics CSCI 5408

Project Report: Choosing a location for establishing a Business

Sogra Bilal Memon
B00786252

Bhavya Chandrappa
B00781097

Table of Contents

1. Summary

In this project we have focused on generating visualizations, based on current demographics that could be helpful to new comers who are aspiring entrepreneurs. For simplicity we have taken a possible business use case and accordingly modeled visualizations to show which city from Nova Scotia has the highest growth potential for this venture. These visualizations of the current population along with the income, age and language distribution gives aspiring business owners an idea of his/her potential customers, their product of interest and their spending budget. We have chosen the idea of starting a senior citizen café and have visualized the current demographics to find the city where this venture has the highest growth potential.

2. Business Idea, Problem statement, Value Proposition

Halifax is growing as a city and with this growth there are many people from other countries who want to migrate to Halifax. With the migration process comes a number of challenges such as visa application, renting a home, finding a school to setting up a business. While google is a great tool for all these searches it does not provide a visualization of all the data available in order for users to take immediate decisions. Users have to rely on their memory and constant google searches to make the right decisions for each of these situations. Our project aims to help newcomers set up a business with ease so that they can focus on other aspects of settling down into a new place. Our project will be working with CENSUS data to give potential business owners all the information they need about their surrounding neighborhood to come to a conclusion about which products will sell in that particular area. The CENSUS data is a record of the number of people living in each location in Canada. Demographic information such as age, gender, marital status and languages spoken are also available for people living in each location. This makes the CENSUS data the perfect dataset for potential business owners to look through before they take a plunge at their new business idea.

We aim to make the CENSUS data available to anyone who is looking to validate the feasibility of setting up their venture at a specific location. Our project will also help users find the best possible location for his or her venture based on the requirements for business success provided by the user. The user will be provided with data visualizations that will provide clear indications on which locations are better for the business as compared to others based on the parameters selected by the user. These visualizations can also be used by the user to develop business proposals and strategies that are not based on intuition but on actual data that can be presented to investors as evidence that the product or service will work.

Starting a business in Halifax is a goal of many Halifax citizens. Many times the users get stuck in locating a place to set up their business. There is no uncomplicated way to make the users' search successful. Our project aides the users to locate the suitable location by using data analytics concepts and tools. The data is and has been available for a long time. This project

aims to make it available to public in a way that it is easy to use and meets the demands of providing some certainty to a startup that is in an uncertain risky environment. Many startups fail every year not necessarily because the venture idea was bad but because it was executed at the wrong location. The location and timing are key to any business. With this solution, users will be able to make better choices based on data and quantitative comparison between locations.[3]

3. Implementation

The implementation of this project has been achieved through a series of sprint. The sprints have been discussing in detail in section 5. In this section we discuss the data set used and the procedures followed to clean, process, visualize and analyze the data.

3.1. Data Source

The Dataset being used in this project has been taken from CENSUS Canada. This dataset consists of all the information gathered by CENSUS in 2016. CENSUS is a survey that takes place every five years. The dataset taken contains information on population, income and language distribution for the entire Canada. These records have been taken for each of the counties in each of the provinces which caused it to occupy 84mb of memory. The data was highly unstructured and required the extraction and transformation of relevant features. The data occupies 84 Mb of memory space and has a basic schema as shown below.

Id	City	Feature	Feature value
1	Halifax	Total population	69084364
1	Halifax	Population over age of 65	14
1	Halifax	Population age range 15 - 64	143536
1	Halifax	Population percentage age 65 and over	40
1	Halifax	Population percentage age range 15 to 64	256678
1	Halifax	Total population with Income	23546576
1	Halifax	Population in income range under 10,000	23465787
1	Halifax	Population that speaks Hindi	23456

Figure 1: Work Structure Breakdown

3.2. Tools used

In this project we have used a variety of tools to store, clean, process, transform and perform visualization on the dataset chosen. Some of the tools and languages used include: Python, AWS EC2 instance, Spark, MySQL and Tablue. The details of how each of the tools that were used has been described in section 3.3.

3.3. Planned Procedures

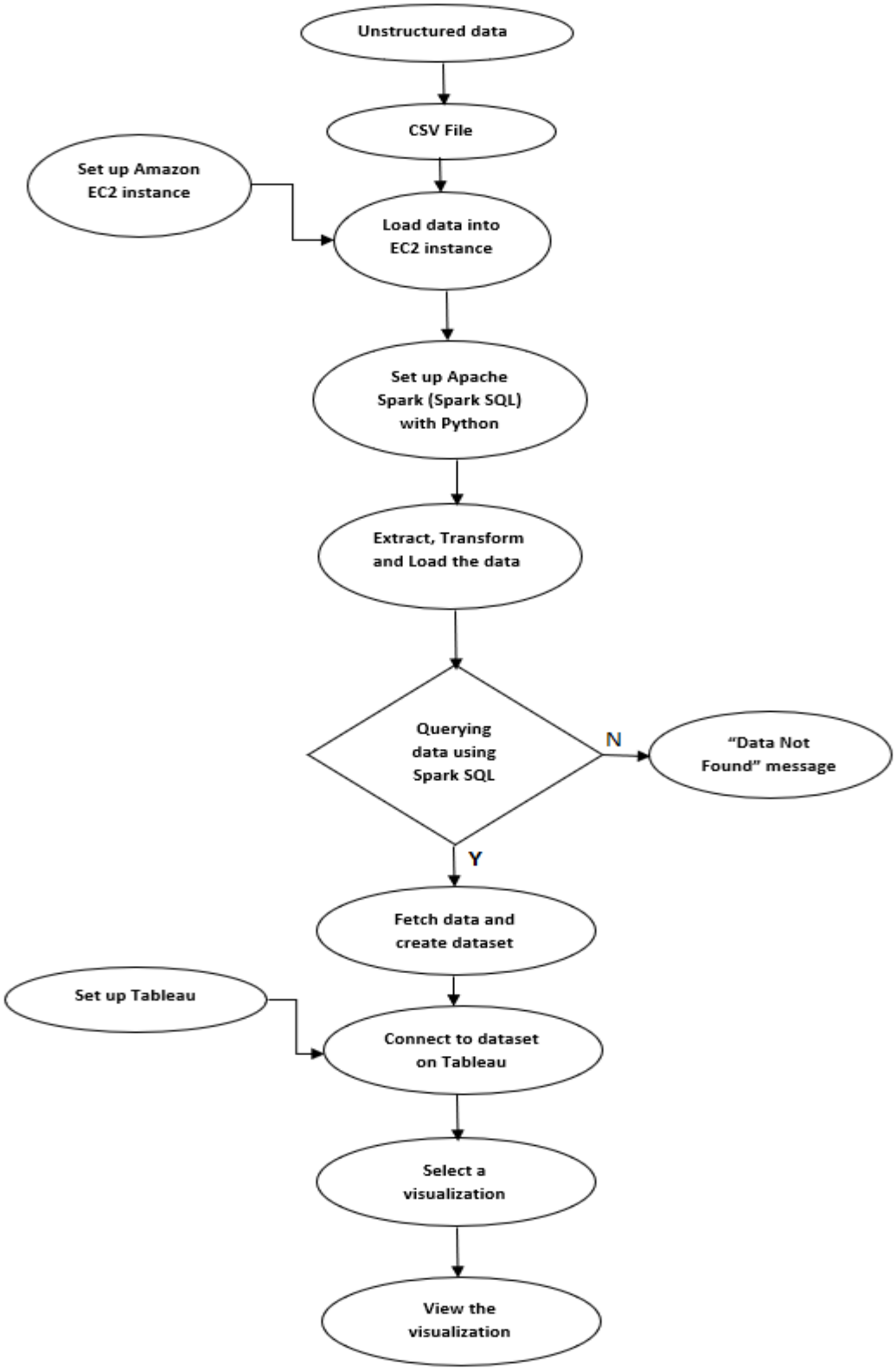
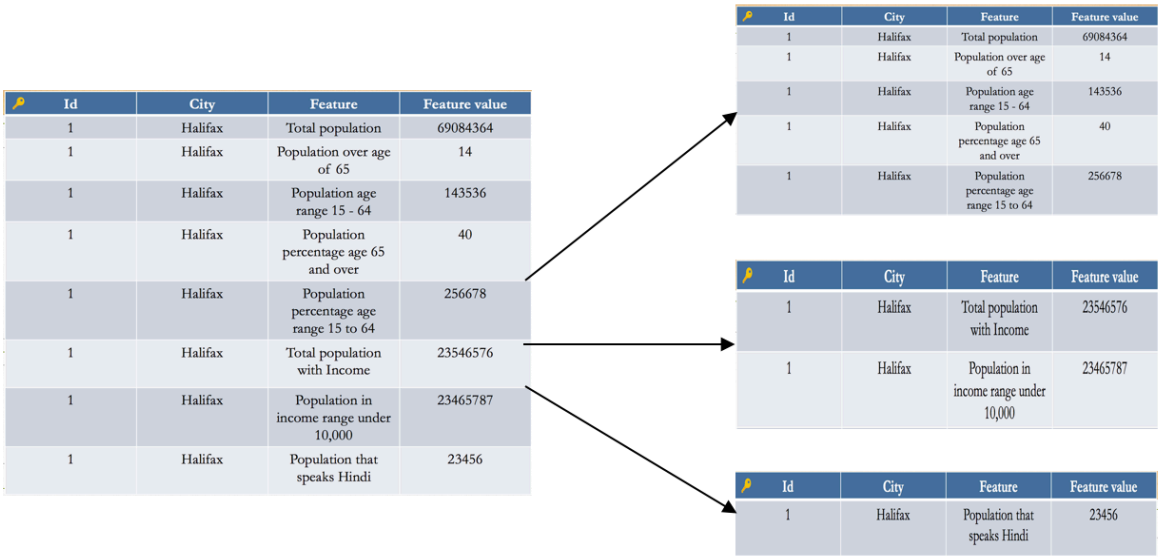


Figure 1: Work Structure Breakdown

In this project we planned to collect, clean and manipulate the CENSUS the following way.

Data Extraction

The CENSUS data seemed to be well structured in the form of tables, rows and columns. However, it is not available in the form of a csv file.



Id	City	Feature	Feature value
1	Halifax	Total population	69084364
1	Halifax	Population over age of 65	14
1	Halifax	Population age range 15 - 64	143536
1	Halifax	Population percentage age 65 and over	40
1	Halifax	Population percentage age range 15 to 64	256678
1	Halifax	Total population with Income	23546576
1	Halifax	Population in income range under 10,000	23465787
1	Halifax	Population that speaks Hindi	23456

Id	City	Feature	Feature value
1	Halifax	Total population	69084364
1	Halifax	Population over age of 65	14
1	Halifax	Population age range 15 - 64	143536
1	Halifax	Population percentage age 65 and over	40
1	Halifax	Population percentage age range 15 to 64	256678

Id	City	Feature	Feature value
1	Halifax	Total population with Income	23546576
1	Halifax	Population in income range under 10,000	23465787

Id	City	Feature	Feature value
1	Halifax	Population that speaks Hindi	23456

Figure 1: Work Structure Breakdown

Data Cleaning

Once the CSV file is ready the data can be cleaned to remove any null values as well as any unnecessary columns. Depending on the quality of the data it may also have to be normalized.

Setting up an EC2 instance and loading the data

This project will run on an EC2 instance which will be setup and configured with an Ubuntu virtual machine and the required software packages such as Python and apache spark will be installed. The CSV files prepared can then be loaded into the EC2 instance.

After ETL processes

Further ETL processes will be carried out on the csv file so that the data can be loaded in apache spark. This will further remove any noise in the data.

