Forest Fire Analysis and Prediction

Major project report submitted in partial fulfilment of the requirement for the degree of Bachelor of Technology

In

Computer Science and Engineering

By

Purav Vashisht

UNDER THE SUPERVISON

OF

Dr Vivek Kumar Sehgal

Department of Computer Science & Engineering and Information Technology

Jaypee University of Information Technology, Waknaghat,
173234, Himachal Pradesh, INDIA

Table of Contents

Declaration

I hereby declare that, this project has been done by me under the supervision of **Dr. Vivek Kumar Sehgal**, (HOD) Department of Computer Science and Engineering, Jaypee University of Information Technology, Waknaghat.

I also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:

Dr. Vivek Kumar Sehgal
Prof. HOD | CSE & IT
Computer Science & Engineering and Information Technology
Jaypee University of Information Technology, Waknaghat,

Submitted by:

Purav Vashisht
(181287)
Computer Science & Engineering Department
Jaypee University of Information Technology

Certificate by Supervisor

This is to certify that the work which is being presented in the project report titled **"Forest Fire Analysis and Prediction**" in partial fulfilment of the requirements for the award of the degree of B.Tech in Computer Science And Engineering and submitted to the Department of Computer Science And Engineering, Jaypee University of Information Technology, Waknaghat is an authentic record of work carried out by Purav Vashisht (181287) during the period from January 2022 to May 2022 under the supervision of Dr. Vivek Kumar Sehgal, (HOD) Department of Computer Science and Engineering, Jaypee University of Information Technology, Waknaghat.

Purav Vashisht
(181287)

The above statement made is correct to the best of my knowledge.

Dr. Vivek Kumar Sehgal
Prof. HOD | CSE & IT
Computer Science & Engineering and Information Technology
Jaypee University of Information Technology, Waknaghat,

Acknowledgement

Firstly, I express my heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the project work successfully.

I'm really grateful and wish my profound my indebtedness to Dr. Vivek Kumar Sehgal, Prof. HOD, Department of CSE Jaypee University of Information Technology, Waknaghat.

Deep Knowledge & keen interest of my supervisor in the field of "Deep Learning and AI" to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts and correcting them at all stage have made it possible to complete this project.

I would like to express my heartiest gratitude to Dr. Vivek Sehgal, Prof HOD, Department of CSE, for his kind help to finish my project.

I would also generously welcome each one of those individuals who have helped me straight forwardly or in a roundabout way in making this project a win.

In this unique situation, I might want to thank the various staff individuals, both educating and non-instructing, which have developed their convenient help and facilitated my undertaking.

Finally, I must acknowledge with due respect the constant support and patients of my parents.

Regards,

Purav Vashisht
181287
CSE
JUI

Abstract

Forest Fire Analysis and Prediction is basic visualization research model which aims to find out different insights and the number of forest fire occured in a particular area. It aims in identifying different kinds of forest fire.

Forest Fire Prediction is a key aspect of forest fire management. This is a prime environmental hassle that creates ecological destruction withinside the form of a threatened panorama of natural sources that disrupts the stableness of the ecosystem, will increase the threat for different herbal hazards, and reduces sources inclusive of water that reasons worldwide warming and water pollution. Fire Detection is a key detail for controlling such incidents. Prediction of forest fire identityentification predicted to lessen the effect of forest fire withinside the future.

Many fire detection algorithms are to be had with exclusive technique in the direction of the detection of fire. In the present paintings approaches the fire affected area is expected primarily based totally at the satellite tv for pc images. To expect the occurrences of a forest fire the proposed gadget approaches the usage of the meteorological parameters inclusive of temperature, rain, wind and humidity had been used. Random forest regression and Hyperparameter tuning the usage of RandomizedSearchCV set of rules we used a numerous sub-samples of dataset on which it suits numerous choice bushes and makes use of averaging to enhance the predictive accuracy and manipulate over-fitting. Based at the evaluation of the fashions with all the chosen meteorological parameters can constitute the wooded area fire events. This paper discusses approximately a comparative observe of various fashions for predicting forest fire inclusive of Decision Tree, Random Forest, Support Vector Machine etc.

Chapter 01: INTRODUCTION

1.1 Introduction

Think Big but start Small, Data is defined as information stored in digital format that can then be used as a basis for implementation analysis and decision making, It’s important to know what you’re looking at when you hear the term data, whether it is traditional data or big data, data is the first step when solving any problem.
Big data is a term reserved for extraordinarily big data and it isn't simply humongous in phrases of volume. This in fact can be in diverse formats, the benefit of having large amount of data doesn’t matter in which format is that we can find patterns hidden in those data by using those patterns in the data we predict the future.
The progress that has been done in this area is remarkable outputs from huge data sets can be retrieved in real time.
This means they can be extracted so quickly that a result could be computed immediately after the source data has been obtained.
Whichever type of data you have on your hands it is your first port of call for problem solving.
The essence of machine learning is to create an algorithm that is then used by a computer to find a model that fits the data as closely as possible and, based on it, makes very accurate predictions and how it differs from traditional methods differs. how to find this model. We offer algorithms that give instructions to the machine on how it can learn for itself.
It’s not important that it is 100 percent accurate, it should not be, Machine Learning is never 100 percent accurate, there will be some errors some missing values which contributes to it’s not having a full accuracy.

1.2 Objective of the Major Project

Forest Fire Prediction and Analysis is all about exploring a particular set of real time data of a particular place which provides us the different datasets which are then evaluated, trained and tested using Machine Learning in Python and then further used to make predictions.

We will explore the data find different correlations using powerful python packages and visualize the data with different graphs like scatterplots, heatmaps and more.
The first step of applying some data science is to analyse the past data that we have acquired, the technology driven tools are involved in the process of analysing understanding and reporting available past data this will result in providing reports or dashboards and will help on our way to making an informed strategic and tactical decisions, Of course the model itself will make decisions our work is to check the accuracy of the predictions and check if model is not overfitting or underfitting.

1.3 Motivation of the Major Project

In view of the recent shocking occurrences of the Amazon wildfires, I decided to conduct a brief descriptive analysis of Brazil fires during the previous 20 years. Media piqued my interest and raised several questions, which I wanted to investigate using real data and an objective perspective.

The dataset I worked on contains data for every state in Brazil from 1998 to 2017, organized by months.

In Brazil, fires are a major issue. "Understanding the frequency of forest fires in a time series can help to take action to prevent them," according to the Dataset description. Knowing where and when that frequency is most frequently noticed should help clarify the scope of our investigation. It is feasible to examine the evolution of fires over time as well as the places where they were concentrated using this information. Acre, Amapá, Pará, Amazonas, Rondonia, Roraima, and parts of Mato Grosso, Tocantins, and Maranho make up the legal Amazon.

1.4 Language Used

1.4.1 Python

Python has been in the programming phase for over twenty years. There were two main reasons Python was used in this project: first, it has several technical

advantages over other programming languages, and second, its practical application covers several industries. Powerful calculation tool when we need to solve a complicated data problem.
It is an open source programming language. We will try break this definition into several parts and try to understand each of these attributes. Open source means it's free. Python has a massive and lively scientific network with get right of entry to the software's source code and contributes to its non-stop improvement and updating consistent with person needs

This is the main reason Python is cross-platform. It is available for all common operating systems. Windows Mac and Linux have the advantage that Python can be intensively applied in any domain of certain languages.
This plays a major role in the popularity of the languages and their flexibility.

1.4.2 Python Packages

There are many powerful packages used in this project for various development ranging from graphical representations to finding correlations and statistical methods
These methods are useful in building a successful model

Packages used are as follows :

1. NumPy
2. Pandas : Python Data Analysis Library
3. Seaborn : Statistical Data Visualization Library
4. Matplotlib

1.4.2.1 NumPy

NumPy is a powerful computational package which provides intensive mathematical functions, random number generator, It provides multidimensional array and matrices which are useful in performing mathematical operations on array from a wide range of trigonometry to statistics.

1.4.2.2 Pandas

Panda is a package that enhances NumPy even further it allows us to organize data in a tabular form and to attach descriptive labels to the rows and the columns of the table not just numbers as it is with NumPy.
In addition Pandas is endowed with a broad Gamma of tools facilitating our work with various data formats and missing data.
Therefore if you want to do data science with Python Pandas is an essential package you will need

1.4.2.3 Seaborn

Seaborn is one of the python visualization library based on matplotlib. It helps in providing a high-level interface for drawing attractive and useful Statistical Graphics.
We will code in Matplotlib but use the beauty and style Seaborn delivers.
It helps in providing a quality visualization of the data which is useful in determining various correlations and regression lines.

1.4.2.4 Matplotlib

Matplotlib is an intensive data visualization library, It helps in visualising various heatmaps, KDE plot, Reg plot and much more.

1.5 Technical Requirements (Hardware)

1.5.1 Hardware Requirements

- Core i3 or higher (cache – 3 MB or 4 MB recommended)
- 4 GB RAM or higher
- GB Hard Drive Space

1.5.2 Software Requirements

- Python 3 or latest version
- Anaconda
- Operating System : Mac/Windows/Linux

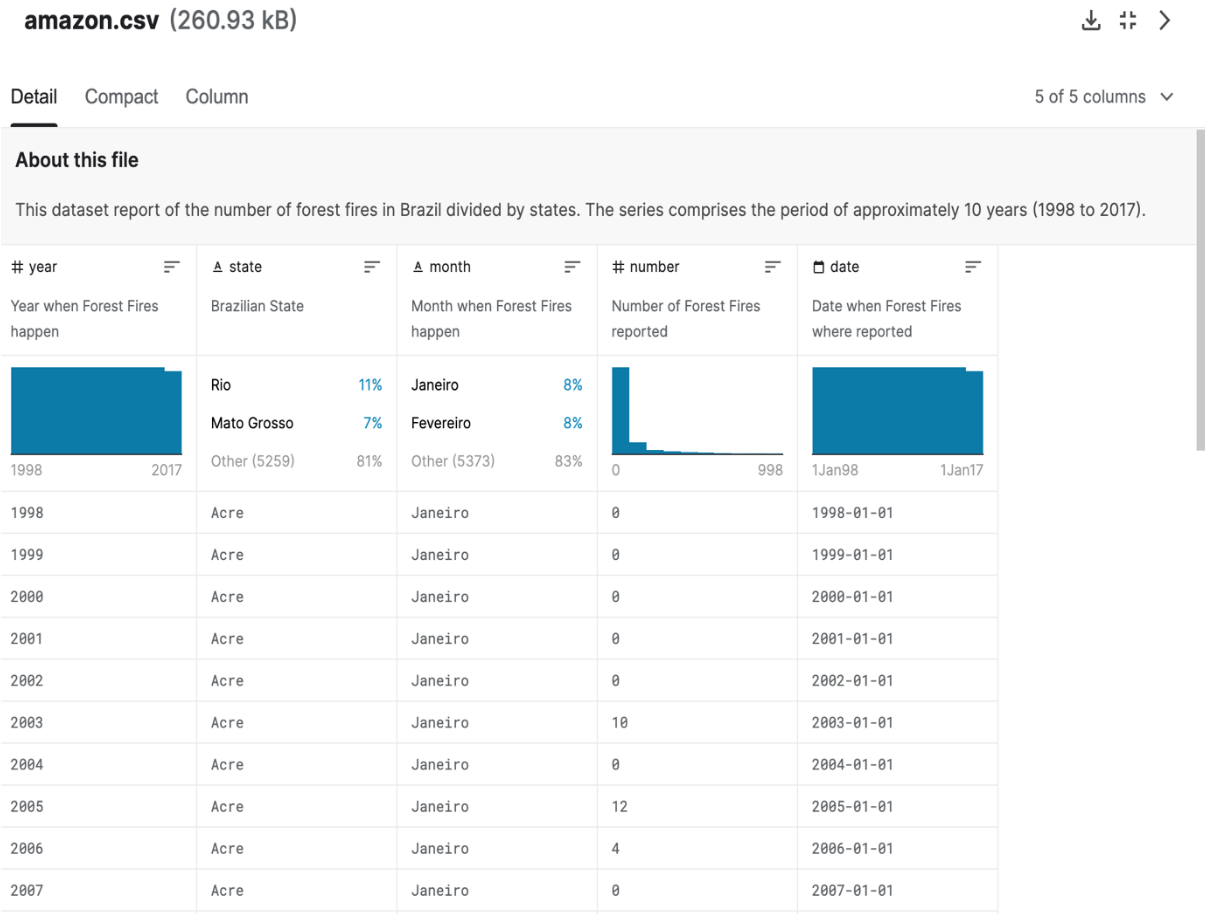
1.6 About Dataset

1.6.1 Context

Forest fires are a significant threat to tropical forest preservation. Understanding the frequency of forest fires over time can aid in taking preventative measures. The Amazon rainforest, located in Brazil, is the world's largest rainforest.

1.6.2 Content

The number of forest fires in Brazil, broken down by state, is reported in this dataset. The series spans around ten years (1998 to 2017). The information was collected from the Brazilian government's official website.



amazon.csv (260.93 kB)

About this file

This dataset report of the number of forest fires in Brazil divided by states. The series comprises the period of approximately 10 years (1998 to 2017).

# year	state	month	# number	date
Year when Forest Fires happen	Brazilian State	Month when Forest Fires happen	Number of Forest Fires reported	Date when Forest Fires where reported
1998 2017	Rio 11% Mato Grosso 7% Other (5259) 81%	Janeiro 8% Fevereiro 8% Other (5373) 83%	0 998	1Jan98 1Jan17
1998	Acre	Janeiro	0	1998-01-01
1999	Acre	Janeiro	0	1999-01-01
2000	Acre	Janeiro	0	2000-01-01
2001	Acre	Janeiro	0	2001-01-01
2002	Acre	Janeiro	0	2002-01-01
2003	Acre	Janeiro	10	2003-01-01
2004	Acre	Janeiro	0	2004-01-01
2005	Acre	Janeiro	12	2005-01-01
2006	Acre	Janeiro	4	2006-01-01
2007	Acre	Janeiro	0	2007-01-01

Figure 1: Dataset overview

1.6.3 Visuals and Understanding

Here is a map of Brazil, just for the reference to get familiar with the geography of Brazil and get a better understanding of the area.

