**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 answer the following research questions.

# Research Questions

1. Prediction of number of Forest Fires (modeling)

--------------------------------------------------------------------------------------

1. Are the number of forest fires increasing or decreasing? (bar chart and time series)
2. What cause is contributing to most of the forest fires? (bar chart)
3. What forest fire location is experiencing increasing or decreasing forest fires? (bar chart)
4. What year(s) are experiencing increasing or decreasing more forest fires? (bar chart)

# 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.

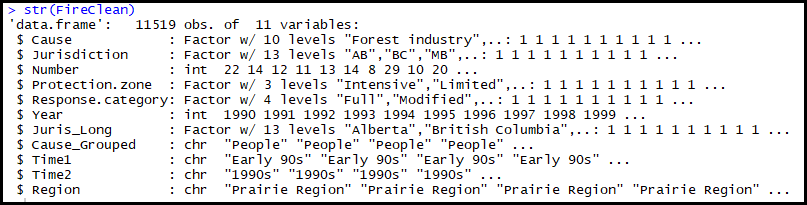
# Dataset Description

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.

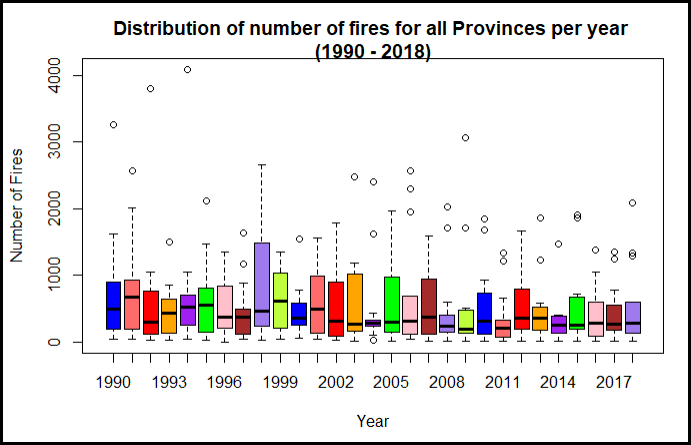
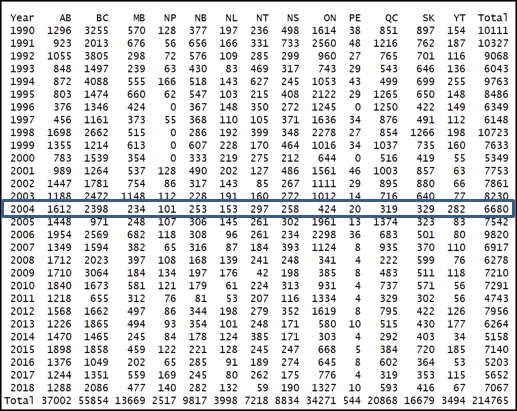
The data set does not have any missing data, but cleaning is required to remove duplicate attributes and duplicate records.

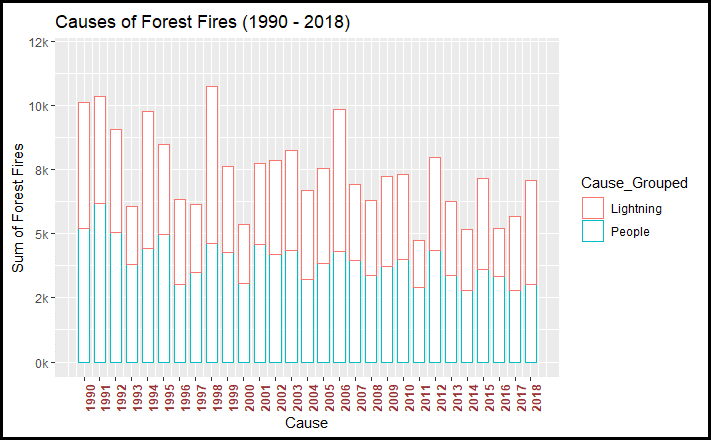
I have also created dummy attributes to compare number of fires from different regions (Pacific vs Central) and different time periods (1990s vs 2000s).

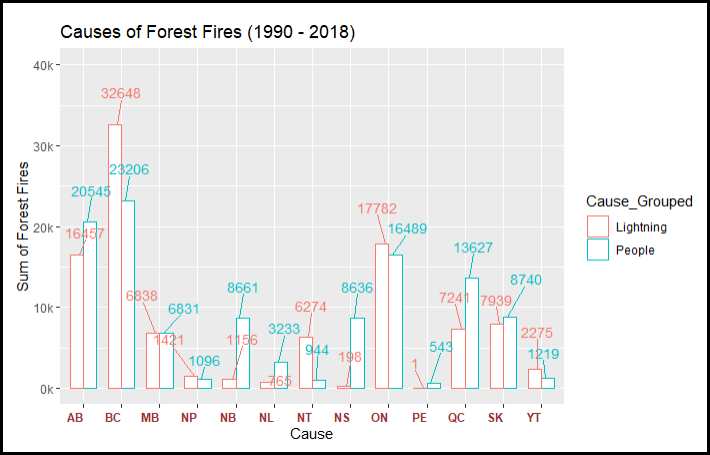
| **Attribute Name** | **Data Type** | **Description** | **Distinct Items** |
| --- | --- | --- | --- |
| Cause | Categorical | Cause of fire | 10 Distinct items:   1. Forest industry 2. Incendiary 3. Lightning 4. Miscellaneous known causes 5. Other industry 6. Railways 7. Recreation 8. Residents 9. 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 |
|  |  |  |  |