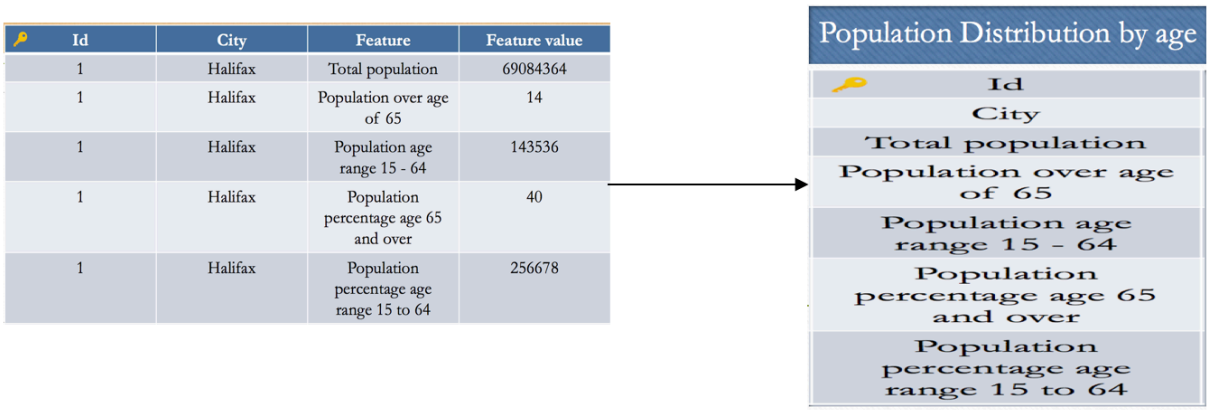


Figure 1: Work Structure Breakdown

Querying the data

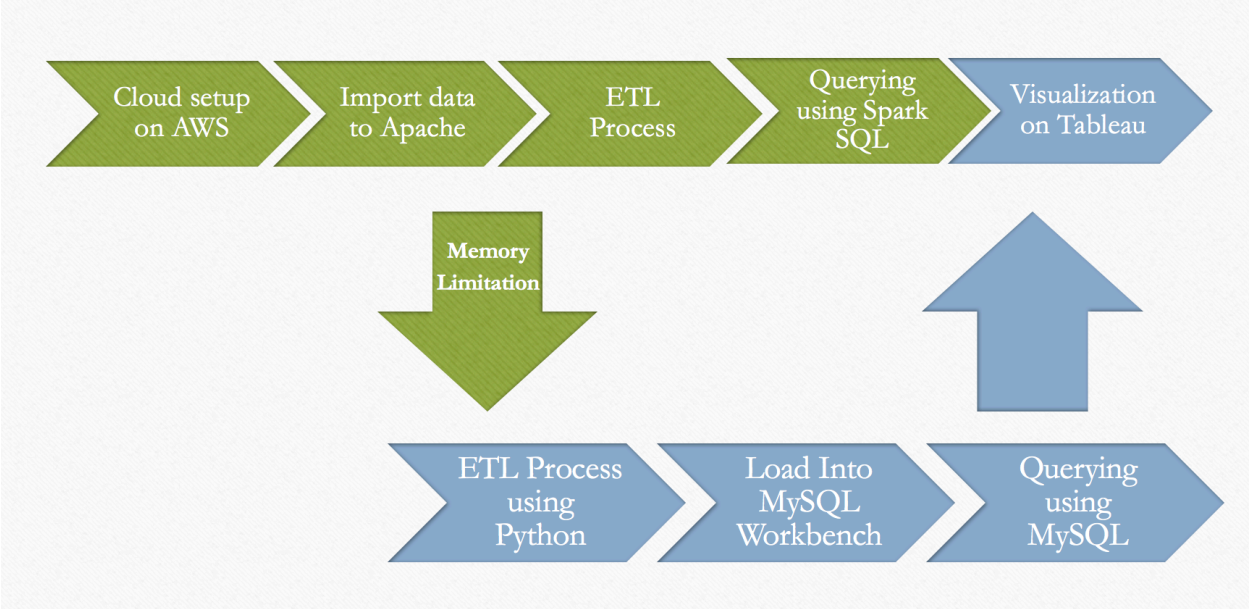
Queries will be generated based on user's requirements. These queries will be executed by Spark SQL to retrieve the required data. If there is no data available that matches the search query an error message will be displayed. on the other hand, if the query is executed successfully the data is fetched and saved.

Connect to tableau

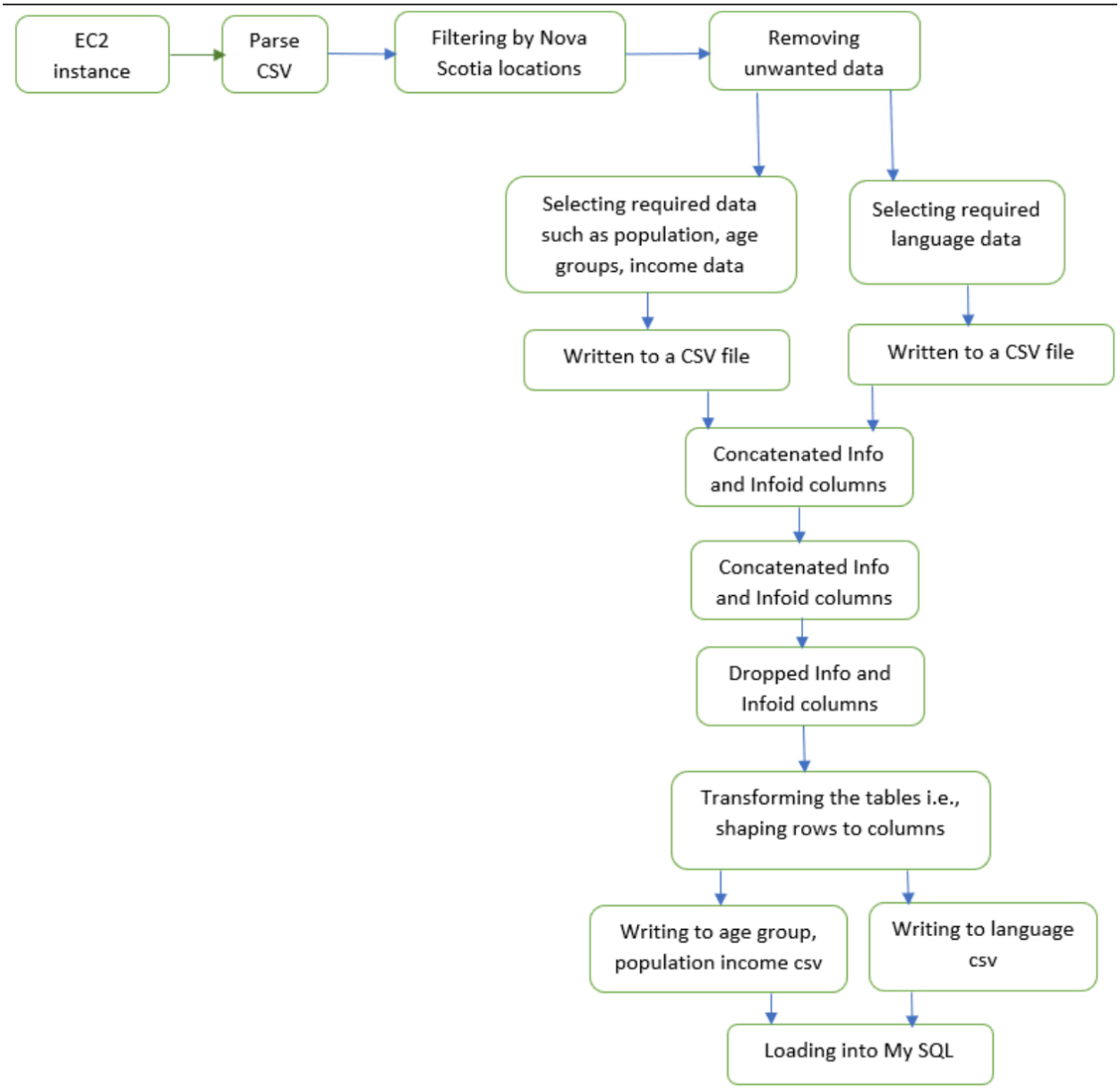
The dataset resulting from the query is connected to tableau and the user is given options to choose the type of visualization he/she would like. Once the type of graph is selected tableau generates and presents the visualization generated.

3.4. Execution of the planned procedures

This project required us to collect, clean and manipulate the CENSUS data to be able to achieve the desired insights. This project will follow the following breakdown structure.



3.5. Implementation Explained in Detail



Data Extraction

The CENSUS data turned out to be more unstructured then we expected. The schema made it very hard to query especially join functions. We had to extract the required data into separate data frames. Where each data frame will have a similar set of data. In the example below we show the complete dataset being split into the population, income and language data frames. We also realized that the data set was too large and decided to only take the records that belonged to Nova Scotia instead of for the entire Canada.