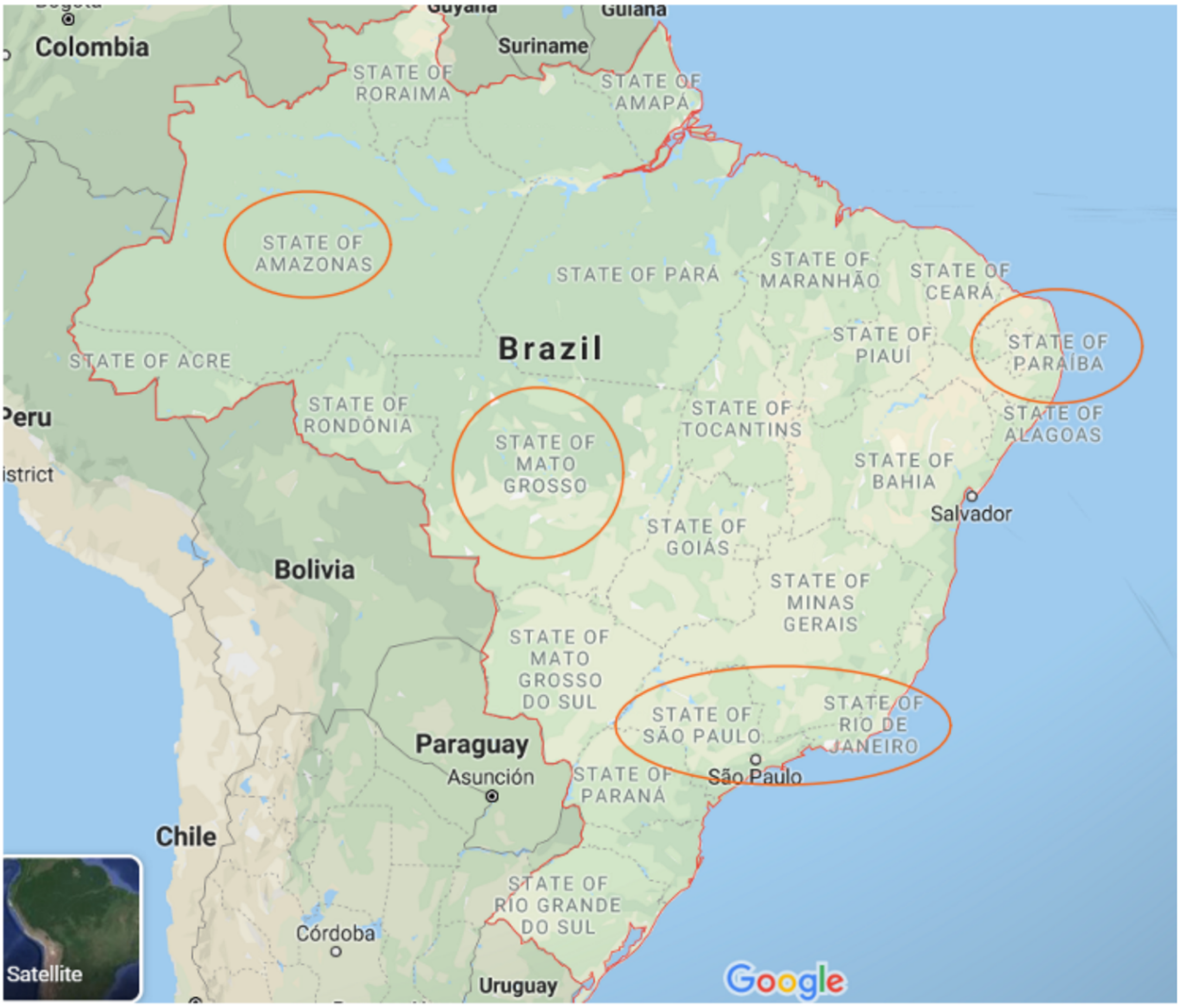


Figure 2: Map of Brazil

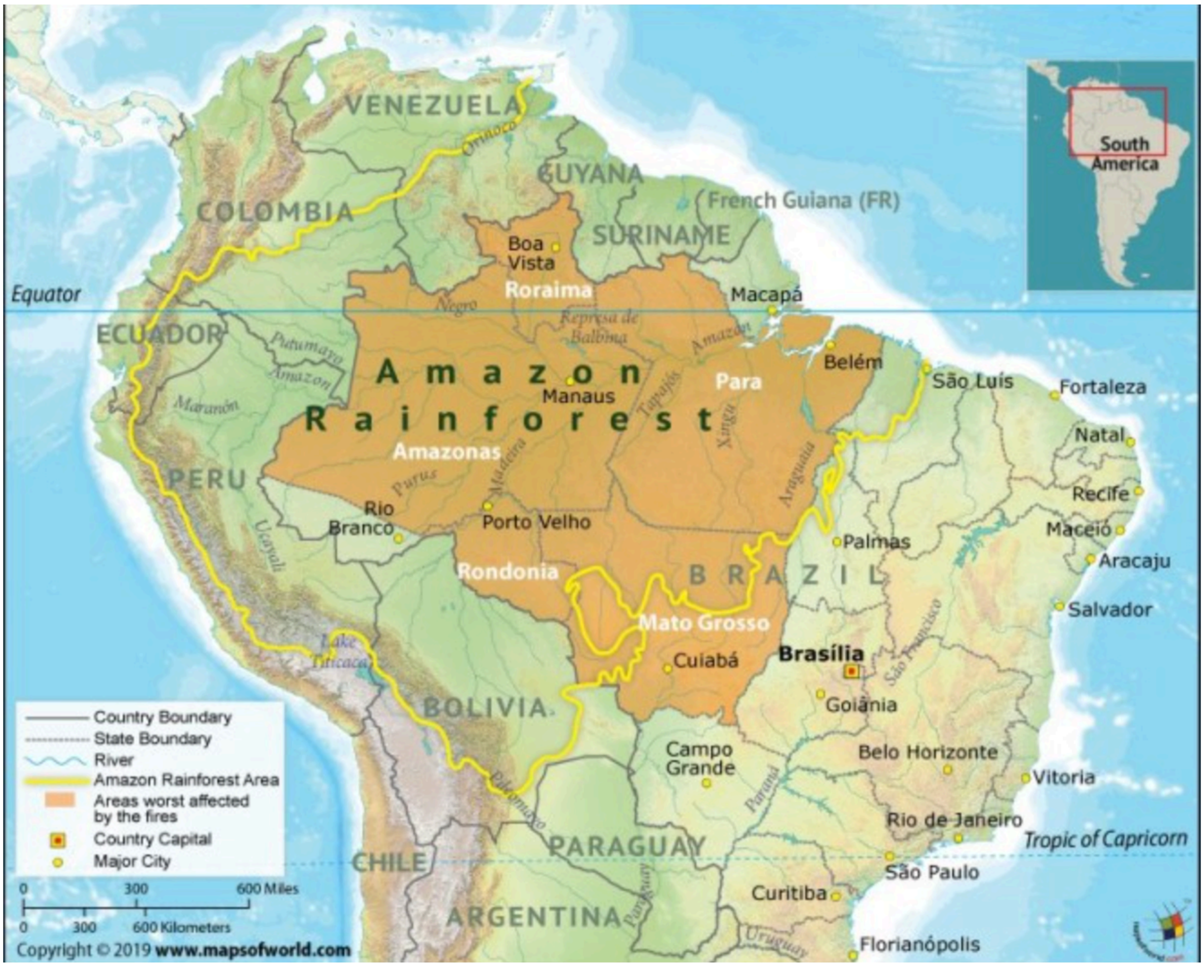


Figure 3: Amazonian Rainforest Spread

1.6.4 Inspiration

It is feasible to examine the evolution of fires over time as well as the places where they were concentrated using this information.
Acre, Amapá, Pará, Amazonas, Rondonia, Roraima, and parts of Mato Grosso, Tocantins, and Maranho make up the legal Amazon.

1.6.5 Acknowledgement

The data set used in this project is from Brazil, I thank the Brazilian system of forest information for providing this dataset,

Chapter 02: Literature Survey

Surapong Surit and Watchara Chatwiriya presented a method for detecting fire in video using smoke detection. This method is based on a digital image processing method that includes static and dynamic characteristic analysis. The proposed method consists of the following steps: the first is to detect the area of change in the current input frame when compared to the background image; the second is to locate regions of interest (ROIs) using a connected component algorithm; the area of ROI is calculated using a convex hull algorithm, which segments the area of change from the image; the third step is to calculate static and dynamic characteristics; and finally, we decide whether or not to use the proposed method. This approach accurately identifies fire smoke, according to the results.

A approach based on the wavelet model and a smoke colour model was proposed by P. Piccinini, S. Calderara, and R. Cucchiara. The suggested method takes advantage of two features: energy variation in wavelet models and a smoke colour model. The decrease of the energy ratio in the wavelet domain between background and current is used to detect smoke. The colour model calculates the colour deviation of the current pixel. To detect smoke, a Bayesian classifier is employed to combine these two features.

Covariance Descriptors, Color Models, and SVM Classifiers were proposed by Osman Gunay and Habiboglu. Video data is used in this system. This approach uses the Spatio-temporal Covariance Matrix (2011), which separates video data into temporal blocks and computes covariance characteristics. This feature is used to detect the fire. To filer fire and fire-like regions, the SVM Classifier is employed. This system only works with clear data and not with blurred data.

R.Gonzalez presented a Wavelet Transform-based approach for detecting fire. The Region of Interest is detected using the Stationary Wavelet Transform. Preprocessing, SWT, and histogram analysis are the three processes in this procedure. During preprocessing, undesired distortions are removed, the image is resized, and the resized image is transformed. SWT is used to remove high frequencies from an image, and inverse SWT is used to recreate the image. The intensity colours that are near to each other are grouped together using image indexing. The various levels of indexation are determined via histogram analysis. Following the analysis, a comparison with a non-smoke frame is done, and non-smoke images are deleted. When these three elements are together, fire is recognised.

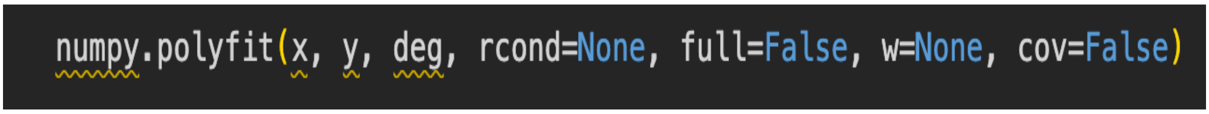
Chapter 03: System Development

3.1 Feasibility Study on Major Project

BACKGROUND: The proposed system includes a fire risk index that ranges from 1 to 4, with 1 indicating the lowest fire danger and 4 indicating the highest fire risk. This index represents the maximum number of fires that could occur on a given day, and can thus be used to estimate the actual number of fires that day. It is necessary to define the parameters or features observed during the day that will be used in the prediction algorithm in order to execute prediction.

3.2 Numpy Polyfit

NumPy.polyfit() assists us by determining the least square polynomial fit. By minimising the sum of squares, the best fitting curve to a given set of points can be found. The user must provide three different inputs: X, Y, and the polynomial degree. The values that we wish to fit on the two axes are represented by X and Y. Let's have a look at its syntax next.



```
numpy.polyfit(x, y, deg, rcond=None, full=False, w=None, cov=False)
```

Figure 4: Syntax of Numpy Polyfit()

The general syntax of our NumPy polyfit function is shown above (). It has three mandatory parameters, as mentioned above, and four optional parameters, each of which has its own effect on the outcome. Following that, we'll go over the many factors that come with it.

3.2.1 Parameters of Numpy Polyfit()

1. X:array_like
   It shows the points that will be displayed on the X-axis

2. Y:array_like
   This parameter is used to indicate all sets of points along the Y-axis.
3. Deg: int
   The degree of the fitting polynomial is represented by this parameter.

4. Rcond: float
   An optional parameter that is used to define the fit's relative number condition. Smaller singular values are overlooked in comparison to the largest singular values.

5. Full: bool
   optional parameter that changes the return value's determining nature. Only the coefficients are returned by default because the value is set to false. The decomposition singular value is also returned if the value is set to true.

6. W:array_like
   This optional parameter specifies the weights to be applied to the sample points' y-coordinate.

7. Cov:bool or str
   If true, this optional argument returns not just an array but also a covariance matrix.

RETURN

1. P: nadarray
   The polynomial coefficient with the highest power is returned first.

2. residuals, rank, rcond

We get this if "full=True" is set. The total of squared residuals of the least square fit is called the residual.

3. V: ndarray
   Only if "full=false" and "cov=true" are set, do we receive this. A covariance matrix of the polynomial coefficient estimate is also provided.

We'll look at an example from the project where we have used poly1d function to create the best fit line.

