The writeup should include 2-4 pages worth of text describing your work, and should address the “why” behind your efforts in addition to the “what”. It’s especially important to articulate the questions about the data you’re hoping to answer through analysis. How you structure the writeup is up to you, but in general you should include:
- The names of all team members, along with a brief overview of how each person contributed
- A description of the data set, including any preprocessing you did to get the data into a usable format
- A short writeup for each task, summarizing the techniques you used, as well as any conclusions you were able to draw
- An overview of the code you wrote and existing tools you used, along with instructions on how to run the code
- Description of challenges you encountered when working with the data, and how you were able to overcome them (or not!)
- Descriptions of any insights into the data or domain that you obtained through your work
- Ideas for future exploration of the data, including interesting questions raised by your analysis

Team members: Leon Ge, Phuong Nguyen

- Leon contributed to the processing of states eviction data, geocoding, times series and project write up
- Phuong contributed to the processing of weather data, visulization and association and project write up

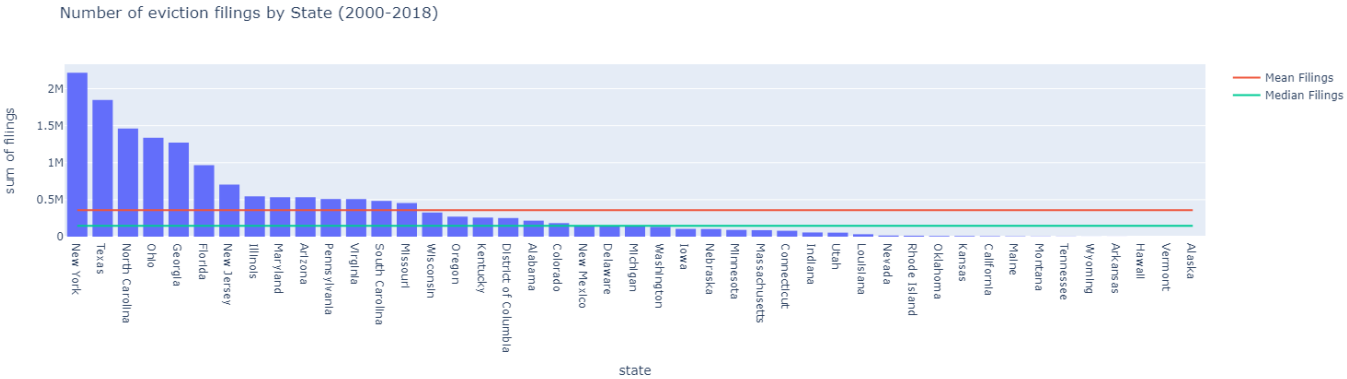
Files Summary:

- WriteUp.rtf: Write up for the whole project.
- leon.ipynb: Leon’s code on Task 1
- phuong.ipynb: Phuong’s code on Task 2

We used the pre-approved eviction dataset as our main data, we picked out specifically data that included county and states, and the number of eviction filings. We also noticed some other interesting datasets available but found it to be difficult to use or inaccurate data, such as estimation. For the weather, we were not able to find historical data on the accuweather API, so we actually needed to download from NOAA website using a portal, which was not hard to work with since it allowed for download into a csv format.

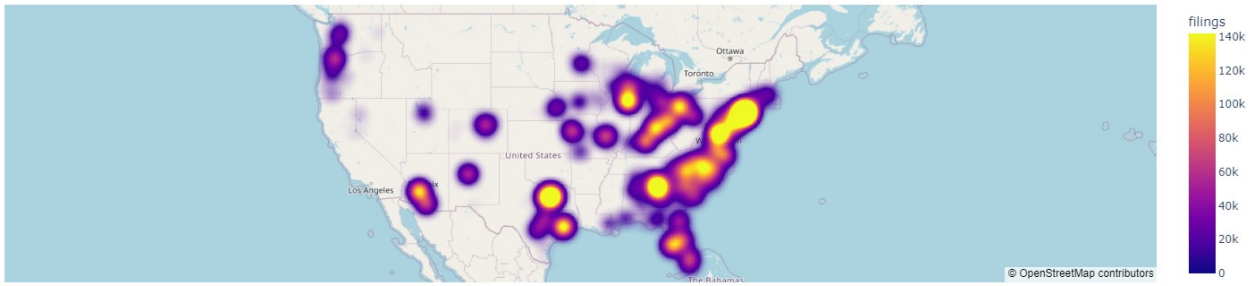
Task 1: Process of States Eviction Data, Geocoding, Time Series

For task one, I wanted to show which states has a higher-than average eviction filings, so one of the easier way to see is to plot a histogram with a line for mean and median.

