**CIND820 – Capstone Project**

**Refined Abstract**

**Katie Schilling**

**501130072**

**Supervisor: Tamer**

**February 14, 2022**

**Vital Events Predicting, Forecasting and Introduction of Possible Causality**

Every year the registrar general collects and releases information on vital events in Canada. Based on research, the vital events data is collected to monitor, and predict future population growth, marriages, deaths and stillborn births. *(https://www.statcan.gc.ca/en/about/relevant/vscc/faq)*

Much of this information can be used to determine what resources will be needed in the future, whether the population is increasing, decreasing, or holding steady, and foresee any public health trends emerging, as we have seen over the past 2 years with the introduction of COVID-19. All of this information is important for understanding and maintain balance of our system as a whole.

In addition to focusing on the births, deaths, marriages and stillborn births, I also want to introduce the number of positive COVID-19 cases, month over month. The reason for this is to see if there is a direct correlation between the increase or decrease of these numbers and COVID-19, and what long term impact it may have on our growth or decline.

**Goals of the analysis:**

looking at the current trends for all vital events data, predict the future population, and trends using time series forecast. Are there any alarming trends, or changes in variables that show an impact on either the growth or decline of these vital events?

When analyzing the vital events data, using forecast for time series, what is the trend, prior to adding in the numbers for COVID-19?

How does the Forecast for 2020-to date, compare to the actual data?

Are we seeing a trend over time of a decrease in stillborn births and deaths as our healthcare system advances, when COVID-19 is not a factor?

Based on history, what numbers did we expect to see in 2020 and 2021? How did those numbers differ with the introduction of COVID-19?

Using the Granger-Causality test in R, what is the causality of each variable when compared to the others?

Are there any specific causality correlations that are import and fall within the null hypothesis?

What do these analysis tell us about the data, and it’s ability to predict future trends, when all the data is used in conjunction to one another?

**Tools and Techniques**

R-Studio

Scatter, bar, and box plots, histograms, to visualize the time-series trends

Decompose in R, to see the breakdown of the time series data

Correlation matrix

Causality test in R using the Granger-causality test

Normalization of the data

Training and test data sets for model validation

Naïve Forecasting Method

Forecasting and linear model prediction

Autoplot to get a glimpse of all variables within the same graph as well as a glimpse at the future predicitons

EDA on each dataset independently, and then another EDA on the final dataset when both have been combined.

**Data Information**

The dataset being used for the vitals information can be found at:

<https://open.canada.ca/data/en/dataset/e100b1d5-b7fb-40fd-b7a0-50289d13c574>

In order to introduce COVID-19 positive cases, to infer causality from this data, I need to pull data from a secondary source, and match the data to the existing dataset. Initially this proved to be a difficult task, as much of the information on COVID-19 is not publicly available or the raw data is coded, and unclear. After research I found the following data:

https://health-infobase.canada.ca/covid-19/epidemiological-summary-covid-19-cases.html

While the timeline on the 2 data sets does not correlate as is, through pre-processing of the data, I will be able to use the totals, per month, to match the data to the vitals event data. The only data required from this dataset will be prname (province name), to which I will only be using the ones listed under Canada as it encompasses all provinces, and the column numtoday (number today) which is the number of new covid cases since previous update. I will also be using the date, so that I can assign the sum of each month to the proper month, to correlate with the vital events dataset. The month of January, 2020 will be excluded as it is an incomplete month and will skew the results.

To ensure I am not removing 90% of the data due to some rows having NA’s, once both datasets have been combined, I will assign any Covid Positive rows that contain NA with 0, and once that is complete, any remaining rows that contain any NA’s will be removed