CMPT 732 Project Report

1. Motivation

Mental Health is one of the most important determinants of the quality of life. It doesn't only have a huge impact on a person's overall well-being and functionality, but it also affects their happiness, resilience, and has an impact on the people around them. Therefore, it is critical to identify the factors behind mental health and utilize this understanding to take targeted steps towards improving the same.

So in this project, we aim to investigate how the most basic socio-economic factors like the level of education and household income affect mental health. We do this analysis province-wise, which also helps us get an overall idea of the Mental Health Landscape of Canada.

1. This will help better our understanding of the current mental health situation and will enable the decision-makers to identify the right leverage in addressing the mental health problems. (to make the mental health situation better)
2. It will empower people with a better understanding of factors that affect mental health so that they can make better decisions that help them improve their wellbeing.
3. It will help direct incoming immigrants with special mental needs, to a better place where they can flourish and get help.

2. Methodology

We aim to find the answers to the following questions:

1. What is the overall mental health landscape of Canada?
2. What is the correlation between household income and mental health?
3. What is the correlation between the level of education and mental health?

For the same, the following methodology has been followed. We analyze the dataset obtained from the Canadian health survey done annually since 2015. This survey involves people giving a binary response to a number of questions like whether or not they consider themselves in the category of - 'Perceived mental health, very good or excellent', 'Perceived mental health, fair or poor, 'Diabetes', 'Fruit and vegetable consumption, 5 times or more per day', 'Perceived life stress, most days quite a bit or extremely stressful', and 21 more such questions, 6 in total related to mental health and rest about general health. The dataset of this survey contains an aggregate of all the responses, based on the income category, education level, and geographical location (province) of respondents, and health characteristics. A row of the dataset (excluding some non-relevant columns) can be seen in Fig 1.

REF_DATE	GEO	Selected characteristic	Indicators	Characteristics	VALUE
2015	Ontario	Household income, first quintile	Perceived health, very good or excellent	Percent	51.4

Fig 1: Sample row from the data set.

First, in order to study the relation of Mental Health with any other factor, we needed to quantify Mental Health. In the original dataset, we had 27 general health-related characteristics scores. So in order to create a 'Mental health score' for a particular province's income group/education group in a particular year, we aggregated the mental health-specific characteristics values, which were the following 6 characteristics:

1. Perceived mental health, very good or excellent
2. Perceived mental health, fair or poor
3. Perceived life stress, most days quite a bit or extremely stressful
4. Mood disorder
5. Sense of belonging to the local community, somewhat strong or very strong
6. Life satisfaction, satisfied or very satisfied

We tried in total 3 approaches to aggregate the above 6 values and find a mental health score. These are as follows:

1. Weighted Sum:

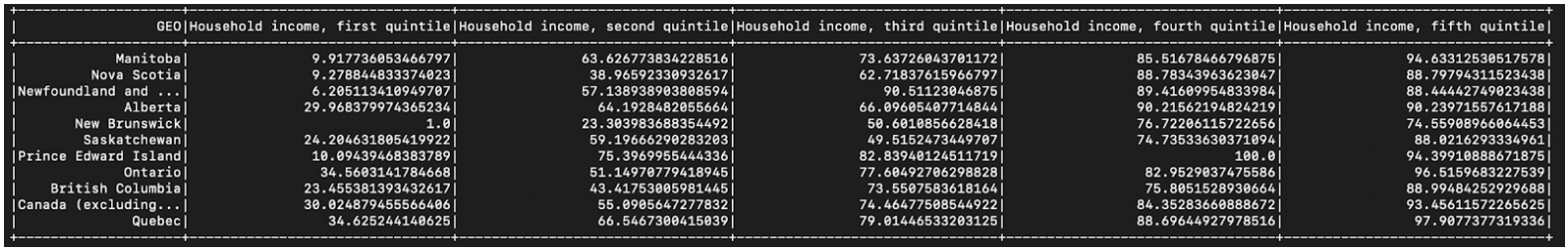
In this approach, we find the weighted sum of all the characteristics above. Weight represents how much a mental health characteristic is important towards good mental health. Out of all these characteristics, 'Life Satisfaction' is considered to have a very strong association with Mental health (Reference: https://www.frontiersin.org/articles/10.3389/fpsyt.2019.00419/full). So in order to find the importance of the other features (i.e their weights), we find the correlation coefficient of all the mental health characteristics with 'Life Satisfaction'. The obtained weighted sum serves as our mental health score.

