*COMP4710 Data Mining Project*

Connor Bean

Claire Chen

Eddy He

James Froese

# Introduction

The novel corona virus of 2019 dubbed Covid-19 has profoundly altered life on earth in 2020. In addition to the human sickness and death associated with a pandemic, widespread economic shutdowns have been implemented across the globe in an effort to limit the spread of the virus. As students, our lives have been directly impacted as international travel has been vastly reduced, classes moved to online delivery, and personal connections limited. This unique situation provides us with an opportunity to study a problem that is hugely relevant to our day-to-day lives, and where results may become immediately useful in the fight to control this disease.

As a result of these unique circumstances, we have decided to mine Canadian Covid-19 patient data in an effort to determine which demographics are most at risk of hospitalization and death and if demographics most at risk of infection change depending on geographic location. Utilizing data provided by federal and provincial government we intend to use a variation of the FP-growth algorithm to determine the most at-risk demographics around the country. Patient data generally includes age, sex, health status, and exposure type which we will mine to determine frequent attributes and hopefully draw some conclusion with regards to high-risk demographics. In addition, in comparing our results province-by-province we will determine if risk has any region-dependency.

Our contributions will be two-fold. First, we will examine high-risk demographics and how they change depending on province. While risk demographics associated with Covid-19 have been studied regionally in Canada already [2], due to the rapidly evolving nature of an on-going pandemic experiencing a second wave of infections, our project will provide more results on a larger set of cases over more provinces. As British Columbia, Ontario, and Quebec contained the vast majority of the initial infections in spring of 2020, most research has been conducted on these specific provinces. At the time of writing (December, 2020) Alberta, Saskatchewan, and Manitoba are all experiencing significant outbreaks which will add a much larger sample to our mining. Second, We will alter the classic FP-growth algorithm to suit our particular needs efficiently. To achieve this, we propose removing the second database scan by representing the full database as a tree in memory, reusing available tree nodes where subsequent transactions have the same prefix. Each path of the tree will then represent a transaction from the database to which the FP-Tree can be constructed. Removing a second pass from the algorithm will reduce computation time and allow for quick data mining as the Covid-19 patient data available continues to grow throughout this pandemic.

# Related Work

The related and previous work considered in completing this project falls broadly under two categories. The first is the related work that characterizes how demographics are differently affected by Covid-19 infection and how that research educates the direction of our project. The second is how our selected data mining algorithm has been applied previously and how and if our selected changes have been implemented in similar cases.

While the study of Covid-19 only began in earnest this year for obvious reasons, there is significant research already completed that we may observe to direct our own goals for this project. One such preliminary study briefly compares the mortality rate between Chinese and Italian Covid-19 patients and reveals significant disparity in fatality rate between equivalent populations [1]. The fact that distributions of similar people in different geographic locations have different resilience to Covid-19 prompts inspection of this question on a finer regional level. While two countries as geographically and ethnically distinct as China and Italy may have considerable differences it may be worth examining if such regional changes in Covid-19 recovery rates exist regionally in Canada.

Examining how demographics react to Covid-19 across Canada has already been attempted in a paper entitled “Demographic Profile of COVID-19 Cases, Fatalities, Hospitalizations and Recoveries Across Canadian Provinces” [2]. This study compares the rates of hospitalization, fatality, and recovery of Canadians across the country by several different demographic characteristics. The authors break risk populations up by sex and age in a given region and found significant differences in the rate of hospitalization and death between provinces. However, this study was published in May of 2020 and is therefore extremely limited to a few select provinces that had substantial outbreaks at the time, namely British Columbia, Ontario, and Quebec. As of the time of writing (December 20, 2020), most provinces have substantial outbreaks ongoing in addition to significant daily hospitalizations and fatalities. With more up-to-date data and a larger sample size, it is likely that we can find more similar results across the country.

A broader study done in Sweden examines finer demographics utilizing data about a subject’s age, net income, education, civil status, and country of birth [3]. The additional personal information allows for a finer examination of exactly what factors might influence a patient’s chances of surviving Covid-19. So while a simple examination of age and gender yields valuable information, a greater knowledge of an individual’s health and socio-economic position will provide much more detailed information on what risk factors are most important to patient death.

Finally, while examining how Covid-19 affects demographics differently is not a novel idea, it may provide novel results given that new data is being received on a daily basis. Moving into our method of mining we can take the lessons learned in previous studies to drive the direction of our data mining efforts.

As our primary goal with this study is to find frequent itemsets that might imply relationships in Covid-19 patient data, our next step is to determine which data mining algorithm to use to mine frequent itemsets from our data. Jeff Heaton in his paper “Comparing Dataset Characteristics that Favor the Apriori, Eclat or FP-Growth Frequent Itemset Mining Algorithms” examines the above three algorithms and their performance as it relates to frequent dataset mining. While he compares the specific frequency density a maximum transaction size to determine best algorithms, he broadly recommends the FP-Growth algorithm as an appropriate technique for frequent itemset mining [4].

