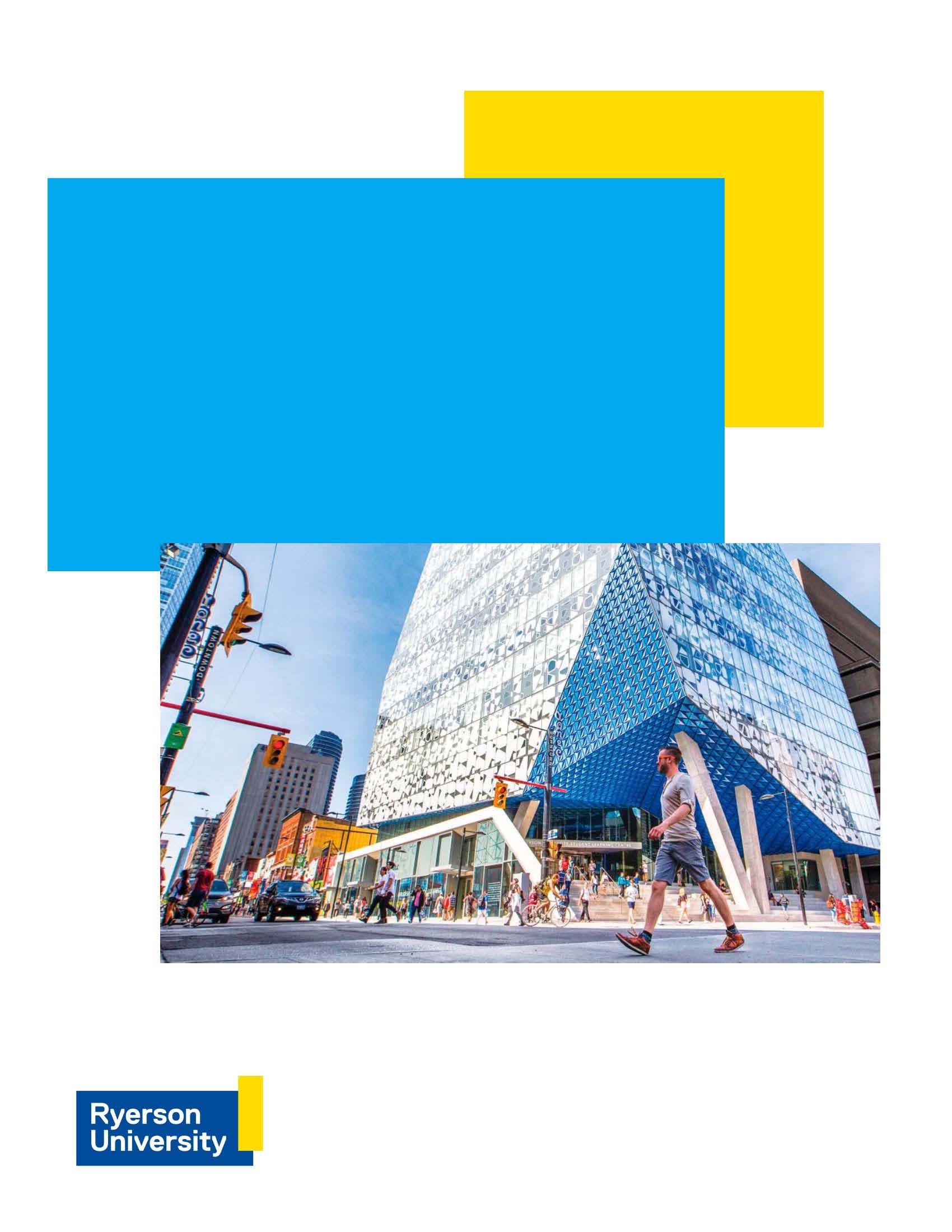
Forecasting Hospital Bed Capacity 

ZhiLei Xiao

501012295

Derya Kici

April 4, 2022

**Table of Contents**

[**Abstract 3**](#_heading=h.gjdgxs)

[Objectives 4](#_heading=h.30j0zll)

[Significance 4](#_heading=h.1fob9te)

[**The Literature 7**](#_heading=h.3znysh7)

[The Working Dataset 10](#_heading=h.2et92p0)

[**Citations 13**](#_heading=h.3dy6vkm)

# 

# Abstract

Public Health Ontario is an online government resource which provides province-wide health data for study. The aim in making the information available is to allow individuals and organizations to conduct scientific analyses which can in turn be used to inform policies and practices to improve citizen’s health.