GEO	Household income, first quintile	Household income, second quintile	Household income, third quintile	Household income, fourth quintile	Household income, fifth quintile
Manitoba	9.24040699005127	67.26025390625	76.69117736816406	85.96367645263672	93.45252990722656
Nova Scotia	3.206308126449585	35.762081146240234	65.84024047851562	92.8214340209961	82.39666748046875
Newfoundland and Labrador	16.043338775634766	62.21135330200195	95.64610290527344	86.7913818359375	85.556640625
Alberta	24.751392364501953	66.07891845703125	66.20958709716797	92.26095581054688	92.02218627929688
New Brunswick	1.0	17.60066795349121	53.2205924987793	74.46809387207031	75.57392120361328
Saskatchewan	30.33203125	59.651649475097656	48.72065734863281	71.63843536376953	87.79559326171875
Prince Edward Island	3.0138282775878906	78.49480438232422	83.52808380126953	95.04623413085938	87.24015045166016
Ontario	30.04503631591797	51.485870361328125	77.85189056396484	84.03160095214844	97.7624740600586
Canada (excluding territories)	29.499849319458008	55.50398254394531	76.21495056152344	85.43017578125	94.67644500732422
British Columbia	21.459341049194336	39.571495056152344	75.4100341796875	74.756591796875	88.85747528076172
Quebec	41.93935012817383	69.13532257080078	83.2826919555664	90.82691955566406	100.0

Fig 2: Mental health scores generated for provinces based on income groups using PCA.

2. PCA:

In this approach, we simply use PCA to reduce the above 6 features (characteristics) to 1 feature which serves as our mental health score.

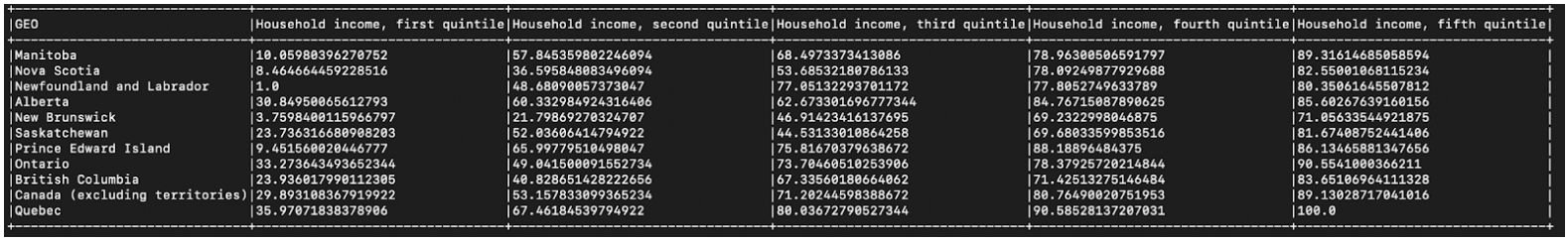


GEO	Household income, first quintile	Household income, second quintile	Household income, third quintile	Household income, fourth quintile	Household income, fifth quintile
Manitoba	9.917736053466797	63.626773834228516	73.63726043701172	85.51678466796875	94.63312530517578
Nova Scotia	9.278844833374023	38.96592330932617	62.71837615966797	88.78343963623047	88.79794311523438
Newfoundland and ...	6.205113410949707	57.138938903808594	90.51123046875	89.41609954833984	88.44442749023438
Alberta	29.968379974365234	64.1928482055664	66.09605407714844	90.21562194824219	90.23971557617188
New Brunswick	1.0	23.303983688354492	50.6010856628418	76.72206115722656	74.55908966064453
Saskatchewan	24.204631805419922	59.19666290283203	49.5152473449707	74.73533630371094	88.0216293334961
Prince Edward Island	10.09439468383789	75.3969955444336	82.83940124511719	100.0	94.39910888671875
Ontario	34.5603141784668	51.14970779418945	77.60492706298828	82.9529037475586	96.5159683227539
British Columbia	23.455381393432617	43.41753005981445	73.5507583618164	75.8051528930664	88.99484252929688
Canada (excluding...)	30.024879455566406	55.0905647277832	74.46477508544922	84.35283660888672	93.45611572265625
Quebec	34.625244140625	66.5467300415039	79.01446533203125	88.69644927978516	97.9077377319336