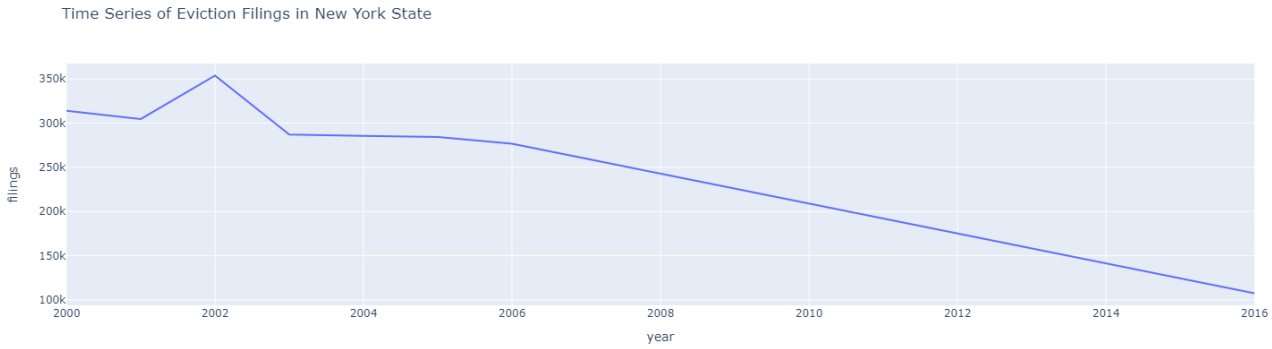


From this, it is easy to see that New York and Texas has the highest number of eviction filings from 2000 to 2018. It is also interesting to note that both the mean and the median is lower than I expected.

Although you can see which states have the highest eviction, I think it might also be interesting to see a visual map of where the evictions are happening, so I also made a heat map based on geo-coded county data.

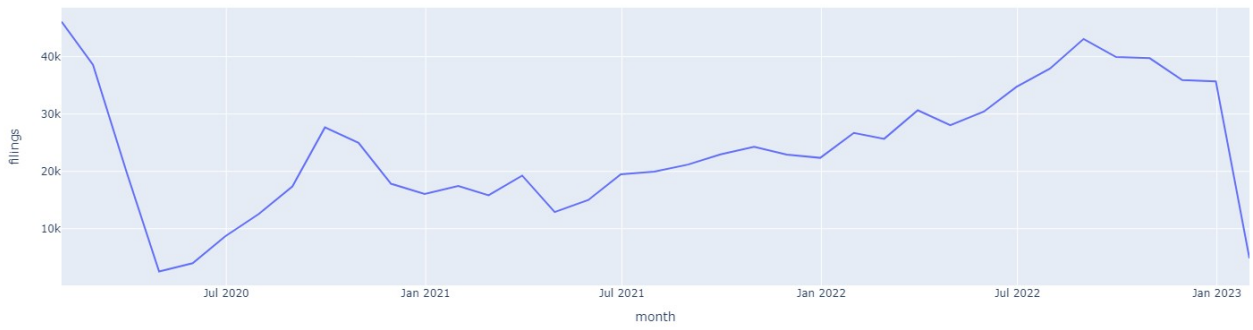


One other thing after this is that I plotted a timeseries of the number of filings over 2000-2018 for the state of New York. It's interesting to see that the number of filings has actually decreased from 2002.



For the code I have written for these tasks, I mostly used plotly and pandas. I also wrote code using the geocoding library. To run the code, running the notebook should work. One of the more challenging part of writing the code is the geocoding, it was difficult for me to figure out how to format the address so that the geocoder and understand then give me a correct result. Other than that, everything else was pretty smooth.

