INFO-B211 Informatics Infrastructure II

Final Project Report

Forest Fire Data Analysis

Thomas Reed and Jace Vayhinger

December 15, 2022

Introduction

Forest fires are a prevalent problem in the Amazon rainforest. That said, these forest fires are often not given the attention they deserve from the public. Data is not often shared in an easy and accessible way. Using libraries and concepts we learned throughout this semester, we hope to present the data in this dataset in a legible and understandable manner for the public to comprehend.

Methodology

The five libraries and their functions that were utilized in this project are listed below:

- Pandas – Data Manipulation
- NumPy – Mathematical Operations/Linear Algebra
- SciPy – Mathematical Operations
- Matplotlib – Data Visualization
- Seaborn – Data Visualization

The dataset we used for this project is one provided by Kaggle, which is a reliable database site that contains many datasets for public use. Using the five libraries listed above we are taking a data visualization-focused approach to enable the public to get a more comprehensive understanding of what the data for these forest fires really look like and why it is an important topic. For prediction, we are using a random forest model that gave us 89.3% accuracy.

Results

As shown in our visualization of the data, these forest fires are a huge issue. Based on these figures we have created; we can see that Brazil's forest fires have been increasing over time. Even if it is slowly, bringing attention to this fact may make it easier to help prevent and even lower rates in the long run. We can also see that as we get into the warmer months of the year, or summer, the number of fires increases significantly and peaks in July. We were able to calculate the total number of fires reported during the 19 years of data collection comes out to be 698,924. Although wildfires are a natural process and can be good for the environment at certain times, it is important to maintain a healthy number of natural fires and reduce the harmful human-caused fires. We used a random forest model to predict the states in which wildfires occurred and we achieved 89.3% accuracy.

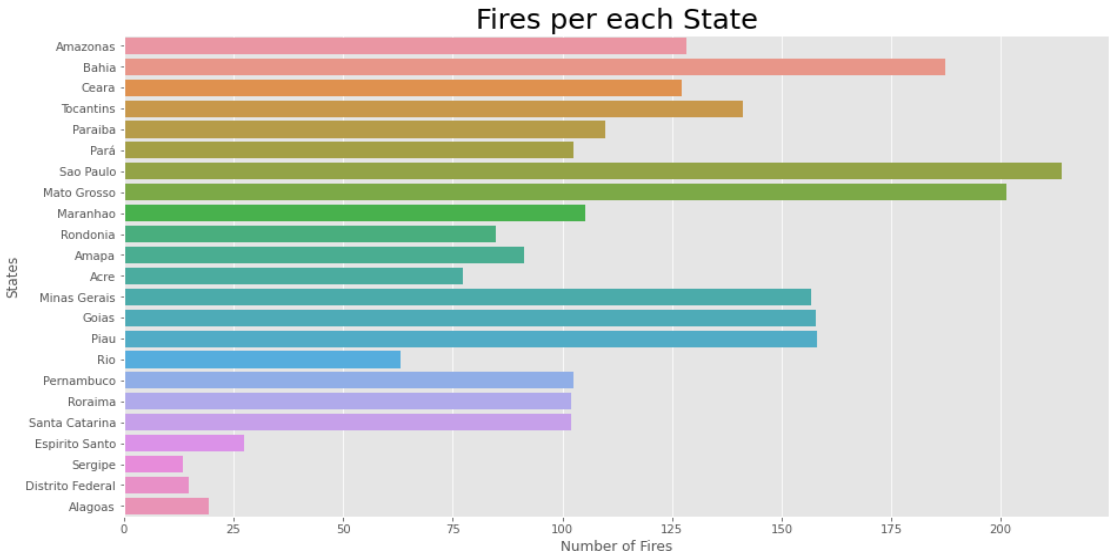

Figure 1: This figure is a bar plot that shows the total number of fires that each state had.