Given our decision to use the FP-Growth algorithm to mine our data for frequent itemsets in Covid-19 patient data, we decided to try and optimize the algorithm to reduce runtime as much as possible. Examining ways in which others have tried to optimize FP-Growth provides a starting point for a new modification.

In a study on oceanographic data, the authors use a parallel method to attempt to optimize their use of FP-Growth to handle large datasets. While this model does result in a faster runtime, the parallelization technique, number of threads used, size of dataset, and itemset characteristics make the ideal parameters to mine with complex. Determining the best optimization using this technique therefore requires significant study and preparation to achieve maximum speedup.

# Mining

The contribution we propose involves modifying the original FP-Growth algorithm in order to reduce the number of transactional database (TDB) scans required and thus, creating a more efficient mining process. However, this does come at the expense of using more memory; a time-space trade-off.

The original FP-Growth algorithm consists of four main steps: completing an initial scan of the TDB in order to find the frequent 1-itemsets, sorting the frequent 1-itemsets according to a set of criteria/heuristic, completing a second scan of the TDB to construct an FP-Tree, and passing this tree off to the actual FP-Growth mining algorithm to find the frequent k-itemsets. Our proposal focuses on removing the second TDB scan, thus reducing the transactions scanned by half.

To achieve this, we first focus on the TDB representation before passing it to the FP-Growth algorithm for mining. This will consist of three crucial steps. First, we perform our one and only TDB scan, scanning each transaction and adding each item to a tree structure as a node represented as *<item, occurrence\_count>*. We will call this tree structure REP-Tree, as it will represent the entire TDB as an in-memory data structure. Each path of the REP-Tree will then represent one transaction from the TDB. To save space, we will re-use tree nodes that have the same prefix as the next transaction from the TDB, incrementing the *occurrence\_count* of said node.

Secondly, our algorithm will traverse the REP-Tree, and create a table consisting of *<item, occurrence\_count>* pairs. In this step, any item whose occurrence count does not meet the user specified minimum support threshold with be pruned from the table. The table will then be sorted according to a heuristic. The heuristic we will be using in our implementation and following example will be based on a frequency descending order. Resolving occurrence count ties based on alphabetical order.

Finally, our algorithm will then convert our REP-Tree into an FP-Tree in preparation for mining. To execute this, we simply need to extract each path in the REP-Tree. Since each path represents a transaction from the TDB, we can remove items that are not present in our header table and add the resulting transaction to an FP-Tree according to the standard FP-Growth algorithm. Once the FP-Tree is constructed it can be passed to FP-Growth for mining. At this point the REP-Tree is no longer required and can be removed from memory.

To illustrate the above description, we will perform a step-by-step example using the data in Table 1. below.

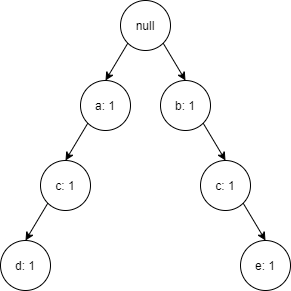
Table 1. Example Transactional Database

|  |  |
| --- | --- |
| **Transaction ID** | **Transaction** |
| t1 | a, c, d |
| t2 | b, c, e |
| t3 | a, b, c, e |
| t4 | b, e |

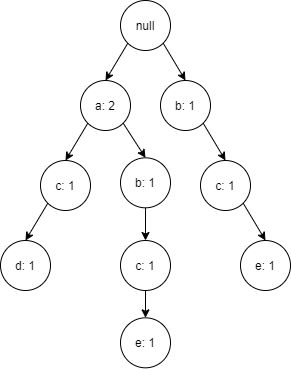
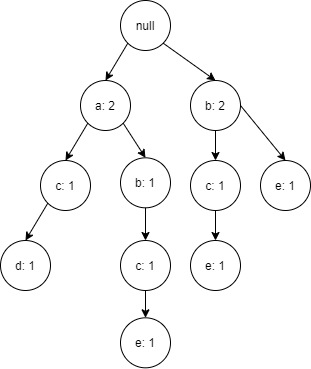
minsup = 2

**Step 1:** Scan the TDB and represent it as a tree.

**Scan t1 Scan t2**

 ****

**Scan t3 Scan t4**

** **

Following the first step, the result is a REP-Tree. Each path of the tree now represents a transaction from the original TDB.