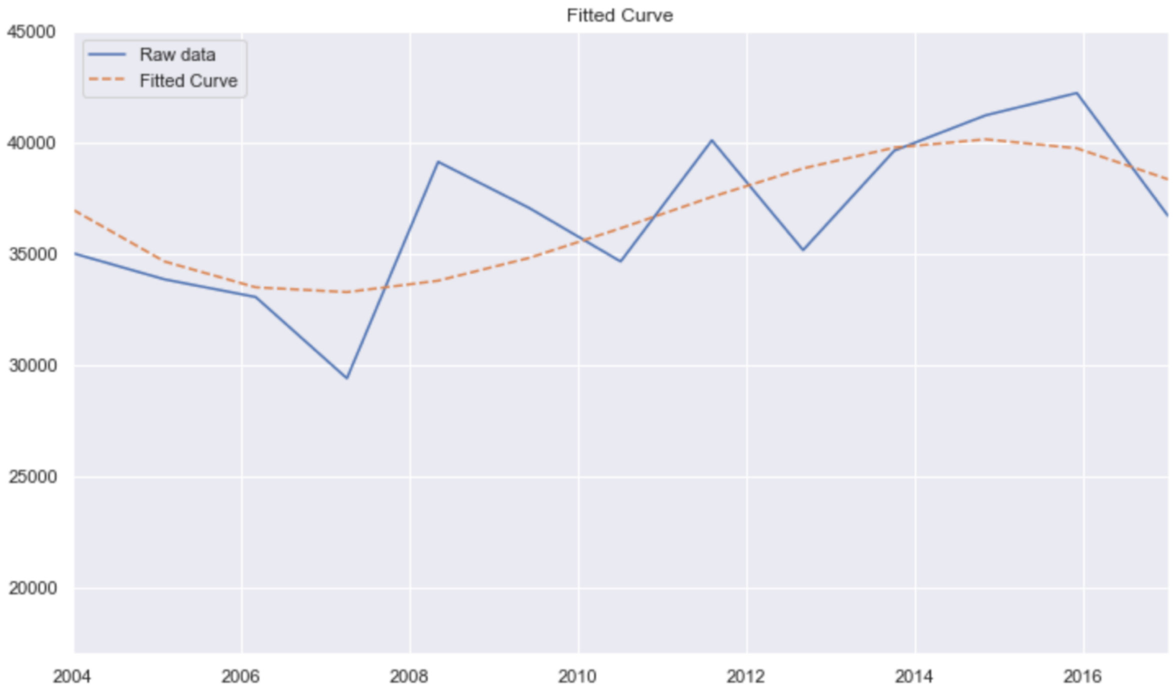
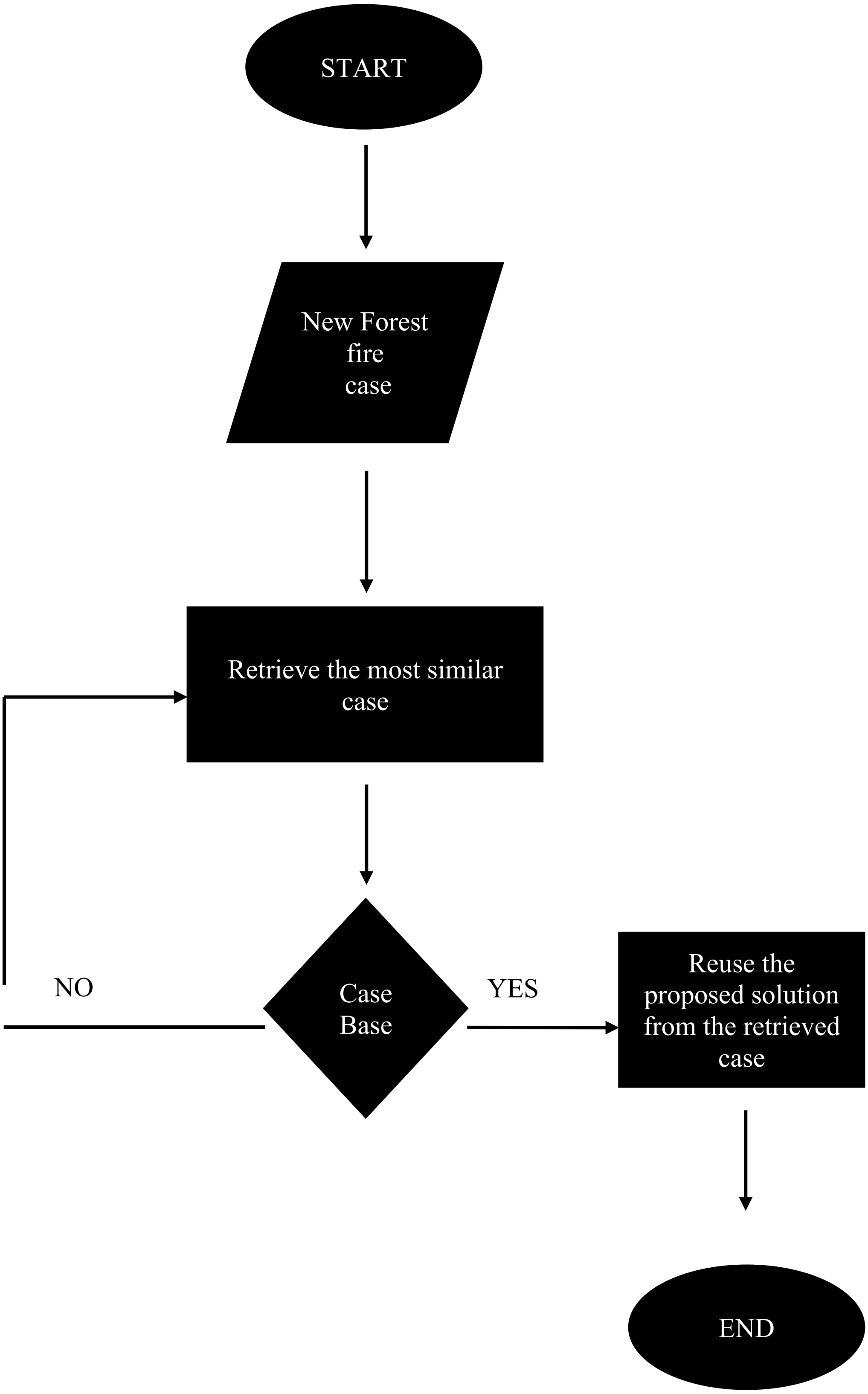


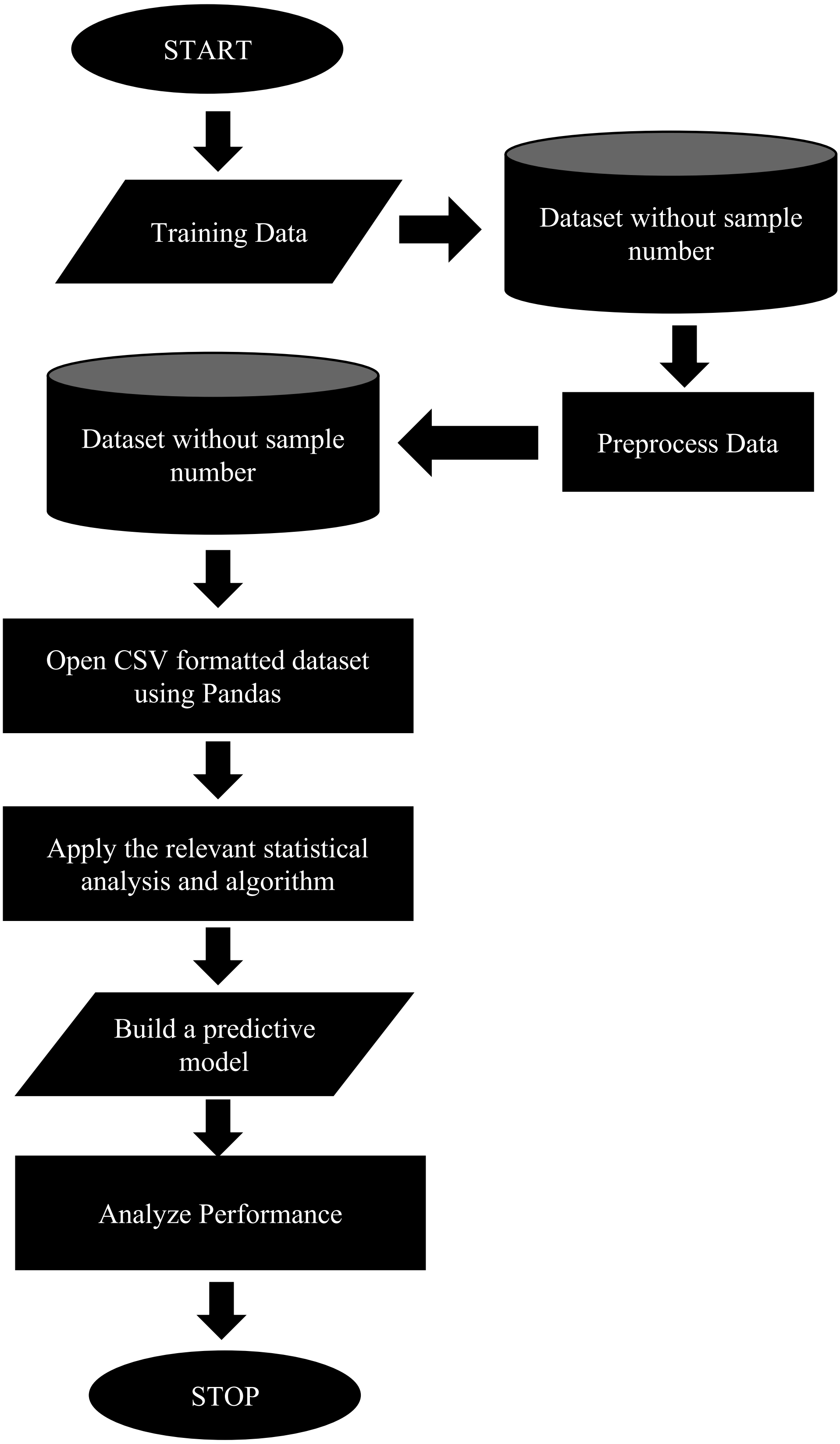
Figure 5: Poly1d Function

To create the best fitting line equation from polyfit, utilise numpy's poly1d function. To draw a line alongside your data, use this equation in plt.plot().

3.3 Case Diagram of the Project


START
New Forest fire case
Retrieve the most similar case
Case Base
NO
YES
Reuse the proposed solution from the retrieved case
END

3.4 DFD Diagram of the Forest Fire Analysis and Prediction


START
Training Data
Dataset without sample number
Preprocess Data
Dataset without sample number
Open CSV formatted dataset using Pandas
Apply the relevant statistical analysis and algorithm
Build a predictive model
Analyze Performance
STOP

Chapter 04: Implementation of the Major Project

4.1 Data Visualization in Python

The field of data visualisation is concerned with the visual representation of data. It depicts data graphically and is a good approach to express data findings.

We may acquire a visual summary of our data by using data visualisation. The human mind has an easier time processing and understanding data when it is presented in the form of drawings, maps, and graphs. Data visualisation is important in the depiction of both small and large data sets, but it is especially useful when dealing with large data sets when seeing all of our data, much alone processing and understanding it, is impossible.

4.1.1 Matplotlib and Seaborn

Python libraries for data visualisation include Matplotlib and Seaborn. They provide modules for plotting various graphs. Seaborn is mostly used for statistical graphs, whereas Matplotlib is used to embed graphs into programmes.

But when should either of these be used? Let's look at this through the lens of comparative analysis. Matplotlib and Seaborn, two well-known Python visualisation libraries, are compared in the table below.

Table 1: Matplotlib and Seaborn

Matplotlib	Seaborn
It's used to make simple graphs like line charts and bar graphs.	It's primarily used for statistics visualisation, and it can handle more complicated visualisations with less commands.
It primarily uses datasets and arrays.	It uses whole Dataset
Matplotlib is capable of working with data arrays and frames. It considers aces and figures to be objects.	Seaborn is far more organised and functional than Matplotlib, because it treats the entire dataset as if it were a single entity.
For exploratory data analysis, Matplotlib is more flexible and works well with Pandas and Numpy.	Seaborn offers a larger number of pre-installed themes and is primarily used for statistical analysis.

4.1.2 Bar Plot

Let’s use the tips dataset in Seaborn next. The dataset consists of :

- Worst hit Sates
- Year
- Number of Fires

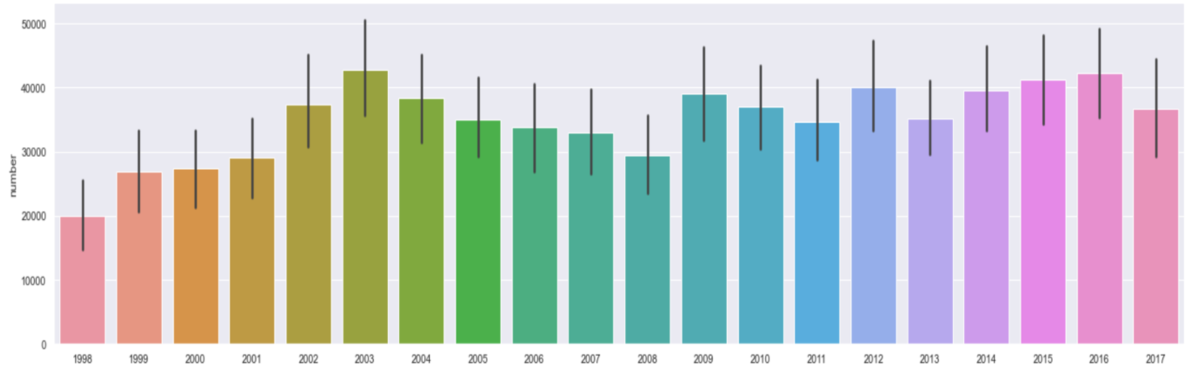


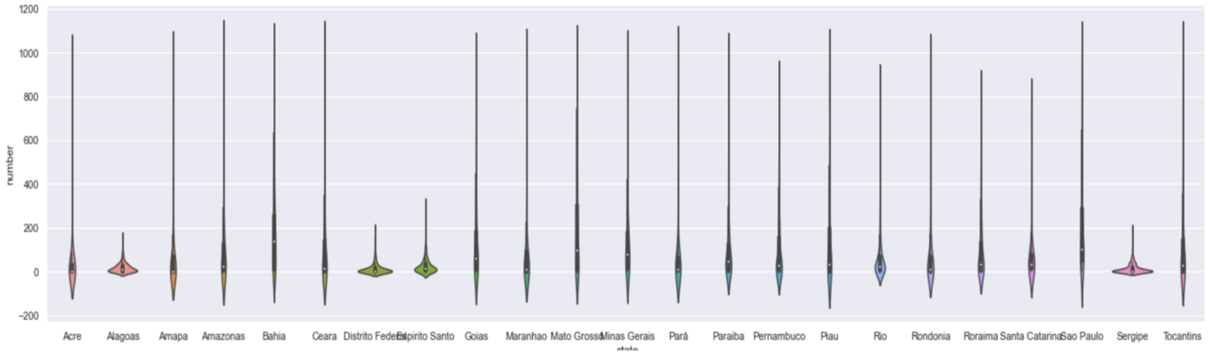
Figure 6: Bar-Plot

We can use a bar chart to depict how the number of forest fires changes over time. This can be accomplished by computing year-by-year averages and then using plt.bar. The Seaborn library also has a barplot function that can compute averages automatically.

