Big Data Analysis and Data Visualization on COVID-19

Thomas Czubryt
Department of Computer Science
University of Manitoba
Winnipeg, Manitoba, Canada
czubrytt@myumanitoba.ca

Pengyu Wang
Department of Computer Science
University of Manitoba
Winnipeg, Manitoba, Canada
wangp1@myumanitoba.ca

Hantong Li
Department of Computer Science
University of Manitoba
Winnipeg, Manitoba, Canada
lih34542@myumanitoba.ca

Zhiyi Chen
Department of Computer Science
University of Manitoba
Winnipeg, Manitoba, Canada
chenz2@myumanitoba.ca

Yue Ma
Department of Computer Science
University of Manitoba
Winnipeg, Manitoba, Canada
may4@myumanitoba.ca

***Abstract*—Analysis of big data has become a major scientific field in the last few years. In big data, researchers aim to discover interesting patterns hidden within massive sets of data. In this paper, we will apply the principles of big data to find interesting associations within a dataset containing information on COVID-19 cases. The dataset will be partitioned into many groups, where each group corresponds to a particular combination of the patient's gender, age range, and the region in which the case occurs. For each group, we will analyze the percentage of COVID-19 cases where the patient dies, exhibits symptoms, and is hospitalized in the intensive care unit. These percentages will be compared to help observe groups of people who are more susceptible to COVID-19 or are more likely to spread COVID-19.**

Keywords—Big data, COVID-19, U-VIPER

I. INTRODUCTION

Against the backdrop of science and technology advances by leaps and bounds, the daily life for people who are living in a world, which is full of data, everything can be transformed to data by high-tech means, such as the frequency of people who access to a website, within one hour how many websites people visited or even what types of websites people visited most, etc. Things like that can be counted as data. Back to Big-Data, the "Big" is the same as "mega", in the literal sense, there is a huge amount of data, which should be processed, speed of data stream is very high and the value of data is rich enough that it can produce high value. There are some characteristics for " Big-Data", " 3Vs" properties: High Volume, High Velocity and Variety [13]. In recent years, there are 2 more properties that have been added to the definition of "Big-Data". There are 5 properties for Big Data, which are named "5Vs"; the two which have been added recently are Veracity and Value [14]. This paper will show the details of 5Vs' property works on Data stream and what results can be mined by data science algorithm.

In the winter of 2020, a pandemic initial outbreak in Wuhan, China, and then spread by modern transportations to the global world. This pandemic has been continuing for almost 2 years. Researchers discovered the information from these big data and transformed them to be valuable, and solid numbers. For example, knowledge was mining from the epidemiological data, the data related to the cases who were infected by viral diseases, for instance, Spanish influenza(H1N1) in 1918- 1919, Severe Acute Respiratory Syndrome(SARS) broke out in 2002-2004, H1N1 broke out in 2009, and the pandemic we are suffering right now, COVID-19 since the end of 2019. The data collected and analyzed by researchers, epidemiologists is getting better to know the disease, from the results such as the distribution of ages and human race between people and the trajectory of propagation. Based on the understanding of the diseases so far, some of the countries invented some apps to help the governments to detect and trace where the headstream of each wave was. For example, in China, the government published a mini program on WeChat and Alipay named Health Code [1] to track each person in China the status of health, and use the QR code shown on the screen as a healthy proof to enter public places. There are three colors that will automatically be shown on the app, Green, Yellow and Red, these three can dynamically show the health status for each user. In Canada, there is a mobile app named "COVID Alert" published by the government of Canada. This is a free COVID-19 exposure notification app [2], it can alert users to possible exposure before they have symptoms. This app works the same as Health Code does. Based on the properties researchers can discover the knowledge, for example the disease between different genders and ages, with the collection of a huge variety of data sources, high volume can be established, such as: Data related to testing location of COVID-19, daily new cases and hospitalization, high value can also be performed, since COVID-19 has continuing for several months, the development of society, and economic activity has frozen for a long time, including accommodation services, food services, traveling, and cultural industries are all affected by it. Since the border is in closure that also induced recession of the activities of other service industries, including providing professional and administrative services for enterprises and consumers[3], this can be called the aftershock led by COVID-19 pandemic said by Statistics Canada. The researchers have noticed that the extent of the spread influenced many fields for human society, so it is urgent for us- Computer scientists to contribute the knowledge we learned to spare efforts on data science to aspects of epidemiology. Therefore, for this paper, we present a data science solution for analyzing big data on COVID-19. Specifically, the main focus on the contents include:

- To show the tendency of covid so far. Estimate what the situation will be like in the future, this can be regarded as a Pre-Warning.
- Analyzes the distribution of people within a region, for instance, age group and gender (Uncertain Data Mining U-Viper [12]).

The work of data science has been gradually revealed to the public. And the characteristics of Big-Data fully play their roles to help epidemiological scientists in their field.