**Step 2:** Scan the REP-Tree, sum item occurrence count, and re-order based on our heuristic.

|  |  |
| --- | --- |
| **Item** | **Occurrence Count** |
| a | 2 |
| b | 3 |
| c | 3 |
| ~~d~~ | ~~1~~ |
| e | 3 |

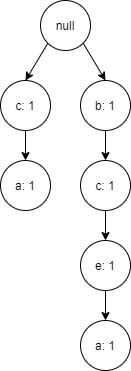
**Re-order in frequency descending order**

|  |  |
| --- | --- |
| **Item** | **Occurrence Count** |
| b | 3 |
| c | 3 |
| e | 3 |
| a | 2 |

**Step 3:** For each path in the REP-Tree, sort in header table order, remove item that do not appear in the header table, and construct an FP-Tree.

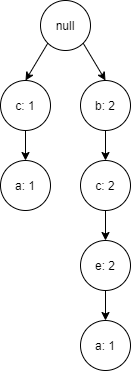
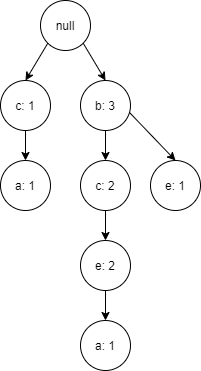
**Path 1 = a – c – d becomes c – a Path 2 = a – b – c – e becomes b – c – e – a**

FP-Tree #1 FP-Tree #2

**Path 3 = b – c – e becomes b – c – e Path 4 = b – e becomes b – e**

FP-Tree #3 FP-Tree #4

Upon completing this construction process, we can pass the resulting FP-Tree to FP-Growth for mining frequent k-itemsets. This requires only one scan of the transactional database at the cost of using more memory. However, in a practical scenario, if FP-Growth is the algorithm being used for mining, the TDB must fit into memory to be represented as an FP-Tree. If this is possible, representation as a REP-Tree should also be feasible.

*Code 1: Pseudocode for main mining:*

OPEN database

for each transaction:

    ADD transaction to REP-Tree

PROVIDE REP-Tree to FP-Growth mining algorithm

*Code 2: Pseudocode for adding a transaction to the REP-Tree:*

ASSIGN the Current Node as the Root Node

for each item in the transaction:

    if the item is in the roots children:

        INCREMENT the nodes count

        ASSIGN this child node as the Next Node

    else:

        CREATE a new node as <item, 1>

        ASSIGN the new node to be a child of the current node

        ASSIGN this new node as the Next Node

    ASSIGN the Current Node as the Next Node

*Code 3: Pseudocode for getting support values of items in a REP-Tree:*

CREATE a dictionary of the form <item, count>

CREATE a queue

ADD the root node to the queue

while the queue is not empty:

    dequeue node from the queue

    if the node contains an item:

        ADD an entry to the dictionary with the nodes item

        and increment the count by its occurrence count

        ADD each child of the node to the queue

RETURN the generated dictionary

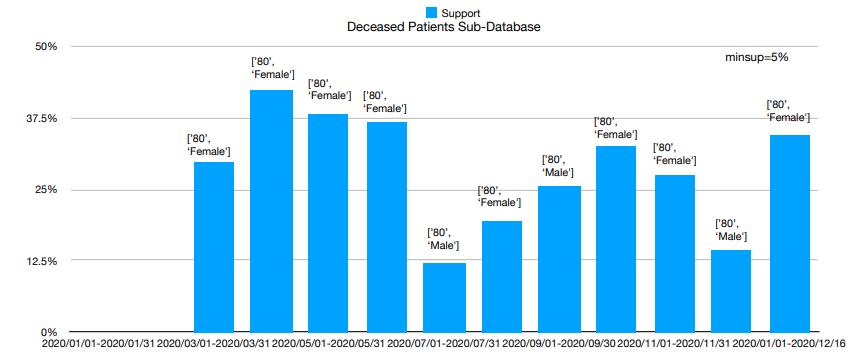
# Results

The dataset used with our modified FP-Growth algorithm has the following relevant attributes: *gender, case\_status, exposure type, age-group,* and  *health\_region*. Given these attributes we have endeavored to mine several things from our data.

*Mortality Risk*

First, we are interested in risk groups for mortality over time. Discovering if the highest-risk groups for death have changed over the course of the pandemic could potentially help to educate our response as the pandemic progresses. See Table 2.

Chart 1. Highest Risk of Death Patients



As can be seen in Chart 1, the patients with highest risk of death due to Covid-19 remains consistently patients in their 80s, both female and male. While these results are not unexpected given the higher prevalence of co-morbidities in older patients, continuing to monitor these numbers as new and various treatments are utilized might inform of changes in infected demographics.

*Exposure Type*

Second, examining the demographics most frequently associated with various forms of infection leads us to the following results, see charts 2 and 3.

