UFO SIGHTINGS

Ali Krakowsky

INTRODUCTION

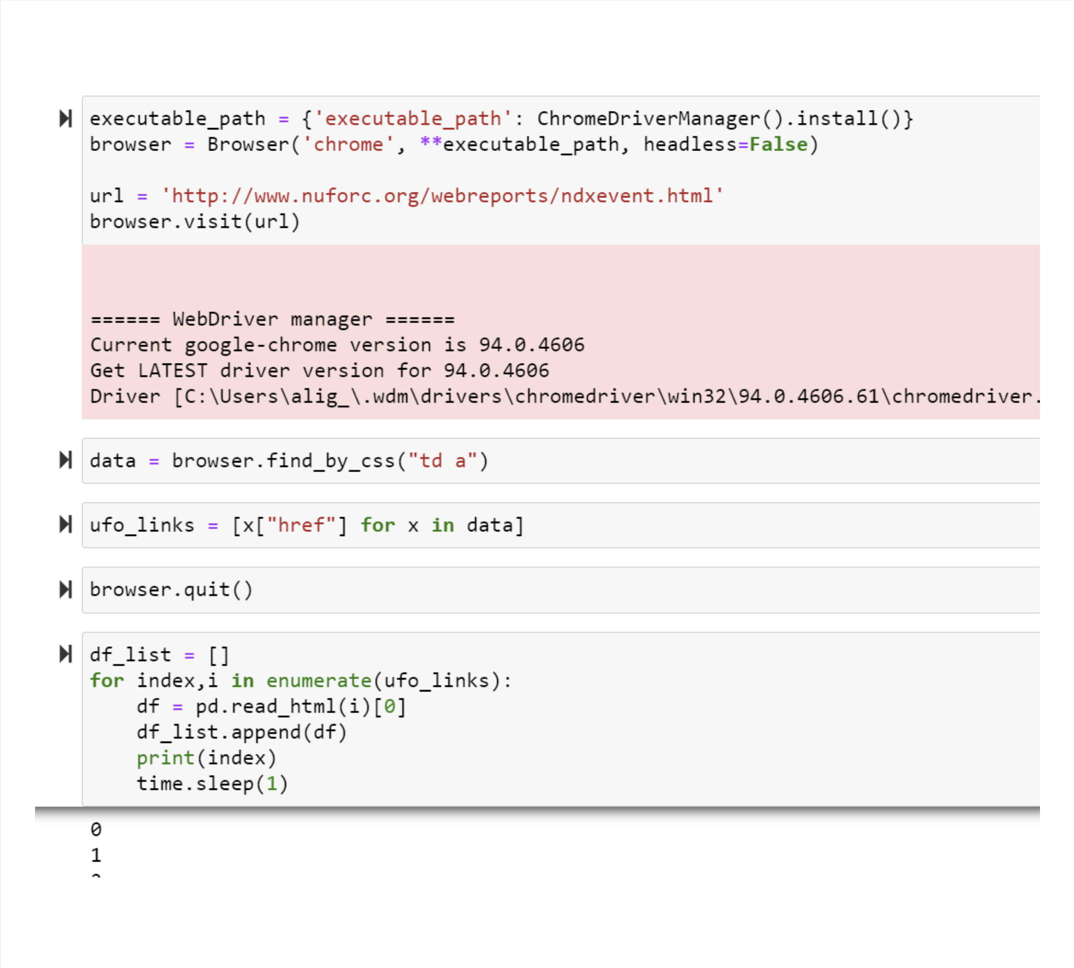
For the final project I'm looking at UFO Sightings again. Here's a quick refresher on the data.

SUMMARY

NUFORC is the National UFO Reporting Center where the reports of UFO sightings are stored. The goal of this project is to use machine learning to see if it's possible to predict the shape of the UFO by location.

PROCESS

- Used Jupyter Notebook to pull data
- Executable path created to search for table
- Looped through each link to create the data frame
- Result = Data pulled from almost 1,000 links



```
executable_path = {'executable_path': ChromeDriverManager().install()}
browser = Browser('chrome', **executable_path, headless=False)

url = 'http://www.nuforc.org/webreports/ndxevent.html'
browser.visit(url)
```

```


====== WebDriver manager ======
Current google-chrome version is 94.0.4606
Get LATEST driver version for 94.0.4606
Driver [C:\Users\alig_\.wdm\drivers\chromedriver\win32\94.0.4606.61\chromedriver.
```

```
data = browser.find_by_css("td a")
```

```
ufo_links = [x["href"] for x in data]
```