One such area of interest is the ability of hospitals to provide adequate care to those who need it. While hospitals require a certain amount of hospital beds to be kept available for patient use, the Ontario Ministry of Health mandates the number of beds offline at any given time (Ontario, 2020). This is especially true after situations such as overfilled hospitals, following the third wave of COVID-19. There is a certain amount of hospital beds that are kept online and thus, available, in a unit at a time, each requiring their own allocation of resources to maintain the status of preparedness for a patient to occupy at a moment’s notice. Hospitals may also have a number of beds offline to ensure efficiency of resource usage so that the cost of maintaining unused available beds do not impede on the function of the unit. What this study seeks to find is the optimal number of beds to keep online in Ontario Intensive Care Units (ICU) during the COVID-19 pandemic. The aim is to achieve this by using known methods of machine learning on historical time series data from the ICU occupancy dataset (Ontario, 2020) gathered by the Ministry of Health Ontario 2020 to predict the days when high volumes of patients requiring critical care emerge. Univariate time series analysis and forecasting using the ARIMA model, and the neural networks inspired by the work of Zhang (2003) will be used to answer the research question of what are the daily occupancy demands of Ontario's critical care units during the covid pandemic.

### Objectives

The objectives of this project are:

* To explore and analyze hospital ICU occupancy during the pandemic in Ontario
* To identify patterns and trends in ICU hospitalizations along the progression of the virus
* To develop machine learning models that predict the daily occupancy demands of Ontario's critical care units using patterns found in earlier stages of the pandemic
* Provide recommendations for hospitals to increase their preparedness for high patient volumes and needs in the future

### Significance

If occupancy patterns of patients for these hospital beds can be tracked, studied, and understood, predictions can be made more accurately for the most efficient number of beds ready to go online at any time. This will be particularly useful when these predictions can be made in advance to anticipate rush times or downtimes to make best use of hospital resources. In this way, machine learning can be useful. Specific data of past historical occupancy rates for patients with COVID, and data during the different stages of the pandemic, can be used in time series data to try and accurately predict the amount available of beds needed overall. In addition, results of this machine learning can help set a stronger precedent for preparations in cases of future epidemics, natural disasters, and extreme cases of hospital overflow.

# The Literature

#### Industry Background

The type of literature referenced is firstly from a field or industry perspective, looking to provide insight on the infrastructure of hospitals and how they approach capacity management, as well as Covid-19 news and reports which can be used to provide more insights on the applications of the research question. From the background provided by the 2016 report of care in Canadian hospitals, the overall state of intensive hospital care has been increasing in necessity due to the aging overall population and an increase in illness severity. While this report may be used to draw a general background understanding it was conducted in 2016 during regular non pandemic times. I expect the ICU data (Ontario 2020) during the COVID-19 period would have an upward shift in volume across the board as well as more dramatic increases of ICU occupancy during surges, behaviors such as plateaus, and decreases as vaccines roll out to the population.

Interestingly the Canadian Institute for Health information's report on December 9th, 2021, regarding COVID-19 impact on hospital services reports add that with each wave two and wave three hospital impatient admissions were actually lower in comparison to the pre-pandemic period. However, the report also states that non COVID-related admissions for procedures were delayed and cancelled with the rising numbers of COVID infections. Respiratory conditions were found to rise in accordance with infection rate as well as ICU admission and ventilator demand increased.

More recent statistics in the Weekly Epidemiological Summary for COVID-19 in Ontario: Focus on January 30, 2022 to February 5, 2022 indicates an overall decrease in outbreaks. Public Health Ontario does warn, however, that the ICU admission and death rates may be under reported due to detection may not have happened in time before death or infection after follow up completion. Public Health Ontario’s January 31, 2022 report *on Omicron in Ontario: Risk Analysis for Approaching Public Health Measures in Winter 202*2 predicts a peak in Omicron infection around 4 weeks, states that currently infection rates are low but warns that there is massive uncertainty regarding the trajectory of the variant in Ontario specifically highlighting higher risk level during the winter of 2022. It is important to examine the wider context of the pandemic in relation to the ICU hospital occupancy in that there is the implication that the trends found in the anticipated overall state and infection rate of the virus would affect ICU rates; during infection outbreaks the volume of patients is expected to rise and along with it, a corresponding increase in ICU beds to meet the demand will be necessary.