Fig 3: Mental health scores generated for provinces based on income groups using Weighted Sum

3. Weighted PCA:

In this approach, we find the weights in the same way as done in the weighted sum approach and use those weights to find a weighted PCA (multiply the features by weights and find PCA to reduce features to 1 feature that serves as mental health score.)



GEO	Household income, first quintile	Household income, second quintile	Household income, third quintile	Household income, fourth quintile	Household income, fifth quintile
Manitoba	10.05980396270752	57.845359802246094	68.4973373413086	78.96300506591797	89.31614685058594
Nova Scotia	8.464664459228516	36.595848083496094	53.68532180786133	78.09249877929688	82.55001068115234
Newfoundland and Labrador	1.0	48.68090057373047	77.05132293701172	77.8052749633789	80.35061645507812
Alberta	30.84950065612793	60.332984924316406	62.673301696777344	84.76715087890625	85.60267639160156
New Brunswick	3.7598400115966797	21.79869270324707	46.91423416137695	69.2322998046875	71.05633544921875
Saskatchewan	23.736316680908203	52.03606414794922	44.53133010864258	69.68033599853516	81.67408752441406
Prince Edward Island	9.451560020446777	65.99779510498047	75.81670379638672	88.18896484375	86.13465881347656
Ontario	33.273643493652344	49.041500091552734	73.70460510253906	78.37925720214844	90.5541000366211
British Columbia	23.936017990112305	40.828651428222656	67.33560180664062	71.42513275146484	83.65106964111328
Canada (excluding territories)	29.893108367919922	53.157833099365234	71.20244598388672	80.76490020751953	89.13028717041016
Quebec	35.97071838378906	67.46184539794922	80.03672790527344	90.58528137207031	100.0

Fig 4: Mental health scores generated for provinces based on income groups using Weighted PCA

Though the results obtained by all these methods were very close (as shown in the screenshots above), we went with the weighted sum approach, since it seemed like the most viable, direct, and intuitive approach which took into account the relative importance of all the characteristics.

All the implementations have been done in PySpark using Dataframes. For correlation and PCA, pyspark.ml package has been used. The computation is being done in real-time on EMR with data and the results obtained stored on S3. A UI has been created using flask. It can be used to start/stop our EMR. It also provides an option to choose from the following 6 queries:

1. Mental Health Analysis based on Household Income
2. Mental Health Analysis based on Highest Level of Education
3. Mental Health Landscape of Canada
4. Mental Health Analysis; variations by Income Groups over the years
5. Mental Health Analysis; variations by levels of education over the years
6. Mental Health trend over the years in Canada

Once a query has been chosen, the corresponding step is added to the already running EMR, where the script is fetched from S3. The EMR writes the output in one CSV on S3. The generated output in the CSV is in the format needed to generate the visualization. The CSV is read onto a pandas data frame in the local server which is used to generate an interactive map, heatmap, and bar graphs using Plotly. These visualizations are displayed to the user on the UI.

Reasons for implementation methodology used:

1. PySpark: Doing all the data pre-processing and querying using PySpark, made the code equipped to scale well to any size of data.
2. AWS: AWS compared to other cloud providers, has a more robust and well-developed python API
3. EMR and S3: EMR with S3 provides an efficient spark job computation platform.
4. Generating graphs at runtime on the server: We chose to create the visualizations at run time on the server instead of generating them on EMR. This was done because the final data frame used to generate the graphs is not Big Data. All the Big Data processing has already been done when EMR generates the final CSV. So we generate the visualization from this CVS on the server and not cluster to prevent unnecessary computation on EMR.