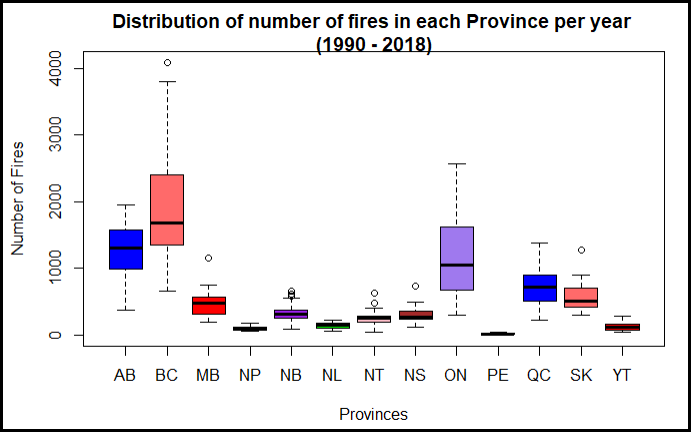


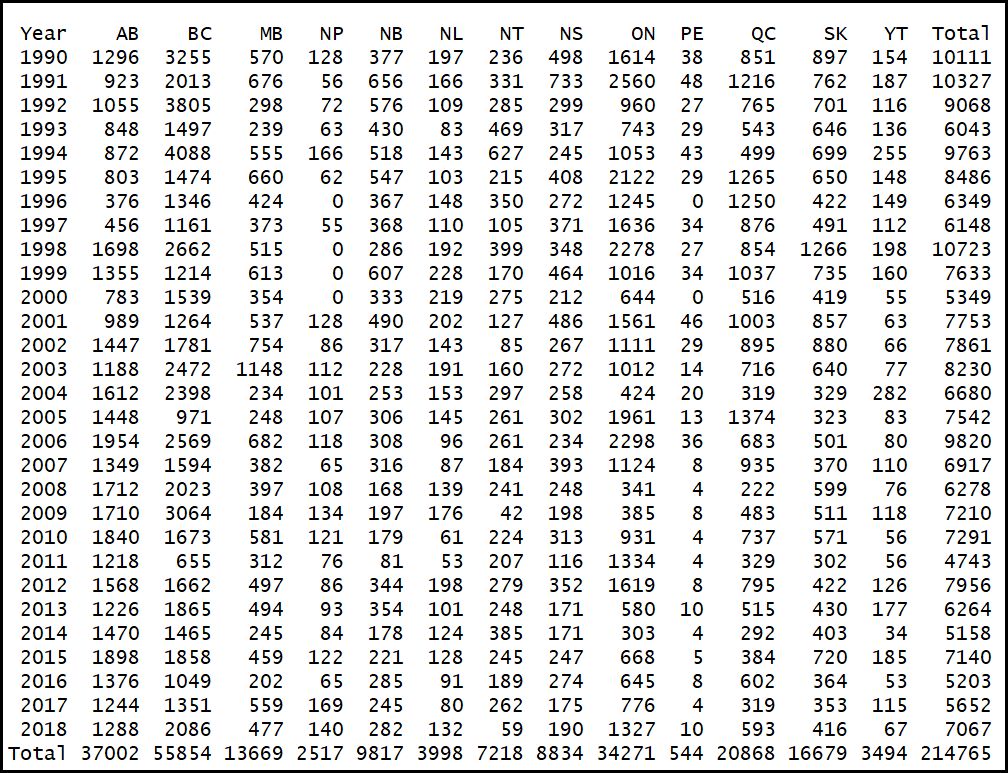
## Dataset Exploration

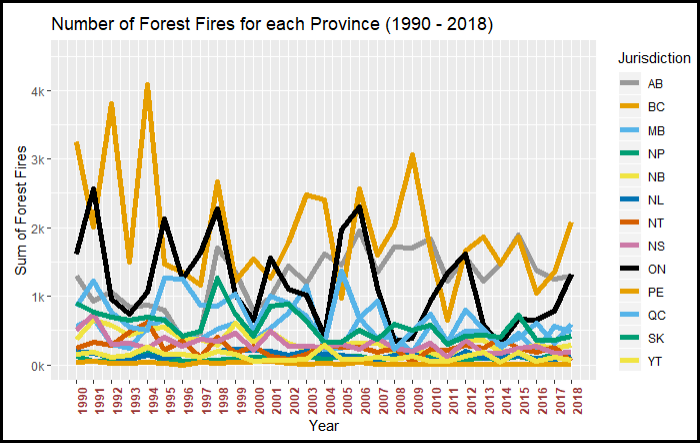
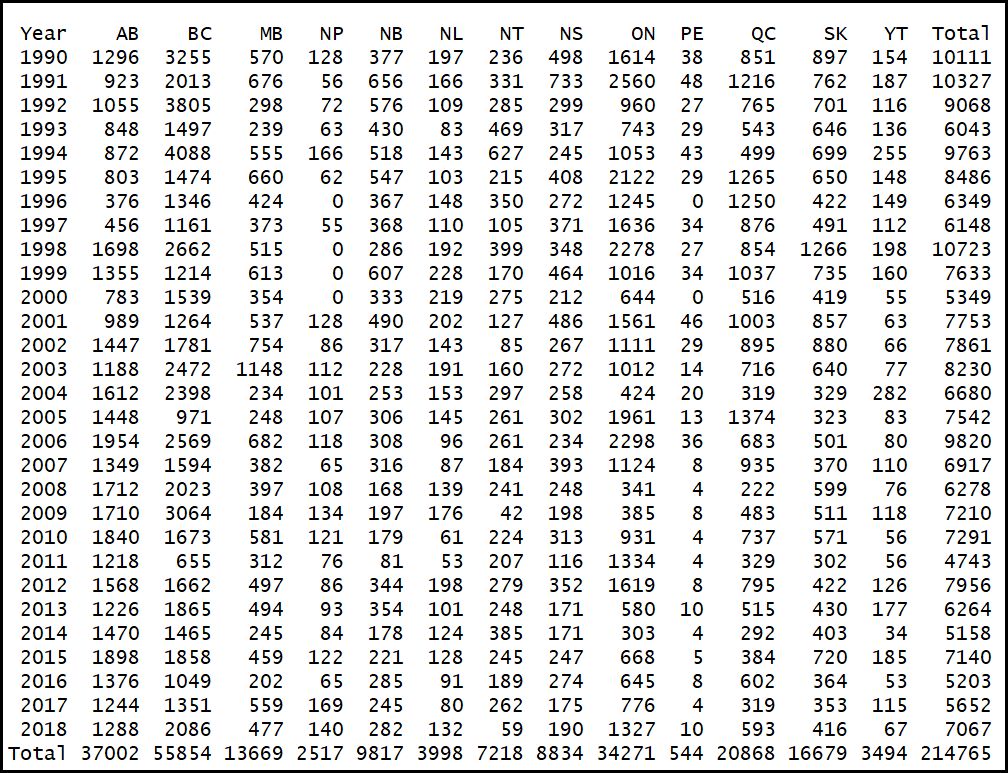
I have created various visualization aids to better understand the data. The boxplot below gives you an idea of the spread per year when all the provinces are combined. A few examples I would like to point out. In 2004, there were 6680 totals fires for all the provinces. The average for 2004 was 514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

The stacked bar chart below shows the total number of fires for each year and the comparison of each forest fire cause for that particular year. There are three years (1990, 1991 and 1998), which had more than 10,000 forest fires. The overall trend is leaning towards decreased number of fires, especially fires caused by people.

The bar chart below shows the total number of fires for each province and the comparison of each forest fire cause for that province. British Columbia has significantly more forest fires than other provinces.

The boxplot and matrix below show the number of fires and spread for each province from 1990 – 2018. An example, 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 substantially more forest fires than the other provinces. The overall trend is decreasing.

