**Analysis and Prediction of Forest Fires**

# Introduction

Every year in Canada, thousands of uncontrolled forest fires destroy millions of hectares of land. These fires destroy everything in their path and cost hundreds of millions of dollars in damages. Not only is the environment damaged, but forest fires are a threat to wildlife and people as well. Forest fires also have a negative impact on the sustainability of our forests.

There are two types of causes for forest fires: human and lightning. Human caused forest fires cause approximately 55% of the fires and are broken down into many types of human causes.

The goal of this project is to forecast the number of future forest fires and where they will most likely occur. The three research questions I will attempt to answer are the following:

# Research Questions

1. **Prediction number of Forest Fires**
2. Are the number of forest fires increasing or decreasing?
3. What cause is contributing to most of the forest fires?
4. What forest fire location is experiencing increasing or decreasing forest fires?
5. What year(s) are experiencing increasing or decreasing more forest fires?

# Literature Review

**D. L. Martell, , S. Otukol, and , B. J. Stocks (1987). A logistic model for predicting daily people-caused forest fire occurrence in Ontario. Canadian Journal of Forest Research, 17(5): 394-401**

In this research paper, the author’s aim is to predict the number of people caused forest fires in Northern Ontario by using a logistic model. Logistic regression analysis techniques were employed to predict the probability of a fire day using a Poisson probability distribution to model daily people caused forest fire occurrence.

Historical forest fire data was studied from the time period, 1965 to 1981 and they relied on various assumptions, such as a fixed fire season from April 15 to October 6. Various plots were produced from the data to determine the average number of fires per day for the different types of human caused forest fires.

Live field tests were performed to evaluate the prediction system. They were only able to conduct field tests from June 10 to August 31, 1984, which is only part of the fire season, April 15 to October 6.

Their field tests results indicated the system generally performed better during the early summer season than it did in the summer. There was subjective data in their equations which might be a reason why the system was not more successful in predicting fire occurrence. Perhaps additional data would improve their model. Also, additional studies would be required to see if additional data and field tests would improve the model.

**Hanes, C., Wang, X., Jain, P., Parisien, M-A., Little, J. and Flannigan, M. 2018. Fire-regime changes in Canada over the last half century. Can. J. For. Res. 49: 256–269 (2019)**

The author of this article was attempting to determine the various fire-regime trends in Canada for two time-periods (1959–2015 and 1980–2015) in Canada. Since 1959, the number of large fires and area burned has substantially increased. The increase in large fires has altered the landscape and one possible reason for the increase of fires is thought to be an increase in lightning strikes. Over the last 5 decades, the fire seasons have started earlier and are ending later. The increased fire season is mainly due to human caused fires. However, human caused fires have shown a decrease, but they are a concern because the vegetation at the beginning and end of the season is more flammable.

The data for the trend analysis was compiled from various government agencies in Canada. However, each government agency has different collection methods and not all the databases contain the same information. The data that was collected was grouped together to calculate statistics for area burned, number of fires, fire cause and fire size. However, the analyses was only performed for select regions. The trend analysis used a nonparametric Mann-Kendall test to detect monotonic trends in time series data by year. The level of significance established used a bootstrapped randomization hypothesis test, the null hypothesis assuming there was no trend in the data.

The result of their study indicate that the area burned has increased since 1959. The increase in area burned is mainly due to increased in lightning strikes. Lightning caused fires are responsible for more area burned because of numerous lightning ignitions occur in clusters and occur in isolated areas. There are some limitations regarding this study. The study covers a large period of time and the more data that is involved, the greater a chance that there are errors in the figures. The analysis of the study does show there is a trend for increased fire activity.

**Burton, Philip & Parisien, Marc-André & Hicke, Jeffrey & Leduc, Alain & Gauthier, Sylvie & Bergeron, Yves & Flannigan, Mike. (2007). Large fires as agents of ecological diversity in the North American boreal forest. International Journal of Wildland Fire. 17. 754-767.**

The journal talks about how forest fires in the boreal forests can have different variability of damage. The varying degrees of damage is the result of many factors, such as differences in climate, terrain and land-uses. Areas that were more fire resistant had a greater number of islands, which prevented the fire from spreading. Lands that had less burn severity were also able to recover at a faster rate, which allowed vegetation to return earlier compared to areas that were more fire prone.

**Canadian Wildland Fire Strategy. A 10-year review and renewed call to action. 2016. Canadian Council of Forest Ministers, Ottawa, Ontario. 15 p. Prepared on behalf of the Wildland Fire Management Working Group established under the Canadian Council of Forest Ministers**

The paper talks about increased wildland fire behavior resulting in threats to life, property, and natural resources. Impact to people and communities across the country are increasing. Canadian jurisdictions are at the limits of existing suppression resources. Effort needs to be made toward increasing capacity.