#### Literature Applications of Time Series Methods in Hospitals

Secondly, academic literature focusing on the applications of predictive data analytics on time series with relevant methodology in the medical field is consulted. The research question of finding the daily occupancy demands of Ontario's critical care units specifically attempts to address patterns in ICU capacity during the pandemic. In some sense, it will help consolidate what hospitals already know and struggle with (in terms of ensuring the correct number of beds available) into relevant data that can be used to make quantitative predictions. It will give insight not only on the rate of patients going in and out of intensive care, but also overall patients’ needs as the coronavirus evolves.

Comparing this proposed methodology with what other practitioners have done in the consulted literature for the purpose of this endeavor, the research pertaining to the specific aim of forecasting ICU bed capacity during a pandemic is unavailable at this time. However, several researchers have conducted related studies, Jones et al. (2009) used the Vector Autoregression (VAR) model to forecast the demands and to analyze the relationship between 8 different resources within the emergency department at three different hospitals. While VAR was shown to be effective in Jones et al.’s study, it would not be as suitable of an approach for this specific research due to the nature of the research question and dataset required. Jones et al. (2009) required datasets which had multiple variables for different hospital resources which could be used to study their relationships among the variables as they change over time unlike the dataset for this project which only focuses on one variable changing upon itself throughout time. However, it does illuminate the option and provides insight on the path of using univariate VAR as a forecasting method which could be applied to this research question. Furthermore, the study by Batal et al. (2001) uses stepwise linear regression to isolate "significant" variables, using which the authors developed their own prediction equation to forecast patient visits depending on weather conditions. In addition to the multivariate differences in Jones et al. (2009) , the methods in Batal et al. (2001) also cannot be applied in the same way, and because a main focus of the topic is to use machine learning, the specifically developed prediction equation would not apply either. There is no doubt, however, that provided with a different dataset with more recorded details on outside variables such as weather, hospital resources, and/or patient information in addition to ICU occupancy it would be beneficial and worth analyzing through employing the methods used in Jones et al. (2009) and Batal et al. (2001) papers as they have shown how to include and consider the influences of multiple variables in their forecasting methodology.

Two studies examining predictive analytics in hospital care units report similar results in their model usage. Kim et al. (2014), and Tandberg and Qualls (1994), both used ARIMA methods for forecasting.

The two main papers used to guide the project are Zhang (2003) and Khashei and Bijari (2010). In these studies, the authors both examine the use of autoregressive integrated moving average (ARIMA) model and the artificial neural networks (ANN) for timeseries forecasting. The study by Zhang (2003) illuminates the strengths in which the ARIMA model has been diploid as well as the strength and circumstances in which ANN have been used and found to be most effective and appropriate. They then combine the two methods in attempt to utilize the advantages in each model to compensate for the other model’s shortcomings. His research and insight on the models state that ARIMA is linear in its prediction of future values and are typically constrained to linear behaviors in past observations. due to the nonlinearity of some datasets, the performance of linear prediction models such as ARIMA may not most effectively capture those nonlinear behaviours in its forecasting. To address the issue of nonlinearity, Zhang (2003) suggests the use of ANN. they assert that artificial neural networks R one type of model suitable for “approximating various nonlinearities in the data.” (Zhang, 2003).

## Data Description

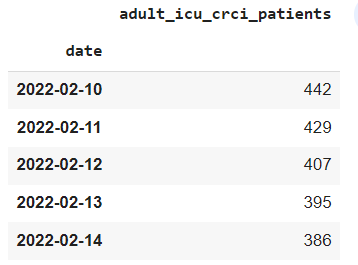
The working dataset is the [*Availability of Adult and Pediatric ICU Beds and Occupancy For Covid-Related Critical Illness (CRCI)*](https://data.ontario.ca/dataset/availability-of-adult-icu-beds-and-occupancy-for-covid-related-critical-illness-crci) obtained from the Ontario Data Catalogue. It is a government site that provides datasets for the whole province of Ontario. The dataset contains observations which were recorded on a daily basis for the period of May 01, 2020 to February 14, 2022.

