Brazil Forest Fires by Sneh Gupta and Gracie Gibbons

Executive summary

1) What are the states in Brazil that have had the highest forest fire frequency in the last two decades?
   Answer: Mato Grosso Do Sul has had the highest number of forest fires between 1998 and 2017.
2) What is the change in forest fire frequency, by location in Brazil?
   Answer: September and October have higher forest fires in most states, annually. There seems to be an increasing trend in forest fires over the last two decades.
3) What is the trend of forest fires in Brazil over the last two decades?
   Answer: The abundance of forest fires in Brazil has on average increased over the last two decades.
4) Can we build a model to predict the future forest fire frequency in Brazil according to the current trend?
   Answer: Yes, we can.

Motivation

Forest fires have become a huge problem lately, especially in Brazil where the Amazon rainforest, which is a large carbon sink, has been burning in acres. The loss of forest by increase of fires due to the effects of global warming not only affects Brazil, but also the rest of the world. It is important to learn more about the trends and frequency of the fires. It is important to note that forest fires are a natural part of forests, but recently their frequency and size has increased due to human driven global warming and deforestation (lack of core moist forest).

Dataset

We are using a dataset of Brazil forest fires - https://github.com/deltalite/Brazil-Forest-Fires/blob/master/data/amazon_with_date.csv
Dictionary - It contains states in Brazil and their associated longitude and latitude (geometry)
https://gist.github.com/ruliana/1ccaaab05ea113b0dff3b22be3b4d637?short_path=b2e63cc#file-br-states-json

Method

1) Q1
   Read in the forest fire csv (using url) and the json geometry file using pandas and geopandas.
   Data cleaning
   - Use external library unidecode to change accented state names into non-accented names so that it matches the geospatial dataframe.

(In the original dataset we were planning to use there was a lot more cleaning, but then we just switched to a cleaned dataset)
- Change the date column format to datetime using pandas.

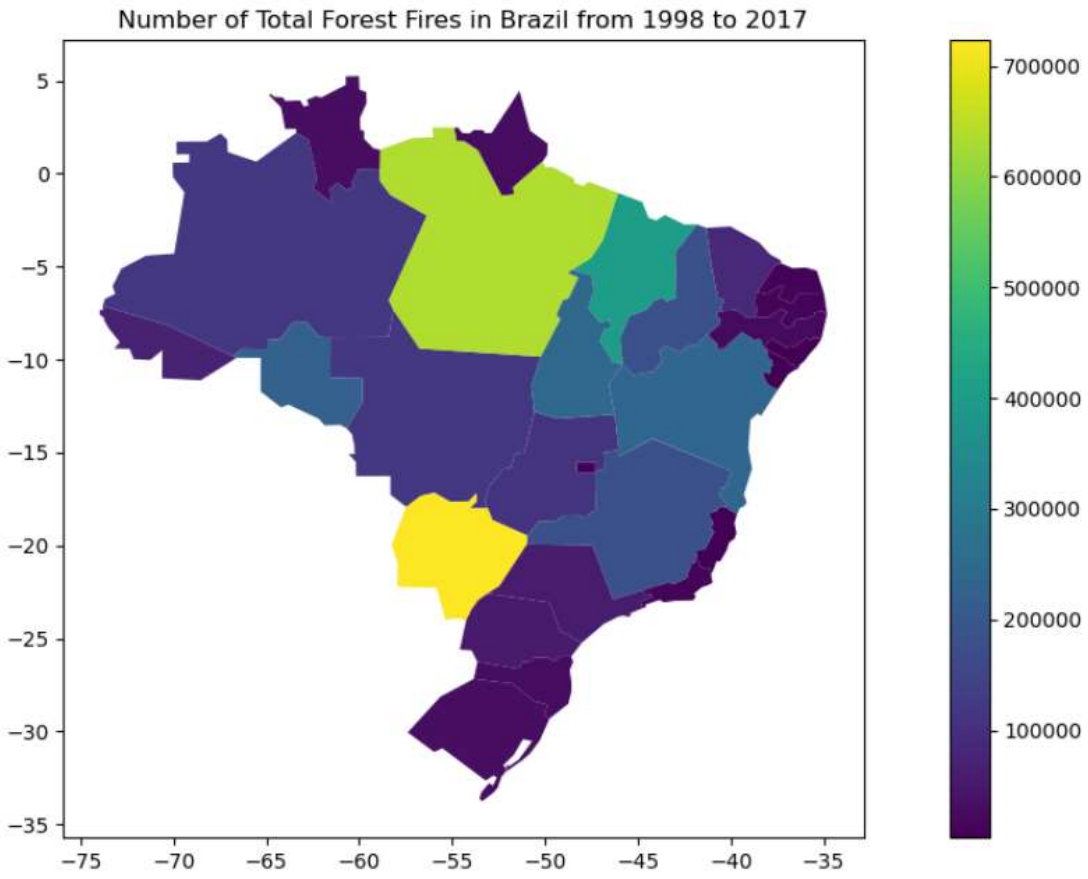
Merging Datasets
- Apply unidecode to names of states in the geometry file.
- Normalize the state names of both files to title format so that they match.
- Merge the forest fire dataset and geometry data using dissolve. Convert the merged dataframe to a geodataframe. Return the merged data and the normalized geometry dataset.