- Due to the tested cases and confirmed cases there is a high volume of data, for instance, there are 48,289,007 tests performed inside Canada as of November 24, 2021[4], from the official website of Government of Canada, scientists can see the distribution of tested performed, confirmed cases and death cases. The data shown on the map and the distribution of confirmed positive cases on the map of Canada, it is not difficult to analyze, the more developed the place, the more positive cases will be tested within 14-days . The scientists then output the stream in data visualization. Either researchers can see the distribution of new cases intuitively and so can the public.
- From the distribution mentioned in the last point, it can also create value either for society or the public. For instance, various industries have been influenced for almost two years [5], in all the provinces of Canada based on the chart of percentage change in operating revenue for selected industries providing professional and administrative services for business and consumers, 2019 and 2020 shown on Statistic of Government of Canada, we can see the impact is especially huge on travel arrangement and accommodation services. So the owner of these industries programs can reduce the cost then switch to other programs. To make the value or profit maximization.

In the next sections we are going to discuss the background and related works. Section 3 is to present the solution of our data scientific approach, then section 4 is to show the results of evaluation. In the last section, it is to draw conclusions of what we discussed.

II. BACKGROUND AND RELATED WORK

A. *COVID-19 Research*

Due to the spread of the Covid-19, there are many researches made by different researchers in the world in various areas.

For predicting and monitoring the coronavirus spread, Big Data analytics was used on HPCC (High- Performance Computing Cluster) system platform to develop a model and track COVID-19 cases [6].

There have been studies focusing on diagnostic testing and vaccination for COVID-19 among different groups of people in Manitoba.(First Nations, Metis and Inuit) [7] and Big Data Science on COVID-19 Data (Leung, Chen, Y., Shang, S., & Deng, D. (2020)) [16].

Besides, there is also a large amount of existing epidemiological data that focused on confirmed positive cases of COVID-19, and daily new cases and hospitalization in Manitoba [15]. Not only to show the cumulative number of cases and how many people have been affected every day, but to distribute all the positive cases into different groups through age, gender, and human race. By comparing and making a statistical analysis of these data, we can discover other useful

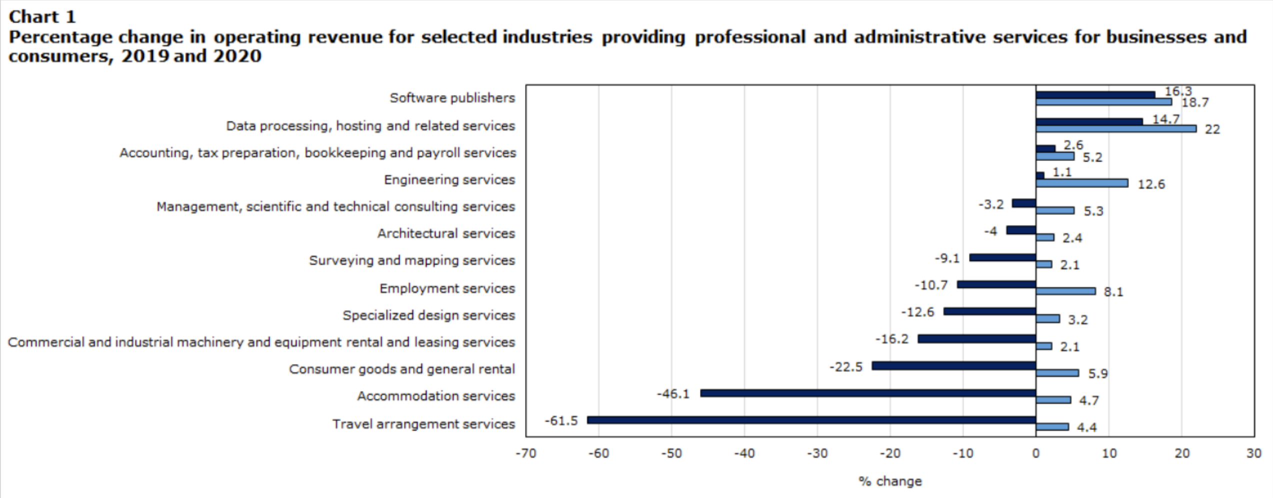


Fig. 1. Percentage change in operating revenue for selected industries providing professional and administrative services for businesses and consumers, 2019 and 2020.

information(e.g. Which age group is susceptible to the New Coronavirus. And what kind of public places have more possibilities of transmission, the tendency of daily increasing cases).

III. Our Data Science Solution

In this part, we will introduce the algorithms we will use and some data processing on COVID-19 epidemiological data.

A. Data Collection

In the data collection part, we need to collect reliable data to analyze the data accurately. In Canada, the government of each province in Canada is obliged to collect relevant data on COVID-19 and then analyze it, also the Public Health Agency of Canada (PHAC) would constantly check it to make sure of the accuracy of the data. Therefore, the availability and feasibility of these data are very high, and the data are relatively complete. We have selected several data with relatively complete attributes from multiple COVID-19 data. The big COVID-19 data contains the following information:

- Personal Information - such as (a) gender, (b) age group, (c) occupation, (d) and region that person lived in.
- The state of the cases of each patient - (a) whether they are alive or not, (b) if they are under the treatments from the hospital, (c) if the cases have been resolved or not, (d) if the patient is asymptomatic or not.
- Transmission - the ways that patients got infected (i.e. Domestic acquisition, international travel).
- Episode - the time would be detailed to which week of the year.