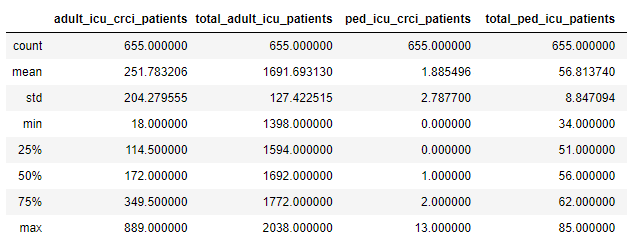
The variables include the date and eight variables measuring the number of persons, adult or pediatric in the following four categories:

* number of patients in ICU/pediatric ICU for COVID-related critical illness (CRCI)
* number of patients in ICU/pediatric ICU for non-CRCI reasons
* number of patients ICU/pediatric ICU beds that are unoccupied
* total number of patients in ICU/pediatric ICU for any reason

A brief examination of the descriptive statistics of the dataset in the attached Jupyter Notebook finds the following Figure 1 which shows the values of the first and five days of the data indicating the period and the starting and ending values. Table

Description automatically generated*Figure 1. First five observations of initial four dataset variables.*

*Figure 2. Last five observations of main variable of interest*

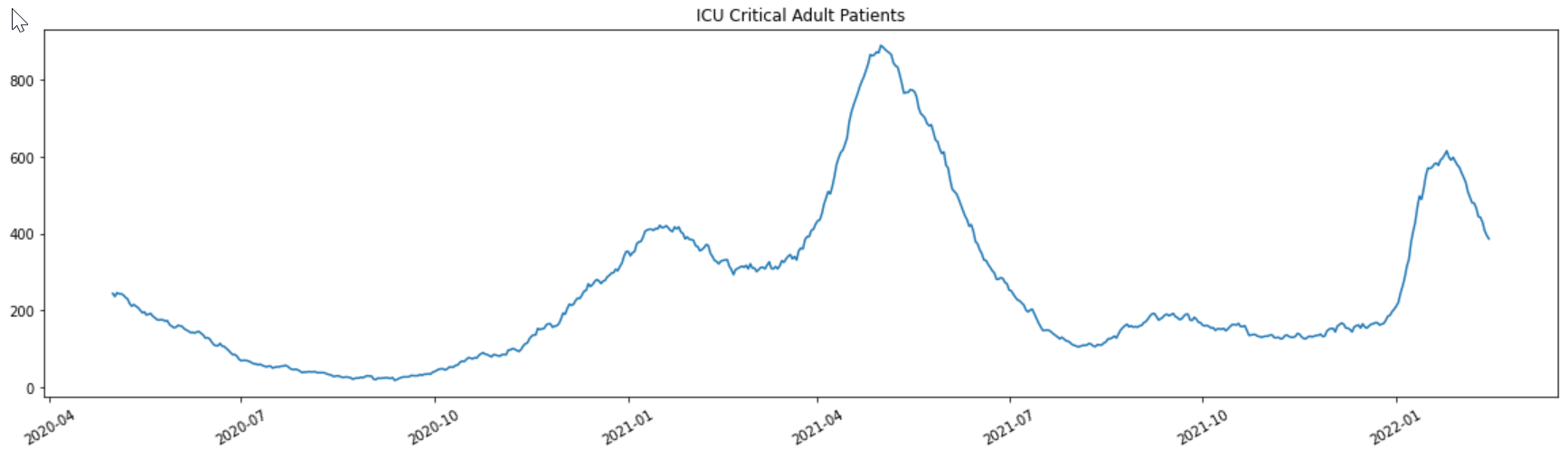
**Exploratory Data Analysis**

*Figure 3. Descriptive statistics table for dataset variables.*

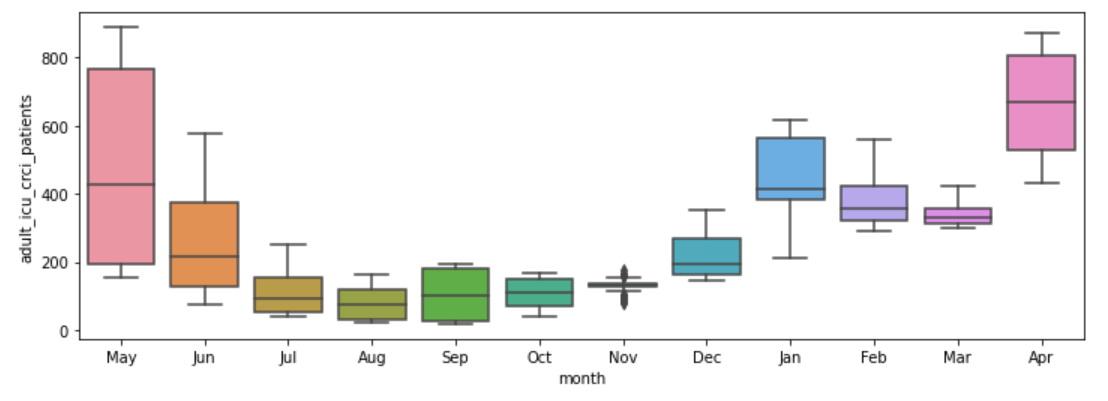
While there are four variables to the dataset, the main category of focus is the number of non-pediatric patients in the ICU for CRCI. This is because in terms of attribute elimination, we will want to focus on the patients in critical conditions rather than amount of beds available or total because it will be more useful to know now many are absolutely necessary to set aside for CRCI patients within the ICU, and the focus will be on adult CRCI patients as the adult population including seniors has been most fatally affected by COVID. Therefore, the main variable of focus will be adult CRCI patient numbers. As the objective is univariate time series analysis, the statistical method means all other attributes can be eliminated. This means the dataset will be subsetted into a univariate time series with the date as the index and the number of non-pediatric patients in the ICU for CRCI as the variable. A brief examination of the total number of patients in ICU for any reason is conducted to contextualize the data in the real world in order to check for province-wide shortages. This is because these best represent the quantity of those who require high resources ventilators in CRCI, and those who still need to be in the ICU regardless of infection status. Figure 2 shows the basic descriptive