## 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 final data analysis:** [**https://github.com/ed209robo/Ryerson**](https://github.com/ed209robo/Ryerson)

## Step 2: Data Cleaning

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 duplicates 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. As per the National Forestry Database, these figures are missing because the data is not available. Not removing these records will produce incorrect calculations that use other attributes.

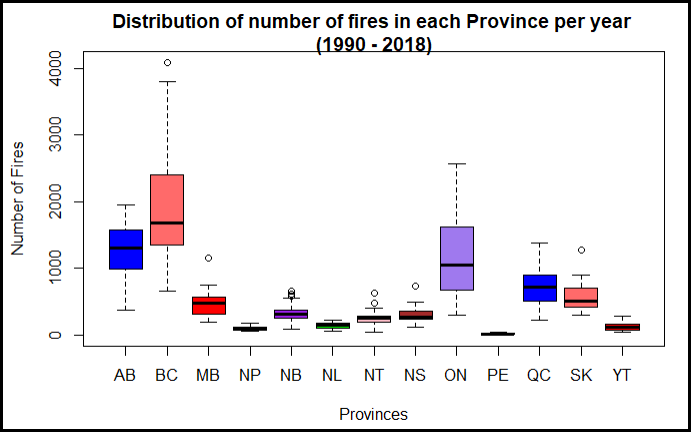
Assigning correct data types: Make sure all the variables have the correct data type. The “Number” attribute and “Year” attribute are required to be numerical data types. All other attributes will be coded as categorical data types.

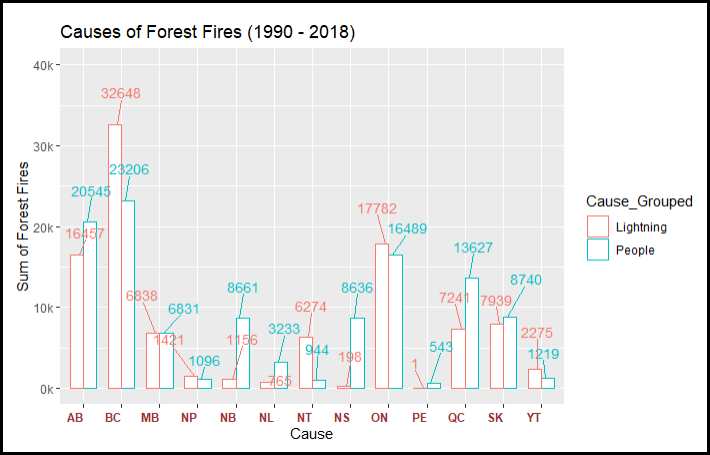
Appropriate column names: Some of the column names could be either too long or too short. Also, the name of the column should relate to the observation in that column. All the column names in this dataset are applicable for the observation type of each column. The column names in the dataset are relevant 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.