Chart 2. Close Contact Exposure by Demographic

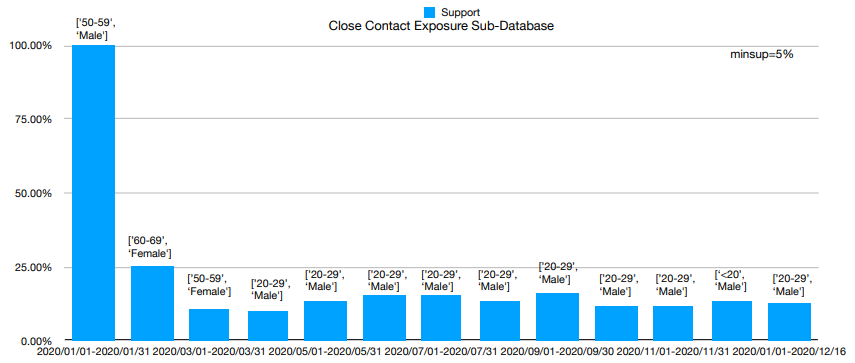
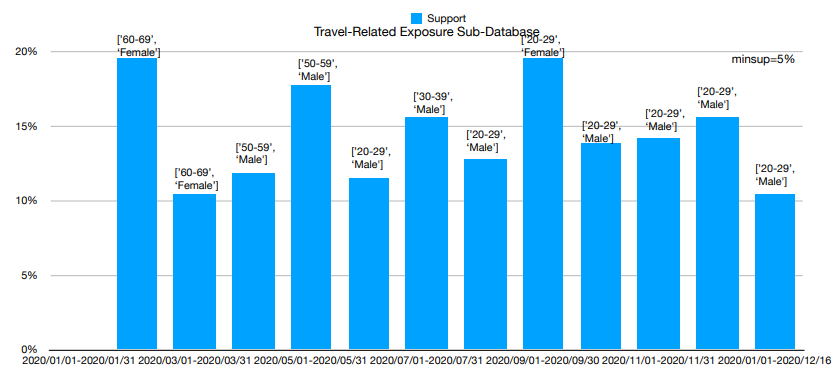


Chart 3. Travel Exposure by Demographic



A look at the above charts shows the demographic most likely to contract Covid-19 through either a close contact or travel. Of particular interest is the shift in at-risk demographic over time seen in Table 4. During the moths of January through May we see older adults being the primary vector for Covid-19 infections through travel. After May however, young adults of both genders consistently account for the most frequent infections through travel.

A simple explanation for this abrupt shift could be the fact that young adults are in a low-risk demographic for serious complication of Covid-19. Since there is relatively little danger of serious personal danger, young adults therefore do not feel the need to limit travel as much as those in older age-groups.

*Age Demographics*

Similarly, a look at Table 3 shows that, aside from at the very start of the pandemic, the demographic most frequently infected through a close contact is young males ages 20-29 years old. A single data point in March 2020 showing females 50-59 most frequently at risk, it is the only outlier to this trend. While this single data point is not enough to illustrate a total shift in frequent close contact infection demographics, presumably due to imposed Covid-19 restrictions, it does support the results discussed in Chart 3. Namely, that government-imposed restrictions since the beginning of the pandemic have been more effective with older adults and less effective at reducing spread through multiple infection vectors with young adults, particularly males.

Chart 4. Cases by Age

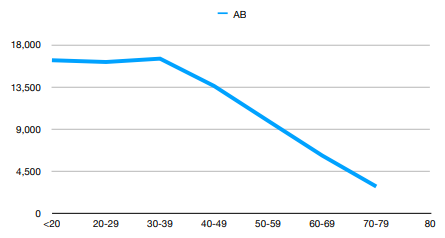
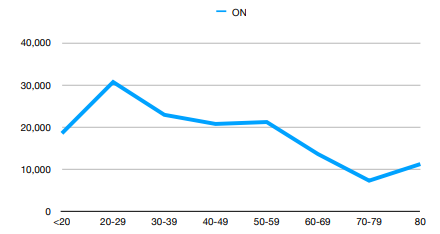
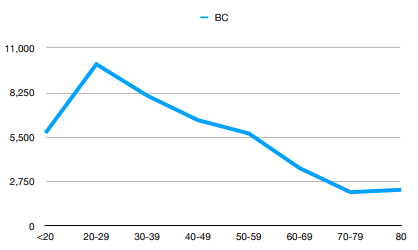
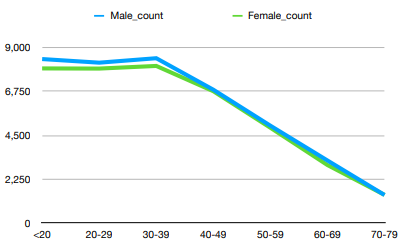
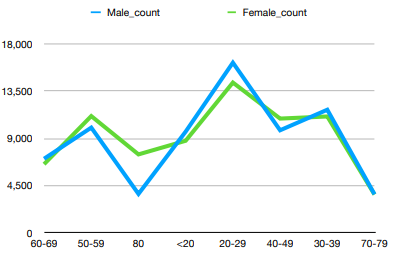
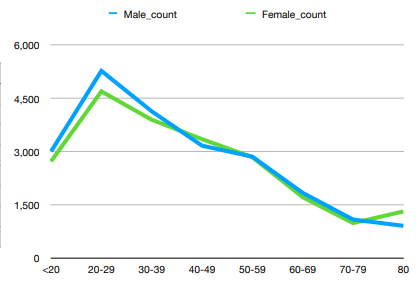