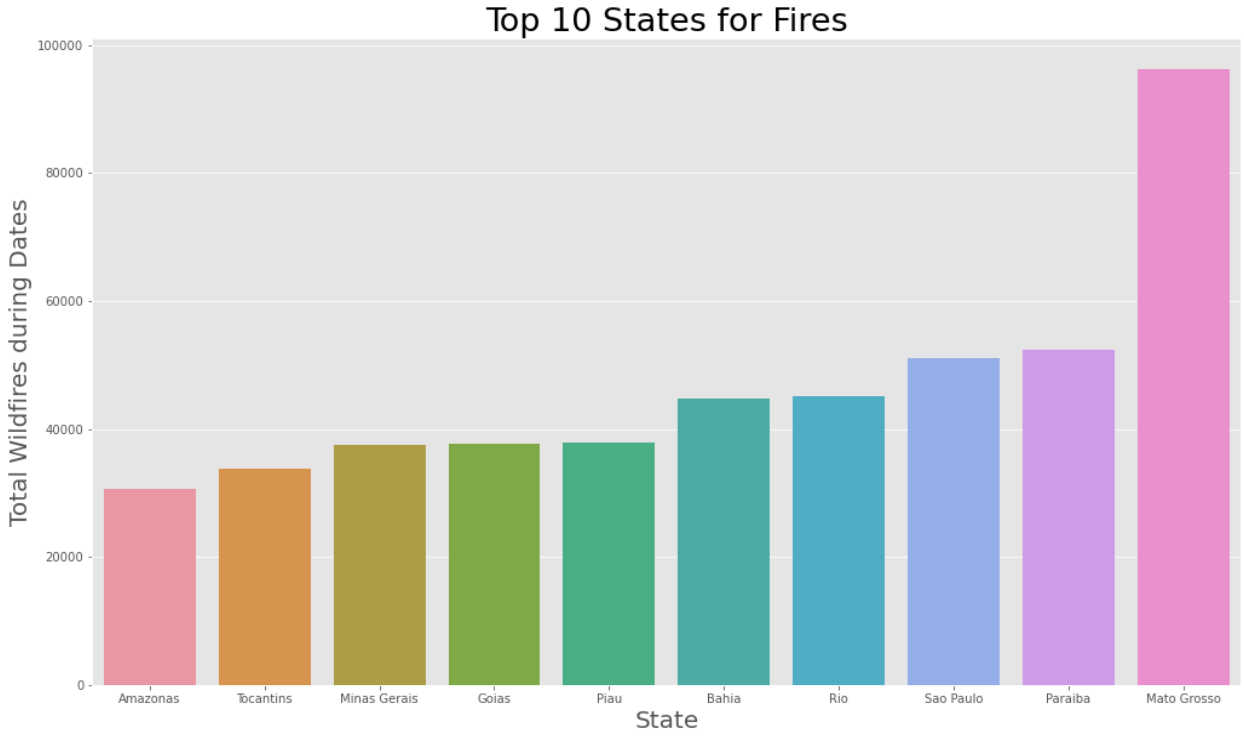

Figure 2: This figure shows the states with the top ten highest total number of fires between the dates of 1998 and 2017.

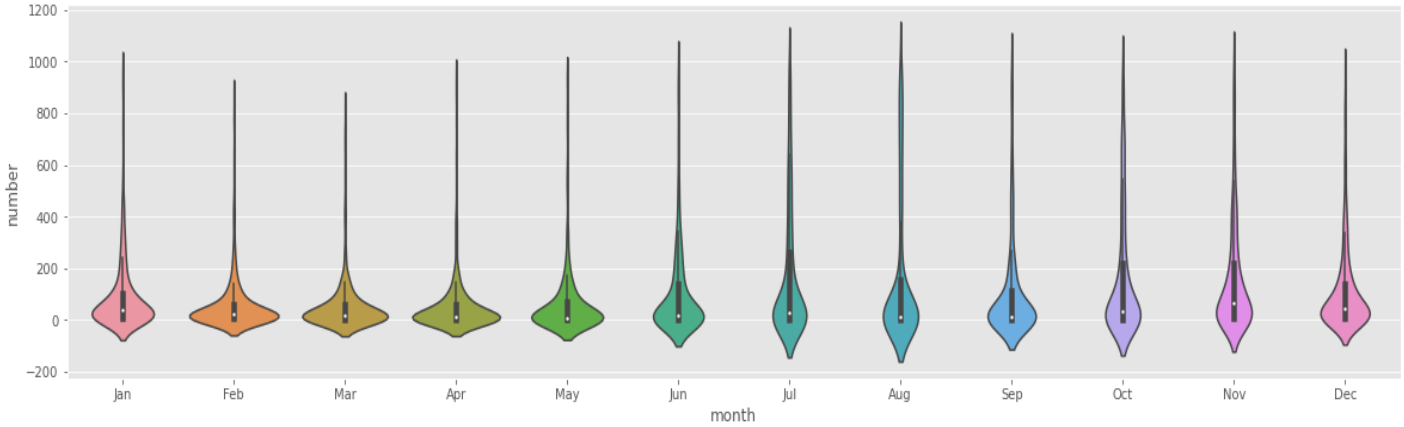

Figure 3: This figure is a violin plot that shows the number of fires based on the months of the year.