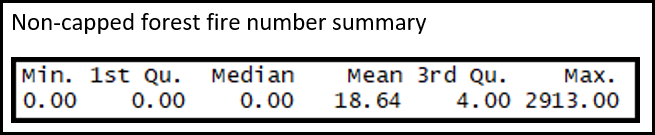
**R code for final data analysis:** [**https://github.com/ed209robo/Ryerson**](https://github.com/ed209robo/Ryerson)

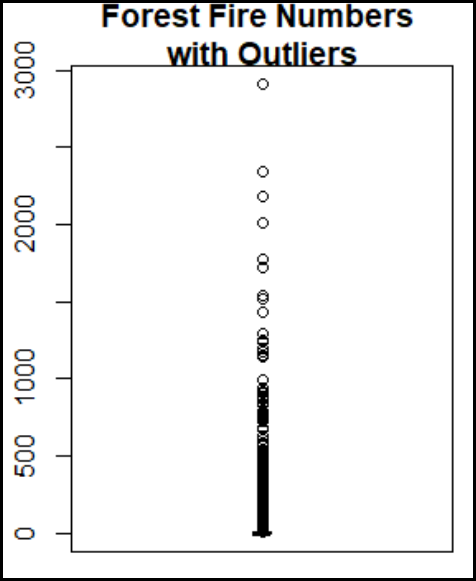
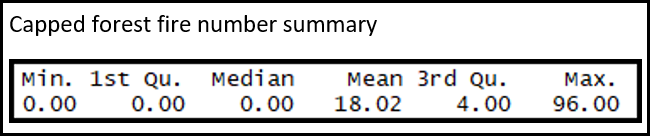
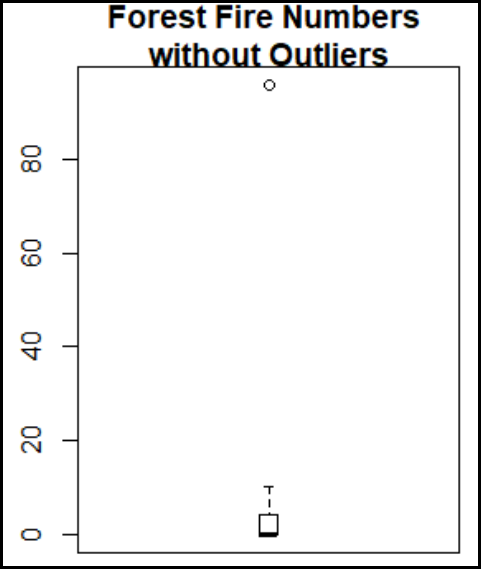
## Step 3: Initial Analysis and Exploratory Data Analysis

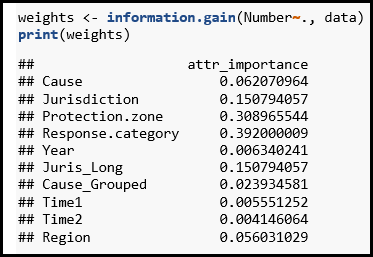
**Check for Outliers**: Outliers can be one of two types. Either a mistake or a true outlier. The forest fire dataset has no outliers due to mistakes. However, the dataset does contain many true outliers for the number of forest fires. Specifically, many of the outliers for the number of forest fires are from the province of British Columbia. This can be seen from the boxplot.

There are several reasons why British Columbia has a significant number of outliers. Firstly, British Columbia has one of the largest forest coverages and the largest reserved forest coverage in Canada. Secondly, British Columbia has more forest fires caused by lightning than any other province. This can be seen from the bar chart below.

In addition, British Columbia and Ontario are the only two provinces that have more forest fires from lightning strikes compared to humans. Many of these fires are in remote areas, which can be difficult to access. Also, the number of lightning strikes continue to increase in British Columbia. The larger forest area in British Columbia and increased lightning strikes will result in more forest fires. Therefore, the number of forest fires in British Columbia is not an outlier from an error but is a true outlier that can be included as a scenario in the statistical modeling. The outliers have been calculated by capping anything above the 95th percentile.

 The number summary below shows the range of number of forest fires before the outliers are capped.

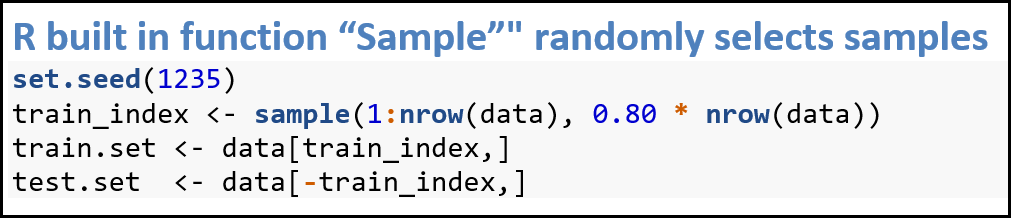
The number summary below shows the range of number of forest fires after the outliers are capped.

**Feature selection**: I used the FSelector package and the information gain function in R to choose the best combination of attributes. The two attributes which scored the highest importance are “Protection.zone” and “Response.category”. I will prepare many different model scenarios with different combinations of attributes.