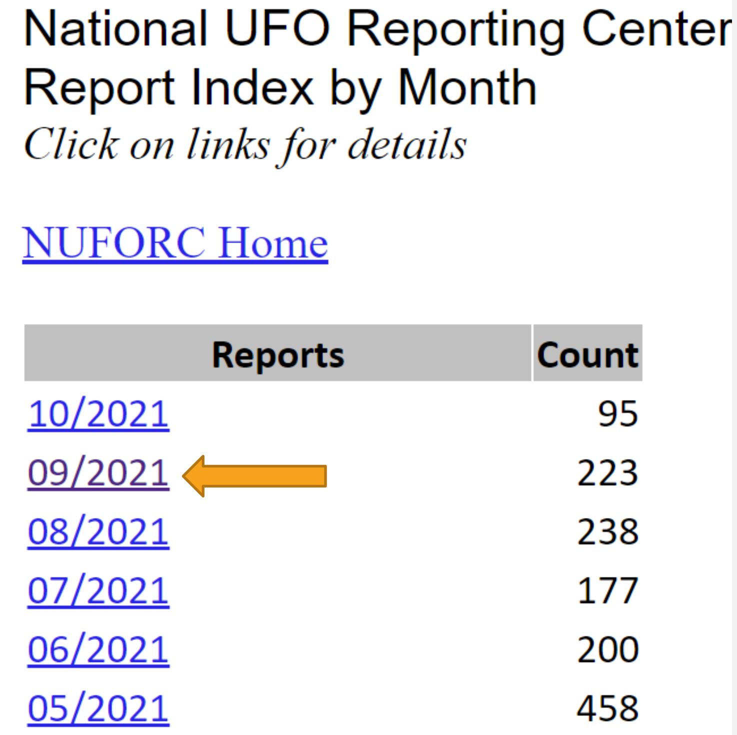
```
browser.quit()
```

```
df_list = []
for index,i in enumerate(ufo_links):
    df = pd.read_html(i)[0]
    df_list.append(df)
    print(index)
    time.sleep(1)
```

```
0
1
```

PROCESS

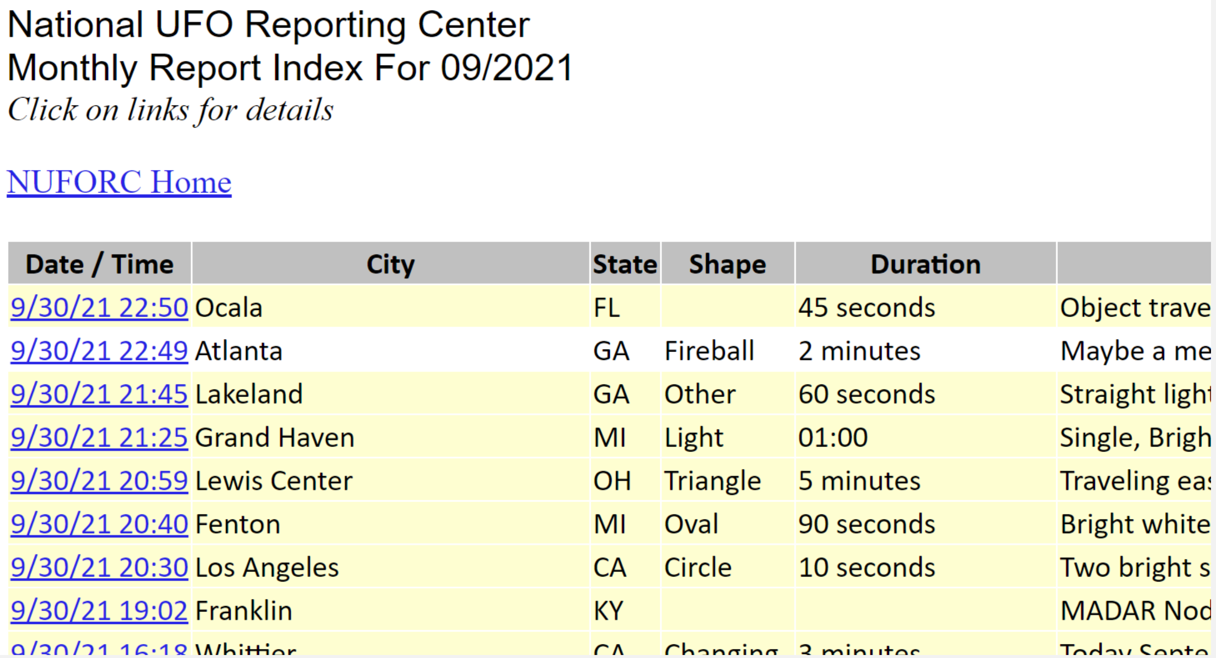
NUFORC Site



National UFO Reporting Center
Report Index by Month
Click on links for details

NUFORC Home

Reports	Count
10/2021	95
09/2021	223
08/2021	238
07/2021	177
06/2021	200
05/2021	458