Chart 4 shows the cases seen in three provinces by age-group. Alberta, Ontario, and British Columbia were displayed as they were the only provinces in the dataset with case numbers in the thousands. While each province shows slight variation in the precise ratio of cases per-age group, the overall trend remains consistent. Overall, younger people represent the most Covid-19 cases in each province. The total number of cases trends downward as age group increases. This affirms the results found with regards to frequent demographics infected via travel and close contact. Examining Chart 5 shows extremely similar distribution of cases by age demographic between sexes. While males are, generally, slightly more likely to be infected across age ranges, the difference is minimal and as a result, the conclusions we draw with regards to at-risk demographics for contracting Covid-19 has more to do with age than gender.

Chart 5. Cases by Age and Gender



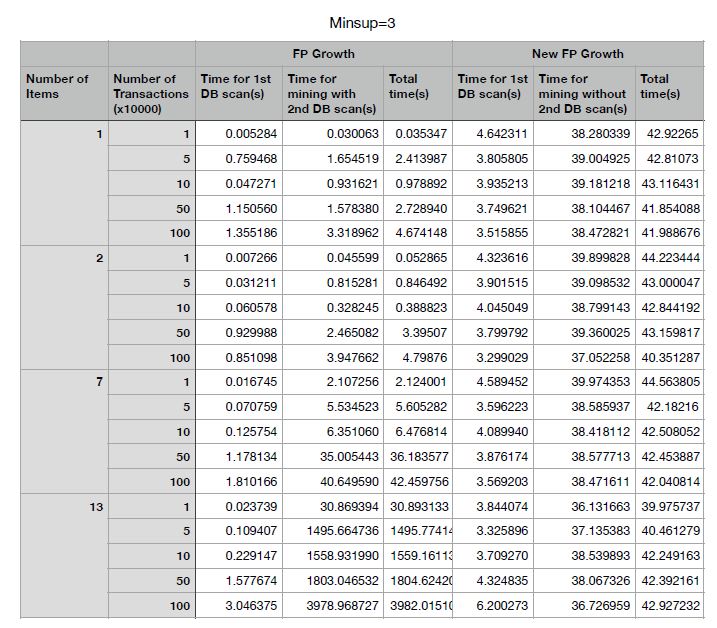
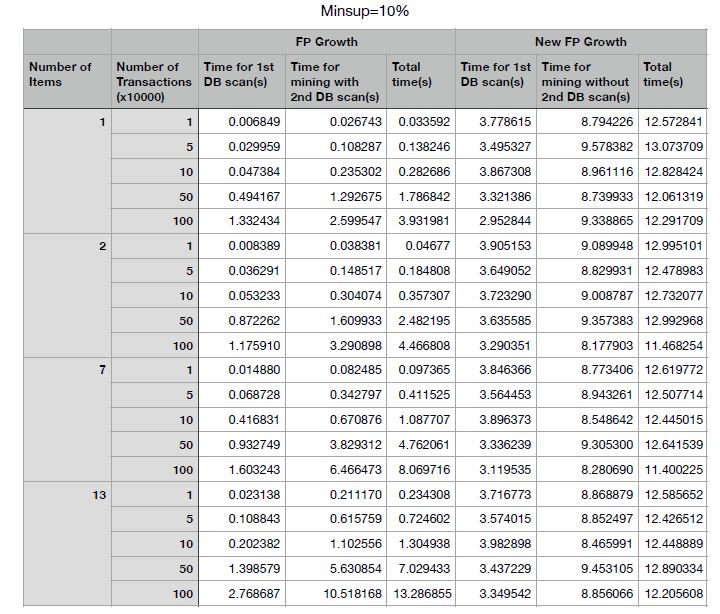


*Modified Algorithm Comparison*

Comparing the original FP-Growth algorithm to our new proposed FP-Growth algorithm, we noticed some interesting performance differences. We compared these two algorithms using a dataset of 1, 2, 7, and 13 unique items on a database size of 10000, 50000, 100000, 500000, and 1000000. We also experimented with a minimum support threshold of 3, and a much larger minimum support threshold of 10%.