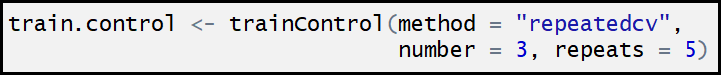
**R code for final data analysis:** [**https://github.com/ed209robo/Ryerson**](https://github.com/ed209robo/Ryerson)

## Step 4: Train and Test Data

## The dataset is split into a training set and test set. I have used the built in “sample” function in R to randomly select samples for the split while ensuring the training set and test set do not have any common data points.

I also used the “set.seed” function in R to make the code reproducible.

The caret package includes a function, “trainControl” that sets the parameters for running models. A few of the parameters are required for the model. One parameter is the “method” value, which is the type of resampling method in the train-control function. I have used “repeatedcv”, (repeated cross-validation), which creates multiple random splits of the dataset into training and validation data. I used the following number of cross-validation and repeats to prevent overfitting. The number field is set to 3, which indicates 3-fold cross-validation. The repeats field is set to 5, which refers to 5 repeated K-fold cross-validation.



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

## Step 5: Model Selection

In R, I have used 7 different regression models from the caret package to predict the number of forest fires. With these 7 models, I have run different scenarios to find the best model and the best set of predictors.

The 7 regression models from the caret package I have used are explained below.

**Linear Regression** **(LM**) (method = 'lm')

Modeling the relationship between a response (dependent variable) and one or more explanatory variables (or independent variables).

**Generalized Linear Model (GLM)** (method = 'glm')

Flexible generalization of ordinary linear regression that allows for response variables that have error distribution models other than a normal distribution.

**The lasso** (method = 'lasso')

Regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the statistical model it produces.

**k-Nearest Neighbors** (method = 'knn')

Algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions).

**Linear Regression with Forward Selection (LF)** (method = 'leapForward')

Forward selection begins with an empty model and adds in variables one by one.

**Linear Regression with Backwards Selection** **(LB)** (method = 'leapBackward')

Backwards elimination begins with a full model and removes variables one by one.

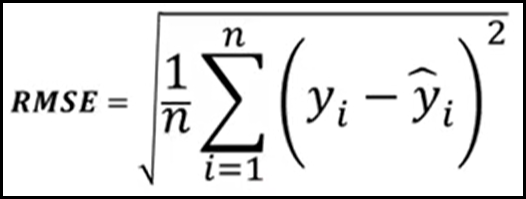
**Linear Regression with Stepwise Selection** (method = 'lmStepAIC')

Stepwise selection begins with an empty model and adds in variables with the largest F-statistic.

**Step 6: Performance Measures for Models**

The performance measures for any type of statistical model is an important step when it come to evaluating the model. I have used regression models for predicting the number of forest fires. Regression analysis involves the process of estimating the relationship between a dependent variable and one or more independent variables.  We are trying to identify the error between the actual and predicted output. It is always better to use at least two different performance measures when comparing different model performance because each performance measure provides a different viewpoint on performance. The three performance measures I have selected below will evaluate the effectiveness of each model by benchmarking the model performance against each other.

**Root Mean Square Error (RMSE)** – standard deviation of the residuals (measure of how well the actual data points measures the difference between predicted values and the actual values). The residuals are a measure of how far from the regression line data points are. We want the value of RMSE to be as low as possible. A lower value of RMSE indicates a better fit, the better the model is with its predictions.

The RMSE is calculated as follows:

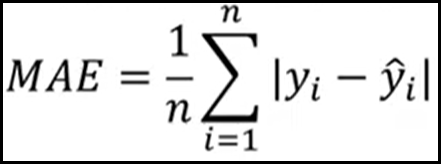
RMSE is calculated by following these steps:

1. Calculate the residual for every data point
2. Calculate the squared value of the residuals
3. Calculate the average of the squared residuals
4. Obtain the square root of the result

Advantages: In the same units as the target variable (number of forest fires).

Disadvantages: RMSE is sensitive to outliers and sensitive to large errors (penalizes large prediction errors more than smaller prediction errors).

**Mean Absolute Error (MAE)** – is obtained by calculating the absolute difference between the model predictions and the true (actual) values. MAE is a measure of the average magnitude of error generated by the regression model.

The mean absolute error (MAE) is calculated as follows:

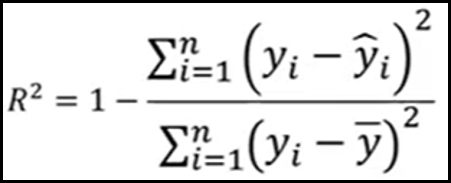
MAE is calculated by following these steps:

1. Calculate the residual of every data point
2. Calculate the absolute value (to get rid of the sign)
3. Calculate the average of all residuals

Advantages: Easy to understand metric since we’re just looking at the absolute difference between the data and the model’s predictions.

Disadvantages: MAE is sensitive to outliers.

**R squared (coefficient of determination)** – R-squared represents the proportion of variance (of y) that has been explained by the independent variable in the model. It gives you an idea of how many data points fall within the results of the line formed by the regression equation. The higher the coefficient, the higher percentage of points the line passes through when the data points and line are plotted.

R-squared is calculated as follows:

Advantages: Easy to understand because the values are between 0 and 1.

Disadvantages: R-squared can be increased by adding independent variables to the model which is misleading since some added variable might be useless with minimal significance.