4.1.3 Seaborn Violin Plot

A violin plot is comparable to a box and whisker plot in terms of function. It depicts the distribution of quantitative data across multiple levels of one (or more) categorical variables, allowing for comparison. The violin plot, unlike a box plot, has a kernel density estimate of the underlying distribution.

This can be a useful and appealing approach to display numerous data distributions at once, but keep in mind that the estimate procedure is impacted by sample size, and violins for small samples may appear misleadingly smooth.



Input data can be sent in a number of different formats, including:

- Data vectors supplied directly to the x, y, and/or hue parameters as lists, numpy arrays, or pandas Series objects.

- The x, y, and hue variables will control how the data is plotted in a "long-form" DataFrame.

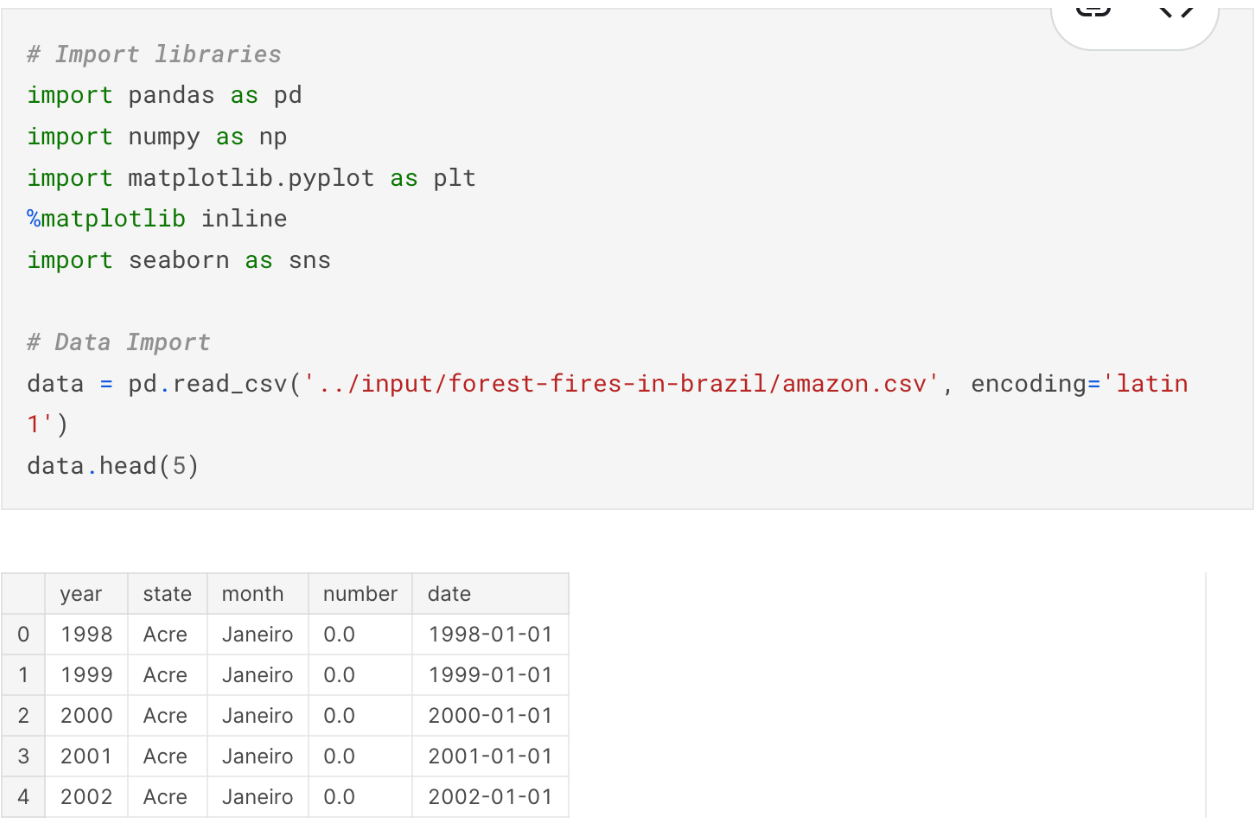
- Each numeric column will be plotted in this "wide-form" DataFrame.

- A list or array of vectors

4.2 Dataset used in the Minor Project

In this Project we'll use the Forest Fire in Brazil Data Set from Kaggle, Feature units furnished withinside the dataset are computed from the number of forest fires in Brazil, broken down by state, is reported in this dataset. The series spans around ten years (1998 to 2017). The information was collected from the Brazilian government's official website.

4.2.1 Data Import



```
# Import libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns

# Data Import
data = pd.read_csv('../input/forest-fires-in-brazil/amazon.csv', encoding='latin1')
data.head(5)
```

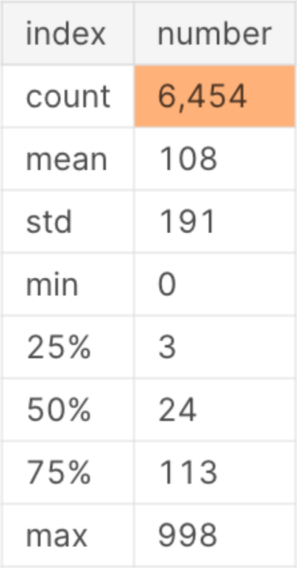
	year	state	month	number	date
0	1998	Acre	Janeiro	0.0	1998-01-01
1	1999	Acre	Janeiro	0.0	1999-01-01
2	2000	Acre	Janeiro	0.0	2000-01-01
3	2001	Acre	Janeiro	0.0	2001-01-01
4	2002	Acre	Janeiro	0.0	2002-01-01

Figure 7: Data Import

4.3 Data Set Features

4.3.1 Types of Data Set

Data Set is in the form of .CSV file which will be use to read in the program using Pandas,
There are total of 6,454 entries present in the data with a mean of 108 and standard deviation of 191.



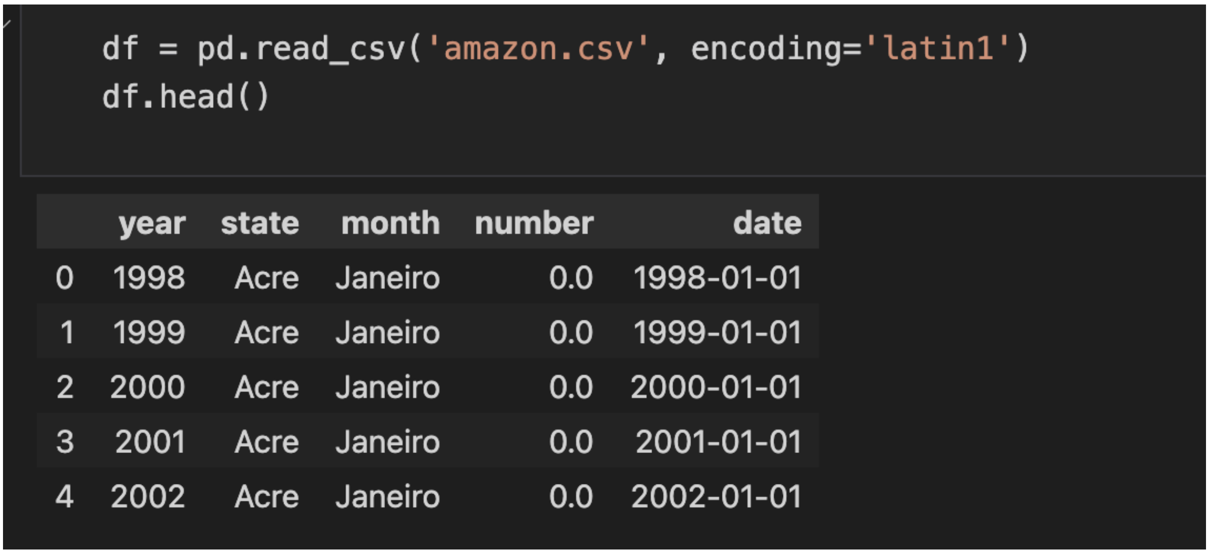
index	number
count	6,454
mean	108
std	191
min	0
25%	3
50%	24
75%	113
max	998

Figure 8: Data Insight

4.3.2 Number of Attributes, fields, Description of the Data Set

Attribute Information:

- year - the year when the fires occurred;
- state - the state in which the fires occurred;
- month - month in which the fires occurred;
- number - the number of fires;
- date - date of fire.



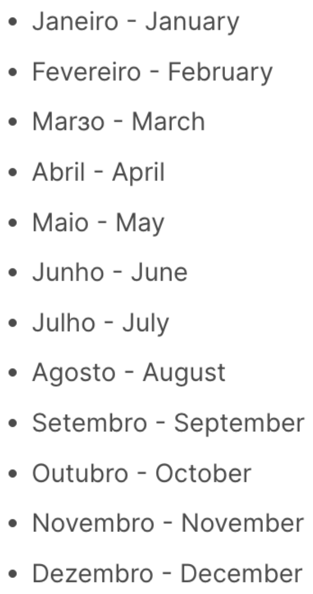
```
df = pd.read_csv('amazon.csv', encoding='latin1')
df.head()
```

```
   year  state  month    number        date
0  1998  Acre   Janeiro     0.0  1998-01-01
1  1999  Acre   Janeiro     0.0  1999-01-01
2  2000  Acre   Janeiro     0.0  2000-01-01
3  2001  Acre   Janeiro     0.0  2001-01-01
4  2002  Acre   Janeiro     0.0  2002-01-01
```

Figure 9: Raw Data

Figure 4 is a Raw data retrieved by using the Pandas framework from the Dataset,

The months are written in Portuguese since the data has been collected from various states of Brazil, and Portuguese is the official language, these months need to be converted to the English language months so as to get a better understanding of the model.



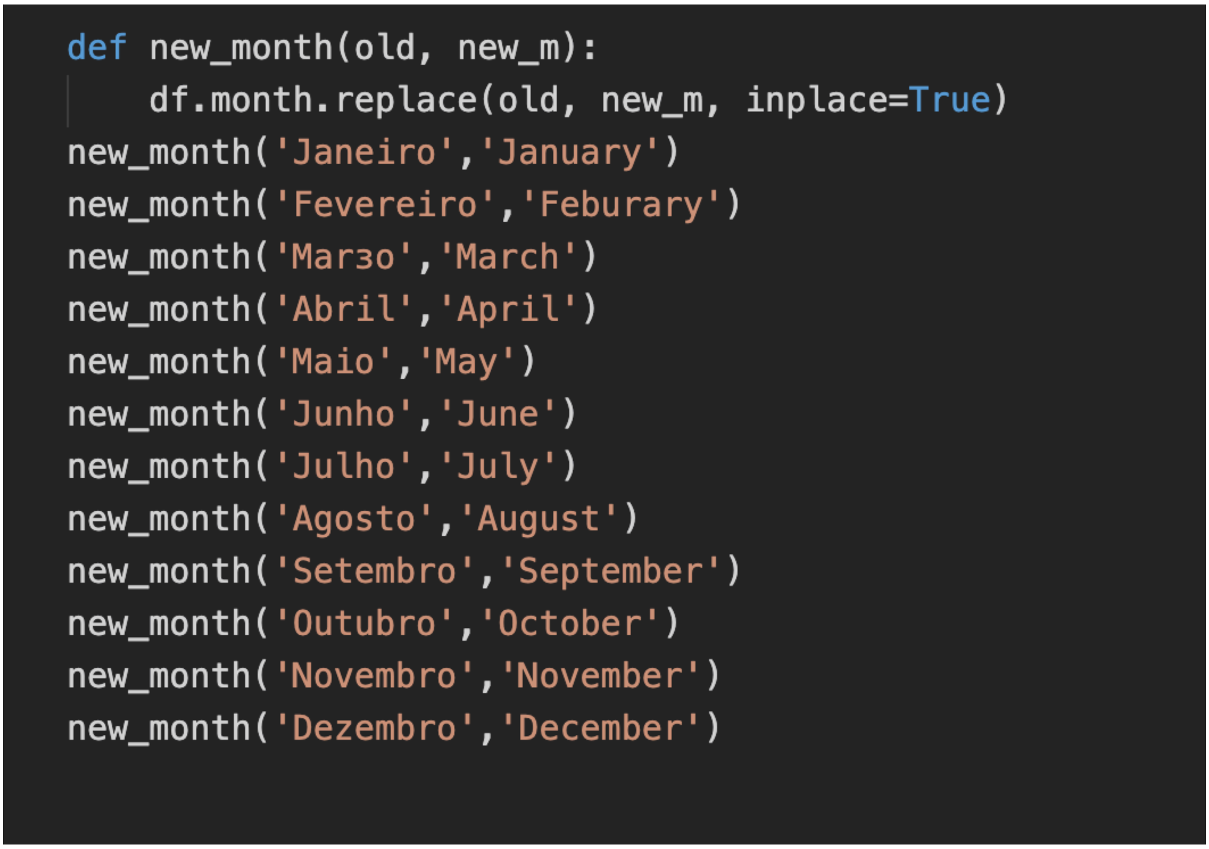
- Janeiro - January
- Fevereiro - February
- Marзo - March
- Abril - April
- Maio - May
- Junho - June
- Julho - July
- Agosto - August
- Setembro - September
- Outubro - October
- Novembro - November
- Dezembro - December

Figure 10:Translating Months

There are many ways to convert the months to English, I have used a very simple yet powerful method which is the replace() function.

The replace() method copies the string and replaces the previous substring with the new substring. The original string has not been altered.