Graph total forest fires in Brazil from 1998 to 2017
- take merged dataset and geometry data as parameters. Sum up the total forest fires by state using dissolve. Plot the geometry data in grey and the total forest fires on the same axes. Include a legend and title

2) Q2
Make a video containing time lapse of the brazil forest fires at each month from 1998 to 2017 - take merged data and geometry data. First plot geometry data in gray and add a colorbar with the scale being 0 to maximum forest fires in a month. Sort the dataset by date. Loop over each month in the dataset - make a new subset in a loop containing the total forest fires by state in that month and plot it on the same axes. Ensure that the color is normalized to the maximum and minimum value in the colorbar. Add text containing the year and month of that time, save the figure as frame.number in a separate file, then remove the text. Combine all the frames to a video using ffmpeg - specify framerate.
3) Q3
To answer question three, use the Pandas groupby() function to group the data set by date and apply the sum function to the Number column. Generate a scatter plot that shows the overall trend of fires in Brazil as a whole. We created a scatter plot in Plotly and used a trend line to visualize the trend over time.
4) Q4
Building off of number 3, use the Pycaret library to create a Machine Learning Model of the trend of fires in Brazil over time and predict the number of fires in Brazil in the future. Firstly, train the model on the training dataset and once you are satisfied with the results, train it on the entire dataset. Generate a graph of the fit using Plotly. Next, using the model, predict the trend of fires and create a line plot that visualizes the prediction using Plotly again. To do this, create a separate dataframe of all the dates you would like to predict and predict the model onto that dataframe. Then you must concatenate the old and predicted data frames on Date. This step will meet the challenge requirements of using machine learning as well as an external library.