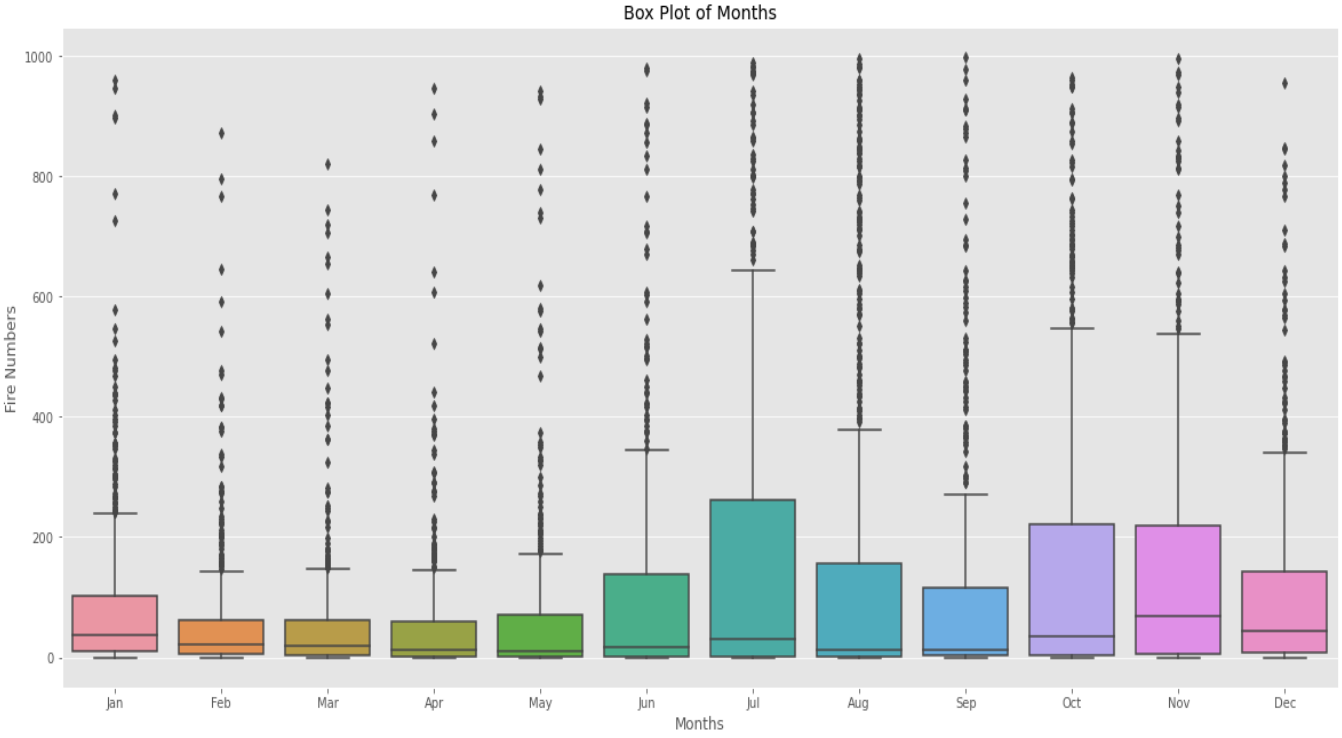

Figure 4: This figure is a boxplot that also shows the fires per month but allows us to interpret.