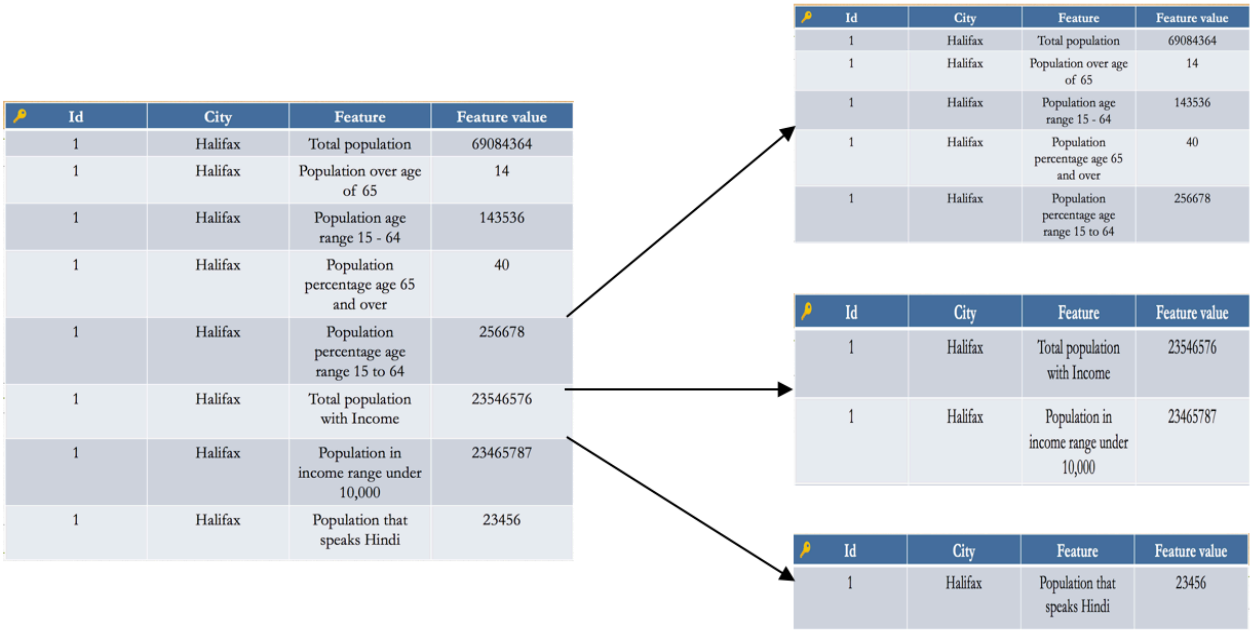
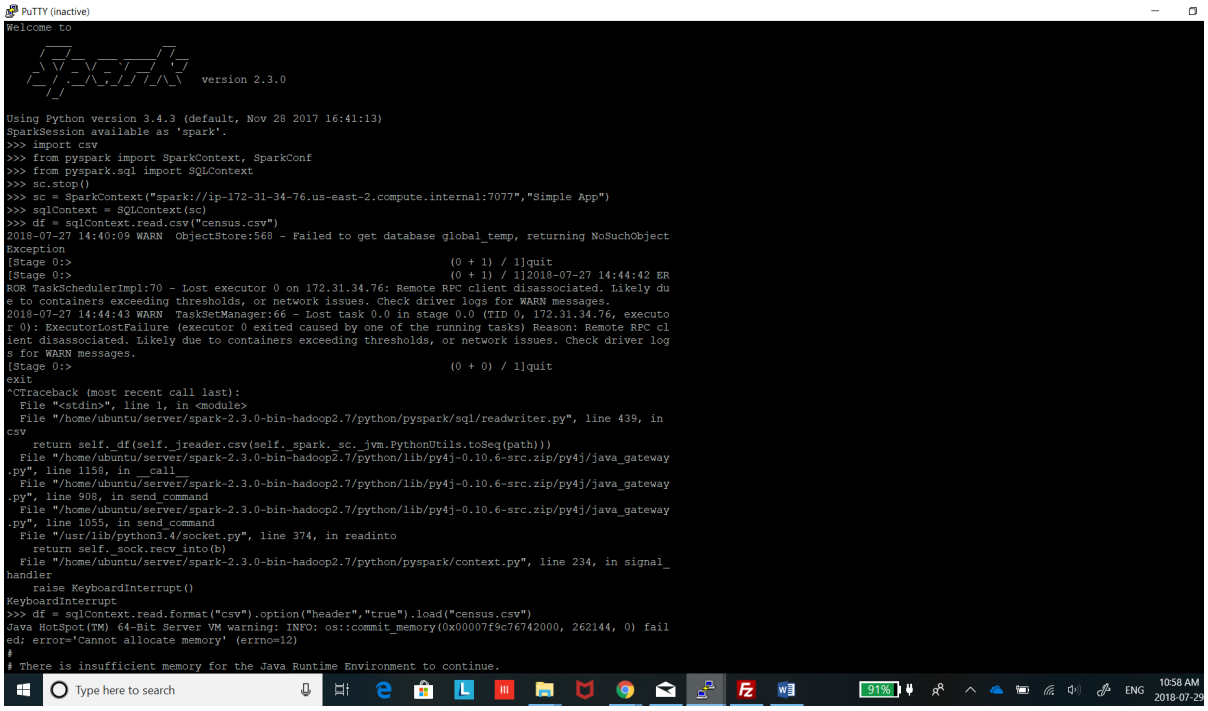
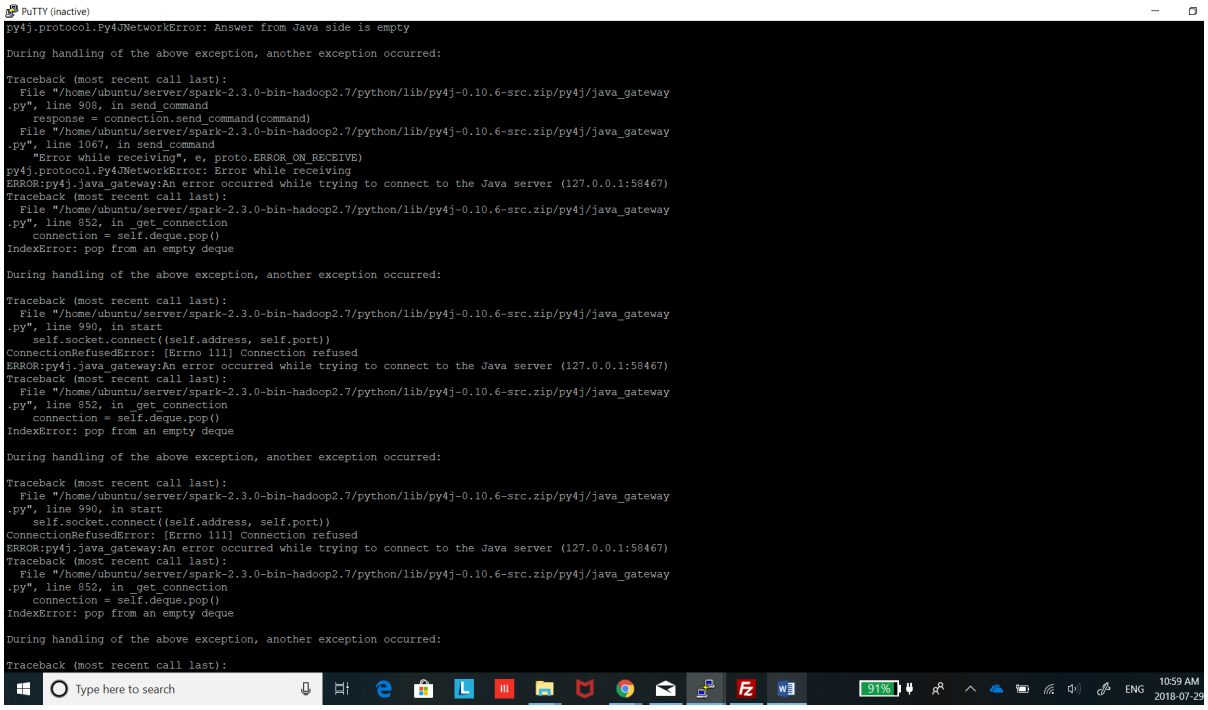


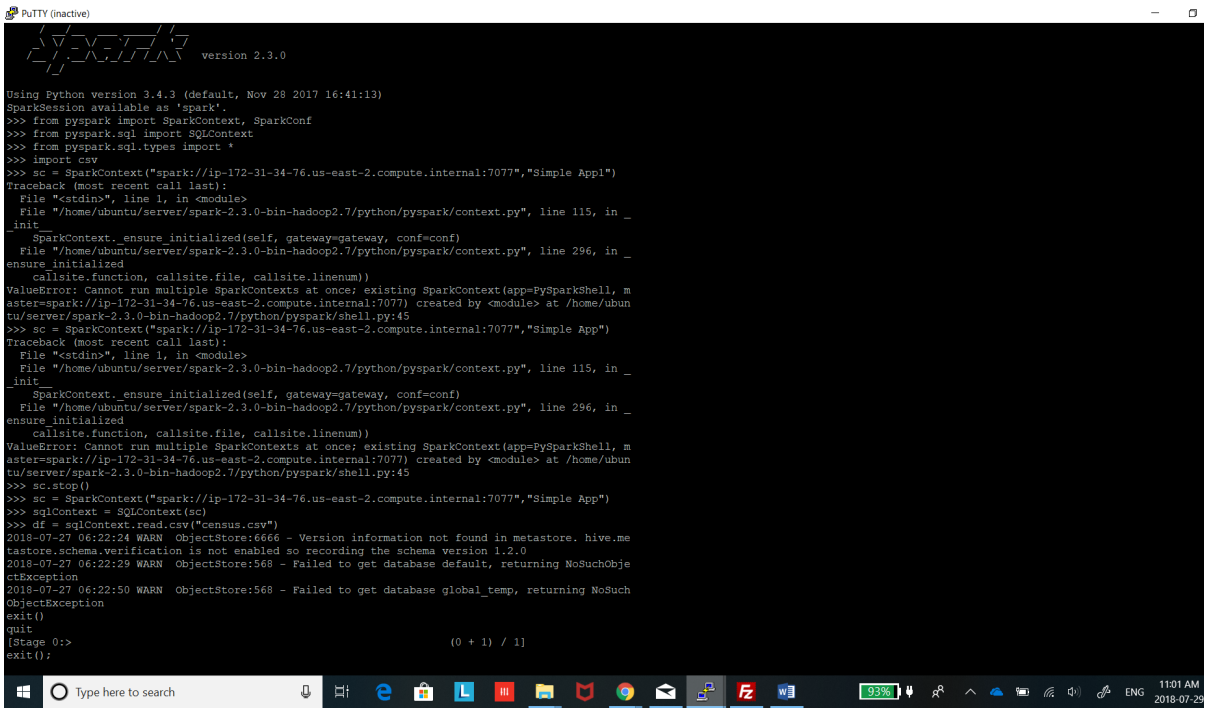
Figure 1: Work Structure Breakdown

A. Data Extraction with Spark

As mentioned in the proposal we planned to use spark for the ETL processes and Spark SQL for querying. However, we ran into memory errors while trying to process our entire dataset. As a result a python scripts where run on the deployed EC2 instance to clean and extract the required data.

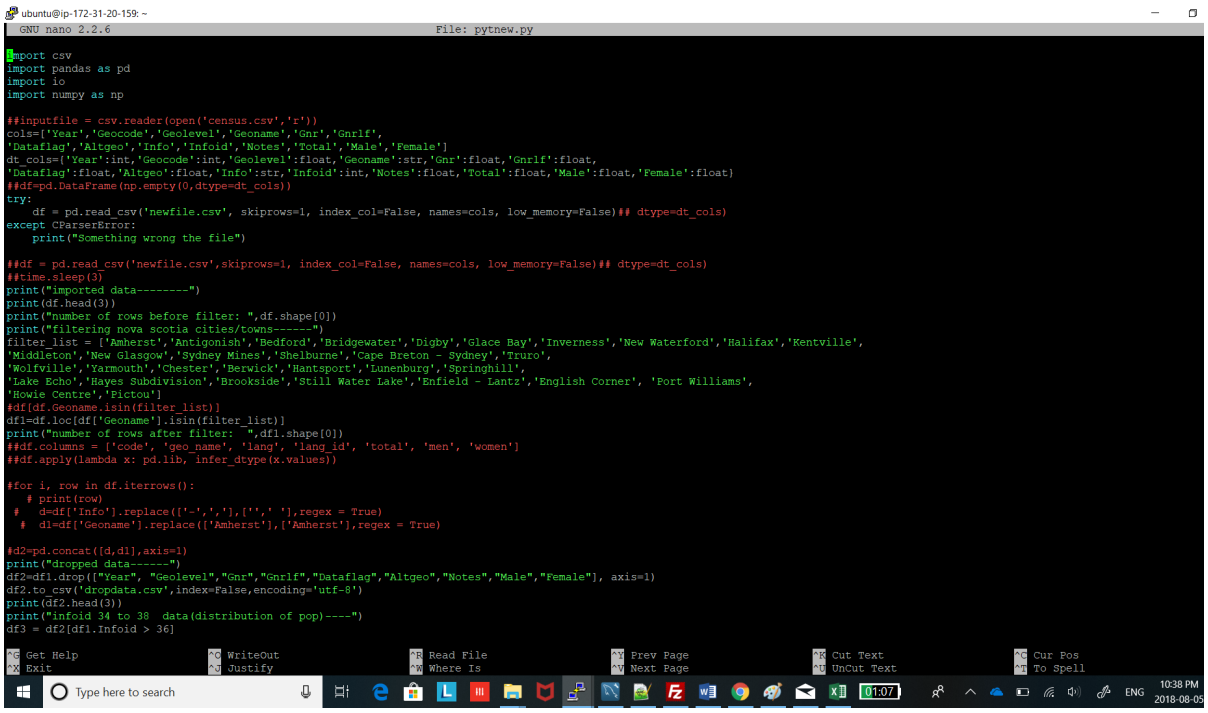

PuTTY (inactive)
Welcome to
version 2.3.0
Using Python version 3.4.3 (default, Nov 28 2017 16:41:13)
SparkSession available as 'spark'.
>>> import csv
>>> from pyspark import SparkContext, SparkConf
>>> from pyspark.sql import SQLContext
>>> sc.stop()
>>> sc = SparkContext("spark://ip-172-31-34-76.us-east-2.compute.internal:7077","Simple App")
>>> sqlContext = SQLContext(sc)
>>> df = sqlContext.read.csv("census.csv")
2018-07-27 14:40:09 WARN ObjectStore:568 - Failed to get database global_temp, returning NoSuchObject
Exception
[Stage 0:> (0 + 1) / 1]quit
[Stage 0:> (0 + 1) / 1]2018-07-27 14:44:42 ER
ROR TaskSchedulerImpl:70 - Lost executor 0 on 172.31.34.76: Remote RPC client disassociated. Likely du
e to containers exceeding thresholds, or network issues. Check driver logs for WARN messages.
2018-07-27 14:44:43 WARN TaskSetManager:66 - Lost task 0.0 in stage 0.0 (TID 0, 172.31.34.76, executo
r 0): ExecutorLostFailure (executor 0 exited caused by one of the running tasks) Reason: Remote RPC cl
ient disassociated. Likely due to containers exceeding thresholds, or network issues. Check driver log
s for WARN messages.
[Stage 0:> (0 + 0) / 1]quit
exit
^CTraceback (most recent call last):
File "<stdin>", line 1, in <module>
File "/home/ubuntu/server/spark-2.3.0-bin-hadoop2.7/python/pyspark/sql/readwriter.py", line 439, in
csv
return self._df(self._jreader.csv(self._spark._sc._jvm.PythonUtils.toSeq(path)))
File "/home/ubuntu/server/spark-2.3.0-bin-hadoop2.7/python/lib/py4j-0.10.6-src.zip/py4j/java_gateway
.py", line 1158, in __call__
File "/home/ubuntu/server/spark-2.3.0-bin-hadoop2.7/python/lib/py4j-0.10.6-src.zip/py4j/java_gateway
.py", line 908, in send_command
File "/home/ubuntu/server/spark-2.3.0-bin-hadoop2.7/python/lib/py4j-0.10.6-src.zip/py4j/java_gateway
.py", line 1055, in send_command
File "/usr/lib/python3.4/socket.py", line 374, in readinto
return self._sock.recv_into(b)
File "/home/ubuntu/server/spark-2.3.0-bin-hadoop2.7/python/pyspark/context.py", line 234, in signal_
handler
raise KeyboardInterrupt()
KeyboardInterrupt
>>> df = sqlContext.read.format("csv").option("header","true").load("census.csv")
Java HotSpot(TM) 64-Bit Server VM warning: INFO: os::commit_memory(0x00007f9c76742000, 262144, 0) fail
ed; error='Cannot allocate memory' (errno=12)
#
There is insufficient memory for the Java Runtime Environment to continue.
Type here to search
91%
ENG
10:58 AM
2018-07-29