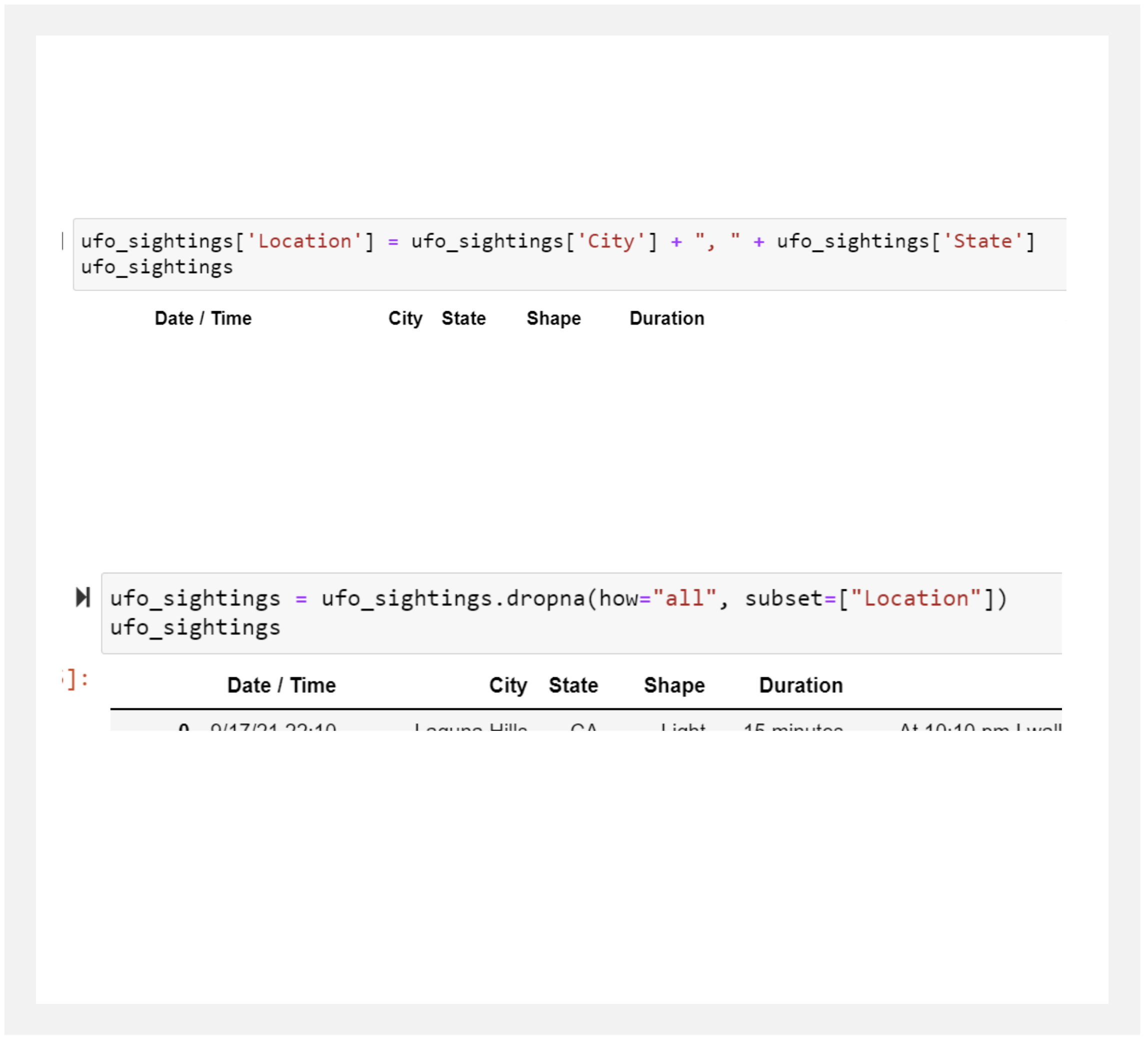
National UFO Reporting Center
Monthly Report Index For 09/2021
Click on links for details

NUFORC Home

Date / Time	City	State	Shape	Duration	
9/30/21 22:50	Ocala	FL		45 seconds	Object trave
9/30/21 22:49	Atlanta	GA	Fireball	2 minutes	Maybe a me
9/30/21 21:45	Lakeland	GA	Other	60 seconds	Straight light
9/30/21 21:25	Grand Haven	MI	Light	01:00	Single, Brigh
9/30/21 20:59	Lewis Center	OH	Triangle	5 minutes	Traveling eas
9/30/21 20:40	Fenton	MI	Oval	90 seconds	Bright white
9/30/21 20:30	Los Angeles	CA	Circle	10 seconds	Two bright s
9/30/21 19:02	Franklin	KY			MADAR Nod
9/30/21 16:18	Whittier	CA	Changing	3 minutes	Today Septe

DATA CLEANUP

- Data frame created
- Prior to merging the csvs
  - The city and state were combined to a new column(Locations)
  - All sightings that were missing the location were dropped
  - Canadian sightings were dropped due to variation in data entry
- After cleaning- over 100,000 rows were left



```
ufo_sightings['Location'] = ufo_sightings['City'] + ", " + ufo_sightings['State']
ufo_sightings
```

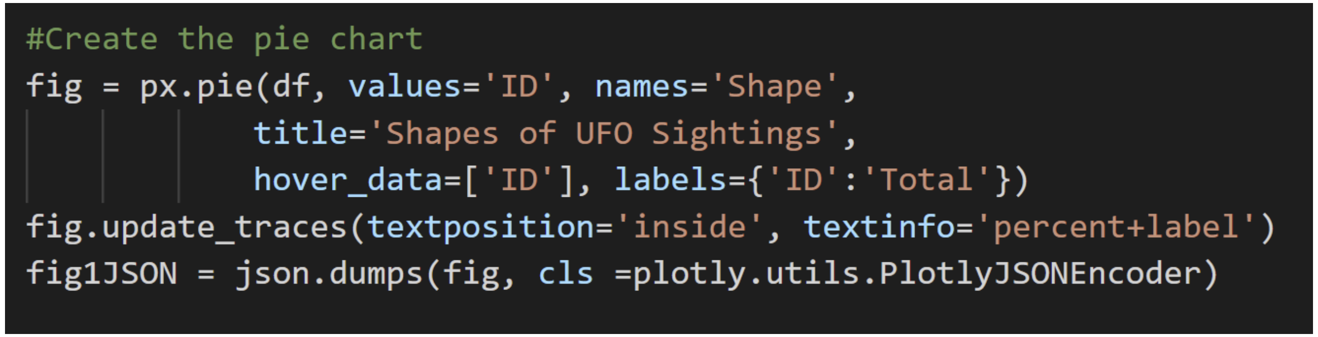
```
        Date / Time          City  State   Shape   Duration
```

```
ufo_sightings = ufo_sightings.dropna(how="all", subset=["Location"])
ufo_sightings
```

```
]:        Date / Time                City   State    Shape    Duration
       0  9/17/21 22:10    Laguna Hills      CA    Light  15 minutes     At 10:10 pm I wall
```