**Stockdale,C. A., Mcloughlin, N., Flannigan,M.,Macdonald,S.E. 2019. Could restoration of a landscape to a pre-European historical vegetation condition reduce burn probability? Ecosphere 10(2)**

This journal is suggesting that restoring forest landscape to historical conditions before the turn of the 20th century could reduce burn probability. Forest regions in Western Canada have experienced an increase in forest canopy closure and expanding forest growth into grasslands. There is evidence this has been a result from climate change and thus possibly increasing the probabilities of additional wildfires. Historical photographs were used to determine the vegetation composition in 1909. From the historical photographs, the study suggests there is a difference in burn probabilities between current landscapes and historical landscapes.

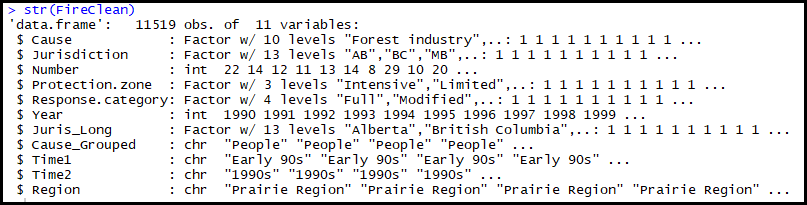
# Dataset Description

This project and dataset will focus on forest fires in Canada from 1990 – 2018. The dataset was generated from the National Forestry Database, <http://www.nfdp.ccfm.org/en/data/fires.php> and has 13 attributes. Half of the attributes are duplicates since the dataset is from a Federal database, where each attribute is listed twice, one for each official language: English and French. One additional attribute, “Data Qualifiers” will be removed as well as the information is not material regarding the analysis. I will use the 6 remaining relevant attributes for analysis.

The data set does not have any missing data but cleaning is required to remove duplicate attributes and will need to be cleaned up for any duplicate records.

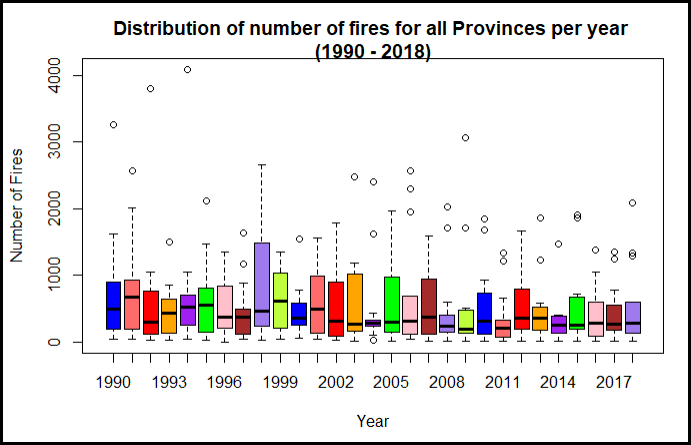
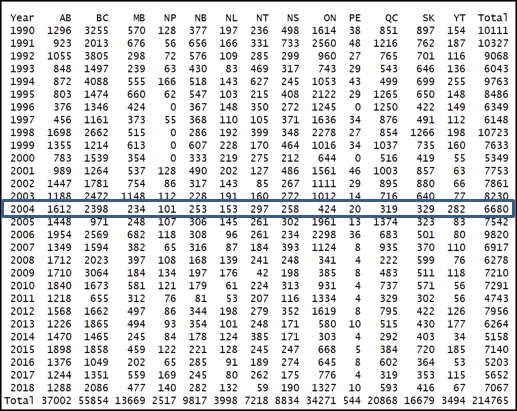
I have also created dummy attributes to compare number of fires from different regions (Pacific vs Central) and different time periods (1990s vs 2000s).