## Step 7: Improve model performance by changing model parameters

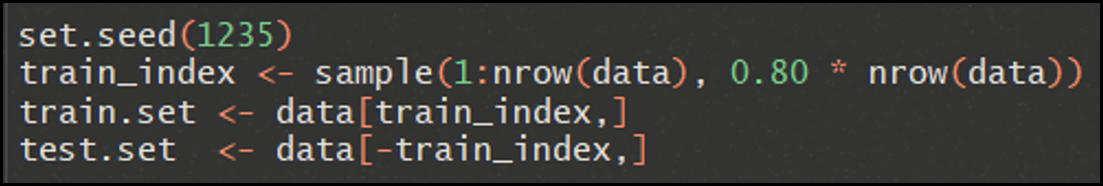
All 7 models were run multiple times with various changes in the parameters. My goal was to adjust and tweak the various parameters to improve the model. Some of these changes resulted in better performance, while others resulted in worse performance. I tracked the performance of the models with three performance measure. In addition, I also calculated the runtime for each model’s training set and test set.

Below is a list of the different parameter tuning techniques I used for the 7 models to improve the performance measures.

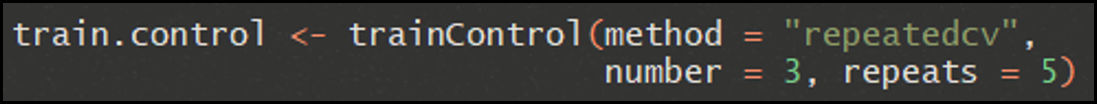
1. adding/removing and trying different combinations of predictors to improve the model

* e.g. running models without “Jurisdiction”

1. changing the ratio of training vs testing

* e.g. different ratios for training/testing: 65/35, 70/30, 75/25, 80/20, 85/15, 90/10

1. changing the number of cross-validation folds and number of repeats.

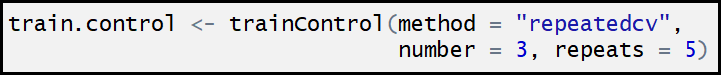
* e.g.

1. running the models with capped outliers and non-capped outliers.

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

## Step 8: Results and Recommendation

The first set of models were run with all predictors in the dataset because I wanted to get a baseline on the performance with the original set of predictors. During the initial testing, I experimented with different values for the parameters, including changing the value of K in the cross-validation as well as the number of repeats. I discovered that changing the level of K did change the results between the “train set” and “test set”. Some of the earlier results seemed to have some overfitting. However, I was able to find a stable performance when using K = 3 and repeats = 5.

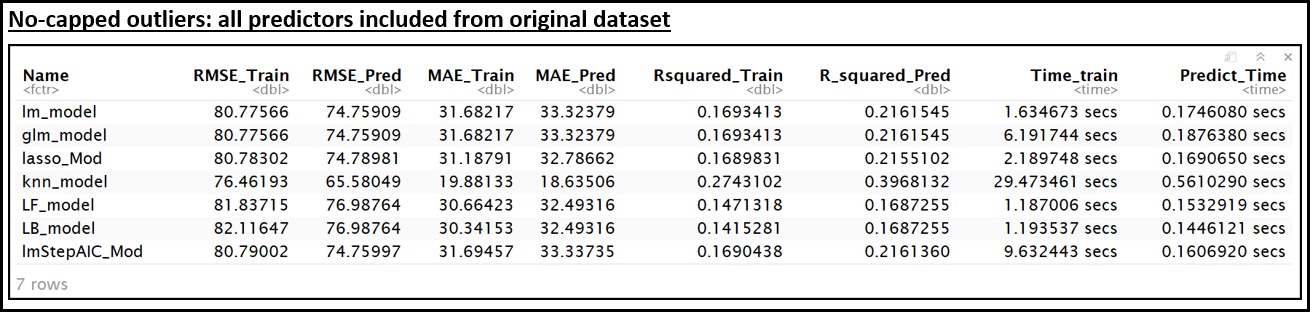


**No-capped outliers: all predictors included from original dataset**

Two of the different performance measures (RMSE, R-squared) improved when predicting on the test set for all 7 models. The MAE only showed improvements on the KNN model, where it was reduced from 19.88133 to 18.63506. The KNN model was the best performing model on both the training set and testing set for all three performance measures (RMSE, MAE, R-squared). The KNN model also had better results on the test set compared to the training set.

The R-squared on the test set was higher for all 7 models. The biggest improvement was on the KNN model where there was an increase from 0.274 to 0.3968, which is an increase of an additional 12% of the variance in the dependent variable that is explained by the independent variables.

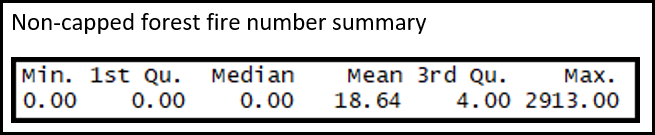
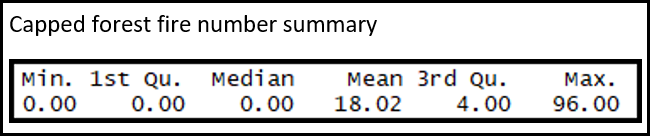
The KNN model had the longest runtime of 29.473461 seconds on the training set because there are more calculations involved.

Overall, when all the original predictors in the dataset were used, the best performing model was the KNN model.