Table 2 and table 3below illustrate the time (in seconds) that it took to execute each algorithm with a minimum support threshold equal to 10% and 3, respectively. These tables are broken down into time it took to complete the first database scan, the time it took to complete the actual mining, and the total execution time for each algorithm.

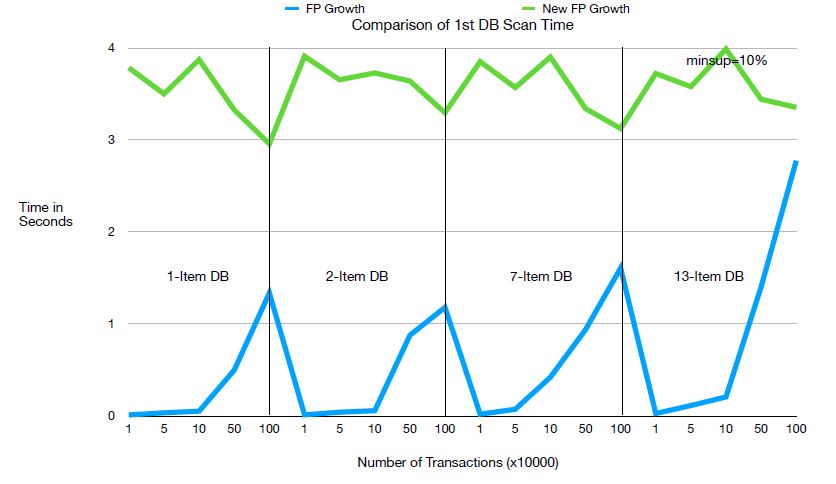
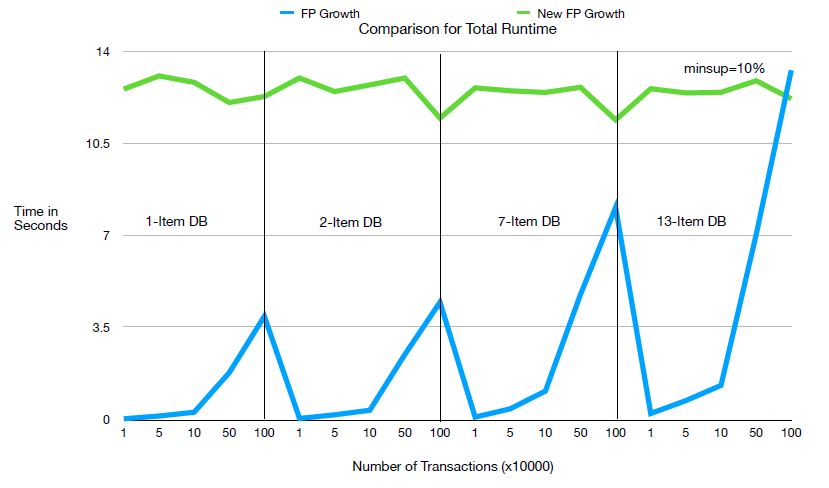
Table 2 Table 3

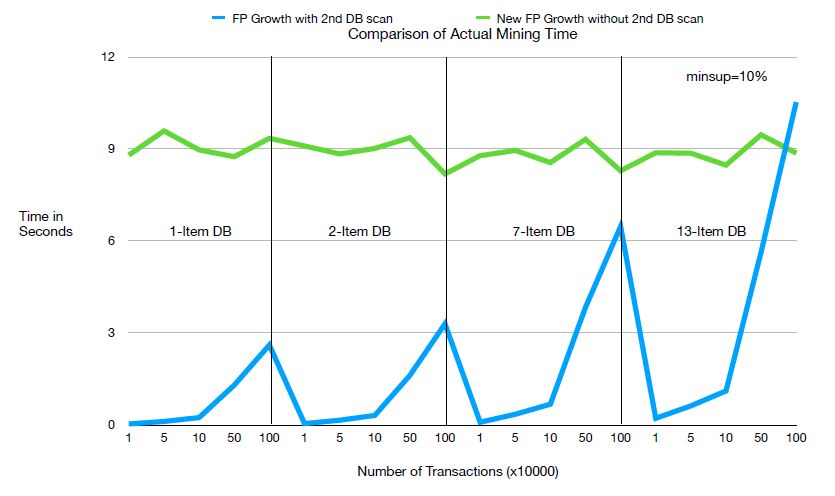


The time for the first database scan in the original FP-Growth algorithm represents the time required to scan each transaction in the database and construct a simple list of transactions. With respect to our new FP-Growth algorithm, this time represents the time taken to scan each transaction in the database and construct our REP-Tree (a tree representing each transaction in the database). Furthermore, the second column of data represents the time taken to complete the second database scan, determine the frequent 1-itemsets, and complete the mining process to determine the remaining k-itemsets with respect to the original FP-Growth algorithm. This column represents the time taken to iterate over the REP-Tree, determine the frequent 1-itemsets, and complete the mining process to determine the remaining k-itemsets for our new proposed FP-Growth algorithm.

