# Alberta Wild Fire Analysis

## Introduction

This project is about analyzing the affect of year and latitude on the size of the wildfires that happens in the month of May between 210 and 2021 in province of Alberta, Canada. The dataset is collected from website of Government of Alberta and is available with other resources in the repository.

Spearman's Rank Correlation Coefficient Analysis is used for analyzing the affect of year and latitude on the size of the wildfires and data visualization is used to give a better grasp of the mentioned correlations. I used Python and libraries such as Pandas, Numpy, Seaborn and Scypi for this Analysis

## Spearman's Rank Correlation Coefficient Analysis

In this correlation analysis, I have to assume the null Hypothesis and come up with the following questions

Null Hypothesis: There is no association between the two variables.

a. in a spearman's rank correlation coefficient analysis where the independent variable is size of the fire and the dependant variable is year, What is the corresponding p-value and rho values?

b. Complete a spearman's rank correlation coefficient analysis where the independent variable is size of the fire and the dependant variable is latitude. What is the corresponding p-value and rho values?

/8

Spearman's correlation coefficient is a statistical measure of the strength of a monotonic relationship between paired data. The assumptions that are required for this analysis are as followed;

- The observations are independent from each other, each fire is not related to another one.

- Data should be ordinal, ratio or interval. Year, area and latitude can be categorized as ratio data.
- There are no extreme outliers (anomalies)

Using a subset of data for cross-validation is not common or necessary for Spearman's correlation coefficient, because this analysis is applied to the whole dataset. However using subsets of data for can be used for tasks such as sensitivity analysis.

By assuming the above assumptions, here's the answers to question parts a and b;

a. Spearman correlation coefficient is 0.0077, and the p-value is 0.61 for this analysis. Meaning, null hypothesis is not rejected and there's no monotonic relationship between the year and size of the fires.

b. Spearman correlation coefficient is 0.416, and the p-value is 4.4 e-184 for this analysis. Meaning, null hypothesis is rejected and there's positive monotonic relationship between the latitude and size of the fires. The analysis shows bigger fires happen at higher latitudes.

## Analysis Interpretation

I will explain more in this section about the result of the Spearman's correlation coefficient. Second analysis (between latitude and size of fire) rejected the null hypothesis and there's positive monotonic relationship between the latitude and size of the fires. The analysis shows bigger fires happen at higher latitudes.

I also believe that the Spearman's correlation coefficient is the best hypothesis testing method for this analysis, because the data fits all the required assumptions for Spearman's correlation coefficient, and the data is numeric, and monotonic relationship between independent and dependent variable gives us a lot of insight about the data. Plus, our data does not have normal distribution which is a criteria for a lot of other analysis but not for Spearman correlation coefficient analysis.

I must also mentions the assumptions about the data for this analysis and I mentioned them below.

- The observations are independent from each other, each fire is not related to another one.

- Data should be ordinal, ratio or interval. Year, area and latitude can be categorized as ratio data.
- There are no extreme outliers (anomalies)