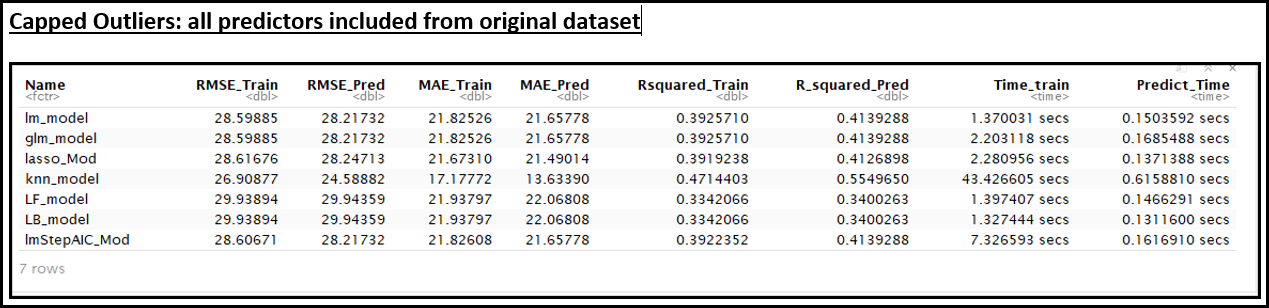
**KNN Model**

* **RMSE: Lowest on test set: 65.58**
* **MAE: Lowest on test set: 18.63**
* **R-squared: Highest on test set: 0.39**

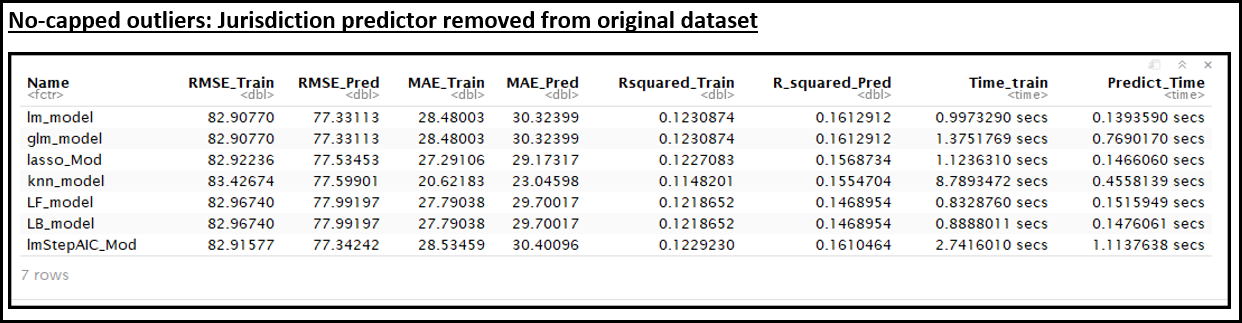
**Capped Outliers: all predictors included from original dataset**

I also wanted to see if there was a performance improvement if I capped the outliers for number of forest fires to the 95th percentile, which would reduce the maximum number of fires from 2913 to 96.

There was a significant improvement for the R-squared value on the prediction, especially for the KNN model. The KNN model’s R-squared value of 0.55496 is higher than the original value of 0.3968132. However, since the outliers are true outliers and certain provinces have significantly more fires, it will be more prudent to keep the original number of fires.



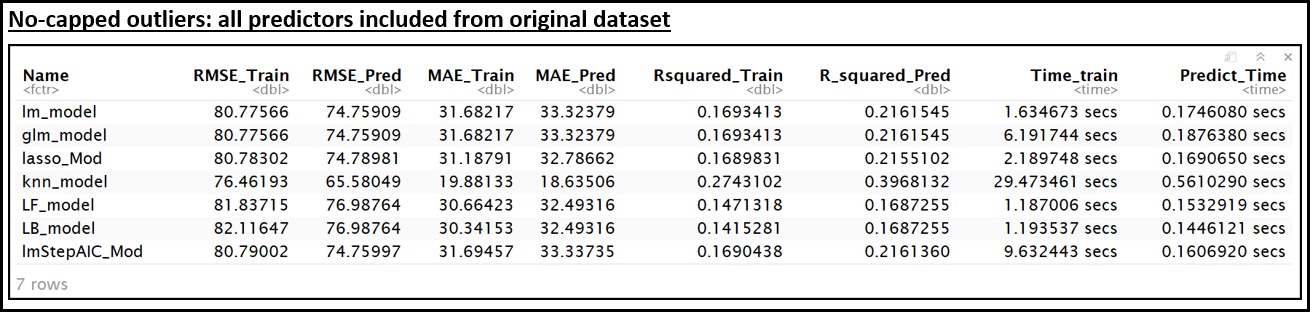
**No-capped outliers: Jurisdiction predictor removed from original dataset**

Running the models with reduced predictors resulted in worse results for all models. I tested all models with varying reduced combinations of predictors. They all performed worse when compared to the original set with all the predictors included. Below is an example of one scenario where I have removed “Jurisdiction” from the model run. The original KNN model with all the predictors had an R-squared value of 0.3968 but was significantly reduced to 0.1554704 when the “Jurisdiction” predictor was removed.

**Best Performing Model**

The KNN model performed the best with all the original predictors. However, as I mentioned above, the runtime was longer than the other models.