statistics drawing attention to the first column which will read the amount of adults in CRCI were recorded 655 across 655 days, mean of 251.78 people, median of 172 people, and standard deviation of 204.27 shows that dispersion of the data relative to its mean is not necessarily very spread out nor very clustered around the mean and the distribution is further explored using boxplot, histogram, and interquartile ranges below.

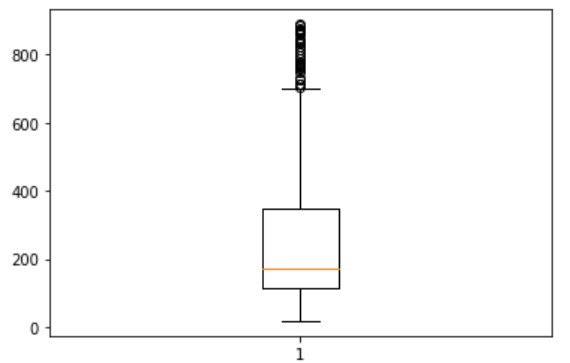
The variable is of numeric integer type. The data is quantitative and discrete as it measures each patient occupancy in terms of beds and there are no half-patient or a fraction of a bed type of occupancy. The finding that there were no zeros in the number of available ICU or pediatric beds at any time during the period, implying that overall across Ontario there the occupancy of ICUs were not exceeded. Despite this, it has been repeatedly reported that hospitals were at maximum occupancy. A possible explanation for this could be that hospitals in more populated locations in Ontario were experiencing shortages while less populated areas had an excess. Including the hospital names or locations in the dataset could have helped combat this constraint and would be more helpful in finding ways to allocate resources throughout the province.

Moving forward in the data cleaning process the check for missing values is conducted and found zero missing values. The check for duplicate observations also concluded that there were no duplicates.



A box plot and histogram is created to analyze the distribution of the observations. The histogram is divided in terms of months to analyze the distribution of data points per month which found that the most and highest points were in May and April with also the largest spread and lowest during August. The interquartile range is 235 with the upper whiskers at 702 and lower at -238, however since there is no such thing as negative occupancy the lower whiskers end at 0. This shows that there is not much variability in comparison with the full range of the data being between 18 and 889 and the data is mainly distributed between 114.5 and 349.5 . There are 35 outliers residing in the upper range and the upper whiskers extend quite high. This reflects the spike in patients experienced during the months of May and April in 2021.