Results -

Question 1



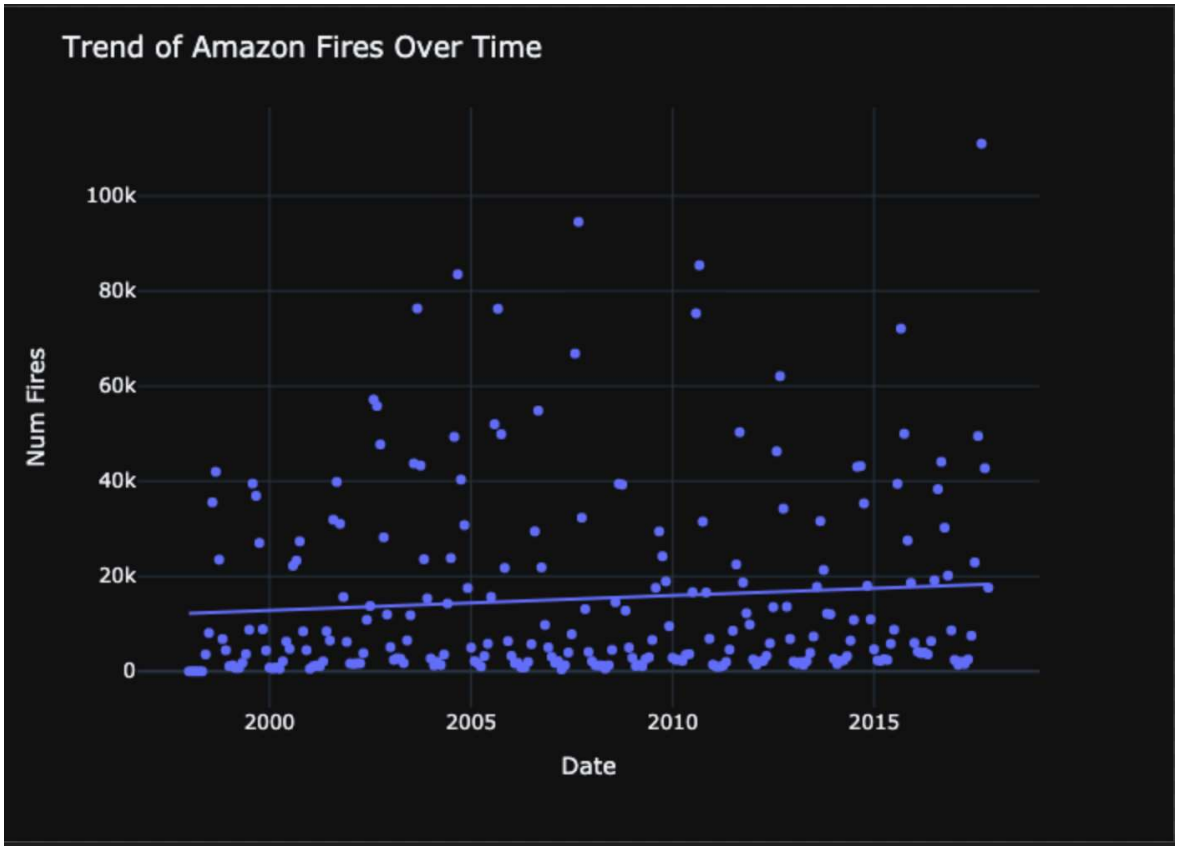
The visualization shows that the State Mato Grosso Do Sul has the highest frequencies of forest fires by a large margin, followed by Para. I was surprised by how many forest fires there were over two decades - about 700,000 in Mato Grosso Do Sul.

This information can be used to enhance forest fire management in states with higher forest fire frequency. But, other factors should be considered like - human settlements across Brazil (human