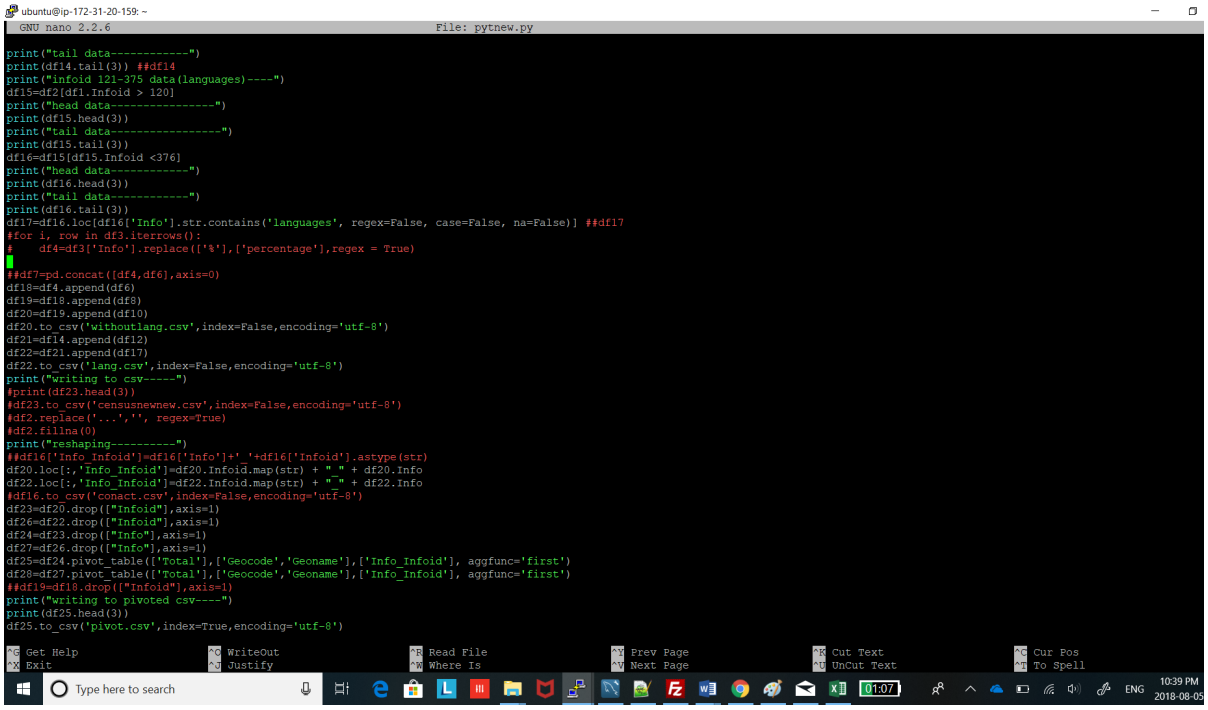

PuTTY (inactive)
py4j.protocol.Py4JNetworkError: Answer from Java side is empty
During handling of the above exception, another exception occurred:
Traceback (most recent call last):
File "/home/ubuntu/server/spark-2.3.0-bin-hadoop2.7/python/lib/py4j-0.10.6-src.zip/py4j/java_gateway
.py", line 908, in send_command
response = connection.send_command(command)
File "/home/ubuntu/server/spark-2.3.0-bin-hadoop2.7/python/lib/py4j-0.10.6-src.zip/py4j/java_gateway
.py", line 1067, in send_command
"Error while receiving", e, proto.ERROR_ON_RECEIVE)
py4j.protocol.Py4JNetworkError: Error while receiving
ERROR:py4j.java_gateway:An error occurred while trying to connect to the Java server (127.0.0.1:58467)
Traceback (most recent call last):
File "/home/ubuntu/server/spark-2.3.0-bin-hadoop2.7/python/lib/py4j-0.10.6-src.zip/py4j/java_gateway
.py", line 852, in _get_connection
connection = self.deque.pop()
IndexError: pop from an empty deque
During handling of the above exception, another exception occurred:
Traceback (most recent call last):
File "/home/ubuntu/server/spark-2.3.0-bin-hadoop2.7/python/lib/py4j-0.10.6-src.zip/py4j/java_gateway
.py", line 990, in start
self.socket.connect((self.address, self.port))
ConnectionRefusedError: [Errno 111] Connection refused
ERROR:py4j.java_gateway:An error occurred while trying to connect to the Java server (127.0.0.1:58467)
Traceback (most recent call last):
File "/home/ubuntu/server/spark-2.3.0-bin-hadoop2.7/python/lib/py4j-0.10.6-src.zip/py4j/java_gateway
.py", line 852, in _get_connection
connection = self.deque.pop()
IndexError: pop from an empty deque
During handling of the above exception, another exception occurred:
Traceback (most recent call last):
File "/home/ubuntu/server/spark-2.3.0-bin-hadoop2.7/python/lib/py4j-0.10.6-src.zip/py4j/java_gateway
.py", line 990, in start
self.socket.connect((self.address, self.port))
ConnectionRefusedError: [Errno 111] Connection refused
ERROR:py4j.java_gateway:An error occurred while trying to connect to the Java server (127.0.0.1:58467)
Traceback (most recent call last):
File "/home/ubuntu/server/spark-2.3.0-bin-hadoop2.7/python/lib/py4j-0.10.6-src.zip/py4j/java_gateway
.py", line 852, in _get_connection
connection = self.deque.pop()
IndexError: pop from an empty deque
During handling of the above exception, another exception occurred:
Traceback (most recent call last):
Type here to search
91%
ENG
10:59 AM
2018-07-29



B. Data Extraction with Python

When Spark started to give memory errors we switched to using python scripts to extract the features we needed and remove any columns and rows with null values. The features we choose to extract where based on the visualizations we wanted to perform to determine the best city to launch our business idea “The Senior Café”. We decided to extract features such as the percentage of the senior citizen population in each city.





Data Cleaning

Once the CSV files were ready the data was cleaned to remove any null values as well as any unnecessary columns. This dataset had a lot of null values that were removed.

Setting up an EC2 instance and loading the data

The EC2 instance was setup opting for the highest memory possible with the free student account. The required programing languages and libraries where installed. Some of the packages installed include Python, pandas library, pip library and MySQL. The dataset was then loaded into the EC2 instance. Python scripts were then written and applied to the loaded dataset to generate extracted csv files.

ETL processes

After Extracting the required data each of the CSV files were transformed using python scripts to change the features recorded in rows to columns. The change in schema is shown below.

Id	City	Feature	Feature value
1	Halifax	Total population	69084364
1	Halifax	Population over age of 65	14
1	Halifax	Population age range 15 - 64	143536
1	Halifax	Population percentage age 65 and over	40
1	Halifax	Population percentage age range 15 to 64	256678

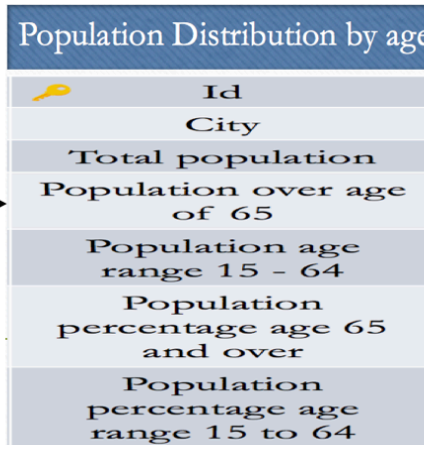
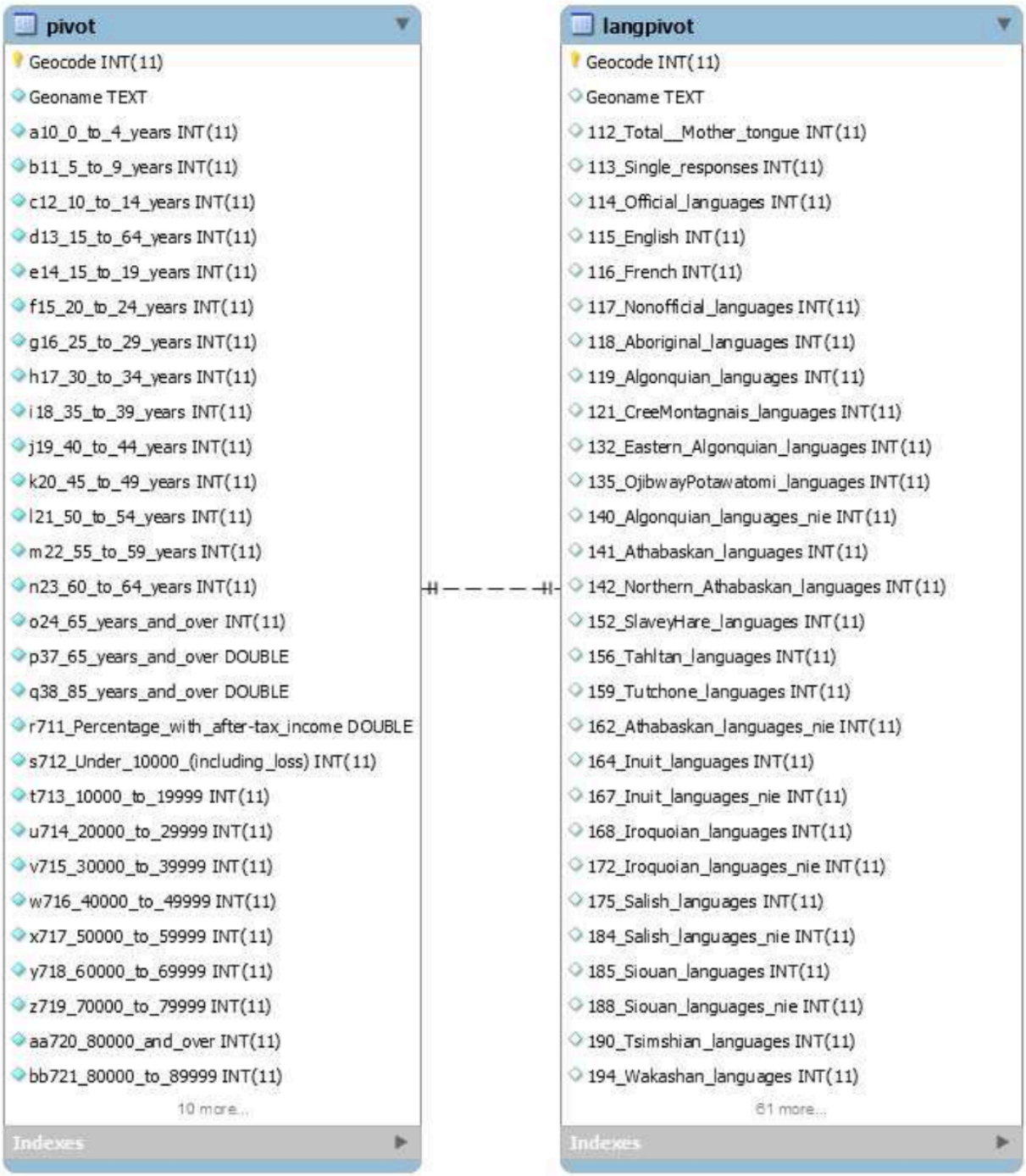


Figure 1: Work Structure Breakdown

Querying the data

The dataset has been transformed from an unstructured dataset to a structured schema. This schema was then loaded into MySQL workbench to allow the data to be queried. The queries were chosen based on our business idea. For the future of this application the queries would be generated dynamically to generate current market visualization for any type of business desired. The query takes a CSV as an input and gives a much smaller more clean csv as a result that is then exported to tableau for visualization. The database schema after ETL processes and loading into MySQL workbench is shown below.