| **Attribute Name** | **Data Type** | **Description** | **Distinct Items** |
| --- | --- | --- | --- |
| Cause | Categorical | Cause of fire | 10 Distinct items:   1. Forest industry 2. Incendiary 3. Lightning 4. Miscellaneous known causes 5. Other industry 6. Railways 7. Recreation 8. Residents 9. Unspecified 10. Unspecified human activities |
| Jurisdiction | Categorical | Location of fire | 13 Distinct items:   1. Alberta 2. British Columbia 3. Manitoba 4. National parks 5. New Brunswick 6. Newfoundland and Labrador 7. Northwest Territories 8. Nova Scotia 9. Ontario 10. Prince Edward Island 11. Quebec 12. Saskatchewan 13. Yukon |
| Number | Integer | Number of fires | 0 is Minimum  2913 is Maximum |
| Protection Zone | Categorical | Level of forestry land value | 3 Distinct items:   1. Intensive 2. Limited 3. Unspecified |
| Response category | Categorical | Level of attempt to control the fire | 4 Distinct items:   1. Full 2. Modified 3. None 4. Unspecified |
| Year | Integer | Year of the fire | 29 Distinct values:  1990 – earliest year  2018 – latest year |
| Cause\_Grouped | Categorical | Dummy variable created to differentiate between the two main causes of forest fires. | Two Distinct items:   1. Lightning 2. Human |
| Time Period 1 | Categorical | Dummy variable created by assigning “Year” values into one of six unique time periods. | 6 Distinct items:   1. Early 90s 2. Late 90s 3. Early 10s 4. Late 10s 5. Early 20s 6. Late 20s |
| Time Period 2 | Categorical | Dummy variable created by assigning “Year” values into one of 3 unique time periods. | 3 Distinct items:   1. 1990s 2. 2000s 3. 2010s |
| Regions | Categorical | Dummy variable created by assigning “Jurisdiction” values into one of 6 unique regions. | 6 Distinct items:   1. Atlantic Region 2. Central Region 3. North Region 4. Pacific Region 5. Prairie Region 6. National Parks |
|  |  |  |  |

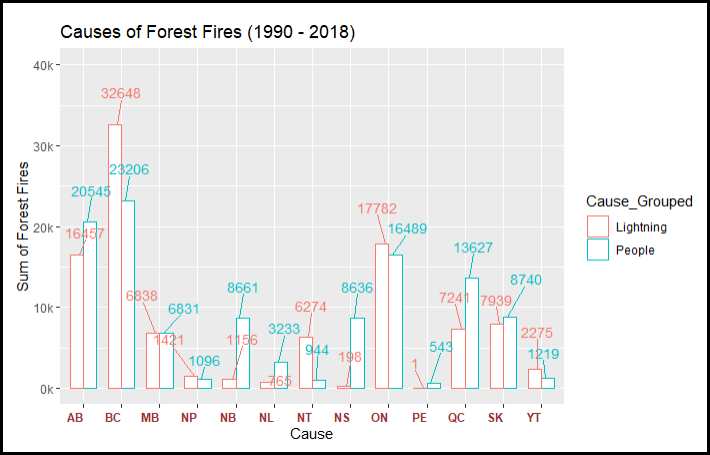
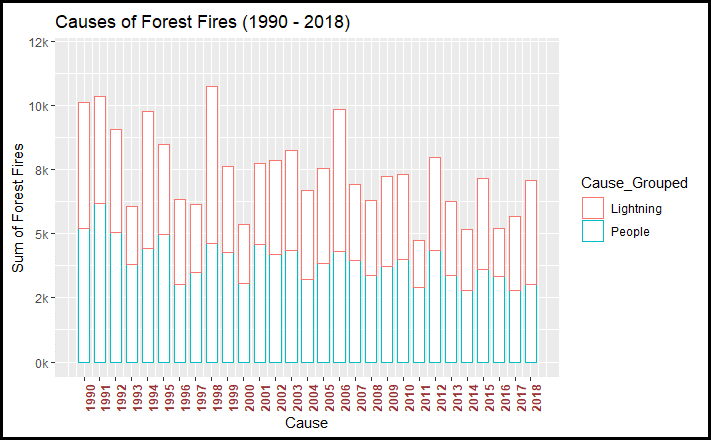
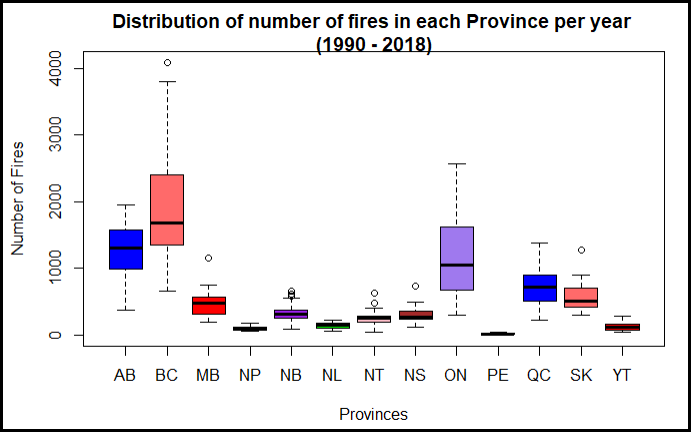
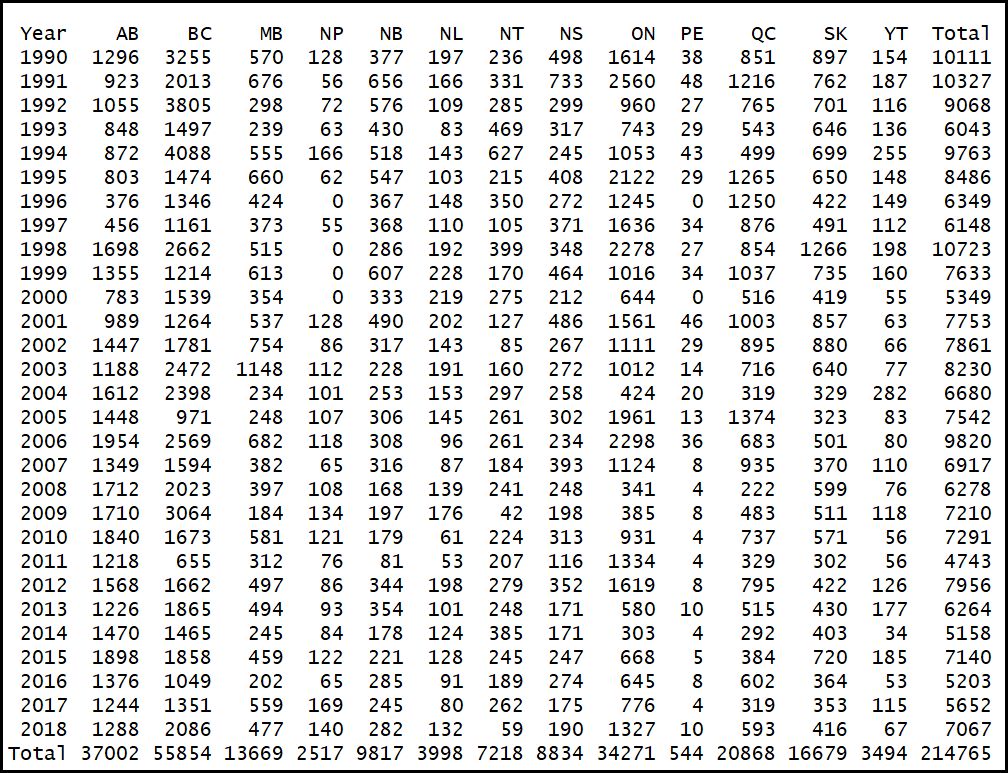