B. Data Preprocessing

1) Data Reduction

By observing the data we have just collected, we found out that some of the data are not given / unknown in the original dataset, which is called "COVID19-eng.csv".In order to avoid all these unrelated datas that will have negative influence on the correctness and accuracy of our analysis and data science solution, the first step that we did is called Data Reduction. For example, we have 1,697,214 patients' information collected, but there are 2,737 of them who do not know their gender. To help the analysis of the data to be more accurate and persuasive, we decided to delete this unknown data. Since 2,737 patients' genders are unknown, we would have 1,694,477 patients left. Since for the gender group of the original source file, "9" represents unknown, "1" equals to male, and "2" equals to female. What we did in this reduction process is to delete all the rows with 9 on the column of COV_GDR. Same method would also be applied to "Age group", "Asymptomatic", "Hospital Status", and "Death".

2) Data Transformation

For the most effective use of our data, we categorize each case into different categories. We use a combination of <gender, age, region> to categorize different categories. For example, if the gender of a patient in a case is male, his age is between 0-19, and living in Manitoba, he will be categorized into the <Male, 0-19, Prairies> combination. When a case is categorized into a combination group, the number of this combination group will be increased by one. From the data we found:

1. Gender: Male or Female.
2. Age group is divided into 8 groups: 0-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80+.
3. Five regions: Atlantic, Quebec, Ontario and Nunavut, Prairies, and British Columbia.

After the data transformation, the final transformed data will only have 80 rows (80 combinations(2*8*5)), and then we will add more columns of data to keep track of how many cases there are in each category, number of cases that have symptoms, the number of cases that died, percentage of symptomatic, percentage of death, percentage of ICU.

C. Uncertain Data Mining (U-VIPER)

As we mentioned in the data processing section, we have transformed the data into a better shape so that we can find the mortality and morbidity of cases in a region. Knowing that people of each age group and different gender respond differently to the disease, our solution studies the common events of each <gender, age, region> combination. That is, for the combination of <male, 0-19, grassland>, morbidity, mortality, ICU rate, and different subsets of these three rates will be recorded. Note that there may be empty attributes in the data. For example, if the combination <Male, 0-19, Prairies> is not dead, the column will be empty. Since we are using U-VIPER [12], the percentages for the elements in each combination would be calculated by (the number of distinct attributes in the specific combination / the amount of this combination). Furthermore, the most attractive feature of the result is that after analyzing different regions, each regional government can take corresponding countermeasures to deal with it.

IV. Evaluation

A. Case Study on Real-Life COVID-19 Data

1) Data Collection

In order to make our data more useful and illustrate the result of the science solution more accurate. We tried several combinations of attributes from epidemiological data of COVID- 19. All of these data are from Canada Statistics(Preliminary dataset on confirmed cases of COVID-19, Public Health Agency of Canada, 2020-2021) [8]. The dataset is collected by provincial and regional public health authorities, which are named as Public Health Agency of Canada(PHAC). As the notes said, this data file is provided to Canadians and researchers to monitor the confirmed cases of Covid in Canada. Since the situation of Covid changes rapidly, these data are based on the result on November 12, 2021 [8]. To process the data that have been collected, we obtain some basic attributes and receive the dataset with the following attributes:

1. Gender of cases (GDR): Male, Female and Not Stated.
2. Age group (AGR): There are 9 groups, and no stated group is removed, which shows " 99" in column, so only 8 groups data are collected, <=19, 20s, 30s, 40s, 50s, 60s, 70s, >=80s, 99 is not stated. These values are

corrected as the Public Health Agency of Canada (PHAC) receives new information.

3. Provinces of different regions inside Canada, and divide them into 5 groups (REG): Atlantic (New Brunswick, Nova Scotia, Prince Edward Island, Newfoundland and Labrador), Quebec, Ontario and Nunavut, Prairies (Alberta, Saskatchewan, and Manitoba) and the Northwest Territories, British Columbia and Yukon.
4. Number of cases.
5. Asymptomatic (ASM): Yes or No.
6. Death (DTH): Yes or No.
7. Hospital status (HSP): There are 4 situations, (a) Hospitalized and in Intensive care unit, (b) Hospitalized but Not in intensive care unit, (c) Not hospitalized, (d) Not stated/Unknown.
8. Percentage/ Possibility of ASM, DTH and HSP: Use Number of ASM, DTH or HSP divides Number of Cases to get the ultimate percentage or possibility of them.

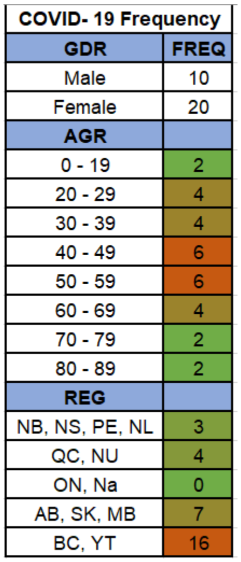
2) *Data Processing*

Of all the attributes we selected, we removed the element which shows Not Stated or Unknown, since it is hard to classify which groups of these elements should be stated, for example, in the Gender column, there are 3 types of elements that could be filled into. However, not stated is difficult for us and scientists to classify. In the biological area, there is only female and male, no other genders in the world. Someone announces they are the opposite gender of their originals, it is on the consciousness, not on the biological area, we are focusing on the episodical field study of COVID-19, so there is no doubt that it should be biological gender. This is the reason why we undocked the cases whose genders are not stated to make the result more accurate. After applying the same method to drop all the "Not Stated" and "Unknown" rows for all the attributes that we listed above, the cumulated number of cases with stated gender, age, and region are 1,315,816. Besides, we also get 80 combinations of <gender, age, region> with the symptomatic distribution for the 5 different regions.