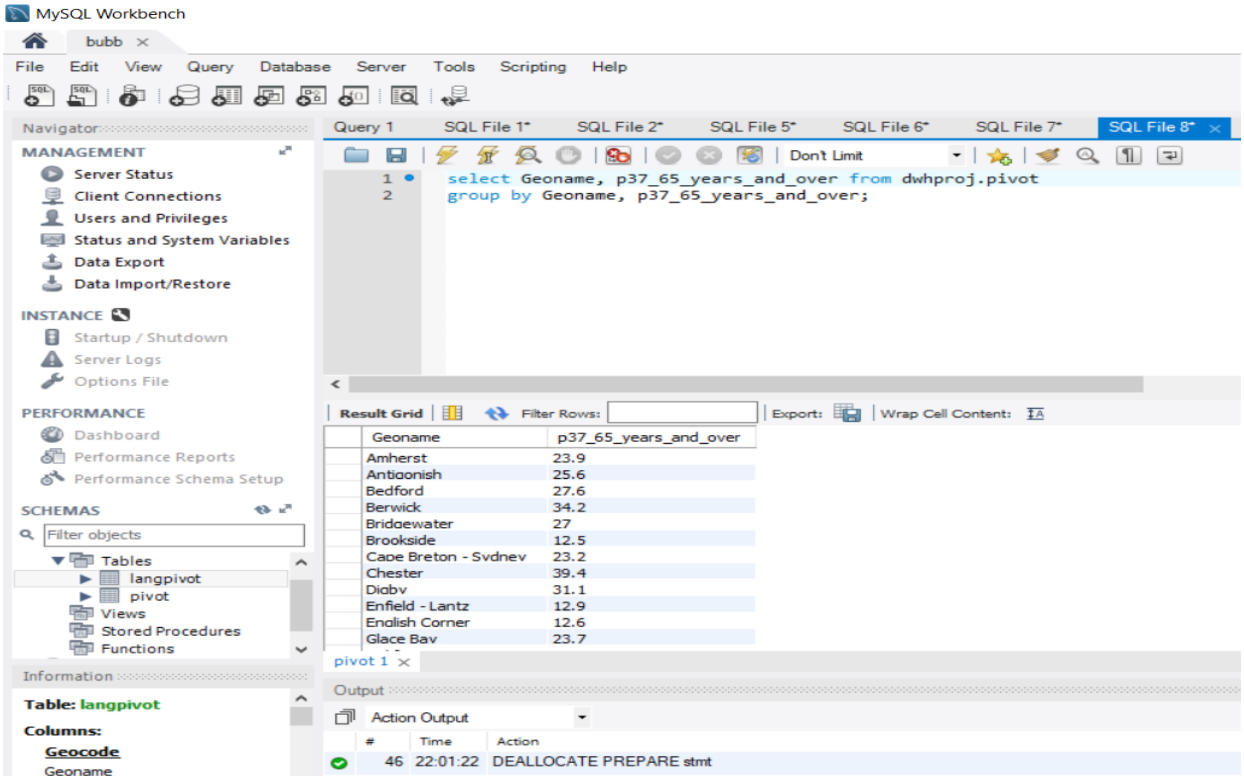


Connect to tableau

The dataset resulting from the query is connected to tableau and visualizations are generated by specifying the parameters used for the visualization.

3.6. Visualization

Query 1: List the population percentage of senior citizens for each city in Nova Scotia



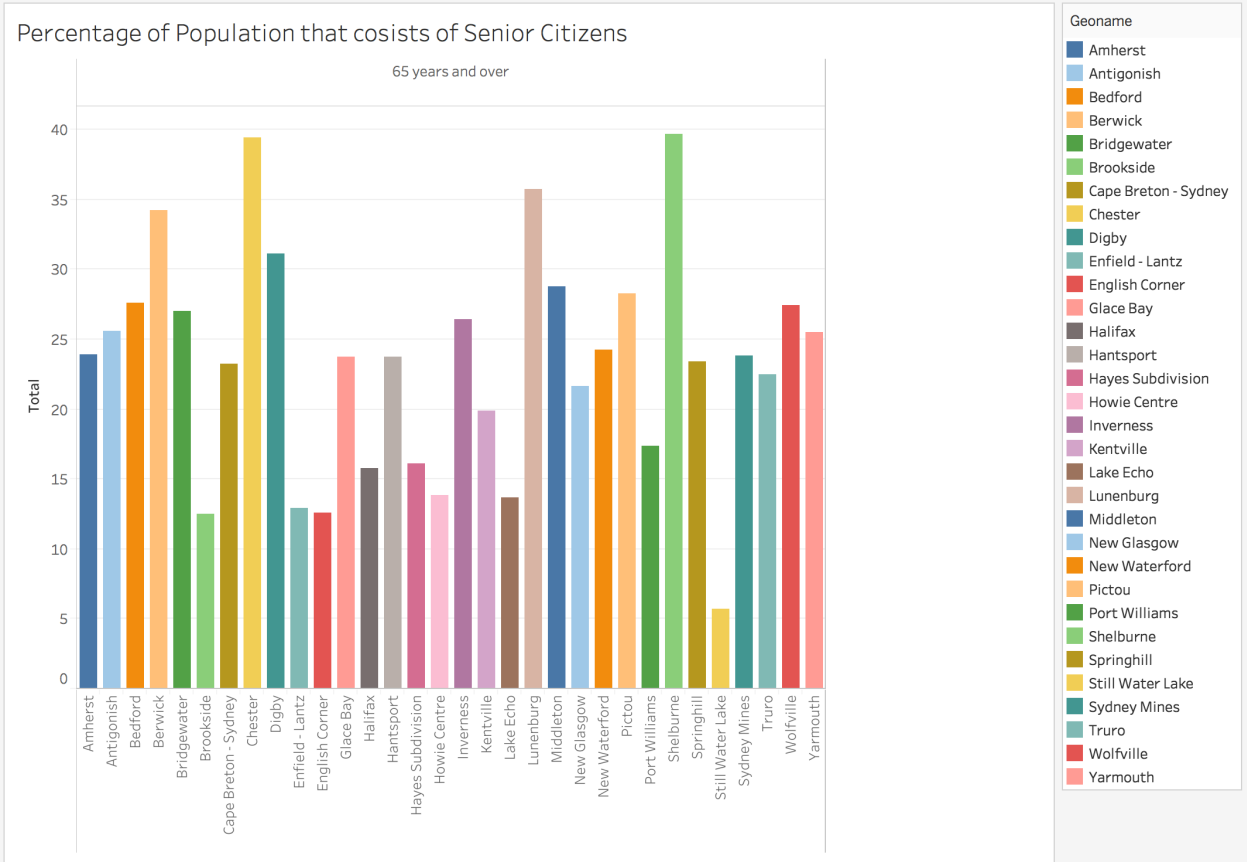


Figure 1: Work Structure Breakdown

According to this visualization Shelburne has the highest percentage of senior citizens. This information would be helpful to someone looking to start a business as they could decide to start in Shelburne based on the given information about the current market scenario. This visualization could be used in a business pitch to support the business owner's decision while presenting to a potential investor.

However, this visualization below shows that even though Shelburne has a higher percentage of the elderly population, The total population of Halifax is higher and hence Halifax has a larger population of senior citizens as compared to Shelburne.

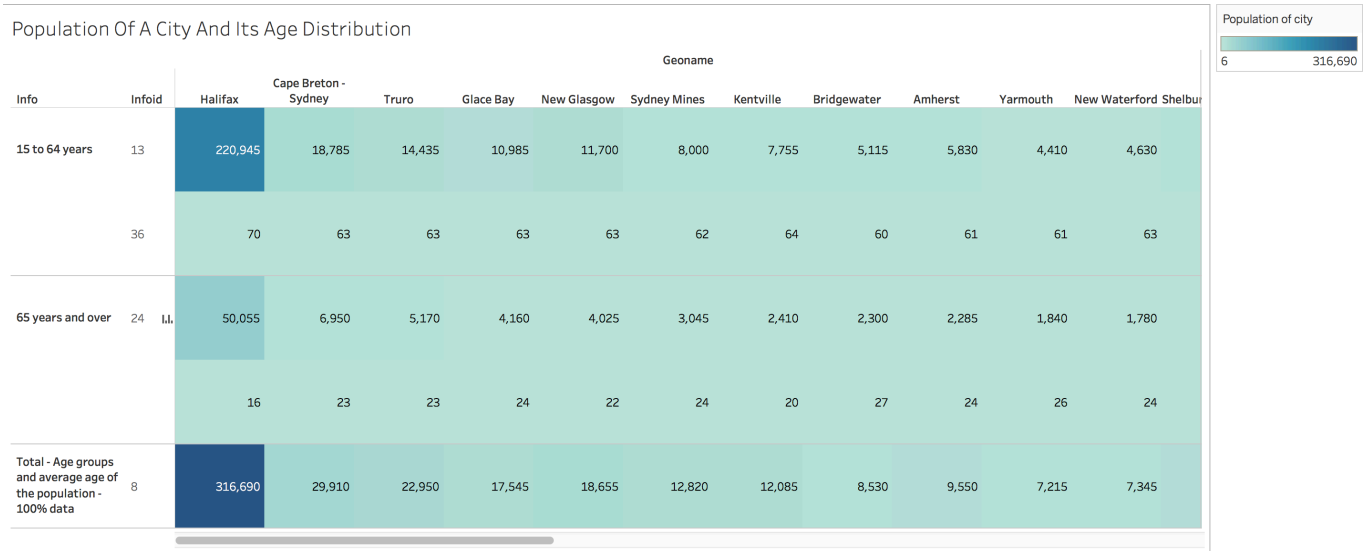
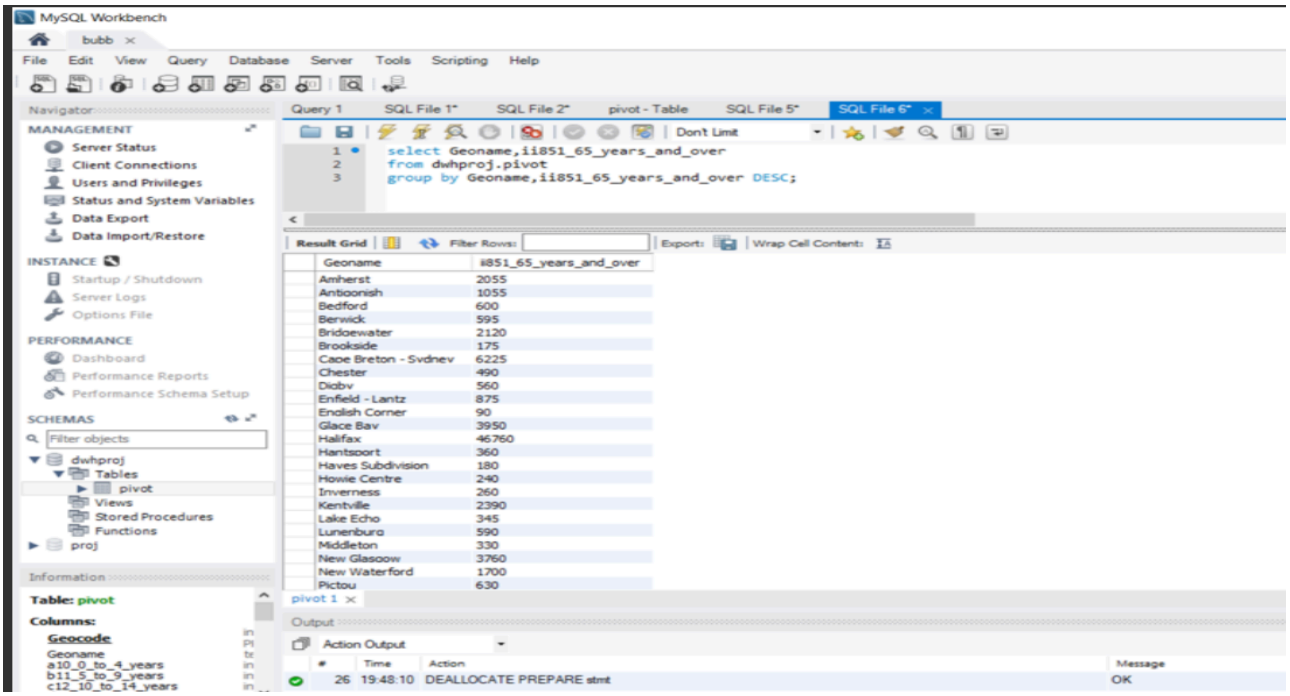


Figure 1: Work Structure Breakdown

Query 2:

The low income of the seniors helps us to decide on where to start a business for the seniors since they should not feel the business is expensive from their perspective. Below, the graph depicts the low income of the seniors versus the location. We can say that Halifax has the highest number of low income seniors who are 65 years and above.