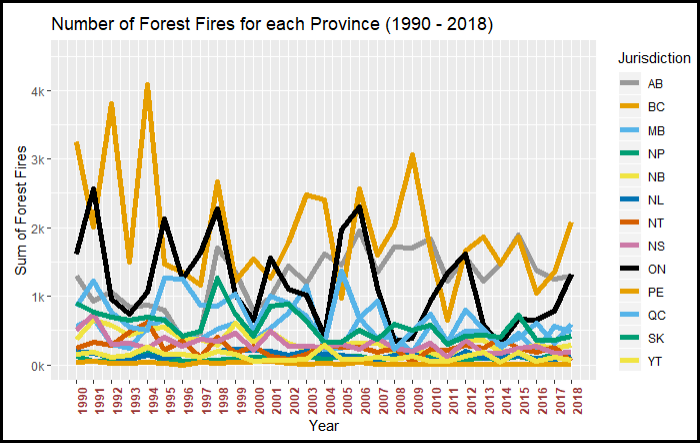
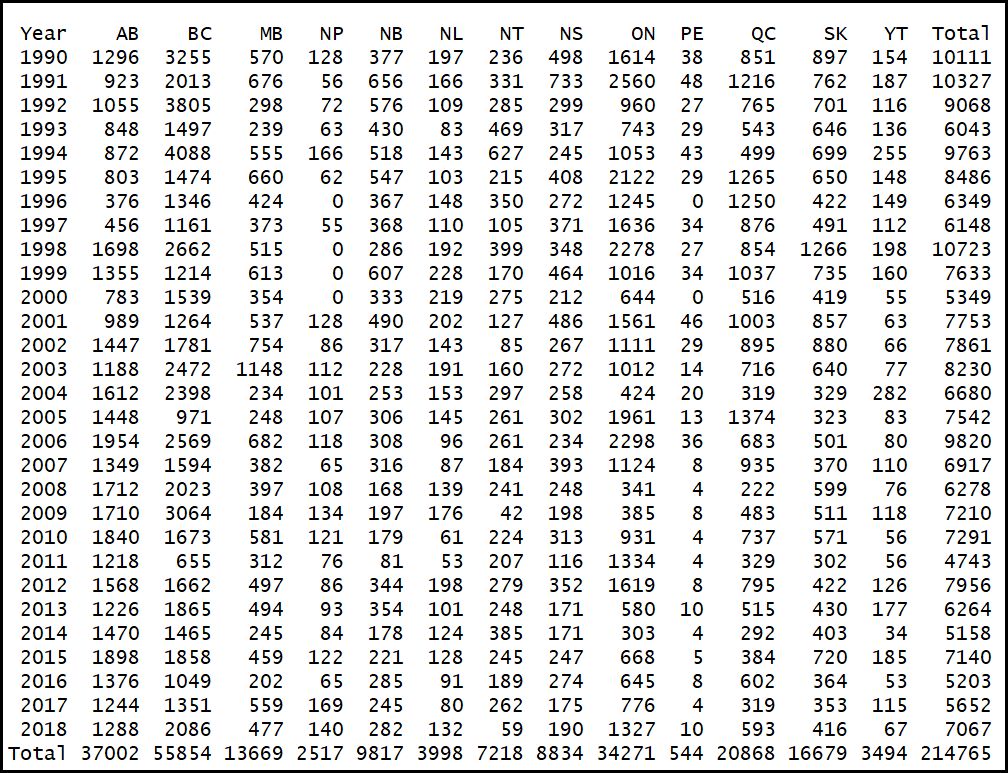