3. Problems faced

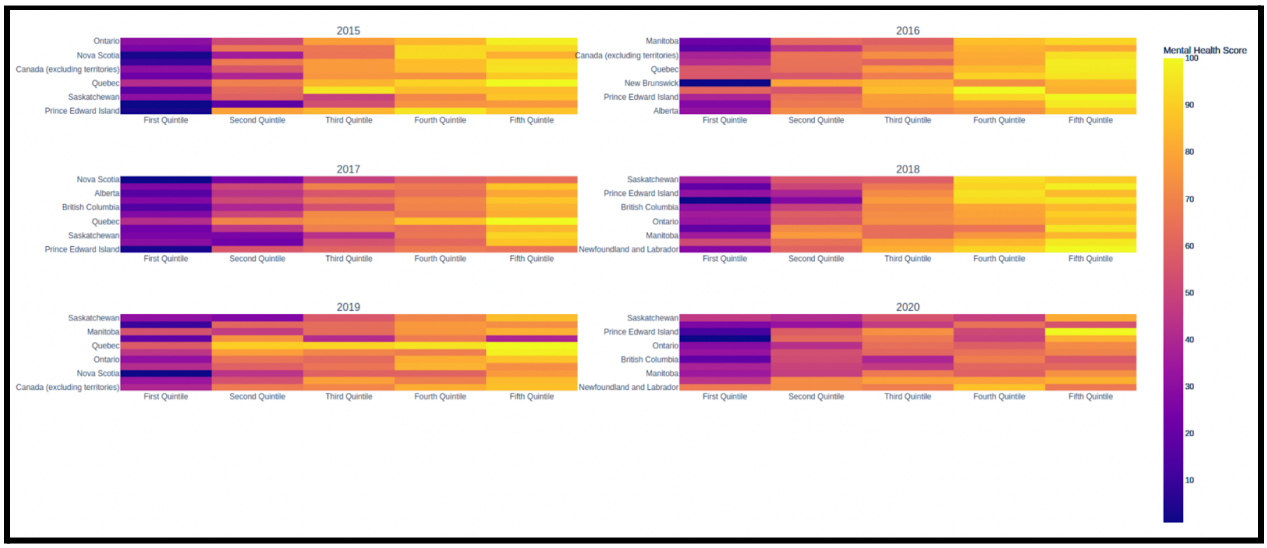
1. Understanding our data: The data obtained from Stats Canada is a combination of data obtained in the years 2015 to 2020. Further, this data exists for all of the Canadian provinces. This data is then classified into two classes, household income and level of education. For these categories, we have the percent value of the population that voted for each of the characteristics. Because of this level of complexity, it was a challenge to understand the data first and obtain it in a format that was possible to work with. Using a pivot operation this problem was solved.
2. Handling the null values: Given the fact that the volume of our data is limited to handle the null values, it became evident that replacing them with averages was the best solution as removing those rows completely from the data would result in a much smaller data set.
3. Quantifying mental health: This was the biggest challenge faced in our project. There is no formal way to define and quantify mental health. In order to solve this problem, we referred to a research article that stated the importance of life satisfaction on mental health. So to overcome the problem of quantifying mental health, we proceeded by using life satisfaction as an indicator and found the importance of the other factors by the method of correlation.
4. Interpreting the results of PCA: After obtaining the results from PCA and weighted sum, it was observed that PCA had contradicting results as compared to the weighted sum. This helped us understand the importance of the magnitude of the features and not the sign (-ve or +ve) since the sign just represents the direction of a feature with respect to the principal component whereas the absolute value gives us the magnitude. After taking the absolute values, the results obtained from PCA were similar to the results obtained from weighted sum and weighted PCA.
5. UI challenges: Since the aim of the project was to perform all the analysis in real-time and on the cloud, it was clear that every time a question was selected, the analysis scripts would need to be triggered on EMR. In order to do this, it did not make sense to have EMR always running. To

solve this challenge, a middleware was written to allow the user to start and stop an EMR right from the UI.

6. Choice of the sum over average: Initially, a weighted average seemed to be the right approach to calculate the mental health score. But, upon closer inspection of the weights obtained by correlation, it became evident that an average would not take into consideration all the negative weights. Therefore, a weighted sum was chosen as a better representation of an overall mental health score.
7. Choice of not proceeding with clustering: Performing a simple clustering could have been an easy and efficient way of quantifying mental health as either good or bad. But, this was not enough for the scope of this project as quantifying the mental health into a score was required to get a correlation analysis of these scores with the factors in question.