Looking at the results for each of these algorithms, we can see that the original FP-Growth algorithm outperforms our new FP-Growth algorithm with large minimum support thresholds (see table 2). This is largely due to the fact that the overhead to construct and iterate over the REP-Tree when a lot of items will inevitably be pruned due to not meeting the minimum support threshold requirement. It should be noted that our new FP-Growth algorithm performs all mining in a relatively constant time due to the tree traversal process. However, we see the opposite result when using a low minimum support threshold (see table 3). In this scenario, the original FP-Growth algorithm must now iterate over almost all the possible itemsets, reading them off disk to construct the FP-Tree. While FP-Growth still outperforms our new FP-Growth algorithm on datasets with few unique items, our proposed algorithm significantly outperforms the original FP-Growth algorithm when we begin mining on large datasets with many unique items.

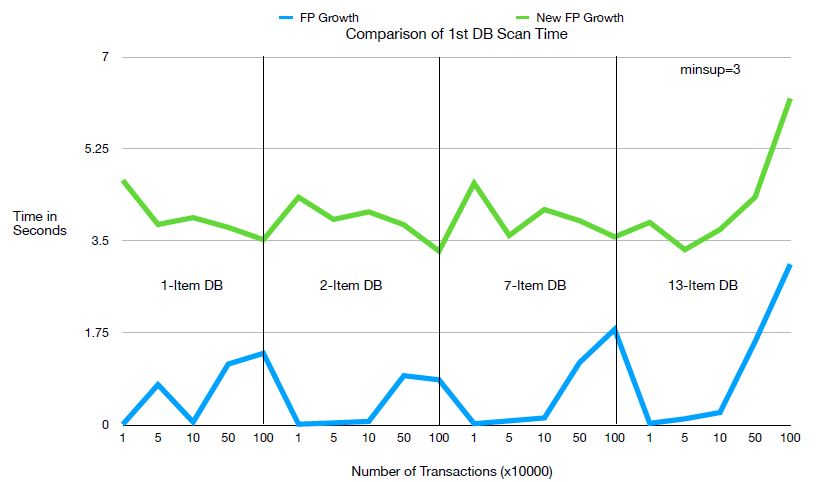
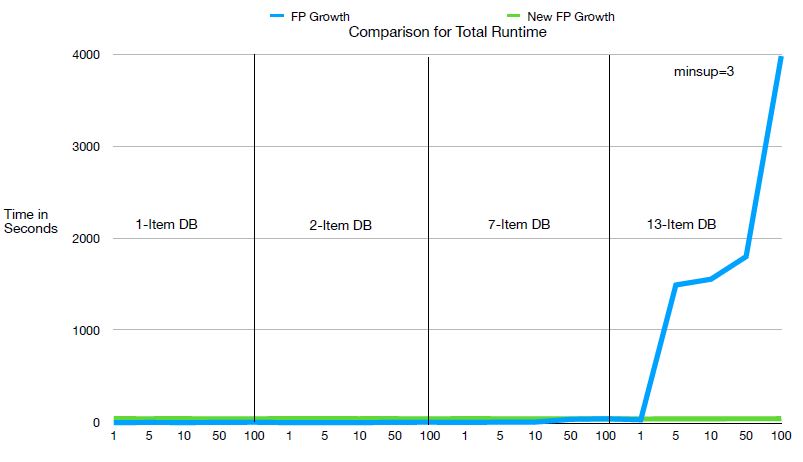
*Chart 6: Comparison for minsup = 10%*