safety), possible under or overreporting in certain areas, and size of forest fires.

Question 2

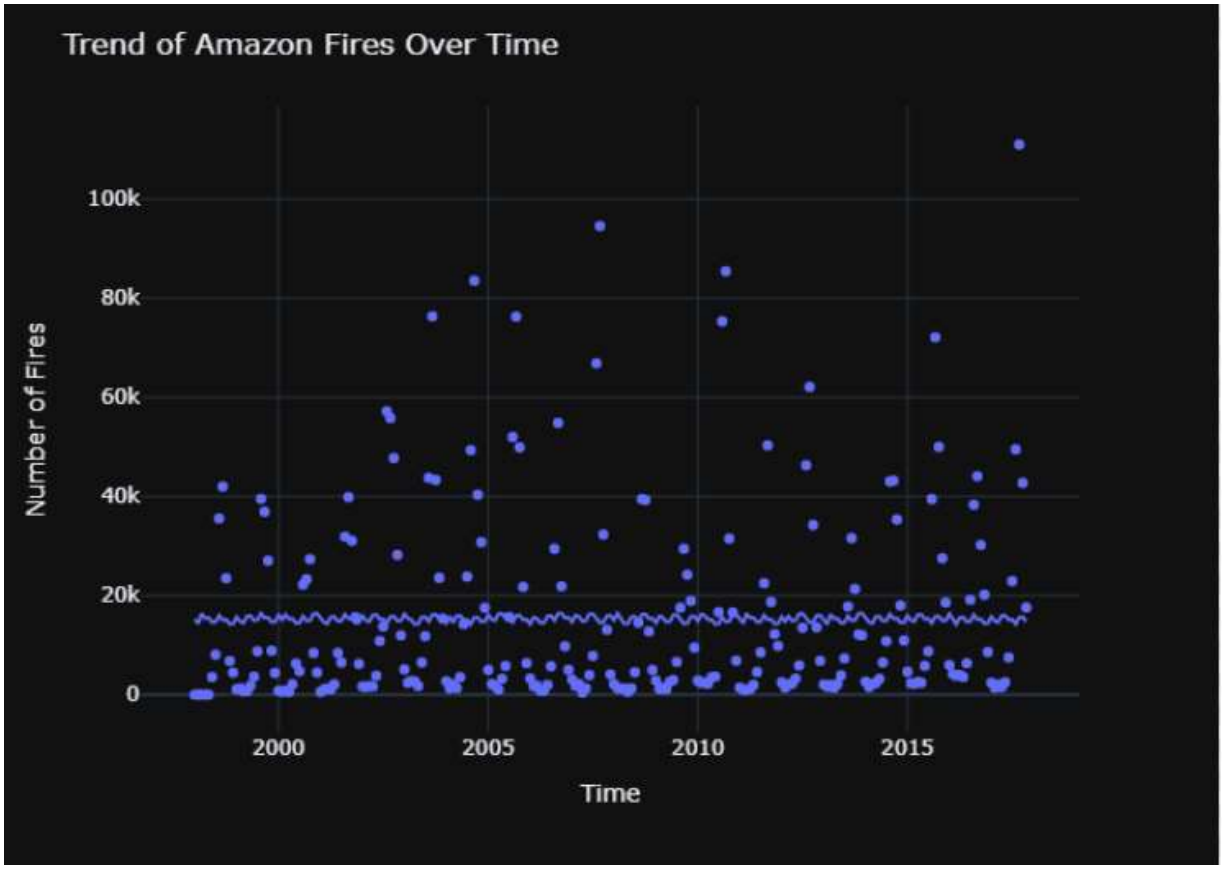
The video shows the total number of forest fires per month, by state, from 1998 to 2017. It is hard to discern long-term trends from just the video, but it is clear that September and October tend to have the highest forest fire frequency. It seems like the forest fire frequency might be increasing over the decades. The following questions give more conclusive results about trends.