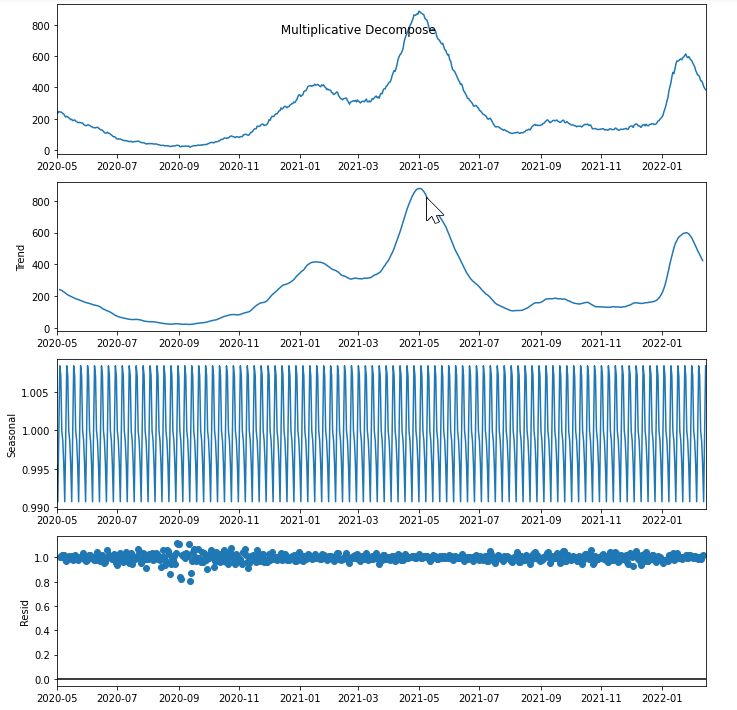
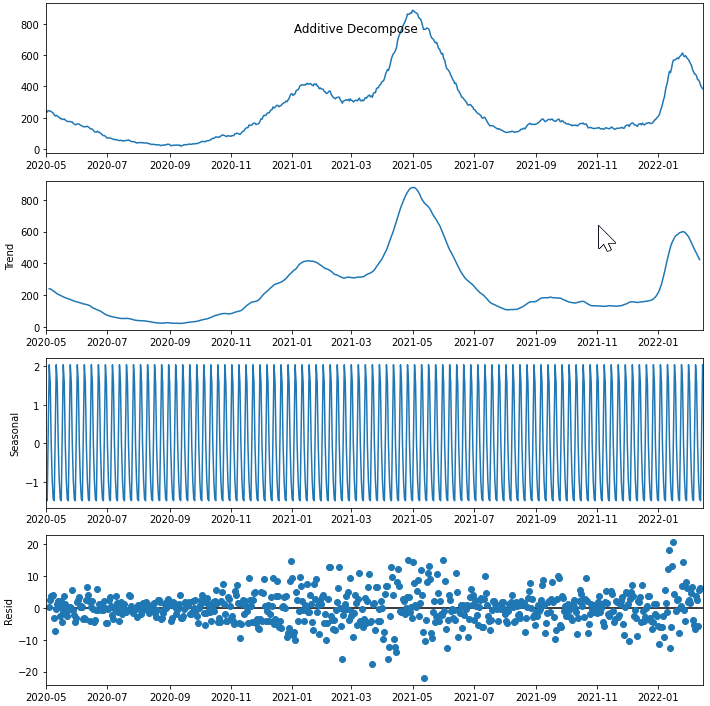




Time series data can be broken down into different components separating its level, trend, seasonality and residual. These components of the time series that have consistency or reoccur in the series could be shown and modeled. The level describes the average value, the trend describes the propensity of the value to increase or decrease over time. Seasonality is the repeating short term cycles found in the time series, and the residual or noise describe the remaining contributors towards the time series which cannot be explained or modeled by the previous components. When these components are added together in order to create the time series it is said that the time series is an additive model whereas when the components are multiplied it is multiplicative. Results show that trend exists in both models, not much seasonality in both, considerable noise in the additive model meaning more variability than in the multiplicative model. These models are expressed and shown in the decomposition of the dataset as follows:

Additive: Time Series = Level + Trend + Seasonality + Residual(also known as Noise)

Multiplicative: Time Series = Level \* Trend \* Seasonality \* Residual(also known as Noise)



As a preemptive measure to the modelling of the time series a stationarity check will be conducted to ensure that the mean and the variance of the dataset is constant, thereby determining whether removing the differences of levels in the dataset or making it stationary is necessary. This is a necessary step for the implementation of the Auto-Regressive Integrated Moving Average (ARIMA) model used further on during the modelling portion of the study.