KNN Model

* RMSE: Training Set: 76.46 Test Set: 65.58
* MAE: Training Set: 19.88 Test Set: 18.63
* R-squared: Training Set: 0.27 Test Set: 0.39

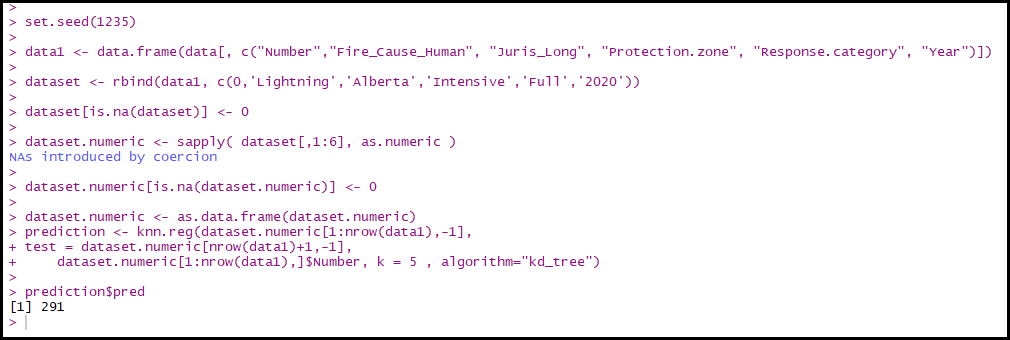
Since the KNN model performed the best, I wanted to see if I could calculate a potential number of fires with specific conditions. I determined that the K=5 for the nearest neighbors and selected some random parameters to calculate a possible number of forest fires. The number of possible forest fires with the conditions below was calculated to be 291 forest fires.

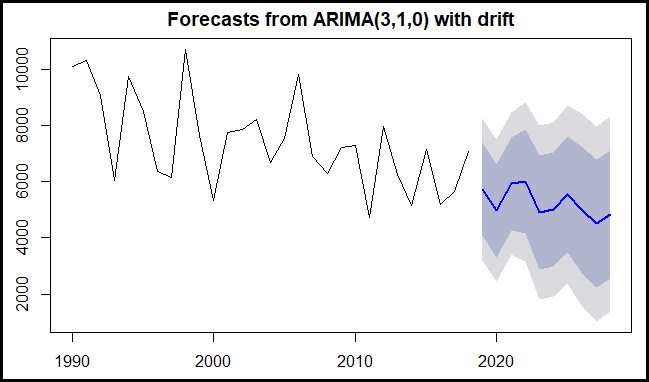
location: Alberta

Cause: Lighting

Protection Zone: Intensive

Response category: Full

Year: 2020

Lastly, I wanted to use some time series analysis to forecast the number of forest fires for all of Canada using the forecast function and ARIMA model. The time series analysis for predicting the number of forest fires indicates that the trendline for the number of forest fires is declining. The light grey area is the 95% interval and the purple area is the 80% interval consisting of the predicted values. Since about 50% of the forest fires are caused by humans, the number of forest fires can be reduced with continued fire prevention education to people about the dangers of forest fires.

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

# Conclusions

In my project, my main goal was to predict the number of forest fires. I used various regression models and time series to predict the number of forest fires. The best performing model was the k-nearest neighbors (KNN) for predicting the number of forest fires. The KNN model produced better results than the other models using the three performance measures (RMSE, MAE, R-squared). However, the runtime was longer for KNN and this could be a setback for future modeling with larger datasets. There is no optimal solution for statistical modelling and further testing is always required to improve the results.

# Limitations

Predicting forest fires is a difficult task even for government agencies that employ hundreds of qualified data scientists and have budgets of millions of dollars. My project is small in comparison and has many limitations. I have listed some limitations below.

* My project only had 7 models. Future analysis could include new models to improve the results.
* I used regression analysis, but perhaps future work could involve classification analysis.
* There are many different performance measures available other than the three I used.
* The dataset I used was a small sample with only 6 main attributes. There are many predictors available such as weather data that could be added to improve my model.
* My dataset is small, and analysis of larger datasets will require more powerful computers.

# References

1. <https://www.nrcan.gc.ca/our-natural-resources/forests-forestry/wildland-fires-insects-disturbances/17598>
2. <https://cwfis.cfs.nrcan.gc.ca/ha/nfdb?type=pnt&year=2018>
3. <https://www.nrcan.gc.ca/our-natural-resources/forests-forestry/sustainable-forest-management/measuring-reporting/forest-classification/13179>
4. <https://www.nrcan.gc.ca/our-natural-resources/forests-forestry/wildland-fires-insects-disturban/climate-change-fire/13155>
5. <https://www.ccfm.org/english/reports_articles.asp>
6. <http://www.nfdp.ccfm.org/en/data/fires.php>
7. <http://www.fao.org/3/xii/0207-b3.htm>
8. <https://www.esquire.com/news-politics/politics/a27558150/canada-wildfire-season/>
9. <https://www.ompe.org/en/causes-et-consequences-of-forest-fires/>
10. Garrett Grolemund, Hadley Wickham, “R for Data Science” <https://r4ds.had.co.nz/index.html>
11. What’s a good value for R-squared? <https://people.duke.edu/~rnau/rsquared.htm>
12. *Max Kuhn* “The caret Package” <http://topepo.github.io/caret/index.html>
13. Dr. Tamer Abdou, CIND 119, 2018, Ryerson University
14. Dr. Ceni Babaoglu, CIND 642, 2019, Ryerson University