Question 3

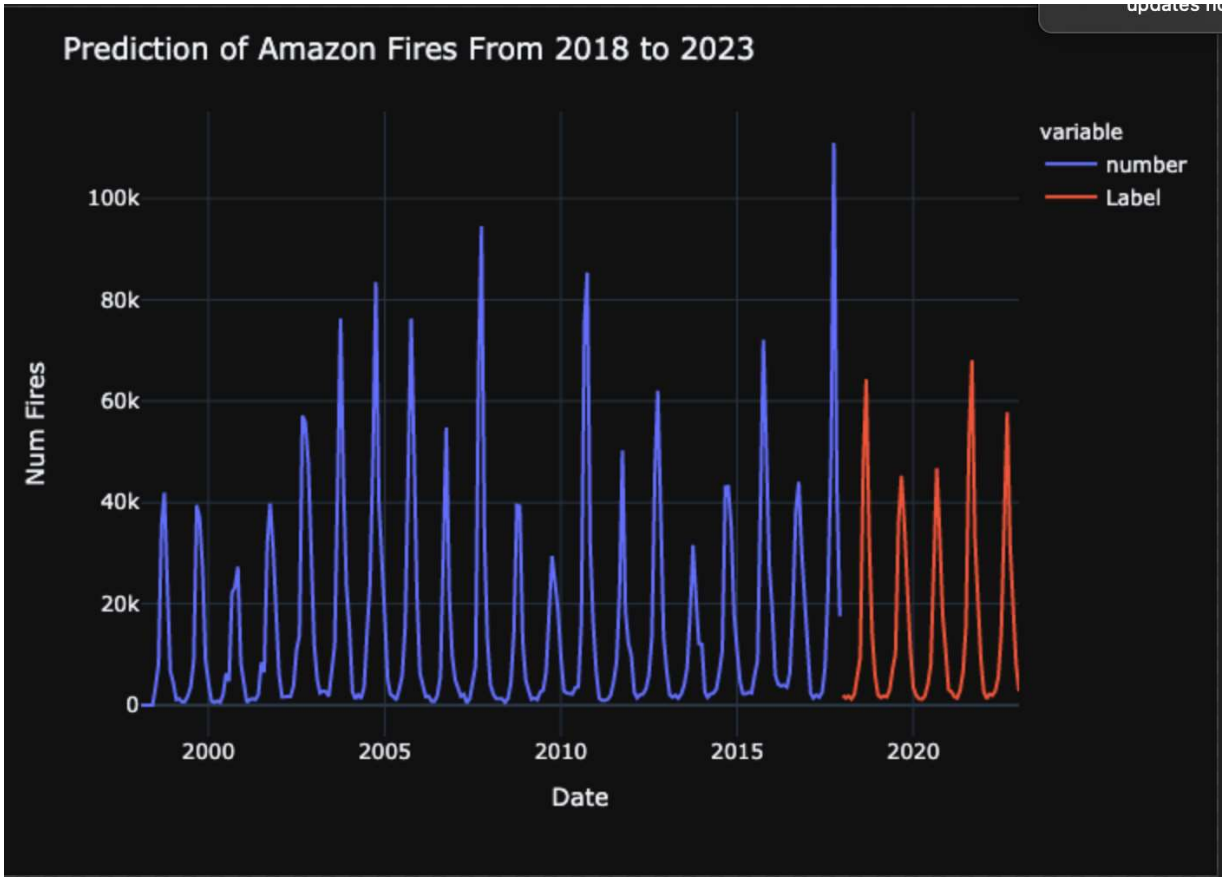


This plot tells us that there is an overall slight upwards trend in forest fire frequency in Brazil each year. Each point on the plot represents a month of fires. Hovering over each point in the Plotly generated graph, which is produced when running the code, will allow the user to see the month, year and number of fires of each data entry. Identifying the cause of this change is already proven: climate change, which is the drastic change in weather patterns due to the emission of greenhouse gases. Because of this, findings from this graph are not necessarily surprising but they are extremely important in that now that we know the trend of fires is increasing. Our data accounts for yearly trends across a span of years, so there are 12 data points per year (one for each month). This means that the averages vary greatly throughout the year, causing the data to have outliers that skew the data. Regardless, the visualization shows that the average total of fires per year increased by a total of 5,000 fires per year, which is a drastic increase over the short span of two decades.

Note: We experienced a variance in the output of this plot, but the trendline consistently stayed under 20,000. We think this error may be due to a difference in Mac and Windows.


Trend of Amazon Fires Over Time
100k
80k
60k
40k
20k
0
Number of Fires
2000
2005
2010
2015
Time

Question 4



Historical data shows the trend of fires throughout an individual year varies greatly with a peak typically in the months October and November, confirming the theory from question 2. The prediction of the trend of average forest fires in Brazil in 2018-2023 follows the same pattern. Before 2017, there was a general upwards trend in annual peaks but as seen in red, the model predicts that trend will not continue, with a few years averaging shorter peaks of around 43,000-47,000 fires. There is a general upwards trend of the forest fire frequency, and the prediction seems to follow this trend between 2020 and 2023.

These findings indicate that the peaks of fires in Brazil are decreasing, but the overall yearly trend is remaining the same. These findings are surprising because as seen from question 2, the general trend of the frequency of fires is increasing. It is possible this prediction is inaccurate due to a lack of data from years prior to 1998 in this dataset.

These findings are important in informing lawmakers and researchers about the environment and to prepare for the next forest fire season. It is important to also investigate further into the size of the forest fires.

Testing

Q1

- the geospatial visualizations match the Brazil map on google maps
- All the results seem to match - I found many bugs related to the dataset using this method, which led to the decision to change datasets to a cleaner dataset.
- Historical - a google search shows that Mato Grosso Do Sul tends to have high number of forest fires/ year - which matches the conclusion for question 1