If the old substring cannot be retrieved, the copy of the original string is returned.



```
def new_month(old, new_m):
    df.month.replace(old, new_m, inplace=True)
new_month('Janeiro','January')
new_month('Fevereiro','Feburary')
new_month('Marзo','March')
new_month('Abril','April')
new_month('Maio','May')
new_month('Junho','June')
new_month('Julho','July')
new_month('Agosto','August')
new_month('Setembro','September')
new_month('Outubro','October')
new_month('Novembro','November')
new_month('Dezembro','December')
```

Figure 11: Replace Function

4.3.3 Visualizing Data

Now we will look more closely at the data and get an overview on which all states have the highest number of forest fires.

We have used a Bar-plot for finding out the top 5 states recording the highest forest fires from 1998 to 2017, It clearly shows the state Mato Grosso has the highest number of forest fires followed by Paraiba, Sao Paulo, Rio and at last Bahia.

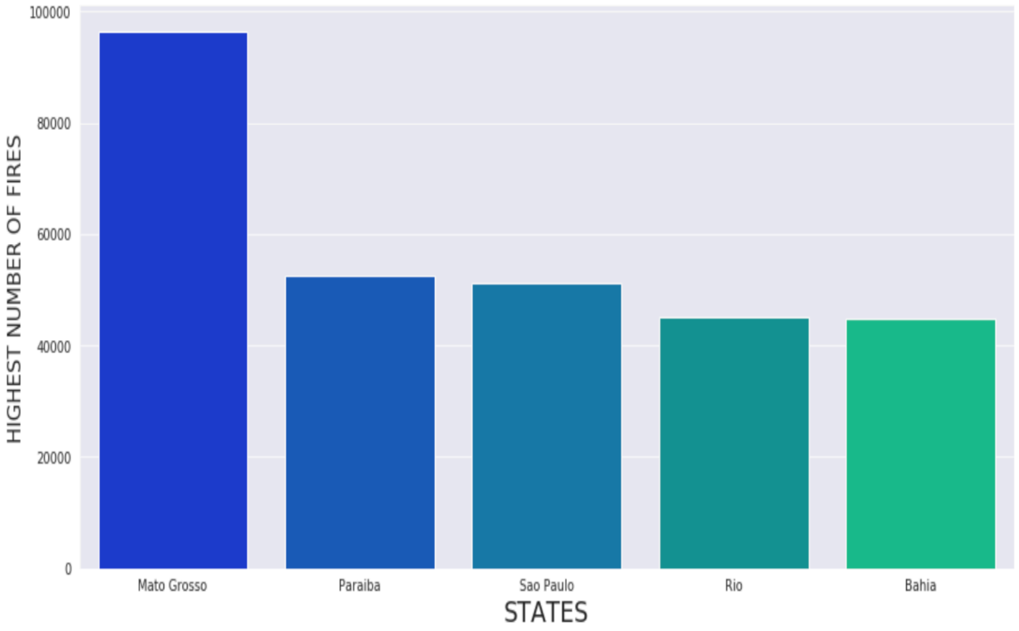


Figure 12: TOP 5 STATES RECORDING HIGHEST FOREST FIRES FROM 1998 TO 2017

The state of Mato Grosso had the most fires, with 96k, and three other states can be identified by the number of fires: Sao Paulo, Rio de Janeiro, and Bahia.

Short insight on Mato Grosso:

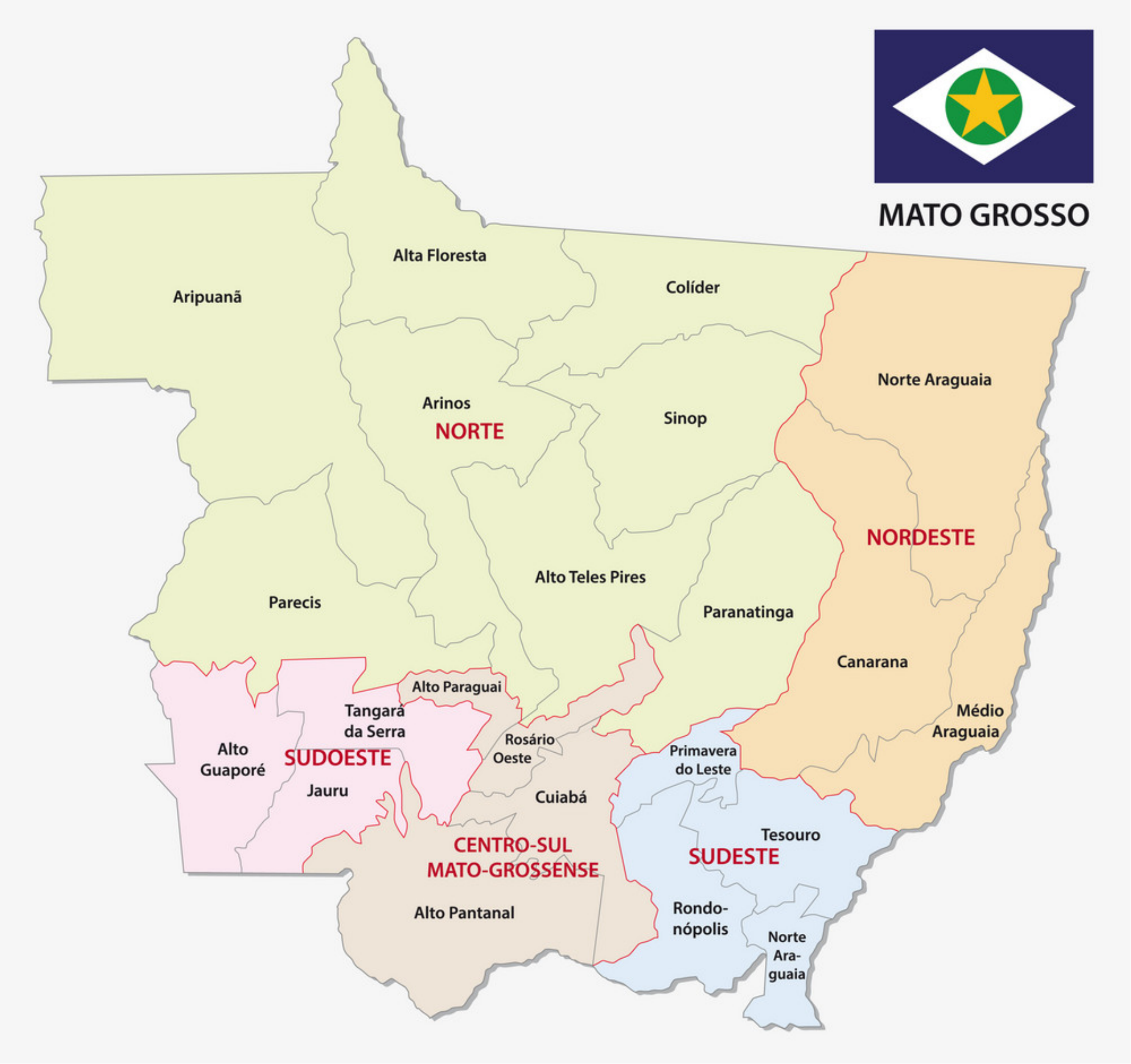
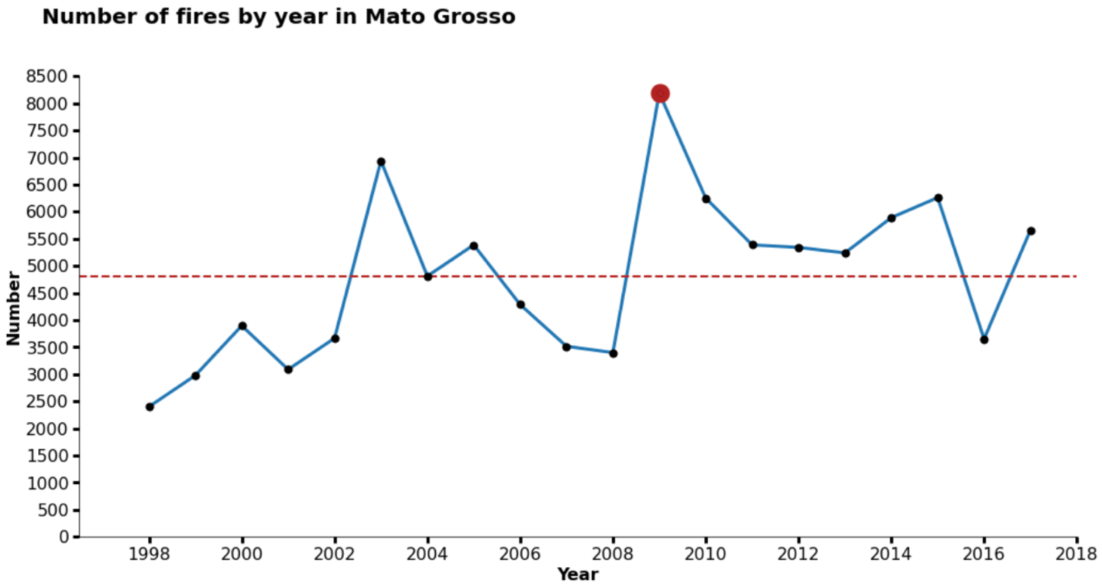


Figure 13: Insight on Mato Grosso

Mato Grosso is the third-largest state in terms of land area, and it is located in the Central-West region. The state is home to 1.66 percent of Brazil's population and accounts for 1.9 percent of the country's GDP. Mato Grosso is a flat state with extensive chapadas and plain landscapes. It is home to three major ecosystems: the Cerrado, the Pantanal, and the Amazon rainforest. Mato Grosso is a flat state with extensive chapadas and plain landscapes. It is home to three major ecosystems: the Cerrado, the Pantanal, and the Amazon rainforest.

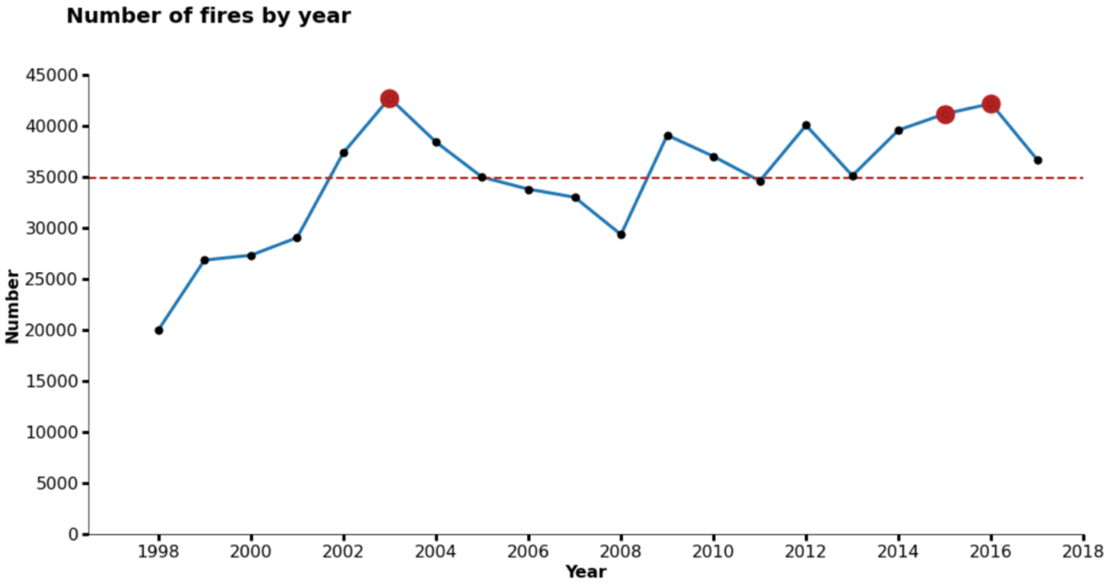
Lets check the values of fire in the state of Mato Grosso



There is a lot of variation. In 1998, there were 2.4 thousand fires in this state. In 2009, 8.2 thousand fires were reported.

There were 4.8 thousand fires per year on an average

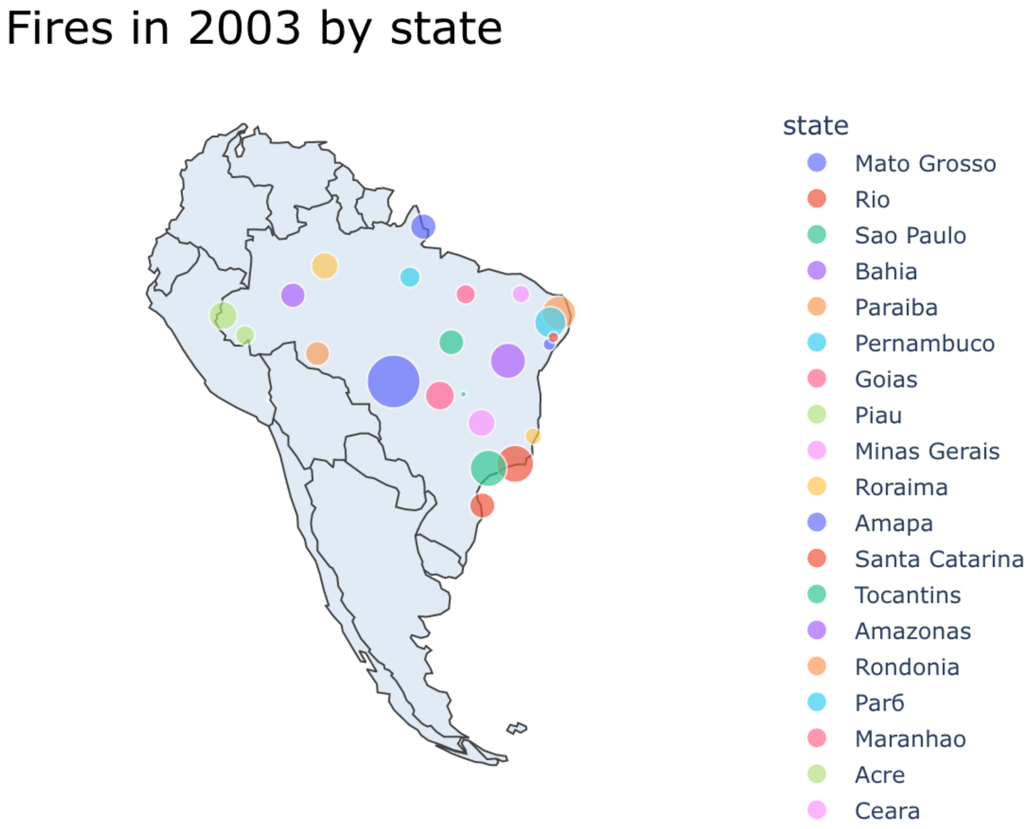
After Visualizing the number of fires by state, now we will look at the number of fires by year.