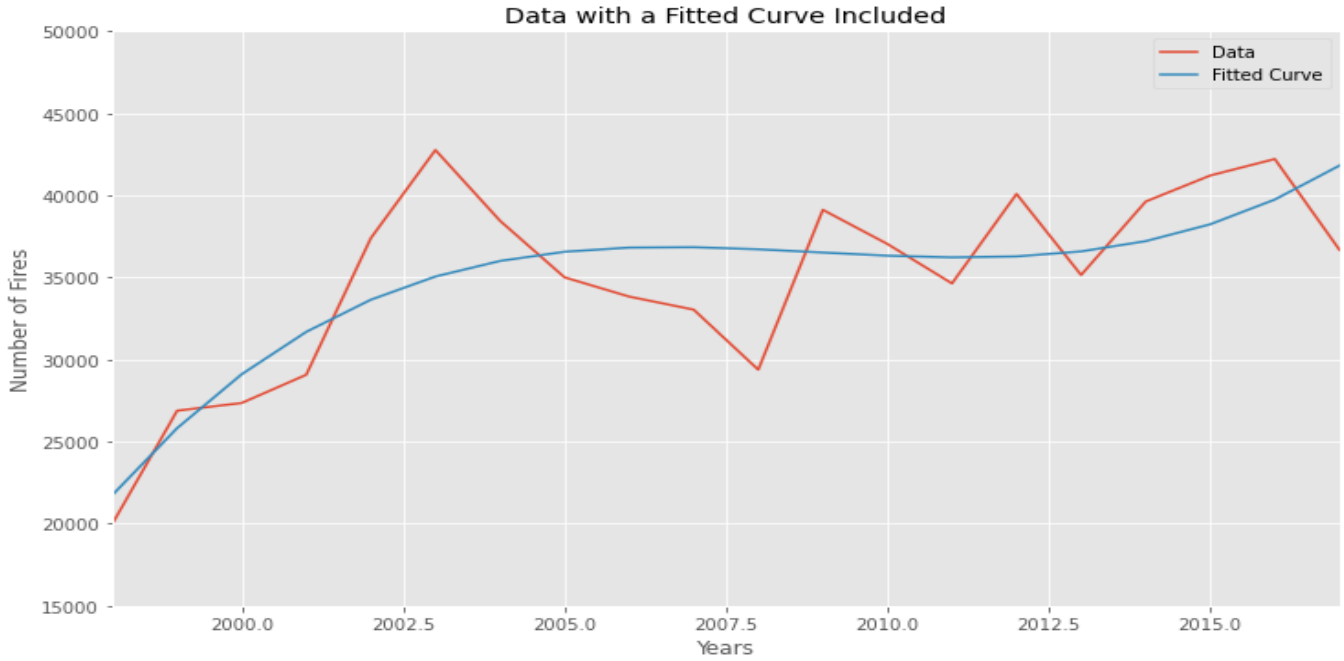

Figure 5: This figure shows the total amount of data over all the years provided in our dataset with a fitted curve.

Conclusion

Through this project, we have been able to accomplish our goal of adequately providing easily understandable and simple visualizations of the forest fire dataset.

Code Snippets

Below have included all code used for the figures shown above, as well as other important pieces used to analyze our data.

Figure 1:

```
#exploratory data analysis showing the states and the amount of fires#
#setting the figure size#
fig, ax = plt.subplots(figsize=(15,8))
#using seaborn#
sns.barplot(data= df.sort_values(by='number', ascending=False), x = 'number', y= 'state', ci=None, ax=ax)
#graph customization#
plt.title('Fires per each State', fontdict={'fontsize':24})
plt.ylabel('States')
plt.xlabel('Number of Fires')
plt.show()
```

Figure 2:

```
#now that we have our new dataframe we can use matplotlib to get some graphs out of it#
#using matplotlib to create a barplot of the top10 dataframe#
#represents the total number of fires in top 10 states from 1998-2017#
plt.figure(figsize = (17, 10))
#using seaborn and setting what the x and y axis are#
graph_10 = sns.barplot(x = top10['state'], y = top10['number'])
#grap customization#
plt.title("Top 10 States for Fires", fontsize = 27)
plt.xlabel("State", fontsize = 20)
plt.ylabel("Total Wildfires during Dates", fontsize = 20)
```

Figure 3:

```
#violin plot using seaborn#
#setting figure size#
violin=plt.figure(figsize=(20,5))
#using seaborn, setting x and y axis#
sns.violinplot(x="month",y="number",data=df)
plt.show()
```

Figure 4:

```
#lets see how this data looks as a box plot#
#setting figure size#
plt.figure(figsize=(20,9))
#using seaborns boxplot function, setting x and y axis#
sns.boxplot(x='month',y='number',data=df)
#graph customization#
plt.title('Box Plot of Months')
plt.ylabel('Fire Numbers')
plt.xlabel('Months')
plt.show()
```

Figure 5:

```
#Researched that a 3rd degree polynomial fitted on the data would be best to see predicted path#
#creating the polynomial#
df_bestfit = df.groupby(['year'], as_index=False).sum()
#using numpy for linear algebra#
p = np.polyfit(df_bestfit['year'],df_bestfit['number'],3)
z = np.poly1d(p)
```

```
years = np.linspace(1998, 2017, 20)

#setting the figure size#
plt.figure(figsize=[12,7])

plt.plot(years, df_bestfit['number'], label='Data')
plt.plot(years,z(years), label='Fitted Curve')
#setting the limits of the graph#
plt.xlim([1998, 2017])
plt.ylim([15000, 50000])
#graph customization#
plt.title('Data with a Fitted Curve Included')
plt.xlabel("Years")
plt.ylabel("Number of Fires")
plt.legend()
plt.show()
```

References:

Modelli, Luís Gustavo. “Forest Fires in Brazil.” *Kaggle*, 24 Aug. 2019, https://www.kaggle.com/datasets/gustavomodelli/forest-fires-in-brazil.

Raut, Shivani. “Information Infrastructure II.” 2022, Indianapolis, IUPUI.