Since there are many combinations of <age, gender , region>, so, our data science solution is going to analyze and choose out of 80 combinations. We extract the top 30 combinations of the result of rates after sorting, first key sorting order by ASM, then DTH, and next HSP. The reason why we do this is that we consider ASM to show whether the case is asymptomatic, because since this pandemic happened, asymptomatic cases are easier to spread viruses. In people's eyes and inherent knowledge, the person who does not show any symptoms, then he/she thinks himself/herself is not affected by the virus. However, for Covid, the situation has changed, the people who do not show any of the symptoms can be the potential patient as well. In the report of Government of Canada, *Individual and community-based measures to mitigate the spread of COVID-19 in Canada*, in the section of Transmission of SARS-CoV-2(alias or scientific name of COVID-19) said infected individuals generate respiratory droplets and aerosols, which can be transmitted to others [9], most of the asymptomatic people do not stay at home because they do not know they have been infected. So this may lead to more serious spread. On the other hand, the risk of transmission by respiratory aerosols is greater in poorly ventilated indoor environments where there is a high density of people and long contact duration said by the report of *National Collaborating Centre for Environmental Health*, "The Basics of SARS-CoV-2 Transmission," [10]. Secondly, we set hospitalization as the second key of sorting due to the speed of spread of the pandemic, most of the hospitals can not hospitalized such a large number of patients within a limited time. At the beginning pandemic effects were seen as collateral health effects on Countries health care systems. And the beds of the Intensive Care Unit (ICU) were few. It is to help us and researchers to dig into the hidden reason, or are there any problems with hidden aspects of management such as hospitalization period (HP) and effects on treatment facilities such as the shortage of beds [11]. The last key is death rate, since the fatality rate is 2.3% in all cases, 70s has 8%, the group of age greater than 80 is 14.8%, so the analysis of death rate is inevitable.

a) *Big data science on percentage of symptomatic cases*

As mentioned before, we extracted top 30 combinations out of 80, here is the frequency of attribute in each group as shown below:



COVID- 19 Frequency	
GDR	FREQ
Male	10
Female	20
AGR	
0 - 19	2
20 - 29	4
30 - 39	4
40 - 49	6
50 - 59	6
60 - 69	4
70 - 79	2
80 - 89	2
REG	
NB, NS, PE, NL	3
QC, NU	4
ON, Na	0
AB, SK, MB	7
BC, YT	16

Fig. 2. COVID-19 Frequency

From the table shown above, there are more Female cases in the gender group. In age group, the relation between ages and frequency is approximately as pyramid shape or stepped appearance, as shown below:

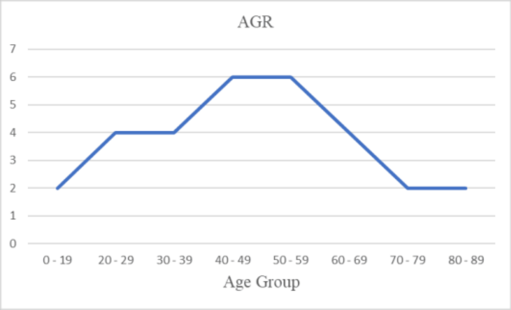


Fig. 3. Frequency of Age Group

This is an interesting finding, based on the death rate that we mentioned above, the tendency of age should be increased as age grows, but in this result it shows stepped appearance. We will talk about this in (2), Big data science on morbidity and mortality of 5 certain regions.