VISUAL #1

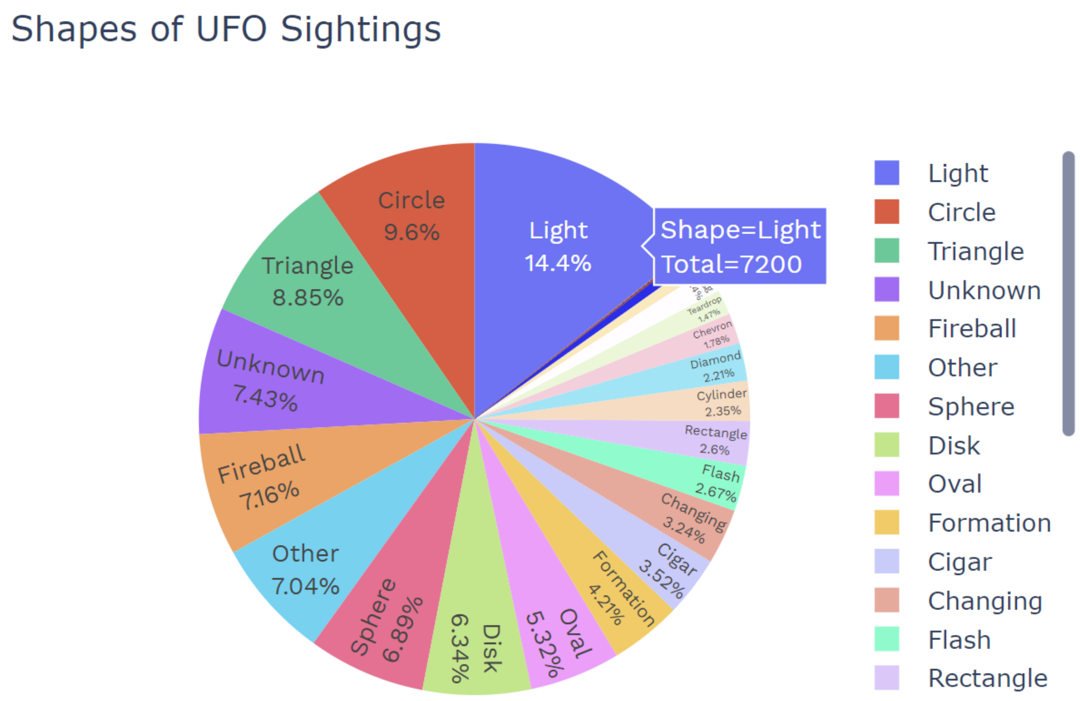
- Grouped the sightings by shapes
- Removed any sightings less than 5
- Created a pie chart with the name and percent inside the wedge
- Json.dumps- creates a trace to pass the data through as html



```
#Create the pie chart
fig = px.pie(df, values='ID', names='Shape',
            title='Shapes of UFO Sightings',
            hover_data=['ID'], labels={'ID':'Total'})
fig.update_traces(textposition='inside', textinfo='percent+label')
fig1JSON = json.dumps(fig, cls =plotly.utils.PlotlyJSONEncoder)
```