Checking for stationarity is conducted through the use of the Augmented Dickey Fuller test (ADF Test) in which the null hypothesis is that the data is not stationary with an alternative hypothesis that the data is stationary. ADF test returned a p-value of 0.137 which is larger than the p-value of 0.05 needed to be able to reject the null hypothesis therefore the data is not stationary.

## Autoregressive Integrated Moving Average Modelling and Prediction

The first part of the ARIMA model is the autoregressive portion in which the variable of interest is forecasted using a linear combination of the variable’s past values. Autoregression describes the regression of the variable against itself.



Autoregressive model of order ***p*** (Hyndman & Athanasopoulos, 2018)

## Feed Forward Neural Network Modelling and Prediction

## Long Short Term Memory Modelling and Prediction

Chart, line chart

Description automatically generated

*Figure 4. Line graph representing the CRCI children in the ICU (Blue) and total children in ICU (Orange)*

**Overall Tentative Methodology Flow**

Diagram

Description automatically generated

*Figure 5. Flow chart depicting the overall tentative direction of data analysis*

**References**

Ontario Ministry of Health. [2020]. *Availability of adult and pediatric ICU beds and occupancy for COVID-related critical illness (CRCI)*. Retrieved from <https://data.ontario.ca/dataset/availability-of-adult-icu-beds-and-occupancy-for-covid-related-critical-illness-crci>

Jneny. (2022).Hospitalcapacity. *Github*. Retrieved from https://github.com/Jneny/Hospitalcapacity.git

CIHI Portal. [2016]. Care in Canadian ICUs. Ottawa, ON: *Canadian Institute for Health Information*. Retrieved from <https://secure.cihi.ca/free_products/ICU_Report_EN.pdf>.

CIHI Portal. [2016]. COVID-19’s impact on hospital services. Ottawa, ON: Canadian Institute for Health Information. Retrieved from https://www.cihi.ca/en/covid-19-resources/impact-of-covid-19-on-canadas-health-care-systems/hospital-services

<https://files.ontario.ca/moh-covid-19-weekly-epi-report-en-2022-02-05.pdf>

<https://www.cp24.com/news/ontario-reports-21-new-covid-19-deaths-icu-occupancy-now-exceeds-400-1.5733238?cache=%3FautoPlay%3Dtrue%3FclipId%3D104062%3FautoPlay%3Dtrue>

<https://toronto.ctvnews.ca/ontario-icu-occupancy-could-hit-200-patients-in-january-as-covid-19-cases-rise-modelling-suggests-1.5663239>

<https://covid19-sciencetable.ca/ontario-dashboard/#effectivereproduction>

Jones, S. S., Evans, R. S., Allen, T. L., Thomas, A., Haug, P. J., Welch, S. J., & Snow, G. L. (2009). A multivariate time series approach to modeling and forecasting demand in the emergency department. *Journal of biomedical informatics*, *42*(1), 123-139.

Batal, H., Tench, J., McMillan, S., Adams, J., & Mehler, P. S. (2001). Predicting patient visits to an urgent care clinic using calendar variables. *Academic Emergency Medicine*, *8*(1), 48-53.

Kim, K., Lee, C., O’Leary, K., Rosenauer, S., & Mehrotra, S. (2014). Predicting patient volumes in hospital medicine: A comparative study of different time series forecasting methods. *Northwestern University, Illinois, USA, Scientific Report*.

Tandberg, D., & Qualls, C. (1994). Time series forecasts of emergency department patient volume, length of stay, and acuity. *Annals of emergency medicine*, *23*(2), 299-306.

Zhang, G. P. (2003). Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing*, *50*, 159-175.

Khashei, M., & Bijari, M. (2010). An artificial neural network (p, d, q) model for timeseries forecasting. *Expert Systems with applications*, *37*(1), 479-489.

Hyndman, R.J., & Athanasopoulos, G. (2018) Forecasting: principles and practice, 2nd edition, OTexts: Melbourne, Australia. OTexts.com/fpp2.