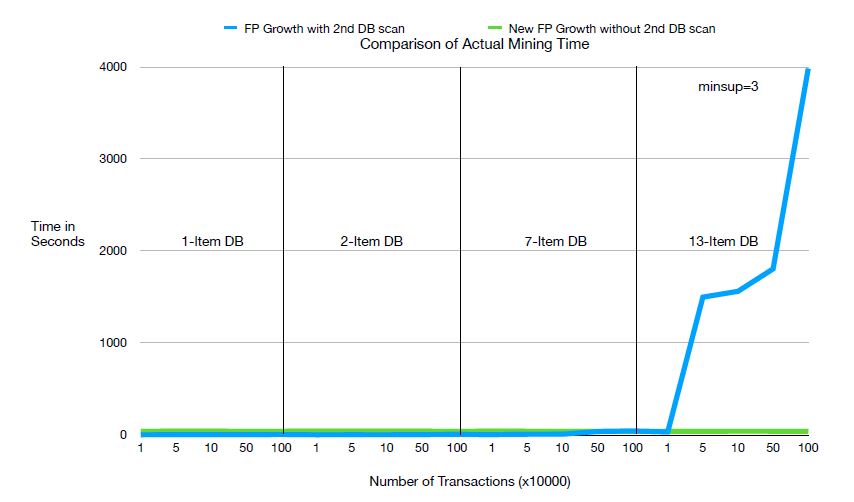




In terms of the practicality of our proposed changes to FP-Growth, these results are promising. Since most interesting mining will be done using large datasets with many unique itemsets, our new algorithm could pose for more efficient frequent pattern mining in certain scenarios.

*Graph comparison for minsup = 3*





# Conclusions

While the results we have found from mining this particular dataset are nothing groundbreaking, there are a few useful conclusions we can draw.

First, the population most at risk of death due to complications related to Covid-19 is consistently those 80 years of age and older. This result might seem obvious but drawing this conclusion through our data-mining process lends credible evidence to what might be seen as common sense.

Second, young people, particularly males between the ages of 20-29 are the most frequent demographic to contract the virus through a close contact or travel. In addition to being to most frequently infected group due to these two specific infection vectors, they also present the most cases per any one demographic. According to Statistics Canada [6], Canadians aged 20-29 make up 13.4% of total population. Comparing to the 30-39 demographic (13.9%), 40-49 demographic (12.8%), and <20 demographic (21.3%), we can see that young adults 20-29 are not the largest demographic in the country, and therefore show a higher than average incident of Covid-19 infection.

The fact that this age group and gender appears to consistently be infected at a greater rate than the general population implies one or more things from the following. First, that either existing messaging to this demographic is ineffective and additional education and restriction might be required to lower infection rates. Second, this demographic may commonly work in industries with naturally higher infection rates. Sectors such as construction and hospitality are disproportionately staffed by young adults, and as a result may influence the high infection rate. Without additional data on each Covid-19 case, it’s hard to draw strong conclusions and is a potential topic for future work.

With regards to our improvement to the FP-Growth algorithm, we can clearly see the benefit of a single database pass on algorithm runtime. As mentioned above, this time improvement comes at a space penalty. While not every data mining situation is appropriate for use of our modified algorithm, the detailed analysis we provide of results provides a good idea of where it may be appropriate.

# Limitations

The primary limitation in our work to determine useful information on various Covid-19 demographics is a result of limited patient information. Even the information included in the database is largely incomplete. For example, our dataset includes an attribute to describe exposure type. A large portion of patients do not have an exposure type reported and as a result, leaves our data incomplete. For us to determine truly novel insights, more attributes per patient are required. For example, including the profession and employment status of each person, their housing situation (ie, house, apartment, rooming house, personal care home, etc), pre-existing medical conditions, etc would provide a much more comprehensive look at the personal situations that surround each Covid-19 case.

Another limitation is the limited quantity of patient data available from a central source. As cases continue to rise throughout the winter at a near exponential rate throughout Canada, more data will become available and will provide either additional insights into Covid-19 infection demographics or additional evidence to support existing points.

# Future Work

Future work will primarily involve the use of more data and more detailed data. Accessing more accurate data from around the country that is regularly updated will provide more insights into how Covid-19 affects demographics differently and can help to educate our response to it.

##### References

1. <https://onlinelibrary.wiley.com/doi/pdfdirect/10.1002/jmv.25860>
2. Simona Bignami-Van, Assche,; for Interuniversity Research and Analysis on Organizations, Center. Demographic Profile of COVID-19 Cases, Fatalities, Hospitalizations and Recoveries Across Canadian Provinces. Montreal, QC, CA: Center for Interuniversity Research and Analysis on Organizations, 2020.
3. Drefahl, S., Wallace, M., Mussino, E. *et al.* A population-based cohort study of socio-demographic risk factors for COVID-19 deaths in Sweden. *Nat Commun* **11,**5097 (2020).
4. J. Heaton, "Comparing dataset characteristics that favor the Apriori, Eclat or FP-Growth frequent itemset mining algorithms," SoutheastCon 2016, Norfolk, VA, 2016, pp. 1-7, doi: 10.1109/SECON.2016.7506659.
5. Jiang, Y., Zhao, M., Hu, C. *et al.* A parallel FP-growth algorithm on World Ocean Atlas data with multi-core CPU. *J Supercomput* **75,**732–745 (2019). <https://doi-org.uml.idm.oclc.org/10.1007/s11227-018-2297-6>
6. Statistics Canada. Table 17-10-0005-01 Population estimates on July 1st, by age and sex,
7. doi: https://doi.org/10.25318/1710000501-eng