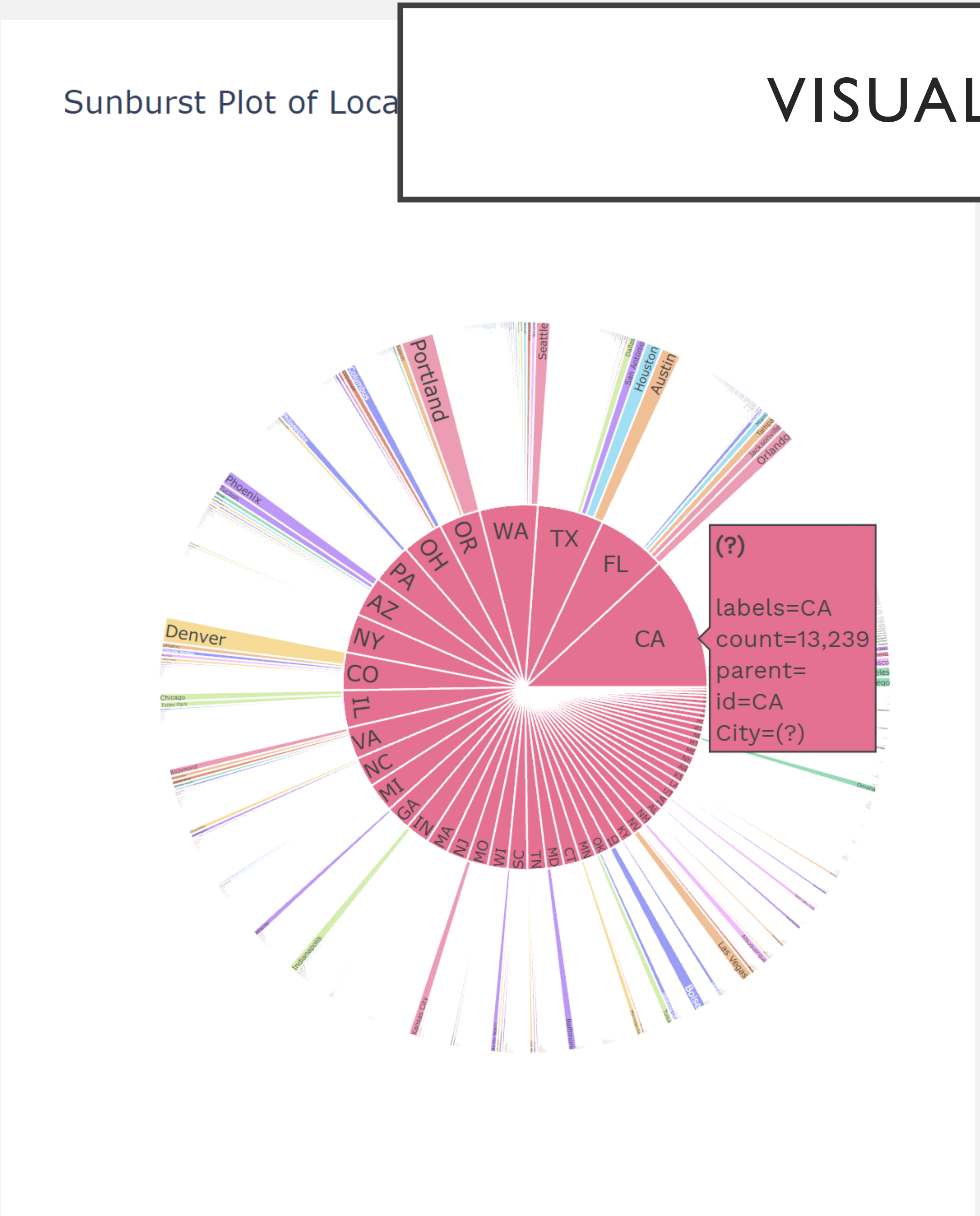
VISUAL #1

- Grouped the sightings by shapes
- Removed any sightings less than 5
- Created a pie chart with the name and percent inside the wedge
- Json.dumps- creates a trace to pass the data through as html



This pie chart is showing the reported shapes of the UFO sightings with the count and percentage. The most popular shape is Light. Sighting reports were created by viewers and submitted as free-text. This causes a variety in the data provided.

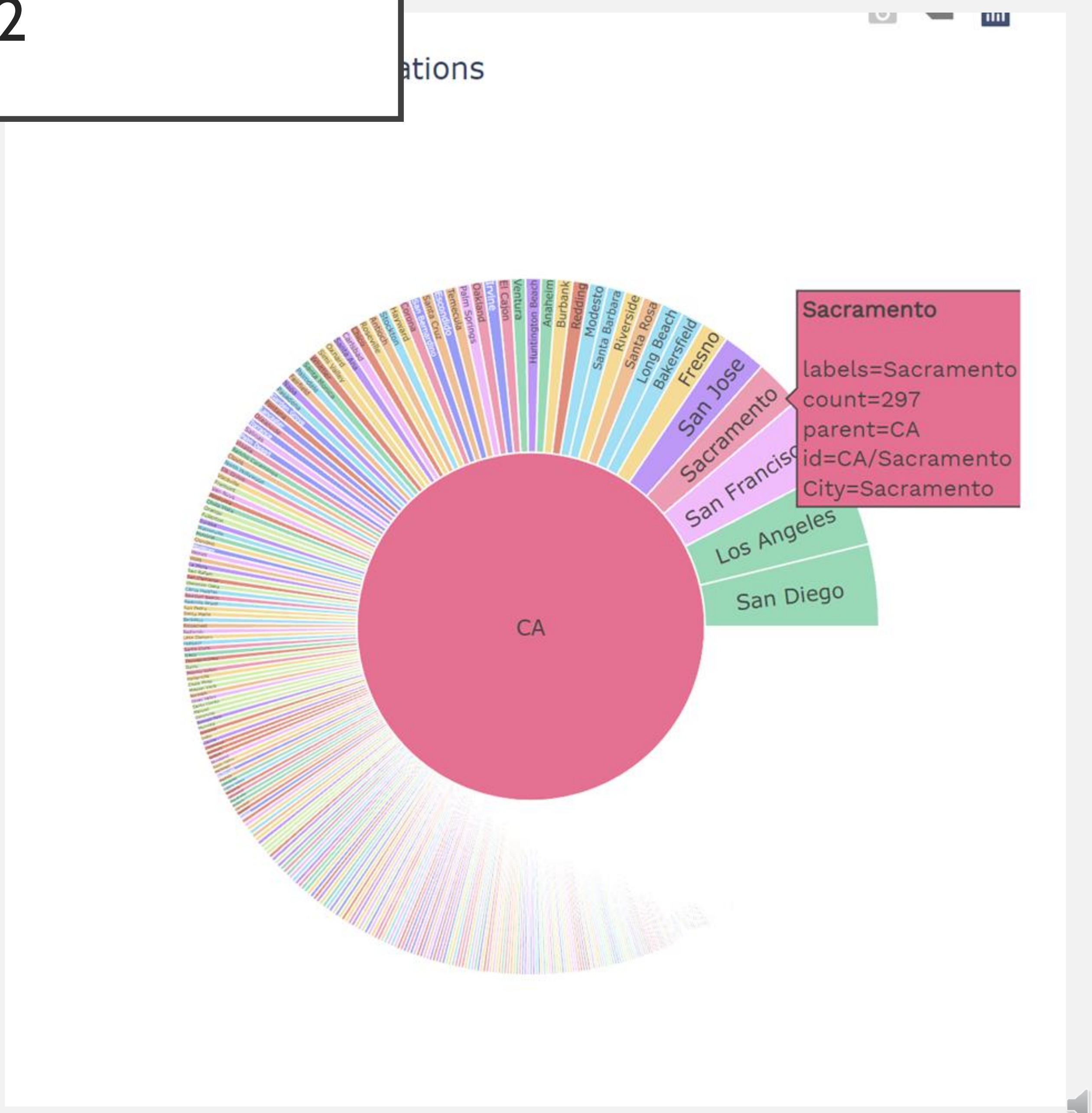
VISUAL #2



Sunburst Plot of Loca

(?)