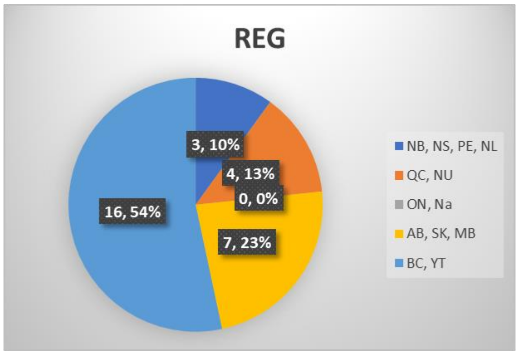


Fig. 4. Distribution of Frequency of Region Group

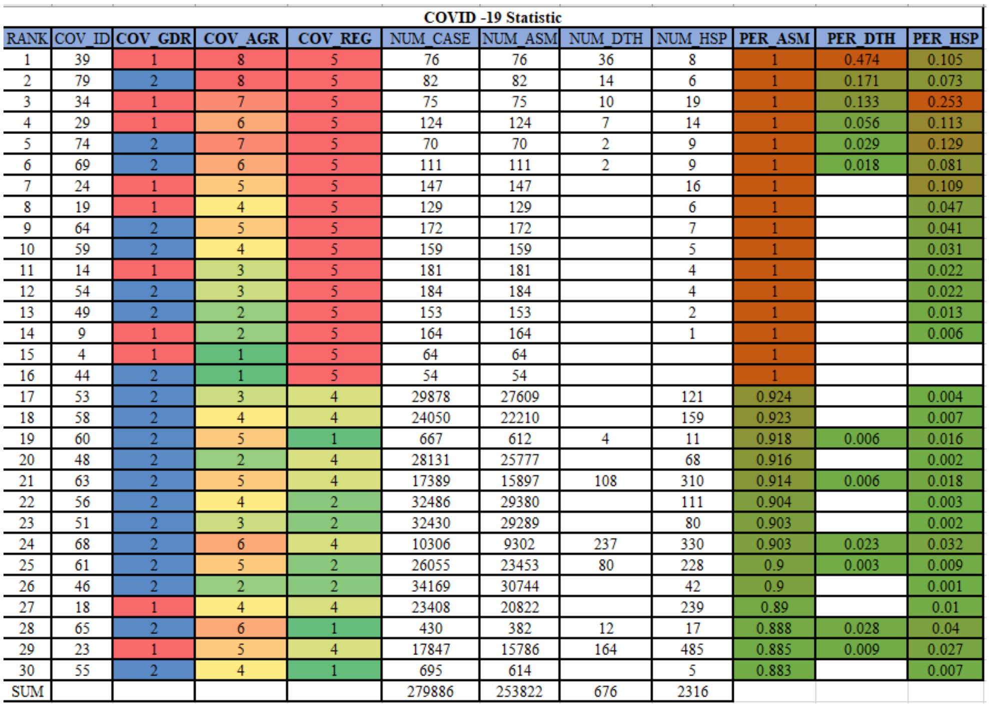
The distribution of regions follows the degree of development, the reason is that, as shown in Fig. 4., BC and YT have the highest proportion of regions, which is 54%. BC has the second largest population in Canada, with 5.2 million in 2021. The high density of population may lead to this, as the transmission we mentioned previously. AB, SK and MB(23%) took the second place of regions' groups, QC, NU and NB, NS, PE and NL with 13% and 10%. However, the weird thing is in ON and NA, ON is the province with the largest population in Canada, there are no occurrences in the top 30, back to rank 80 table, the first ON combinations occurs at 42, so it is possible that the region factor is not the cause of infection or death. Moreover, in the top 30, BC and YT take over half of the table.

b) Big data science on morbidity and mortality of 5 certain regions

In addition to analyzing and examining the total cases of COVID-19, our data science solution also focuses on finding the morbidity and mortality of 5 certain regions in Canada among the 80 combinations of <gender, age, region>. Since we have five pictures about the distribution of death rate among five regions, we plan to focus on Region #3, which represents Ontario and Nunavut. This is because the cumulative number of cases in this region is 519,902. The cumulative number of cases for the rest regions 1,2,4,5 are 10,438/ 411,543/ 321,988 and 1,945 respectively. So we choose region 3 as the number of cases of this region is highest among all the five regions so that our data science solution can be more representative.

See Fig.5 for the death rate distribution of Ontario and Nunavut (Region 3).

Table 1. COVID-19 Statistics Rank 30



COVID -19 Statistic											
RANK	COV_ID	COV_GDR	COV_AGR	COV_REG	NUM_CASE	NUM_ASM	NUM_DTH	NUM_HSP	PER_ASM	PER_DTH	PER_HSP
1	39	1	8	5	76	76	36	8	1	0.474	0.105
2	79	2	8	5	82	82	14	6	1	0.171	0.073
3	34	1	7	5	75	75	10	19	1	0.133	0.253
4	29	1	6	5	124	124	7	14	1	0.056	0.113
5	74	2	7	5	70	70	2	9	1	0.029	0.129
6	69	2	6	5	111	111	2	9	1	0.018	0.081
7	24	1	5	5	147	147		16	1		0.109
8	19	1	4	5	129	129		6	1		0.047
9	64	2	5	5	172	172		7	1		0.041
10	59	2	4	5	159	159		5	1		0.031
11	14	1	3	5	181	181		4	1		0.022
12	54	2	3	5	184	184		4	1		0.022
13	49	2	2	5	153	153		2	1		0.013
14	9	1	2	5	164	164		1	1		0.006
15	4	1	1	5	64	64			1		
16	44	2	1	5	54	54			1		
17	53	2	3	4	29878	27609		121	0.924		0.004
18	58	2	4	4	24050	22210		159	0.923		0.007
19	60	2	5	1	667	612	4	11	0.918	0.006	0.016
20	48	2	2	4	28131	25777		68	0.916		0.002
21	63	2	5	4	17389	15897	108	310	0.914	0.006	0.018
22	56	2	4	2	32486	29380		111	0.904		0.003
23	51	2	3	2	32430	29289		80	0.903		0.002
24	68	2	6	4	10306	9302	237	330	0.903	0.023	0.032
25	61	2	5	2	26055	23453	80	228	0.9	0.003	0.009
26	46	2	2	2	34169	30744		42	0.9		0.001
27	18	1	4	4	23408	20822		239	0.89		0.01
28	65	2	6	1	430	382	12	17	0.888	0.028	0.04
29	23	1	5	4	17847	15786	164	485	0.885	0.009	0.027
30	55	2	4	1	695	614		5	0.883		0.007
SUM					279886	253822	676	2316			

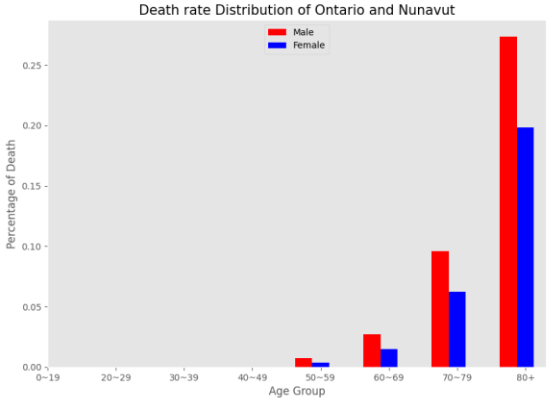


Fig.5 Distribution of death rate for Ontario and Nunavut region.