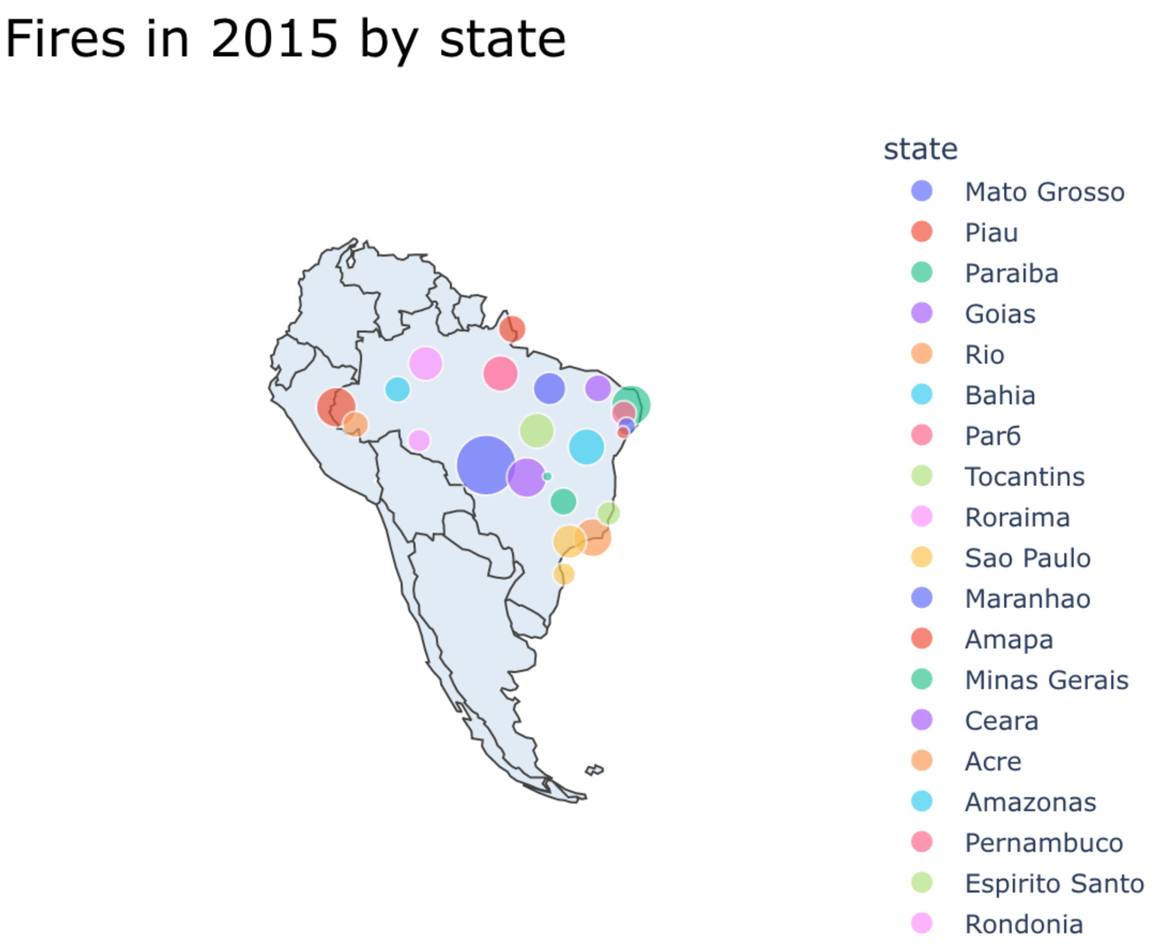
The research spans 20 years, from 1998 to 2017. Almost 700,000 fires occurred over this time span. The years with the most fires were 2003, 2015, and 2016.

There were about 35k fires every year on average.

Let's look at the amount of fires that occurred in each state in 2003, 2015, and 2016.



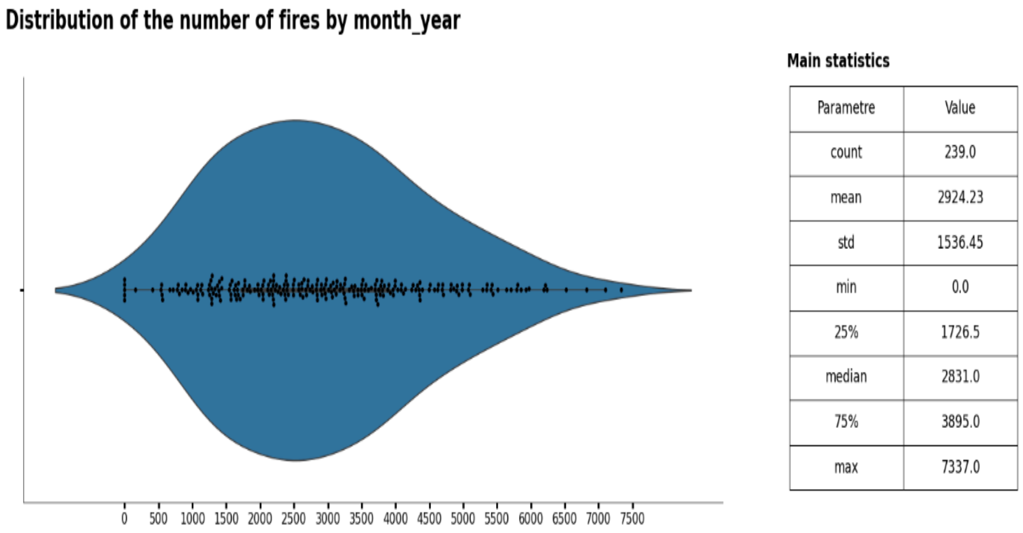
Mato Grosso had the most fires (7k), Rio had 3.4k, and Sao Paulo had 3.3k.



Fires in 2015 by state

The states with the most fires were Mato Grosso (6.2%), Piau (2.8%), and Paraiba (2.7%).

Let's look at the amount of fires per month. But first, let's look at how the number of fires is distributed.

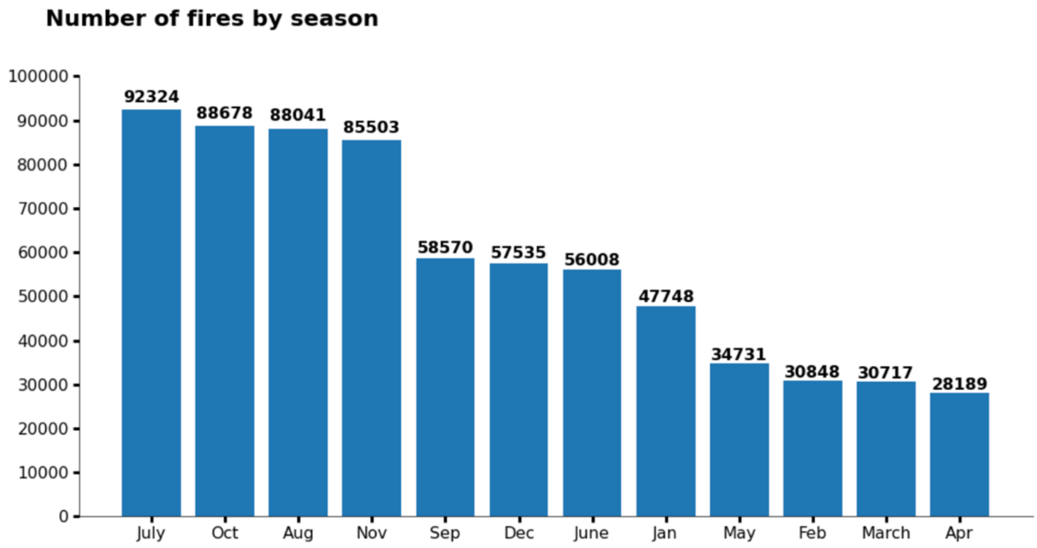


Main statistics

Parametre	Value
count	239.0
mean	2924.23
std	1536.45
min	0.0
25%	1726.5
median	2831.0
75%	3895.0
max	7337.0

- The number of months studied - 239;
- There is a slight displacement of the data to the right, which is caused by the outlier. The largest number of fires by month and year ~ in the range 1.2k - 4.3k;
- Average number of fires per month - 2.9k;
- Minimum number of fires - 0;

- The maximum number of fires - 7.;
- We also observe one pronounced outburst.

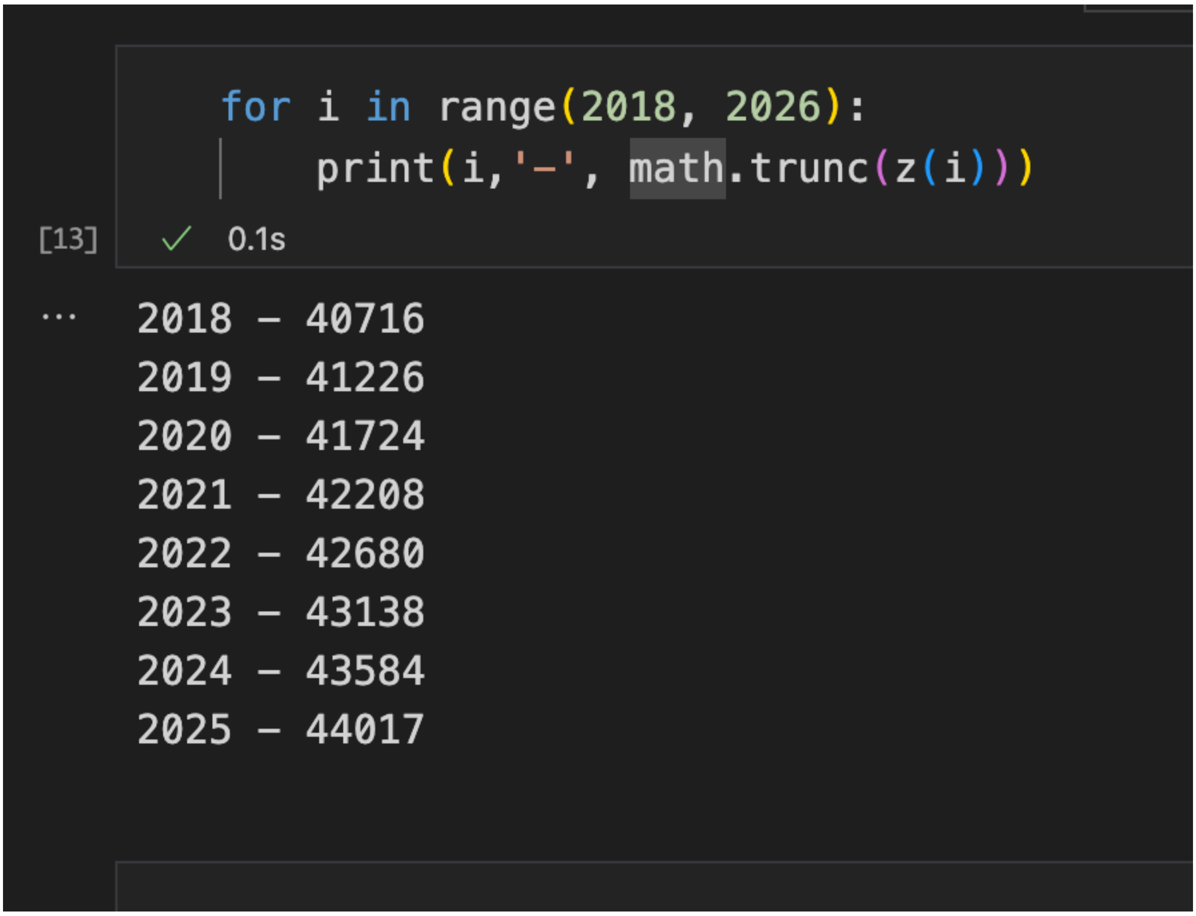


July had the most flames, but there were also several fires in October, August, and November.

4.4 Prediction

Finally we have used the Numpy's Polyfit() function to do some basic prediction by creating the best fit line and later using the Trunc() function.

When the supplied input is positive, Python trunc() behaves like a floor(); when the given input is negative, it behaves like a ceil(). The floor() and ceil() functions are used to round down to negative infinity and up to positive infinity, respectively.

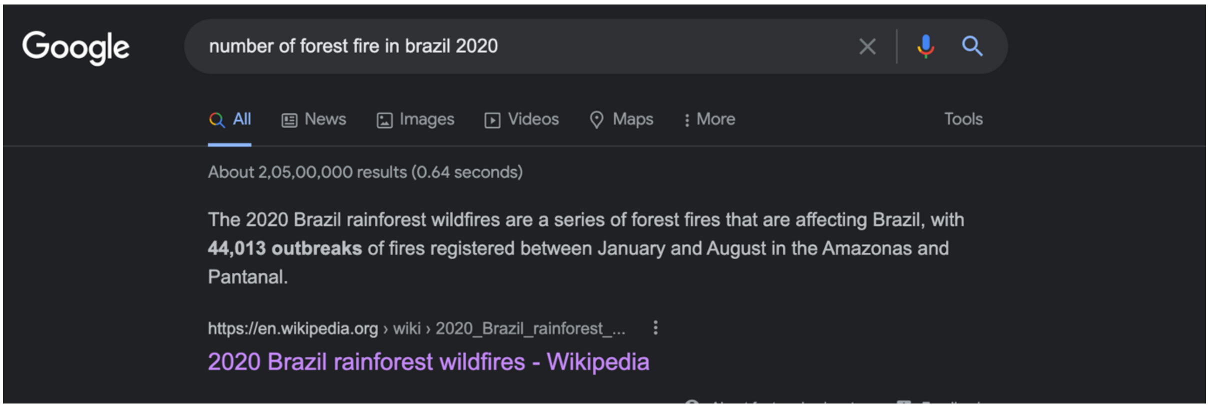


```
for i in range(2018, 2026):
    print(i,'-', math.trunc(z(i)))
```

```
2018 - 40716
2019 - 41226
2020 - 41724
2021 - 42208
2022 - 42680
2023 - 43138
2024 - 43584
2025 - 44017
```

Figure 14: Predicted Data

Let’s look at the real data and compare it with our predicted data for the year 2020.



number of forest fire in brazil 2020

About 2,05,00,000 results (0.64 seconds)

The 2020 Brazil rainforest wildfires are a series of forest fires that are affecting Brazil, with **44,013 outbreaks** of fires registered between January and August in the Amazonas and Pantanal.