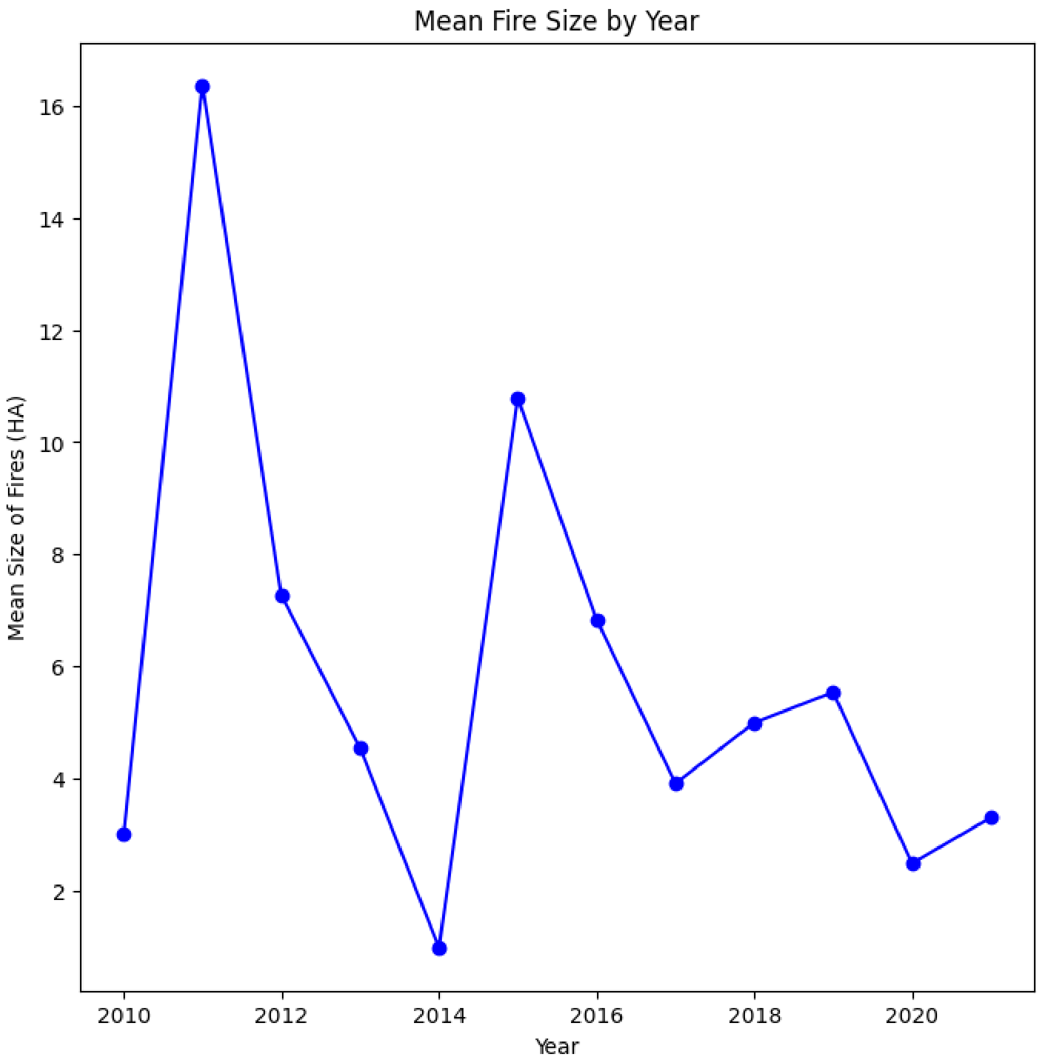
fire agency in alberta is concerned with fires that get out of control and have the potential to impact people, local communities and or infrastructure. The Calgary Forest area records a high number of human-caused fires because they document abandoned campfires – these are fires within an enclosed fire pit that don't normally cause any concern. How does this information change the understanding about the statistical analysis? Should I keep these fires in the data when completing further statistical analysis – explain why or why not.