This bar chart above shows that (a) only the patients' age over 50s may die due to the COVID-19. (b) as patients' age increases, the death rate of cases increases at the same time. These two situations mean that young people and mid aged patients may have a higher resistance, and the COVID-19 is more harmful for old people's life. (c) no matter in the 50s,60s,70s, or even 80s age group, male have a higher percentage than females. For instance, the percentage of death rate for males in their 70s is 0.0958, but the percentage of death rate for females in their 70s is 0.0626 that is less than 0.0958.

Next, we analyzed the morbidity among all the cases. See Fig.6 for the Symptomatic Distribution of Ontario and Nunavut. This chart below reveals that (a) except for the age group of 80+(PER_ASM is 0.6268 in female and PER_ASM is 0.6901 in male) the rest seven age groups' morbidity of females is slightly higher than male. The symptoms of COVID-19 are less likely to appear in male. (b) there is an increasing tendency from young people(0-19) to middle aged people(40s) and then the tendency declines as the age increases.

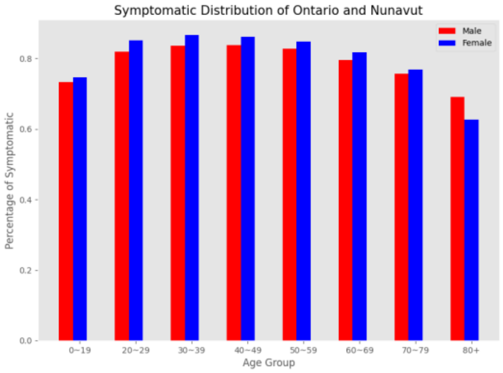


Fig.6. Distribution of Symptomatic for Ontario and Nunavut region.

c) Big data science on three different percentages of hospitalization

Our solution also examines the percentage of hospitalization, the hospitalization rate of symptomatic patients who are in intensive care units, and the mortality of hospitalized patients who are in intensive care units. First of all, see Fig.7 for the ICU Distribution of Ontario and Nunavut.

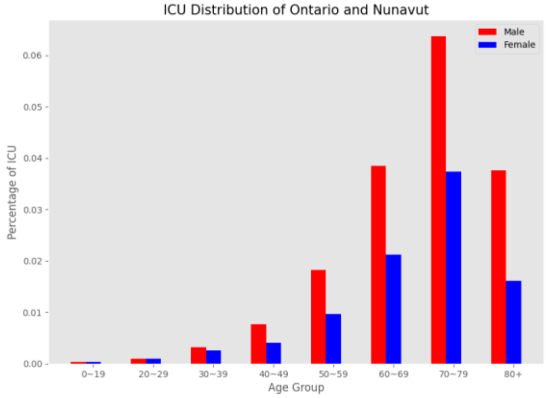


Fig.7. ICU Distribution of Ontario and Nunavut

As the graph above reveals that (a) as the age increases there is a rising trend for the percentage of cases in ICU from the first age group to the 70s. Once patients' are among 80+, there is a significant decrease. We can see that the percentage for male and female is 0.0638 and 0.0374 respectively, but in 80+ the PER_HSP for male and female is 0.0377 and 0.0162 respectively. By checking Fig.4 we find that the death rate in 80+ is the top of all the age groups. This illustrates why there is a huge decline in the percentage of hospitalization as the age increases to 80+. Therefore, this finding means that young people who are infected by COVID-19 have a less need for hospitalization. (b) Among all the age groups, male have a higher percentage of hospitalization than females, which is shown by all the red pillars being higher than blue pillars except the age groups in 0-19, and 20s. Then, we examine the hospitalization rate of symptomatic patients who are in intensive care units. See Fig.8 for ASM and ICU Distribution of Ontario and Nunavut.

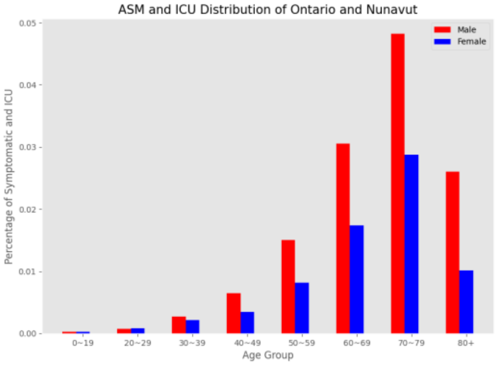


Fig.8. ASM and ICU Distribution of Ontario and Nunavut.

This chart typically focuses on symptomatic cases rather than ignoring whether the patient is symptomatic or not. For those symptomatic cases, we can see that there is also an upward tendency for PER_ASM_HSP among 0-7 , and then has a sharp decrease from 0.0483 to 0.0260 in male and from 0.0356 to 0.0101 in female when the age reaches 80+.