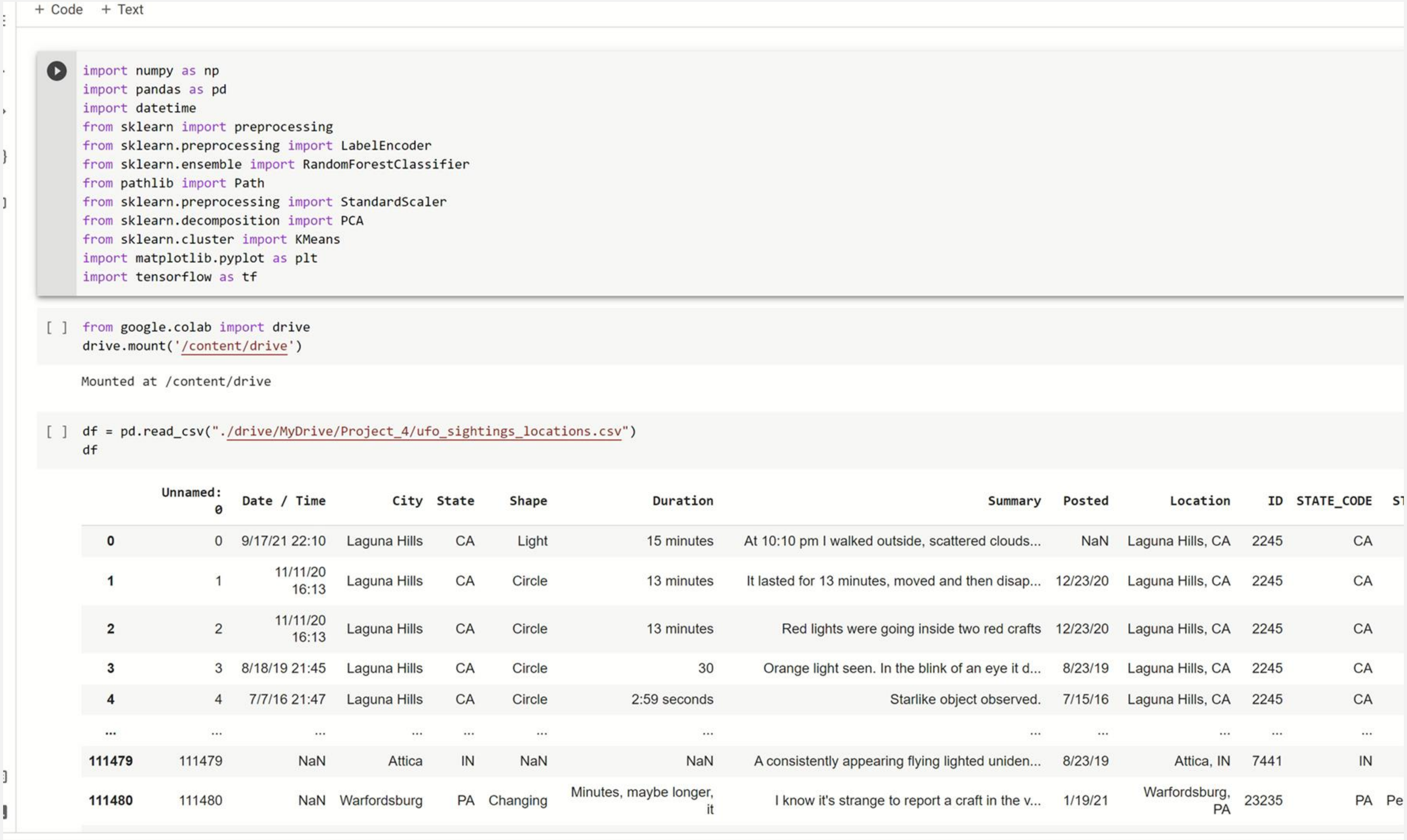
labels=CA
count=13,239
parent=
id=CA
City=(?)