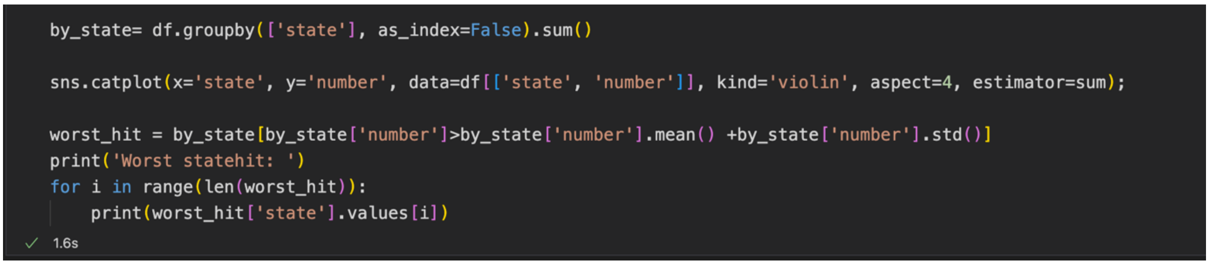
https://en.wikipedia.org › wiki › 2020_Brazil_rainforest_...

2020 Brazil rainforest wildfires - Wikipedia

Figure 15: Real Data

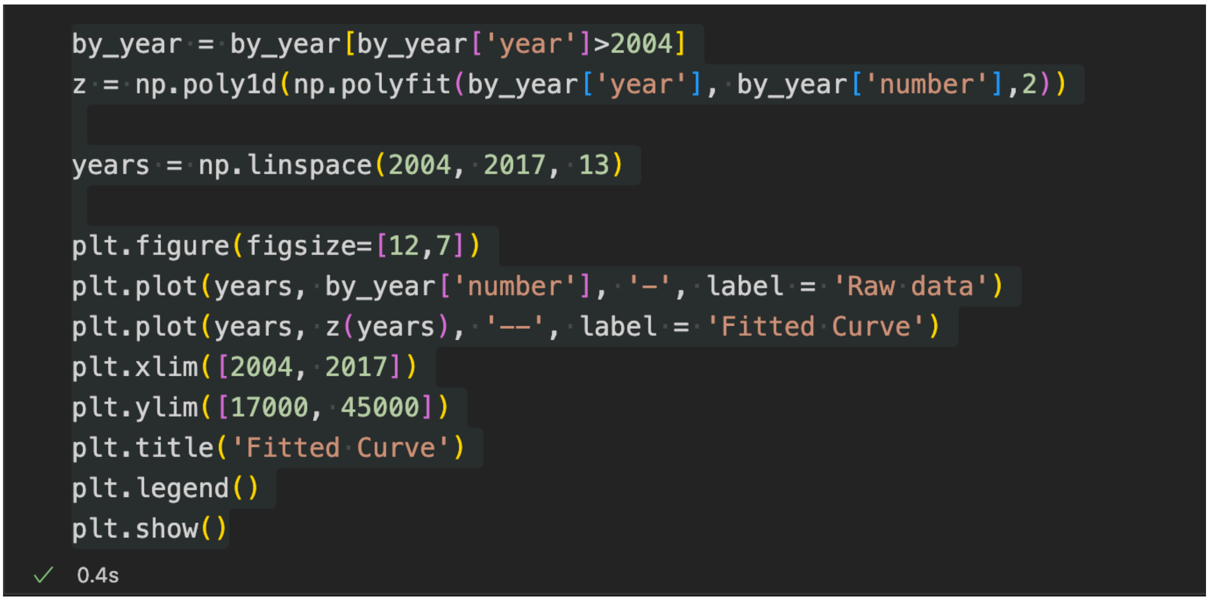
As we can see we are only few thousand less than the real data, our model predicted 41,724 fires for the year 2020.

4.5 Algorithm / code of the Project Problem



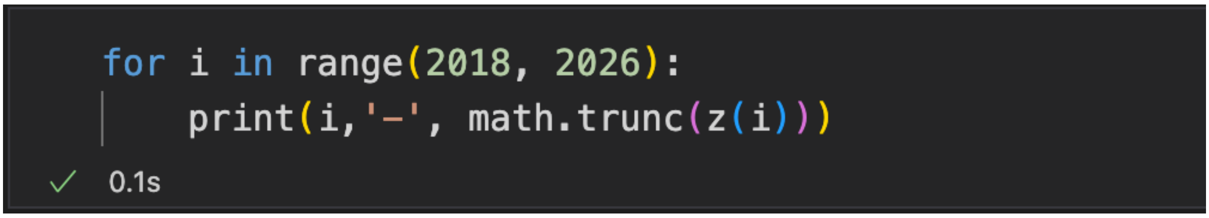
```
by_state= df.groupby(['state'], as_index=False).sum()

sns.catplot(x='state', y='number', data=df[['state', 'number']], kind='violin', aspect=4, estimator=sum);

worst_hit = by_state[by_state['number']>by_state['number'].mean() +by_state['number'].std()]
print('Worst statehit: ')
for i in range(len(worst_hit)):
    print(worst_hit['state'].values[i])
```

✓ 1.6s



```
by_year = by_year[by_year['year']>2004]
z = np.poly1d(np.polyfit(by_year['year'], by_year['number'],2))

years = np.linspace(2004, 2017, 13)

plt.figure(figsize=[12,7])
plt.plot(years, by_year['number'], '-', label = 'Raw data')
plt.plot(years, z(years), '--', label = 'Fitted Curve')
plt.xlim([2004, 2017])
plt.ylim([17000, 45000])
plt.title('Fitted Curve')
plt.legend()
plt.show()
```

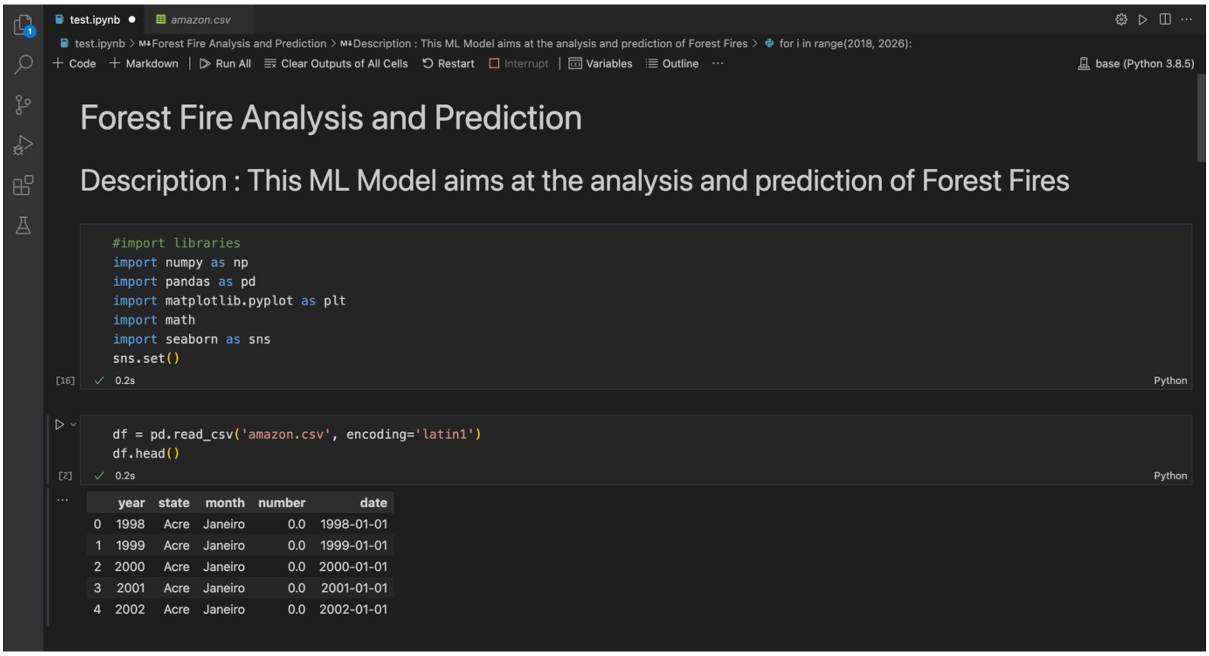
✓ 0.4s



```
for i in range(2018, 2026):
    print(i,'-', math.trunc(z(i)))
```

✓ 0.1s

4.6 Screenshot of the various stages of the Project



test.ipynb
amazon.csv
test.ipynb > M↓Forest Fire Analysis and Prediction > M↓Description : This ML Model aims at the analysis and prediction of Forest Fires > for i in range(2018, 2026):
+ Code
+ Markdown
Run All
Clear Outputs of All Cells
Restart
Interrupt
Variables
Outline
base (Python 3.8.5)

Forest Fire Analysis and Prediction

Description : This ML Model aims at the analysis and prediction of Forest Fires

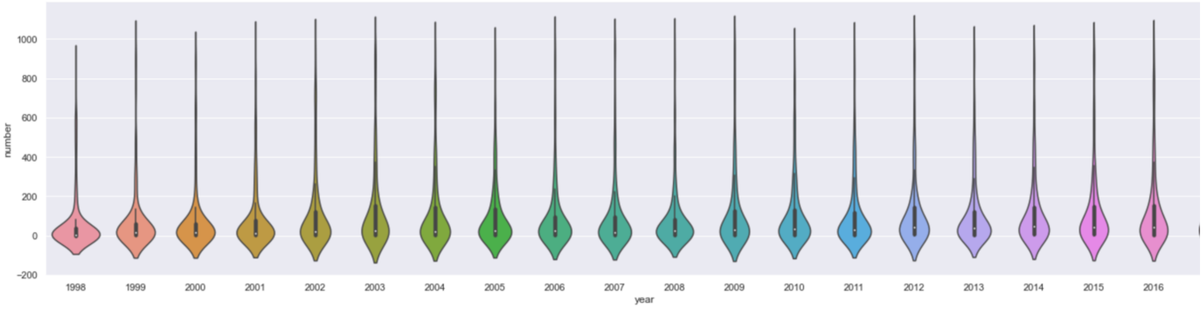
```
#import libraries
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import math
import seaborn as sns
sns.set()
```

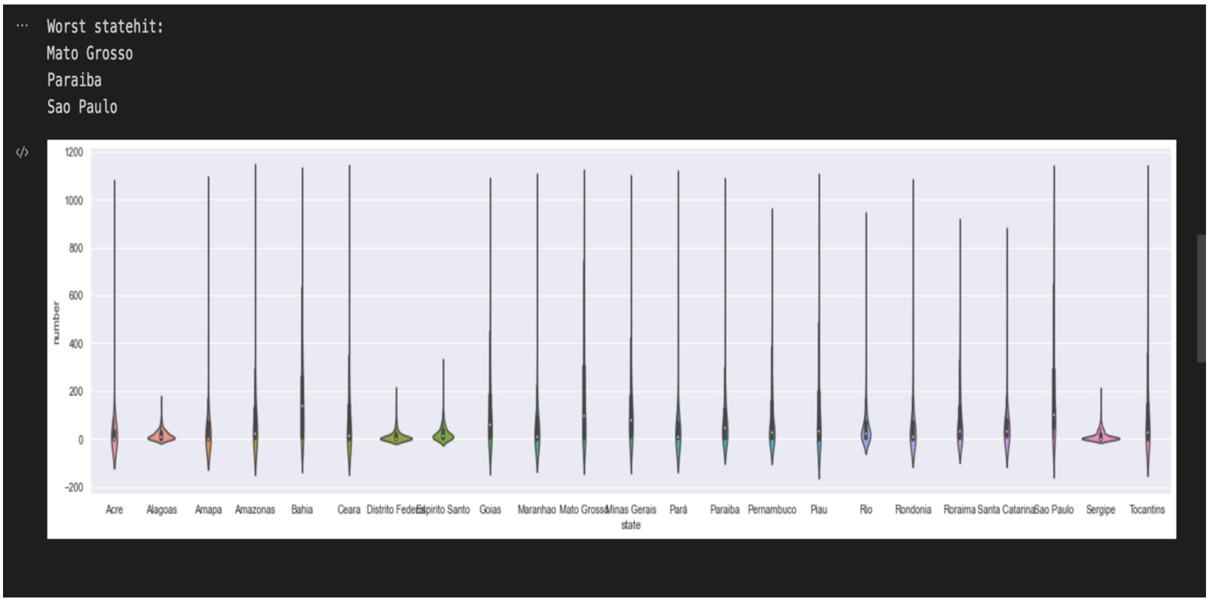
[16] ✓ 0.2s Python

```
df = pd.read_csv('amazon.csv', encoding='latin1')
df.head()
```