Q2

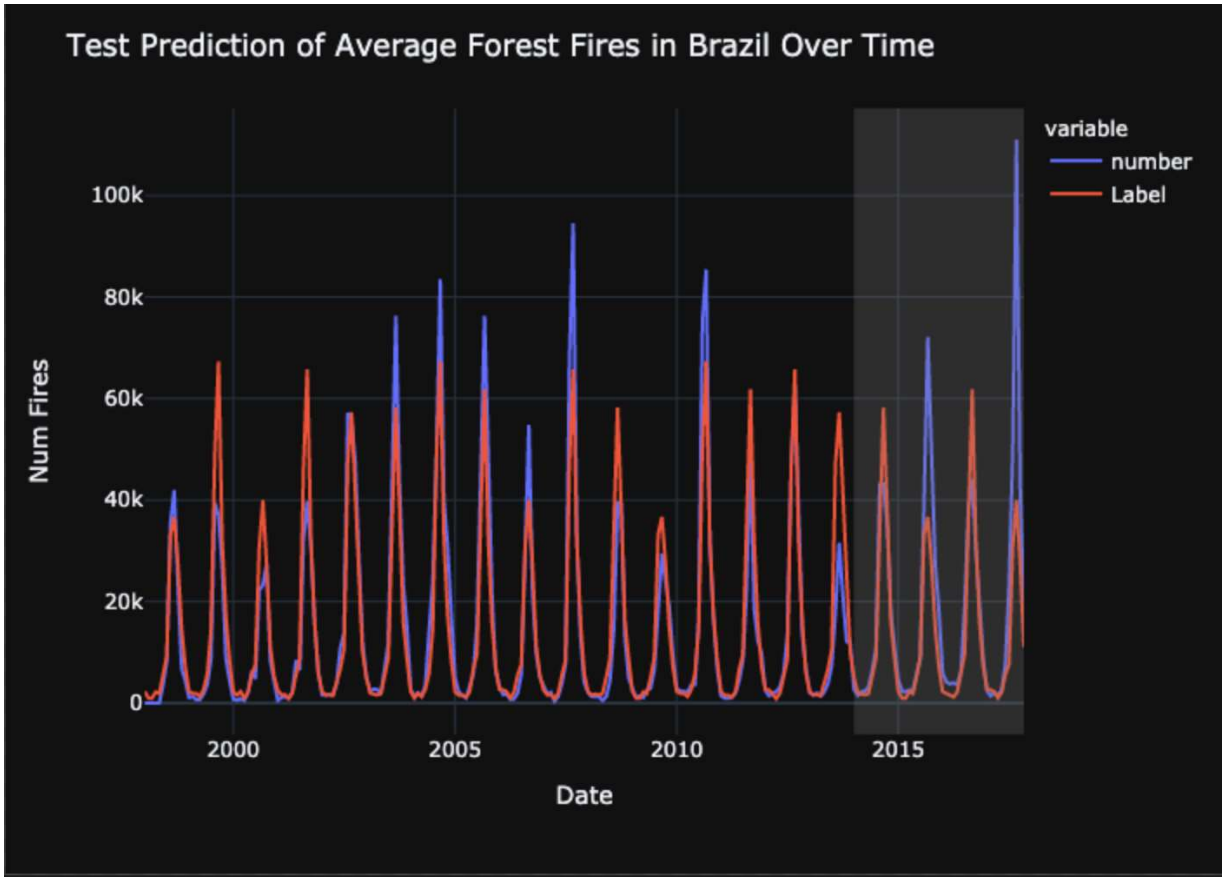
- the geospatial visualizations match the Brazil map on google maps
- All the results seem to match - including the maximum number of forest fires.
- I manually checked the values for one month, a couple states and checked visually for question 2 to verify that my numbers were correct.
  February 2017 - Acre - 1 fire, Mato Grosso - 165, Mato Grosso Do Sul - 337
  December 2000 - Mato Grosso - 139, Acre - 0, Mato Grosso Do Sul - 98
  I found a bug using this method related to the plot not corresponding to the legend and fixed that.

Q3

Qualitative assessments -

- Using a random number generator, I chose random data point entries to manually check to see that the graph was accurately representing the data. I found that all the data points I checked were correct. The code for this is in the ml.py file.
- The R Squared value for the trendline is 0.007908, which indicates that 0.7% of forest fire frequency can be explained by date. This is considered a very week correlation, but this may be because there are so many outliers in the data due to the annual trend of fires. The value of increase indicates that the trend is likely still significant.

Q4



Qualitative assessments -

- To assure that my prediction model was accurate, I split the dataset into a testing and training set. The data was split into a training and testing set and I tested the model on the section of data colored grey. * As depicted in the graph and confirmed visually, the model is adequately accurate. Due to these findings, there was enough confidence in the model to move onto the next stage of the process: predicting.
  *NOTE: I was unable to find the MSE in Pycaret due to a version error.

Collaboration

Other resources - a lot of googling, but mainly referred to the following -

- https://medium.com/tech-carnot/time-lapse-choropleth-map-visualization-using-geopandas-8adb77a7d14
  Time lapse video
- http://blog.gregzaal.com/how-to-install-ffmpeg-on-windows/#:~:text=If%20you%20try%20that%20right,and%20it'll%20understand%20us.
  https://www.gyan.dev/ffmpeg/builds/
  Downloading ffmpeg for making video out of frames

- https://towardsdatascience.com/time-series-forecasting-with-pycaret-regression-module-237b703a0c63
  For the ml model and regression line
- https://towardsdatascience.com/time-series-forecasting-with-pycaret-regression-module-237b703a0c63
  ML library information