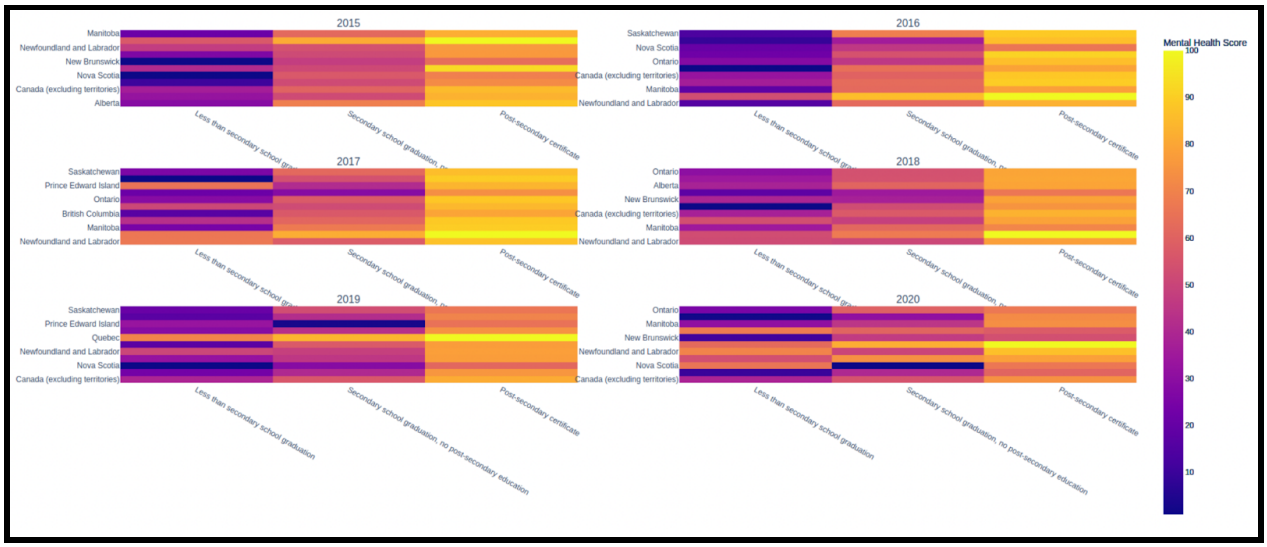
4. Results: Outcomes, Data Analysis, and learnings from the project

Some of the results that we obtained can be seen below. Each image is followed by the inference that can be drawn from the graphs. All the results that we obtained can be seen here.

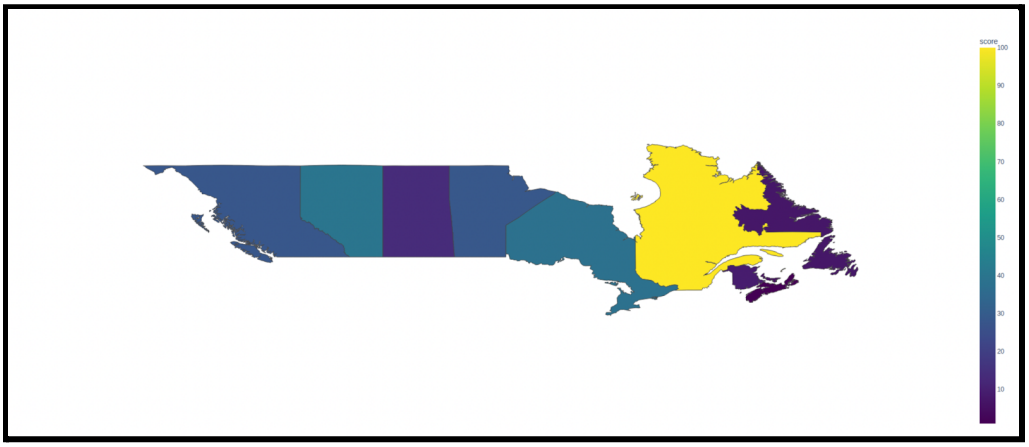


In this image, we see the heat maps of different years. On each heat map, we have provinces on the y-axis and income categories on the x-axis. The color represents the mental health score, with yellow representing the highest score and darker colors representing lower scores.