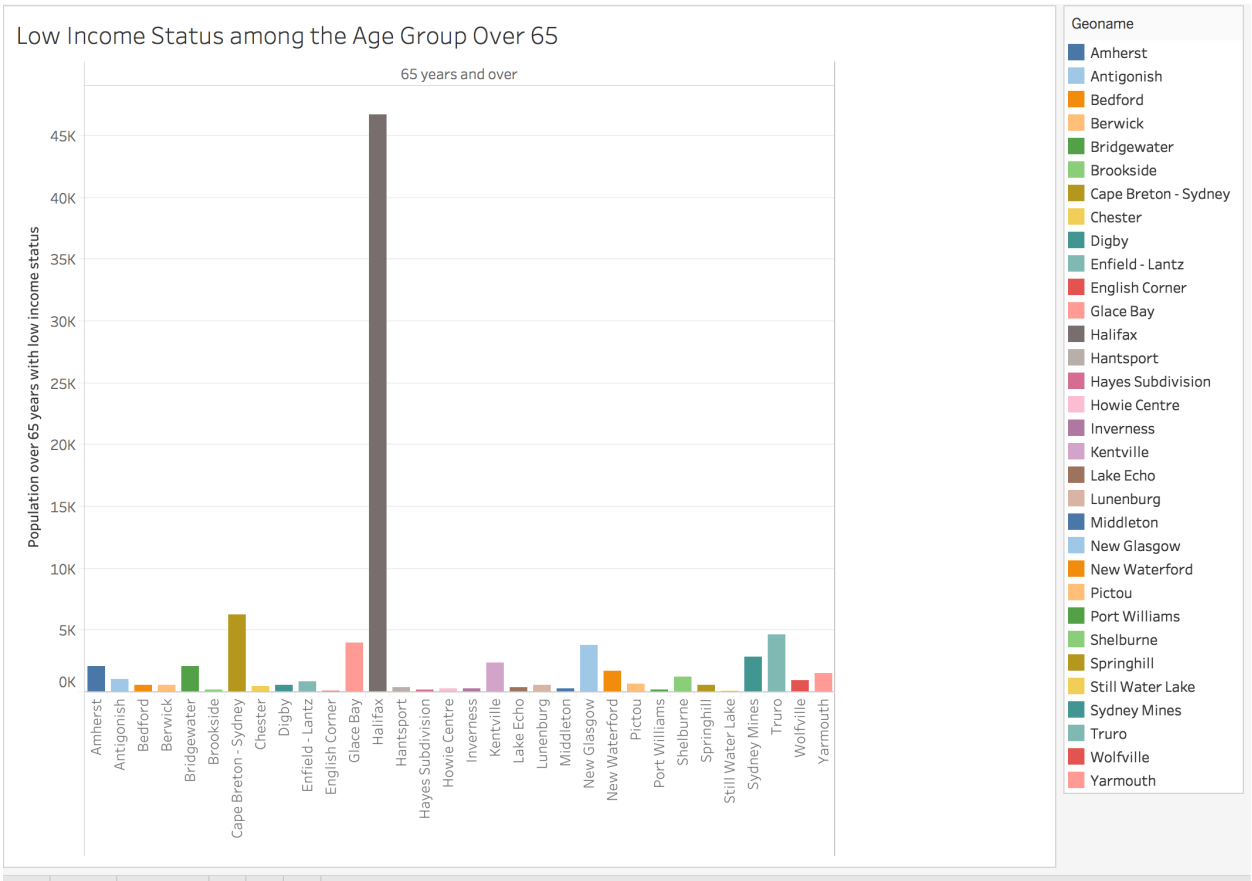


Figure 1: Work Structure Breakdown

Languages:
Languages play an important role in our lives. Though English and French are the most spoken by the population all over Canada. There are people who still converse in mother tongue. Conversing in mother tongue makes one feel comfortable. Starting a business for seniors and letting them talk in the language they want to improves the business. They might find more people who speaks their language on daily basis. The business may also run on a specific theme where a specific ethnicity has been followed by more people in a location compared to other locations.

Below are the graphs that conveys most number of languages being spoken by Halifax population. The first graph shows the non-aboriginal languages being spoken in the Halifax where the value is represented by its size. Indo-European is spoken by most and Afro-asiatic languages come in second. If the business is targeted towards these specific languages, the number of customers who speak the same language may like the idea of meeting more people in the same location. The second graph depicts the aboriginal languages such as Algonquian languages, Easter Algonquian languages and Mi'kmaq languages