ations

Sacramento

labels=Sacramento
count=297
parent=CA
id=CA/Sacramento
City=Sacramento



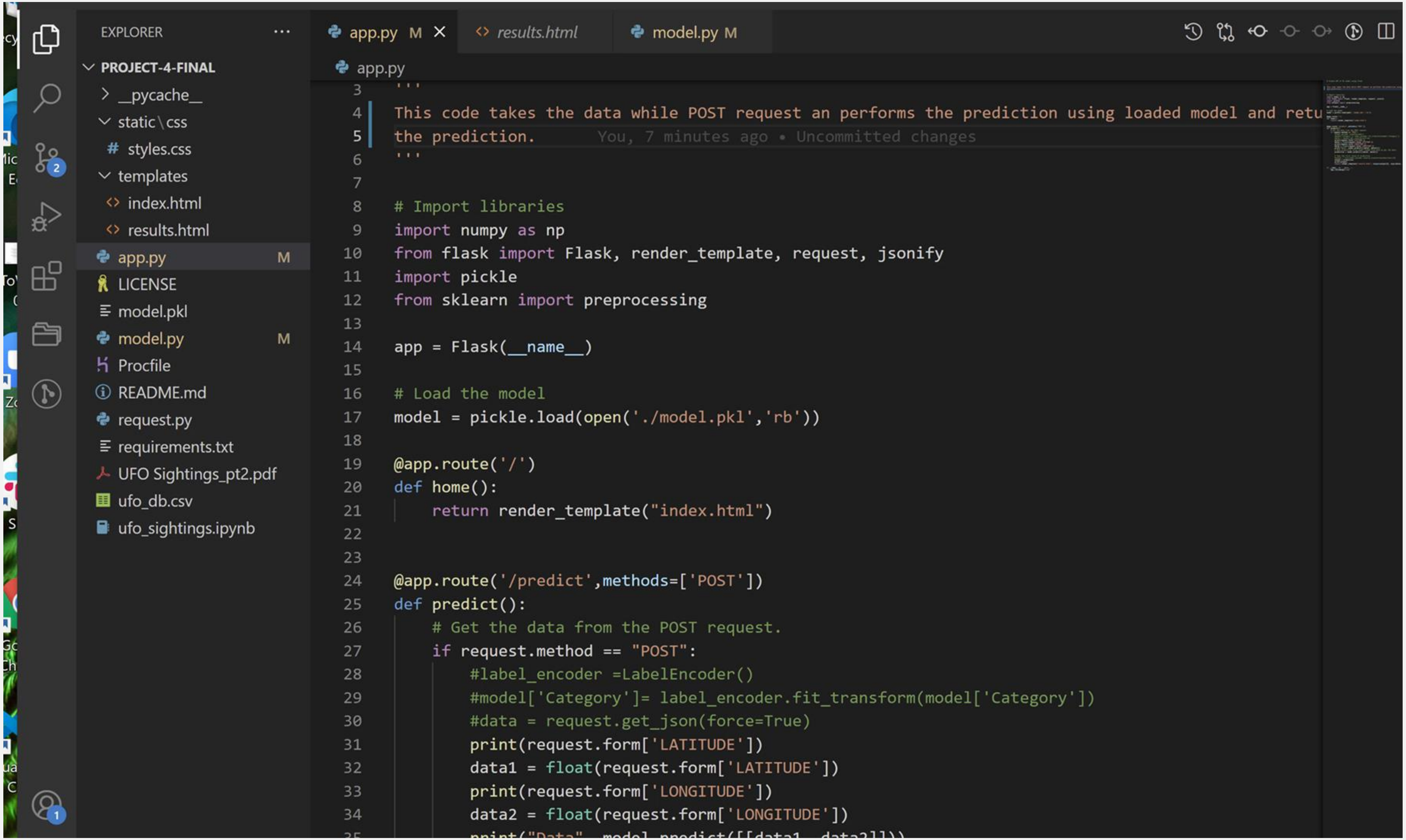
```python
import numpy as np
import pandas as pd
import datetime
from sklearn import preprocessing
from sklearn.preprocessing import LabelEncoder
from sklearn.ensemble import RandomForestClassifier
from pathlib import Path
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA
from sklearn.cluster import KMeans
import matplotlib.pyplot as plt
import tensorflow as tf
```

```python
from google.colab import drive
drive.mount('/content/drive')
```