The third situation that we examine is the mortality of hospitalized patients who are in ICU. Let us see Fig.8 for Death and ICU Distribution of Ontario and Nunavut.

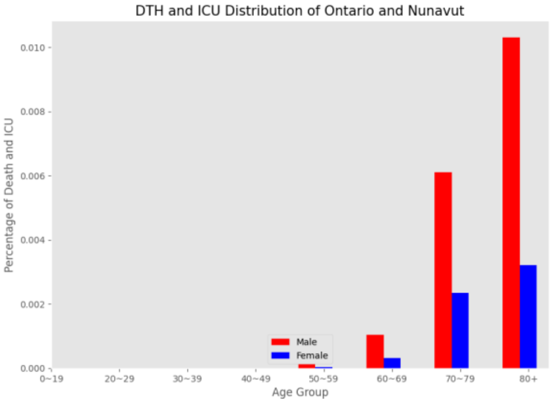


Fig.9 Death and ICU Distribution of Ontario and Nunavut.

Fig.9 shows that (a) for both two genders, the death rate of cases in ICU has an upper trend from 50s-80+.(b) the PER_DTH_HSP of male is higher than the PER_DTH_HSP of females. For instance, 74 (0.6106%) of 12060 COVID-19 cases for males in their 70s is larger than 29 (0.2342%) of 12150 for females in their 70s. Even if the total cases for male is smaller, the death rate for male is still higher than females. This means male will suffer more risks when they are in ICU due to the COVID-19.

d) Big data science on morbidity and mortality in Different Distribution

Besides, Ontario and Nunavut, there are some interesting percentages shown in the charts, for example in British Columbia and Yukon, the Symptomatic Distribution in these two provinces of Male and Female are equal in all the age groups.

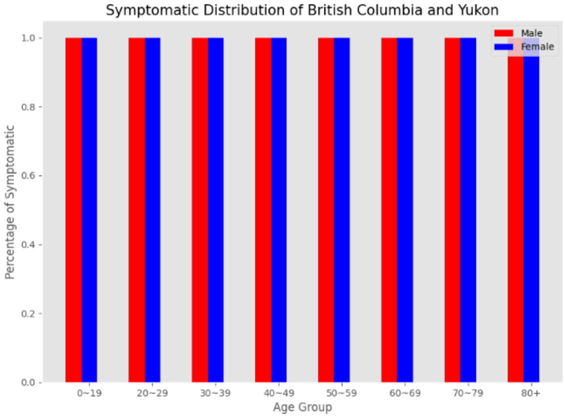


Fig. 10. Symptomatic Distribution of British Columbia and Yukon

Throughout all the charts of Symptomatic Distribution in these five regions the percentage of Female and Male are approximately equal, but those two provinces we mentioned above, British Columbia and Yukon are totally equal.

For Death Rate Distribution of those five regions, throughout all the distributions we can see blue bar almost all lower than red bar, in other words, the death Rate distribution of female is lower than male, for instance, when we see the bar chart shown in Quebec in Fig.11, as we can see all the percentage of female are lower than male, especially, in 80s age group the death rate of female is two thirds of male. But the group of age goes younger the differences between female and male are going less.

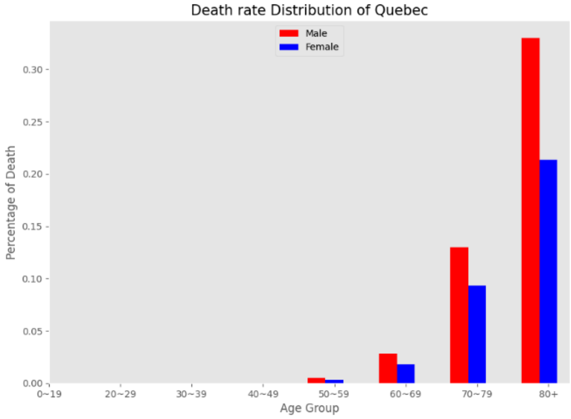


Fig. 11. Death rate Distribution of Quebec

As we can see, the ICU Distribution of these five regions presents that, in group of 70s either male or female shows the highest percentage of all the range of ages, and an interesting thing shows again, almost the height of bar of male are higher than female, in Prairies and the Northwest Territories clearly shows the finding in Fig. 12.

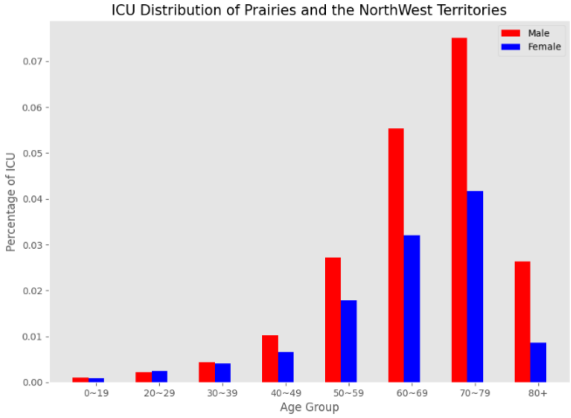


Fig. 12. ICU Distribution of Prairies and Northwest Territories

As the findings we discussed in this section, different distributions show different percentage in the chart, the differences of female and male which means gender does exist. And different types of distributions named ASM, DTH and ICU shows different results.