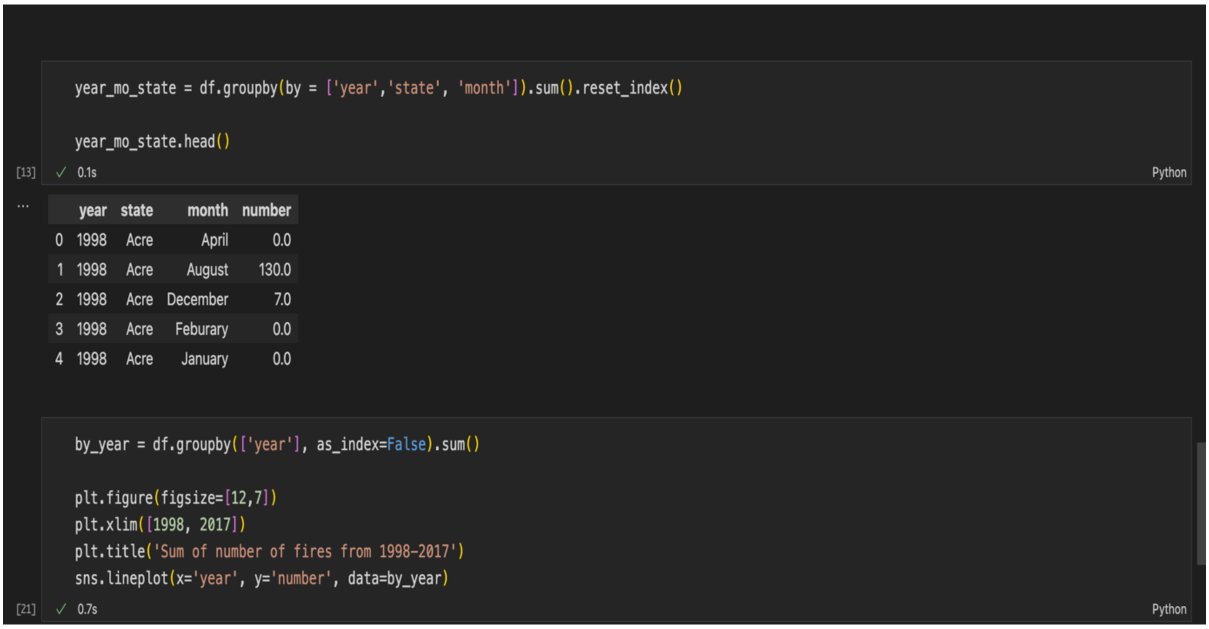
[2] ✓ 0.2s Python

```
   year  state  month    number  date
0  1998  Acre   Janeiro  0.0     1998-01-01
1  1999  Acre   Janeiro  0.0     1999-01-01
2  2000  Acre   Janeiro  0.0     2000-01-01
3  2001  Acre   Janeiro  0.0     2001-01-01
4  2002  Acre   Janeiro  0.0     2002-01-01
```





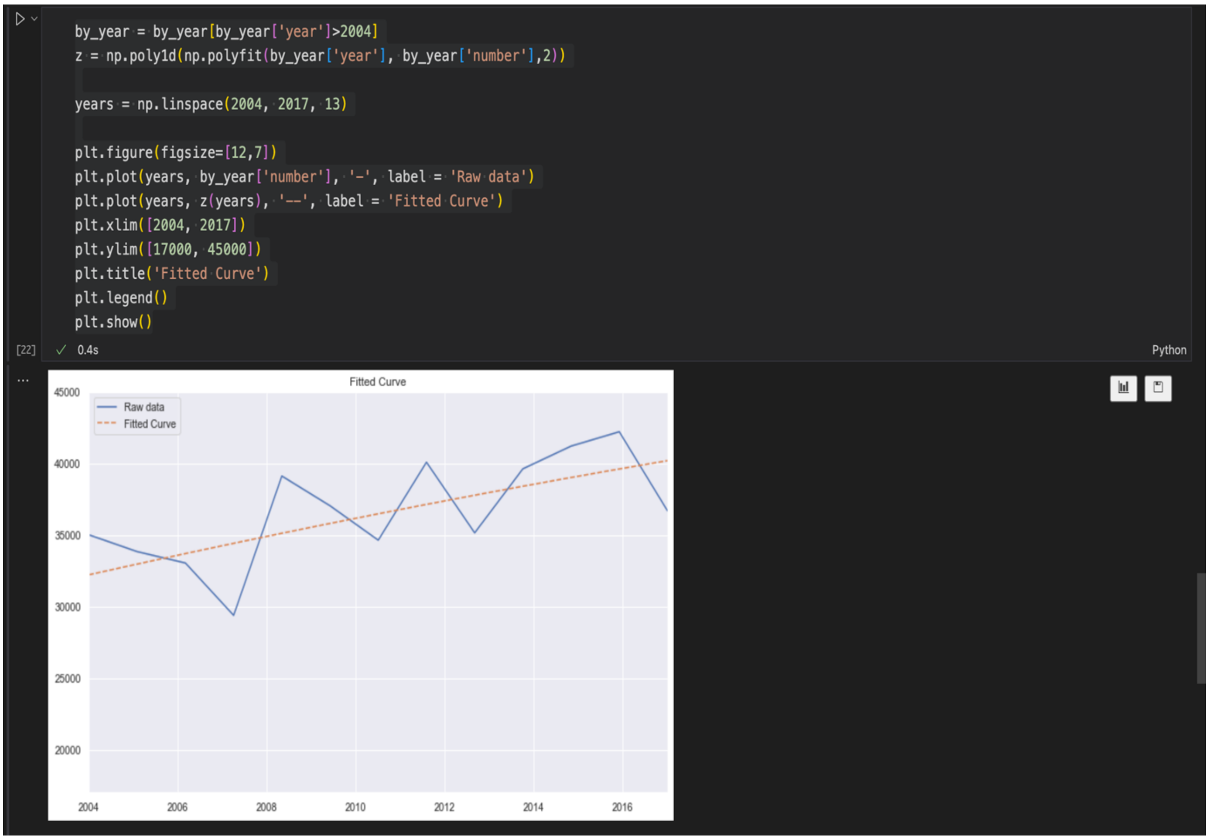
```
Worst statehit:
Mato Grosso
Paraiba
Sao Paulo
```



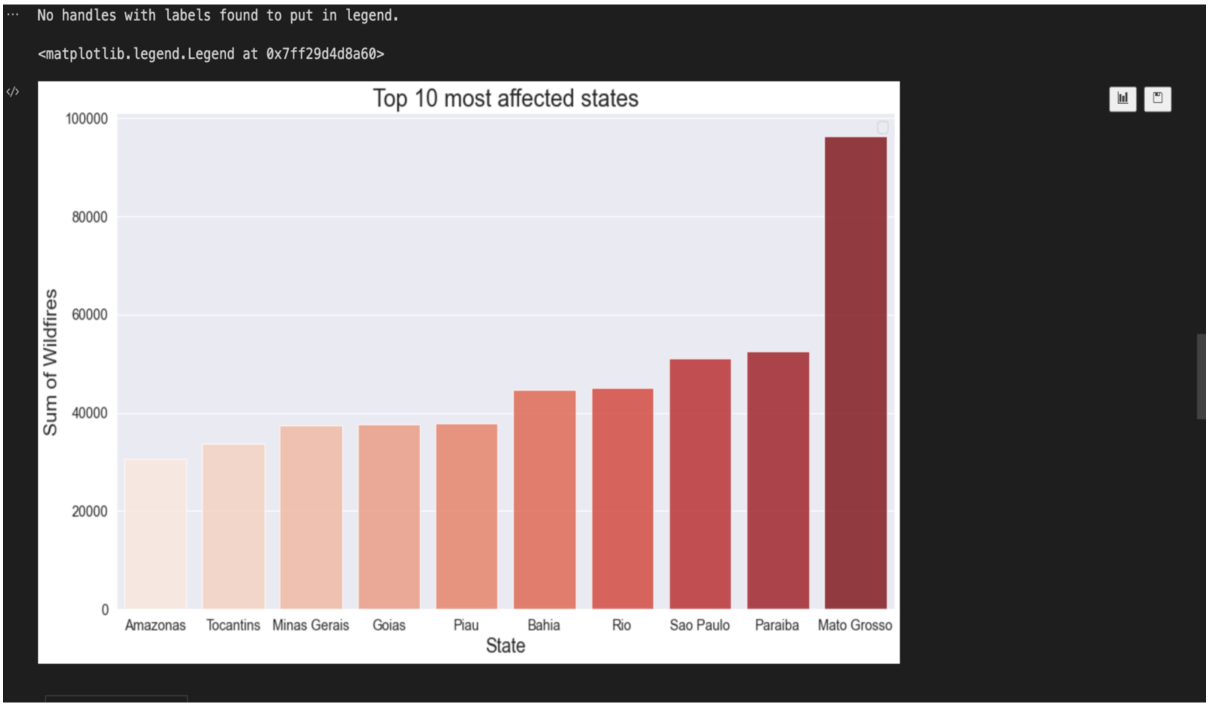
```python
year_mo_state = df.groupby(by = ['year','state', 'month']).sum().reset_index()

year_mo_state.head()
```

	year	state	month	number
0	1998	Acre	April	0.0
1	1998	Acre	August	130.0
2	1998	Acre	December	7.0
3	1998	Acre	Feburary	0.0
4	1998	Acre	January	0.0

```python
by_year = df.groupby(['year'], as_index=False).sum()

plt.figure(figsize=[12,7])
plt.xlim([1998, 2017])
plt.title('Sum of number of fires from 1998-2017')
sns.lineplot(x='year', y='number', data=by_year)
```



```
by_year = by_year[by_year['year']>2004]
z = np.poly1d(np.polyfit(by_year['year'], by_year['number'],2))

years = np.linspace(2004, 2017, 13)

plt.figure(figsize=[12,7])
plt.plot(years, by_year['number'], '-', label = 'Raw data')
plt.plot(years, z(years), '--', label = 'Fitted Curve')
plt.xlim([2004, 2017])
plt.ylim([17000, 45000])
plt.title('Fitted Curve')
plt.legend()
plt.show()
```



```
No handles with labels found to put in legend.

<matplotlib.legend.Legend at 0x7ff29d4d8a60>
```

Chapter 05: Results

5.1 Discussion on the Results Achieved

After reviewing these data, I'd want to point out that I had never undertaken an analysis on this topic before, and that any statistics, figures, and information I knew only from the news, social media, and networks. As a result, the findings I received were unexpected. I wasn't expecting to see so many fires. Work's main outcomes:

- State with the most fires - Mato Grosso;
- The largest number of fires was in 2003, 2015 and 2016, during these years there were different top 3 states by the number of fires;
- The largest number of fires occurred in August 2006;
- Almost 2/3 of the fires occurred in the second half of the year, the top 5 states, depending on the half of the year, are different;
- The greatest number of fires was in summer and autumn;
- The highest number of fires was in July.
- Predicted value of number of forest fire in the year 2020 is 41724 and the real data value is 44,013

5.2 Future Work

Machine Learning or ML is becoming a new way of predicting the forest fires and damage caused by forest fires, however there may be nonetheless plenty of room for development and capability for additional models to be built. Existing predictive models nonetheless face limitations due to the shortage of records pre-processing steps and immoderate differences withinside the choice of sample characteristics and troubles associated with validation and promotion. Model overall performance nonetheless desires to be optimized and other obstacles addressed.

Researchers and wildlife professionals need to connect to reality, carefully select a model, use the model in scientific exercise after review, and apply rigorous layout and validation methodologies with a massive pattern of high quality studies data primarily based totally on previous results. The applicability and predicament of those models must be carefully evaluated in order to enhance the level of precision for prediction.

It is suggested that complete methodological records consisting of lacking cost processing, outlier processing, magnificence imbalance processing, hyperparameters optimization, feature selection, variable significance rating processing, version scoring and validation, and the machine learning model's overall performance.

5.3 End Note

Finally, it's always a good idea to be more environmentally conscious. It doesn't take much, just a little more thought when you're shopping (look at the label and ask yourself: is this recyclable?). Is it environmentally friendly? Is it causing environmental damage?). You can also start turning off the water more frequently or only keep the lights on in the room where you sleep. Minimize your possessions.

References

1. Sakr, George & Elhajj, Imad & Mitri, George & Wejinya, Uche. (2010). Artificial intelligence for forest fire prediction. IEEE/ASME International Conference on Advanced Intelligent Mechatronics, AIM. 1311-1316. 10.1109/AIM.2010.5695809.
2. A. Alonso-Betanzos, O. Fontenla-Romero, B. Guijarro-Berdi˜nas, E. Hern´andez-Pereira, M. Inmaculada Paz Andrade, E. Jim´enez, J. Luis Legido Soto, and T. Carballas, “An intelligent system for forest fire risk prediction and fire fighting management in Galicia,” Expert systems with applications, vol. 25, no. 4, pp. 545–554, 2003.
3. N. Aronszajn, Introduction to the theory of Hilbert spaces. Stillwater, Oklahoma: Reasearch [sic] Foundation, 1950.
4. D. R. Brillinger, H. K. Preisler, and J. W. Benoit, “Risk assessment: A forest fire example,” in Statistics and Science: a Festschrift for Terry Speed, ser. IMS Lecture Notes Monograph Series, D. Goldstein and T. Speed, Eds. Beachwood, OH: Institute of Mathematical Statistics, 2003, vol. 40, pp. 177–196.
5. T. Cheng and J. Wang, “Applications of spatio-temporal data mining and knowledge for forest fire,” in ISPRS Technical Commission VII Mid Term Symposium, Enschede, 2006, pp. 148–153.
6. ——, “Integrated Spatio-temporal Data Mining for Forest Fire Prediction,” Transactions in GIS, vol. 12, no. 5, pp. 591–611, 2008.
7. K. Clarke, J. Brass, and P. Riggan, “A cellular automaton model of wildfire propagation and extinction,” Photogrammetric Engineering and Remote Sensing, vol. 60, no. 11, pp. 1355–1367, 1994.
8. P. Cortez and A. Morais, “A data mining approach to predict forest fires using meteorological data,” in Proceedings of the 13th Portugese Conference on Artificial Intelligence, 2007, pp. 512–523.
9. A. Dunn and G. Milne, “Modelling wildfire dynamics via interacting automata,” Lecture notes in computer science, vol. 3305, pp. 395–404, 2004.
10. J. Han, K. Ryu, K. Chi, and Y. Yeon, “Statistics Based Predictive Geo-spatial Data Mining: Forest Fire Hazardous Area Mapping Application,” Lecture notes in computer science, pp. 370–381, 2003.
11. L. Iliadis, “A decision support system applying an integrated fuzzy model for long-term forest fire risk estimation,” Environmental Modeling and Software, vol. 20, p. 613, 2005.