/10

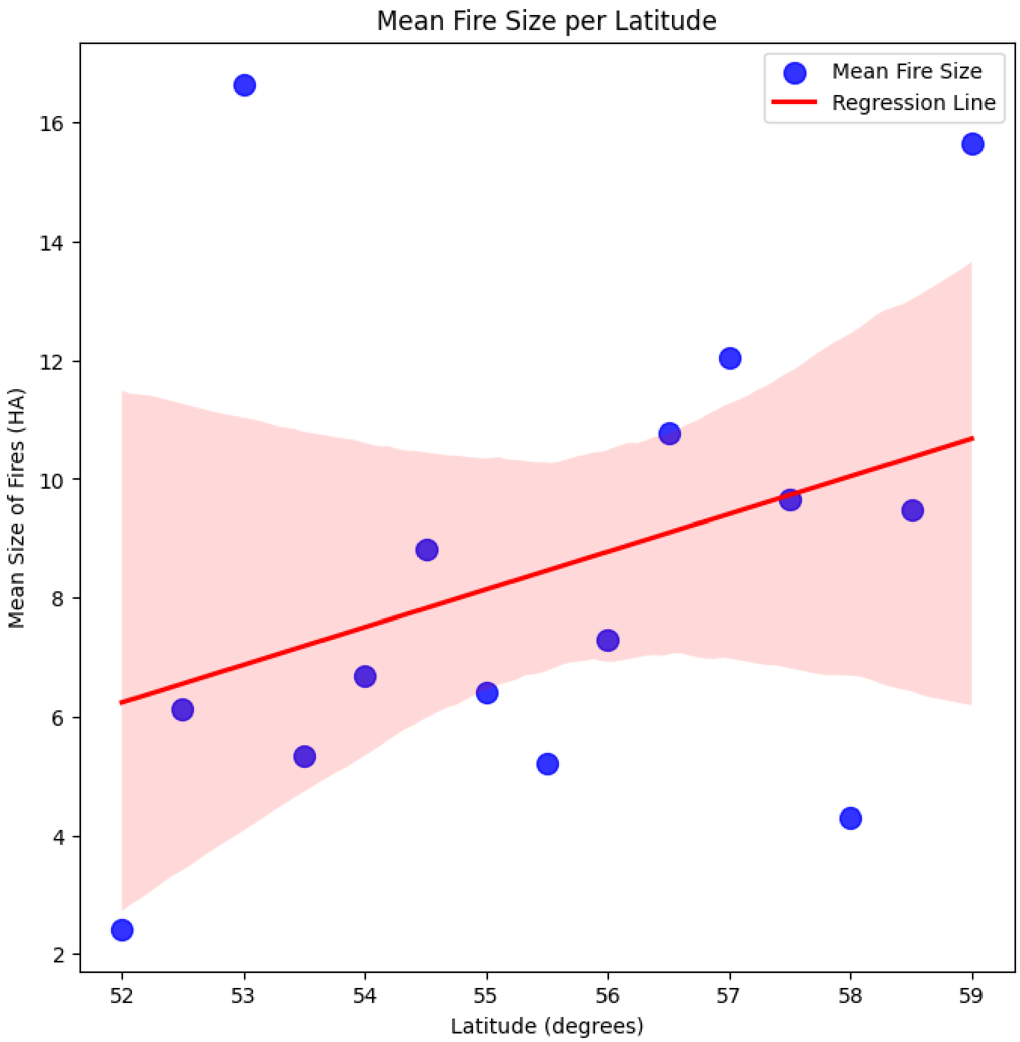
I can do sensitivity analysis on the Calgary Forest area data and see how it is affecting our overall analysis and maybe include factors such as proximity to populated areas in our analysis. I definitely will keep the data in furthered analysis.

## Data Visualization

I included the mean size of the fires vs year and latitude in the images below and described each figure.



This figure is plotted by Python and using Seaborn library. It shows the mean of size of all the fires that occurred in each year. Please note we considered only fires that are smaller than 1000 hectares, because most of the fires were smaller than 1000 hectares. As you can see, the size of the fires fluctuate without any trend as the years pass so there’s no correlation between size of the fires and the year they occur.



This figure is plotted by Python and using Seaborn library. It shows the mean of size of all the fires per latitude. Please note we considered only fires that are smaller than 1000 hectares, because most of the fires were smaller than 1000 hectares. As you can see, the size of the fires tend to increase as the latitude increases and the regression line that fit the data tend to show that as well. The analysis shows bigger fires tend to happen at higher latitudes.

This is happening because of multiple reasons. First, large cities are scarce as you go north, as a result there are less artificial obstacles for fires such as roads, and deforestation and putting the fire out will be harder because of remoteness and having less infrastructures to help the firefighters. As a result, fires can become larger. Second, Alberta becomes wider as we go north and there are more flat lands with winds in the north so the fires have more space to grow in the north.