## Dataset Exploration

I have created various visualization aids to better understand the data. The boxplot below gives you an idea of the spread per year when all the provinces are combined. A few examples I would like to point out. In 2004, there were 6680 totals fires for all the provinces. The average for 2004 was 514 fires. The two outliers are Alberta and British Columbia, which had 1612 and 2398 fires respectively.

The boxplot and matrix below show the number of fires and spread for each province from 1990 – 2018. British Columbia has had more fires than any other province.



The trend line below shows the change in the number of fires for each province from 1990 – 2018. British Columbia, Alberta and Ontario have a substantially more forest fires than the other provinces.

## Approach

## Step 1: Import and review data

Download the data (csv file) from the National Forestry Database, <http://www.nfdp.ccfm.org/en/data/fires.php>.

Import the data into R and review the structure.

**R code for initial data analysis:** [**https://github.com/ed209robo/Ryerson**](https://github.com/ed209robo/Ryerson)

## Step 2: Data Cleaning and Initial Analysis

Delete unnecessary columns: This dataset was obtained from a federal database and has duplicate columns, one in English and the other in French. Therefore, we can remove the duplicate because they are not required for analysis.

Remove missing values: There are missing values for the “Number” attribute and these records must be removed before analysis. Not removing these records will produce incorrect calculations that use other attributes.

Assigning correct data types: Make sure all the variables have the correct data type. The “Number” attribute and “Year” attribute are required to be numerical data types. All other attributes will be coded as categorical data types.

Appropriate column names: Some of the column names could be either too long or too short. Also, the name of the column should relate to the observation in that particular column. All the column names in this dataset are applicable for the observation type of each column. The column names in the dataset are relevant to the observations and no change was required.

Appropriate categorical observation names: Some of the categorical observations may have long names that can be shortened. Shortening observation names has the benefit of making it easier to interpret any type of chart. I have created an extra column and shortened the Province name to its two-letter abbreviation.

Check for Outliers: Check for outliers in the data. BC is an outlier can be seen from the boxplot. Why is BC is an outlier? because of the amount of forests. Weather produces lightning. Not an outlier in the sense of an error.

Feature selection: We will use all features of the data. We will Explore various feature so the best attributes will produce a better predictive model.

**R code for initial data analysis:** [**https://github.com/ed209robo/Ryerson**](https://github.com/ed209robo/Ryerson)

## Step 3: Exploratory Analysis

Exploring the data with various boxplots, bargraphs and correlation to better understand the data. And answering questions 2-5.

**R code for exploratory data analysis:** [**https://github.com/ed209robo/Ryerson**](https://github.com/ed209robo/Ryerson)

## Step 4: Train and Test Data

I will use different classification algorithm to train and test the data.

Will use three models.

1. Regression for predicting number of forest fires per year.
2. Using time series analysis for predicting number of forest fires per year.
3. Poisson regression with time series data (GLARMA model).

## Step 5: Model Selection

I will use the model that performs the best to determine the number of forest fires.

Model used for the train and test data are

* 1. Regression
  2. Time series
  3. Poisson regression with time series data model.W

## Step 6: Results and Recommendation

We will use the model with the best accuracy as the recommended model.