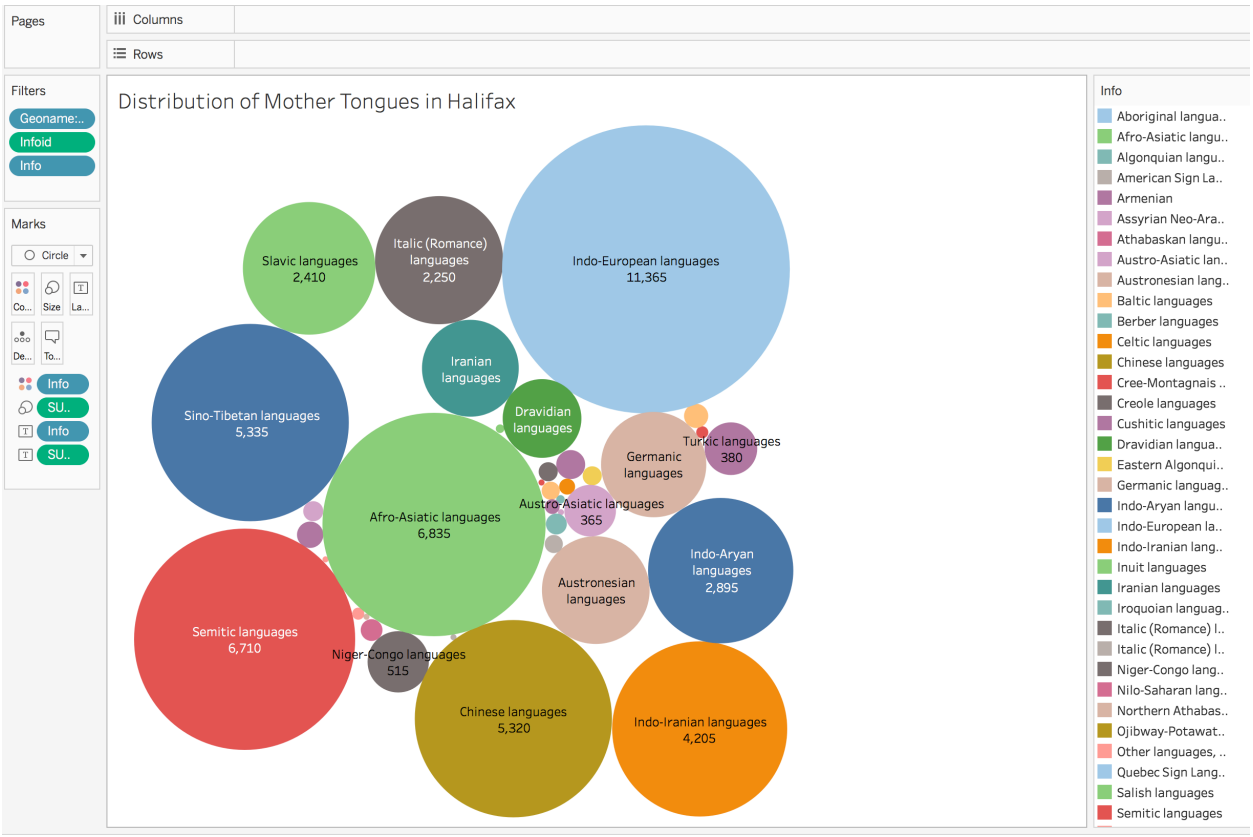


Figure 1: Work Structure Breakdown

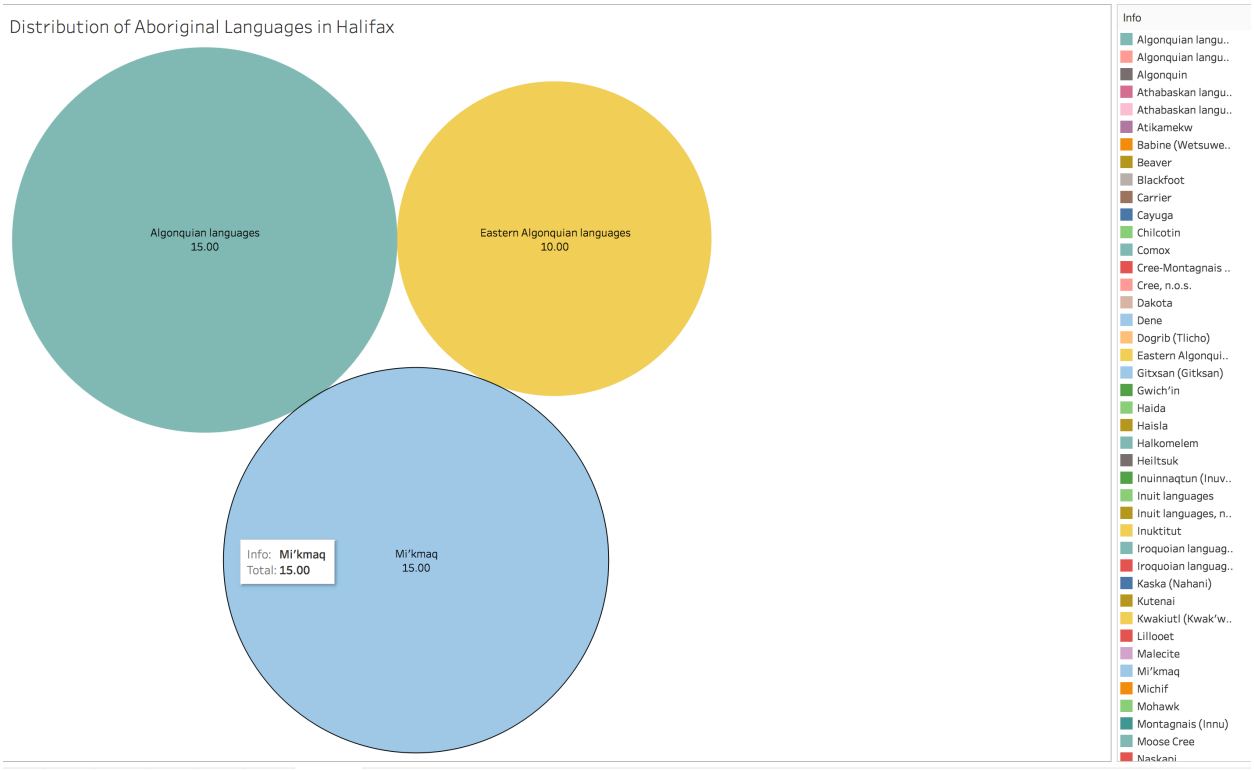
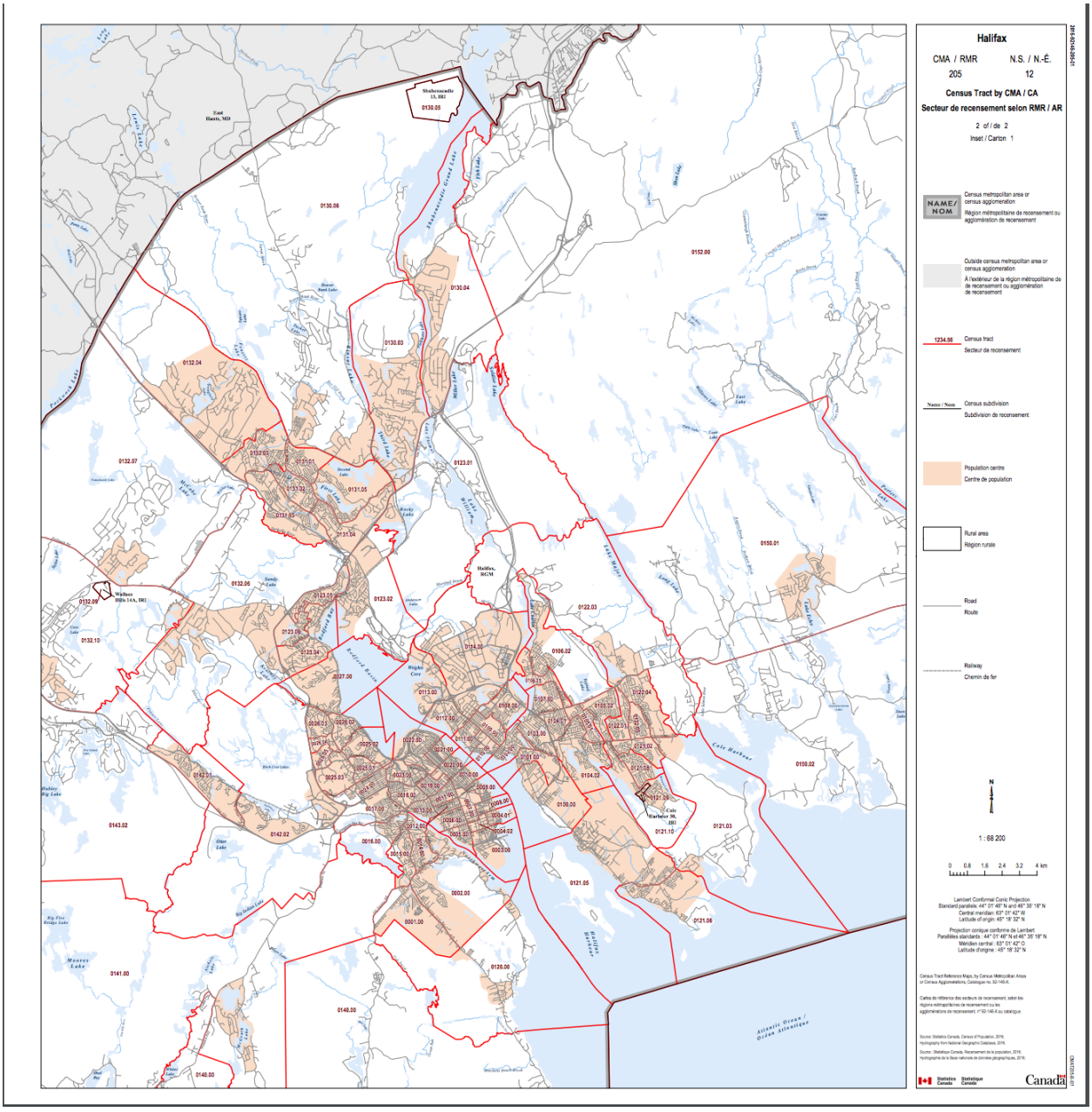


Figure 1: Work Structure Breakdown

As we can see from the above visualizations that Halifax is the most optimal location for a business in the hole of Nova Scotia. We decided to analyze the location within Halifax in more detail. We extracted another dataset specific to regions within Halifax. The below map shows the area division and the associated regional codes below.



The visualization below shows the number of refugees that have move into specific regions of Halifax over the time period between 1980 – 2016.

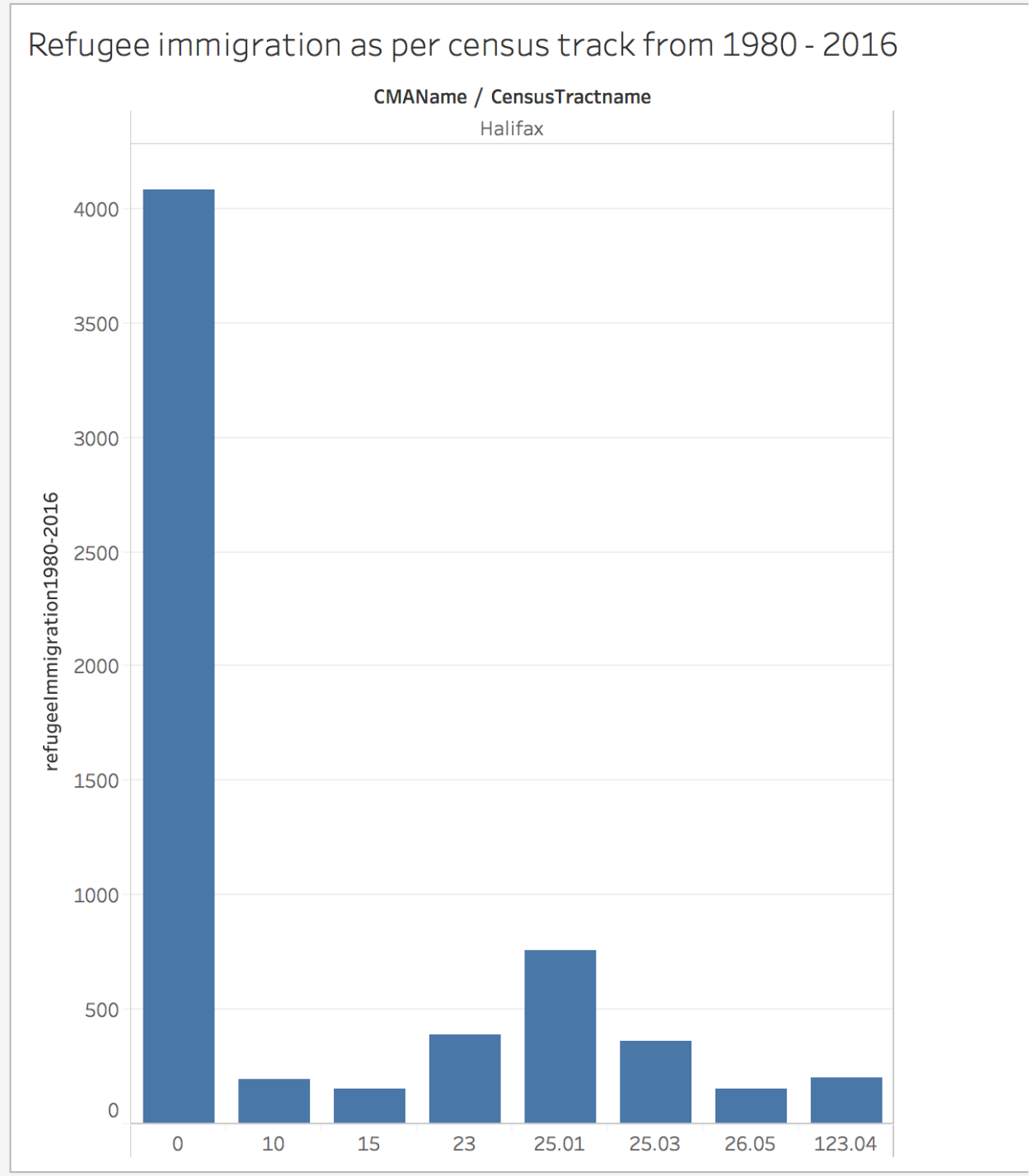


Figure 1: Work Structure Breakdown

These track numbers correspond to regions with in Halifax. The zoomed map below shows one of the regions displayed in the bar graph above.

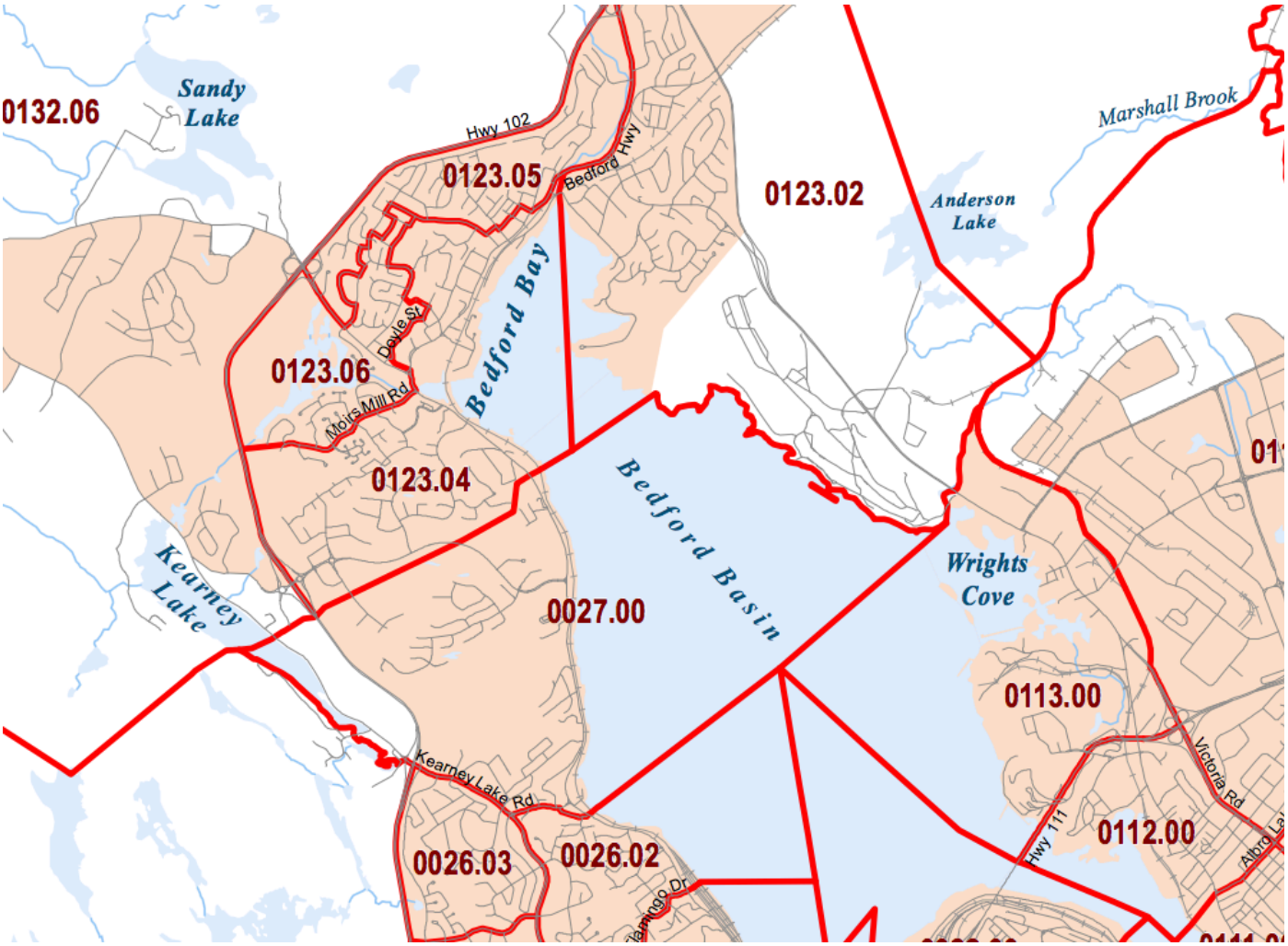
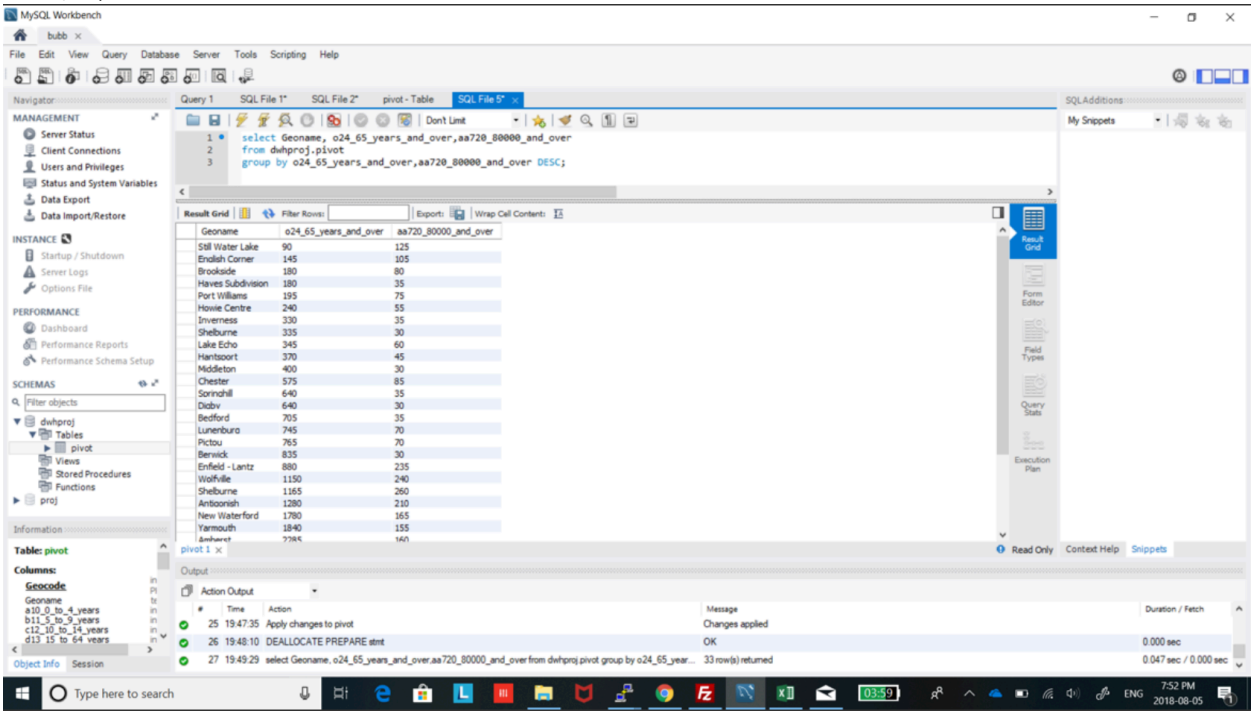