V. Conclusion

According to this paper, we used Frequent Pattern, and Uncertain Data Mining's U-Viper [12] to be the data science solution. There have been so many researches for the relation with COVID-19 and some attributes, compared with other researches, we focus on the attributes of gender, age and region, then based on the results of 80 groups of data, we got the possibilities of Asymptomatic, Death and Hospital Status. Eventually, we have pie and bar charts of the rate or possibilities of those attributes, by these charts we implemented data visualization. Data visualization helps the public and researchers better to know the facts visually. Then they can analyze those data and charts to come up with better strategies, such as the data of ASM, DTH and HSP. At the beginning of work, we stuck at finding suitable algorithms to solve the answers of questions that we are looking for. Since we want to find the relations with different attributes, finally we decided to use U-Viper [12] to solve, because U-Viper sets different possibilities of different attributes, and we need to use several combinations to get the possible possibilities and find out what factors affect people to get infected mostly. From the findings we have discussed and explained in the section of evaluation, the result and efficiency of algorithms showed their strength and reliability in solving problems. We can use the result to contribute to epidemiology. As we mentioned in the start of the paper, for data scientists we need to use what we learned to make contributions to this world. For COVID-19, such a pandemic has been leading to many problems to modern society, the original social order has been messed up by this. We need to push everything back to normal.

References

[1] Is it so hard to be serious across the country? https://www.chinanews.com.cn/gn/2020/04-28/9170116.shtml

[2] Government of Canada Official Website. https://www.canada.ca/en/public-health/services/diseases/coronavirus-disease-covid-19/covid-alert.html

[3] Government of Canada: Coronavirus disease (COVID-19): Outbreak update. https://www.canada.ca/en/public-health/services/diseases/2019-novel-coronavirus-infection.html

[4] Government of Canada: Statistics Canada. https://www150.statcan.gc.ca/n1/pub/45-28-0001/2021001/article/00020-eng.htm

[5] "Chart 1" of "Fallout from the COVID-19 pandemic: A look back at selected industries in the service sector in 2020" by Marie-Christine Bernard, Graeme Fell and Vivian Lin. https://www150.statcan.gc.ca/n1/pub/45-28-0001/2021001/article/00020-eng.htm

[6] Villanustre, F., Chala, A., Dev, R. et al. Modeling and tracking Covid-19 cases using Big Data analytics on HPCC system platform. J Big Data 8, 33 (2021). https://doi.org/10.1186/s40537-021-00423-z

[7] Nickel NC, Clark W, Phillips-Beck W The COVID Equity Team, et alDiagnostic testing and vaccination for COVID-19 among First Nations, Metis and Inuit in Manitoba, Canada: protocol for a nations-based cohort study using linked administrative dataBMJ Open 2021;11:e052936. doi: 10.1136/bmjopen-2021-052936. https://bmjopen.bmj.com/content/11/9/e052936.citation-tools

[8] Preliminary dataset on confirmed cases of COVID-19, Public Health Agency of Canada, 2020-2021, 2021Nov12 version. https://www150.statcan.gc.ca/n1/en/catalogue/13260003 https://doi.org/10.25318/132600032020001-eng

[9] Individual and community-based measures to mitigate the spread of COVID-19 in Canada, Government of Canada. https://www.canada.ca/en/public-health/services/diseases/2019-novel-coronavirus-infection/health-professionals/public-health-measures-mitigate-covid-19.html

[10] National Collaborating Centre for Environmental Health, "The Basics of SARS-CoV-2 Transmission," 21 January 2021. [Online]. Available: https://ncceh.ca/documents/evidence-review/basics-sars-cov-2-transmission [Accessed March 4 2021]

[11] Mohammad Sarmadi, Samaneh Kakhki, Maryam Foroughi, Tahere Sarboozi Hosein Abadi, Somayyeh Nayyeri, Vahid Kazemi Moghadam, Mahsan Ramezani, Hospitalization period of COVID-19 for future plans in hospital, British Journal of Surgery, Volume 107, Issue 10, September 2020, Pages e427–e428, https://doi.org/10.1002/bjs.11871

[12] Leung, C.K.S., Tanbeer, S.K., Budhia, B.P. and Zacharias, L.C., 2012. Mining probabilistic datasets vertically. In *Proceedings of the 16th International Database Engineering & Applications Sysmposium* (pp. 199-204).

[13] Sagiroglu, Seref, and Duygu Sinanc. "Big data: A review." In *2013 international conference on collaboration technologies and systems (CTS)*, pp. 42-47. IEEE, 2013.

[14] Anuradha, J., 2015. A brief introduction on Big Data 5Vs characteristics and Hadoop technology. *Procedia computer science*, *48*, pp.319-324.

[15] Government of Manitoba COVID-19 Cases: https://www.gov.mb.ca/covid19/index.html

[16] Leung, Chen, Y., Shang, S., & Deng, D. (2020). Big Data Science on COVID-19 Data. *Proceedings - 2020 IEEE 14th International Conference on Big Data Science and Engineering, BigDataSE 2020*, 14–21. https://doi.org/10.1109/BigDataSE50710.2020.00010