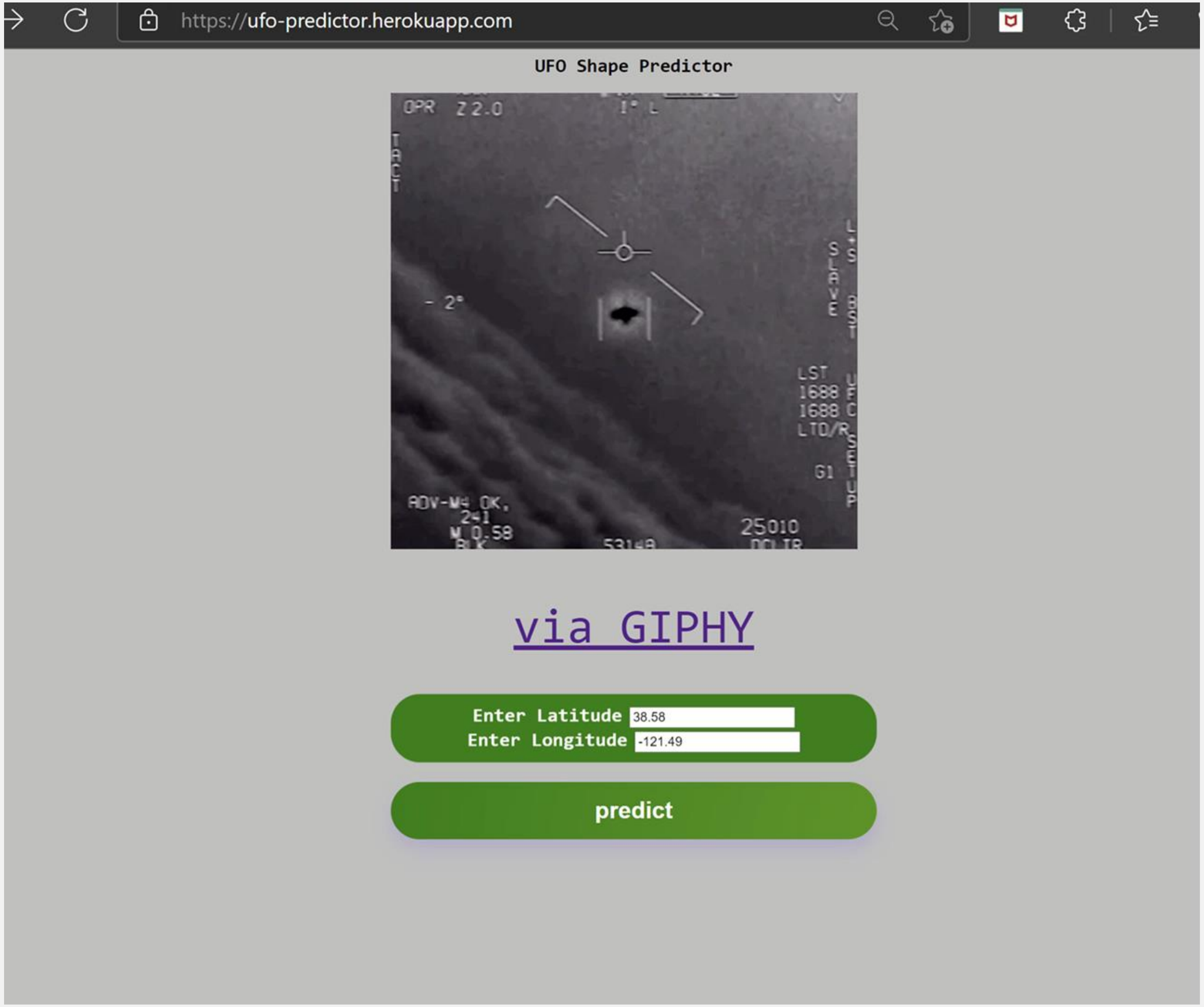
```
Mounted at /content/drive
```

```python
df = pd.read_csv("./drive/MyDrive/Project_4/ufo_sightings_locations.csv")
df
```

	Unnamed: 0	Date / Time	City	State	Shape	Duration	Summary	Posted	Location	ID	STATE_CODE	ST
0	0	9/17/21 22:10	Laguna Hills	CA	Light	15 minutes	At 10:10 pm I walked outside, scattered clouds...	NaN	Laguna Hills, CA	2245	CA	
1	1	11/11/20 16:13	Laguna Hills	CA	Circle	13 minutes	It lasted for 13 minutes, moved and then disap...	12/23/20	Laguna Hills, CA	2245	CA	
2	2	11/11/20 16:13	Laguna Hills	CA	Circle	13 minutes	Red lights were going inside two red crafts	12/23/20	Laguna Hills, CA	2245	CA	
3	3	8/18/19 21:45	Laguna Hills	CA	Circle	30	Orange light seen. In the blink of an eye it d...	8/23/19	Laguna Hills, CA	2245	CA	
4	4	7/7/16 21:47	Laguna Hills	CA	Circle	2:59 seconds	Starlike object observed.	7/15/16	Laguna Hills, CA	2245	CA	
...	...	...	...	...	...	...	...	...	...	...	...	
111479	111479	NaN	Attica	IN	NaN	NaN	A consistently appearing flying lighted uniden...	8/23/19	Attica, IN	7441	IN	
111480	111480	NaN	Warfordsburg	PA	Changing	Minutes, maybe longer, it	I know it's strange to report a craft in the v...	1/19/21	Warfordsburg, PA	23235	PA	Pe



```python
'''
This code takes the data while POST request an performs the prediction using loaded model and retu
the prediction.
'''

# Import libraries
import numpy as np
from flask import Flask, render_template, request, jsonify
import pickle
from sklearn import preprocessing

app = Flask(__name__)

# Load the model
model = pickle.load(open('./model.pkl','rb'))

@app.route('/')
def home():
    return render_template("index.html")


@app.route('/predict',methods=['POST'])
def predict():
    # Get the data from the POST request.
    if request.method == "POST":
        #label_encoder =LabelEncoder()
        #model['Category']= label_encoder.fit_transform(model['Category'])
        #data = request.get_json(force=True)
        print(request.form['LATITUDE'])
        data1 = float(request.form['LATITUDE'])
        print(request.form['LONGITUDE'])
        data2 = float(request.form['LONGITUDE'])
        print("Data", model.predict([[data1, data2]]))
```



https://ufo-predictor.herokuapp.com

UFO Shape Predictor

via GIPHY

Enter Latitude 38.58

Enter Longitude -121.49

SETBACKS

- Submissions are dependent on how the user enters the data- this created a wide variety of data types that needed to be cleaned
- The varied entries limited the number of categories that could be used
- Initially running the machine learning crashed the notebook due to RAMs being used(too many columns)
- There is a very weak correlation between location and shape prediction
- Prediction is currently overfitting

DISCUSSION

The column Category was created by combing the shapes into 2 groups: Light and Dim. The shapes placed in each group was decided by me. This was to help with the accuracy of the training as having even the limited 9 shapes as the class gave poor predictions.

GOING FORWARD

- Interesting to try other categories such as date
- Require more time cleaning the data and making it uniform

QUESTIONS?