Figure 1: Work Structure Breakdown

Query 3: Find the city with the highest population of senior citizens and a population of people with income higher than $80,000



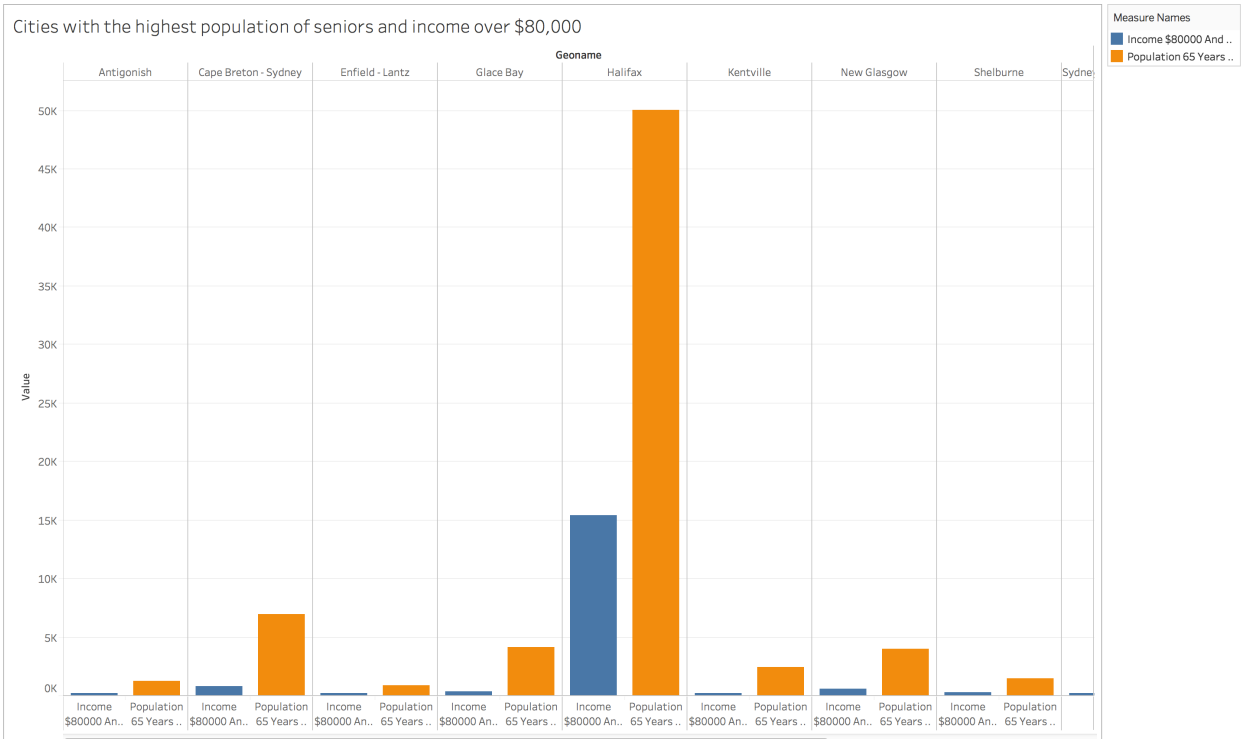


Figure 1: Work Structure Breakdown

This visualization is important to understand which city has the highest growth potential for the senior café. This shows that Halifax has the highest population of both seniors and high income population.

4. Role based work and distribution

Name	Banner ID	Role
Bhavya Chandrappa	B00781097	Data Scientist
Sogra Bilal Memon	B00786252	Data Engineer

Table 1: Team member roles

The roles taken up by Bhavya and Sogra are Data Scientist and Data Engineer respectively. The data was extracted, transformed and loaded to the My SQL Workbench by the Data Scientist. The Data Scietist then queried the results of the resulted dataset. The Data Engineer then took the project forward by analyzing the queries and building graphs to help the customers to look into their searches in a creative view. The Data Engineer also tested the whole process of the project for the user to provide a good user experience.

5. Milestones/Sprints

	Planned	Executed	Tasks
Milestone 1	July 4 - July 7	July 4 July 10	Data gathering and Loading data into EC2 instance
Milestone 2	July 8 - July 14	July 11 – July 20	ETL Process
Milestone 3	July 15 - July 21	July 21 – July 26	Spark SQL query

Milestone 4	July 22 - July 28	July 27 – Aug 1	Tableau Visualization

Table 2: Milestones and sprints

6. Limitations of Current work

The limitation of current work includes the fact that the data source is static. Furthermore, the CENSUS data is collected and made available every 5 years this makes it hard to make the dataset dynamic. Another limitation is the memory space. This issue arises because the dataset records have been recorded for the entire Canada. For this pro

7. Future Work

For the future work of this project we would like to build a front end where users can select the columns they would like to visualize based on the requirement of their business and the user interface would generate the resulting query and a visualization automatically. In this way a normal user can have access to visualizations that can be shown as part of business pitches and decisions being made without having to know the details about ETL processes or about Data Science.

8. Github link

The github link: https://github.com/SograMemon/DataWarehouse-A2.git

9. Critical Review

The project was designed to help newcomers to find a location to start up their business. The project is quite straightforward as the data is in the same manner. We couldn't experiment with the data. If only there were more required data in the census file, the data generated could have been more creative. We made lots of trials and errors on how to process the data, in the end we were convenient to visualize our results as it made us to improvise for each visualization onwards. We wanted to display the geolocations on the map for the final result as the Geolocations data weren't available in the census file, we stuck to what we have.

10. References

[1] "Tableau Spark SQL Connector Demo - YouTube." [Online]. Available: https://www.youtube.com/watch?v=OKcIf6UdK7c. [Accessed: 03-Jul-2018].

[2] "Spark SQL & DataFrames | Apache Spark." [Online]. Available: https://spark.apache.org/sql/. [Accessed: 03-Jul-2018].

[3] "90% Of Startups Fail: Here's What You Need To Know About The 10%." [Online]. Available: https://www.forbes.com/sites/neilpatel/2015/01/16/90-of-startups-will-fail-heres-what-you-need-to-know-about-the-10/#5b657bfa6679. [Accessed: 03-Jul-2018].

[4] "Tableau Help | Tableau Software." [Online]. Available: https://www.tableau.com/support/help. [Accessed: 03-Jul-2018].

[5] "Answer questions as fast as you can think of them with Tableau | Tableau Software." [Online]. Available: https://www.tableau.com/trial/tableau-software?utm_campaign_id=2017049&utm_campaign=Prospecting-CORE-ALL-ALL-ALL-ALL&utm_medium=Paid+Search&utm_source=Google+Search&utm_language=EN&utm_country=USCA&kw=tableau&adgroup=CTX-Brand-Core-E&adused=221441171506&m. [Accessed: 03-Jul-2018].

[6] "Date Functions." [Online]. Available: https://onlinehelp.tableau.com/current/pro/desktop/en-us/functions_functions_date.html. [Accessed: 03-Jul-2018].

[7] "Descriptive Analytics with Tableau - Creating Correlation Matrix and Boxplot Chart - YouTube." [Online]. Available: https://www.youtube.com/watch?v=eQnH_IqUfEA. [Accessed: 03-Jul-2018].

[8] Shanelynn, "The Pandas DataFrame – loading, editing, and viewing data in Python," *Shane Lynn*, 18-Dec-2017. [Online]. Available: https://www.shanelynn.ie/using-pandas-dataframe-creating-editing-viewing-data-in-python/. [Accessed: 06-Aug-2018].

"Fixing Column Names in pandas," *The New, Interactive Singles Map.*[Online]. Available: http://jonathansoma.com/lede/foundations/classes/pandas columns and functions/fixing-column-names-in-pandas/. [Accessed: 06-Aug-2018].

“Multiple Criteria Filtering,” *ritchieng.github.io*. [Online]. Available: https://www.ritchieng.com/pandas-multi-criteria-filtering/. [Accessed: 06-Aug-2018].

[9] “Pandas read_csv: AttributeError: 'NoneType' object has no attribute 'dtype',” *Stack Overflow*. [Online]. Available: https://stackoverflow.com/questions/30383478/pandas-read-csv-attributeerror-nonetype-object-has-no-attribute-dtype. [Accessed: 06-Aug-2018].

[10] “Frequently Asked Questions (FAQ)” *pandas: powerful Python data analysis toolkit - pandas 0.22.0 documentation*. [Online]. Available: http://pandas.pydata.org/pandas-docs/stable/gotchas.html#support-for-integer-na. [Accessed: 06-Aug-2018].

[11] “Convert Pandas column containing NaNs to dtype `int`,” *Stack Overflow*. [Online]. Available: https://stackoverflow.com/questions/21287624/convert-pandas-column-containing-nans-to-dtype-int. [Accessed: 06-Aug-2018].

[12] “Select rows from a DataFrame based on values in a column in pandas,” *Stack Overflow*. [Online]. Available: https://stackoverflow.com/questions/17071871/select-rows-from-a-dataframe-based-on-values-in-a-column-in-pandas. [Accessed: 06-Aug-2018].

[13] “pandas dataframe - select by partial string,” *Stack Overflow*. [Online]. Available: https://stackoverflow.com/questions/11350770/pandas-dataframe-select-by-partial-string. [Accessed: 06-Aug-2018].

[14]“Combine two columns of text in dataframe in pandas/python,” *Stack Overflow*. [Online]. Available: https://stackoverflow.com/questions/19377969/combine-two-columns-of-text-in-dataframe-in-pandas-python. [Accessed: 06-Aug-2018].

[15] “pivot_table No numeric types to aggregate,” *Stack Overflow*. [Online]. Available: https://stackoverflow.com/questions/39229005/pivot-table-no-numeric-types-to-aggregate. [Accessed: 06-Aug-2018].