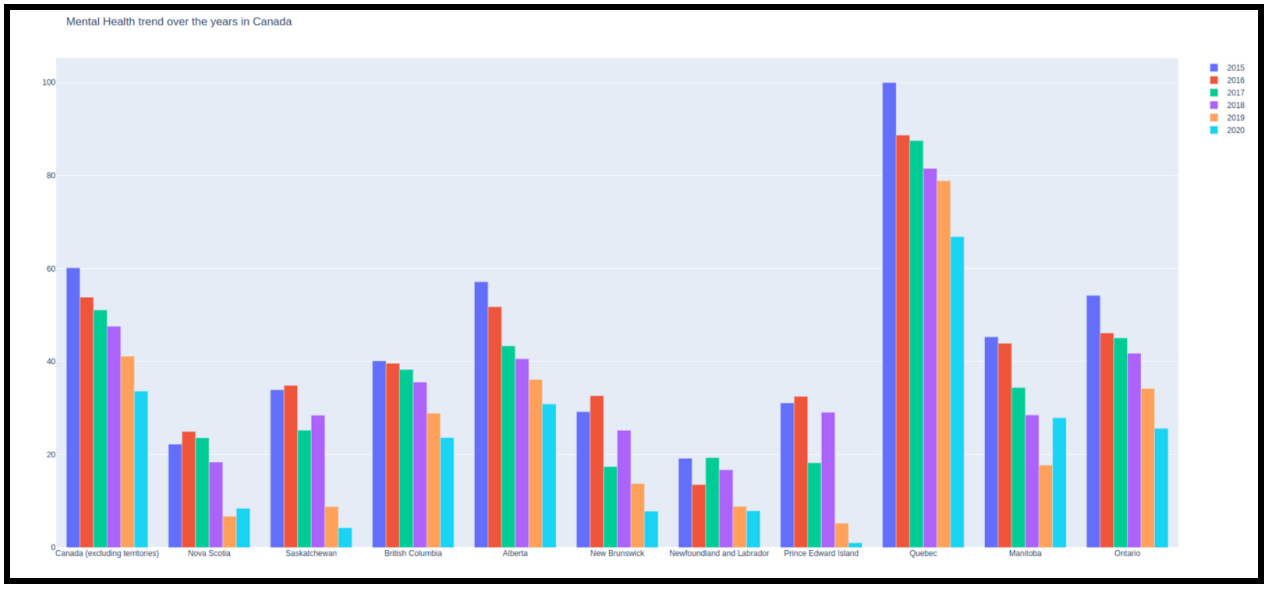
In all the heatmaps it can be clearly seen that the yellow color concentrates towards the right, which is the higher income group, representing a higher mental health score for them. This trend is evident in all years for all provinces, suggesting that income has a positive influence on mental health



These heatmaps show the education category on the x-axis with education levels low to high from left to right. It can be clearly seen that the yellow color concentrates towards the right, which is the higher education level group, representing higher mental health scores for them. This trend is evident in all years for all provinces, suggesting that education level also has a positive influence on mental health



This represents the mental health landscape of Canada with the color of each province representing their mental health score



In the bar graph above, it can be seen that mental health reduces for each consecutive year for almost every year.

Learnings from the implementation:

1. We got to learn that income and education positively affect mental well-being and life satisfaction very strongly.
2. For some reasons, the overall mental health of almost all the provinces has been reducing for all consecutive years.

5. Project Summary

1. **Getting the data**: Acquiring/gathering/downloading. **1**
2. **ETL**: Extract-Transform-Load work and clean the data set. **3**
3. **Problem**: Work on defining the problem itself and motivation for the analysis. **2**
4. **Algorithmic work**: Work on the algorithms needed to work with the data, including integrating data mining and machine learning techniques. **4**
5. **Bigness/parallelization**: Efficiency of the analysis on a cluster, and scalability to larger data sets. **3**
6. **UI**: A user interfaces to the results, possibly including web or data exploration frontends. **3**
7. **Visualization**: Visualization of analysis results. **3**
8. **Technologies**: New technologies learned as part of doing the project. **1**