I also contributed to the task involving correlating number of eviction filing to weather data. I made a time series plot of all the eviction filings data available from 2020 to 2022. Now it seems like the number of evictions dipped very hard during covid, which is also interesting to see.



Task 2: Processing Weather Data, Visualize Weather Data and Analyze the Association of Temperature and Conviction Filings Count

Data set overview:

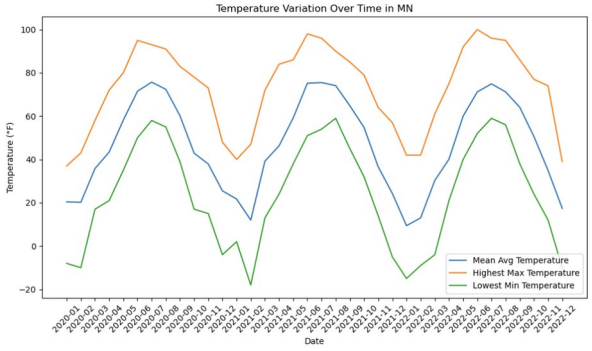
- Conviction_data: This data set consist of the following columns
  - State: Name of state. There are 9 unique states in this data set.
  - Type: Type Census Tract
  - GEOID: the GEOID number of the specific location in the state
  - Month: Month and Year Recored
  - Filings: number of filings for the specific GEOID location and in specific month
  - filings_avg: average of filings out of 100
  - last_updated: store the last updated date
- Temperature Data: 9 separate data set of temperature recorded by month (from January 2020 to Decemeber 2022) for 9 states: Connecticut, Minnesota, Pennsylvania, Virginia, New Mexico, Indiana, Missouri, Wisconsin. Each data set consisted of the following columns:
  - Date: Month and Year recorded
  - MeanAvgTemperature: recorded average temperature (in Fahrenheit)
  - HighestMaxTemperature: recorded maximum temperature in the month (in Fahrenheit)
  - LowestMinTemperature: recorded minimum temperature in the month (in Fahrenheit)

Processing Weather Data and Conviction Data:

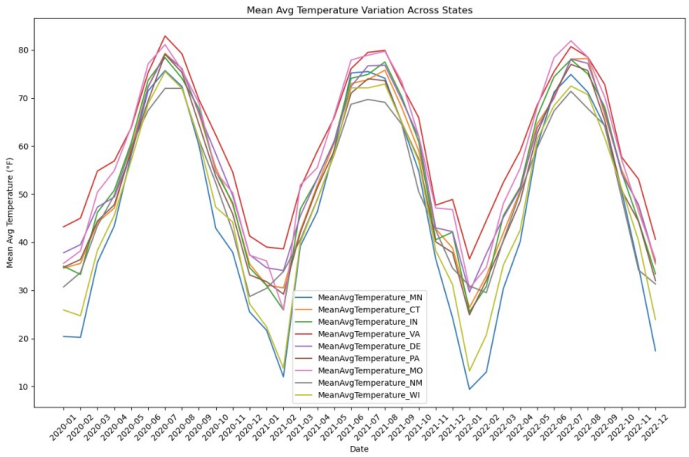
- Weather Data: As weather on the NOAA website has multiple options for me to choose, I spend a while to choose what data I should get and how should I arrange the dataset in the first place. As there are 9 different datasets from 9 different states, sometimes it's mundane to do a step repetitively 9 times. Something I'd try to learn more after this project is how to do the same task to similar datasets.
- Conviction Data: I need to modify the original dataset because it currently shows filing counts by GEOID (geographical identifier) per month. However, I require filing counts aggregated by state per month. Therefore, I must sum up all the filing counts that belong to the same state within the same month.
- Combincation of Weather Data and Conviction Data: the part where aligning these two data sets are not too difficult as they are both represented specific numbers for specific month. However, their format of date are not the same, and conviction data have extra months compare to weather data. However, these minor differences are nor hard to solve.

Visulizing Weather Data:

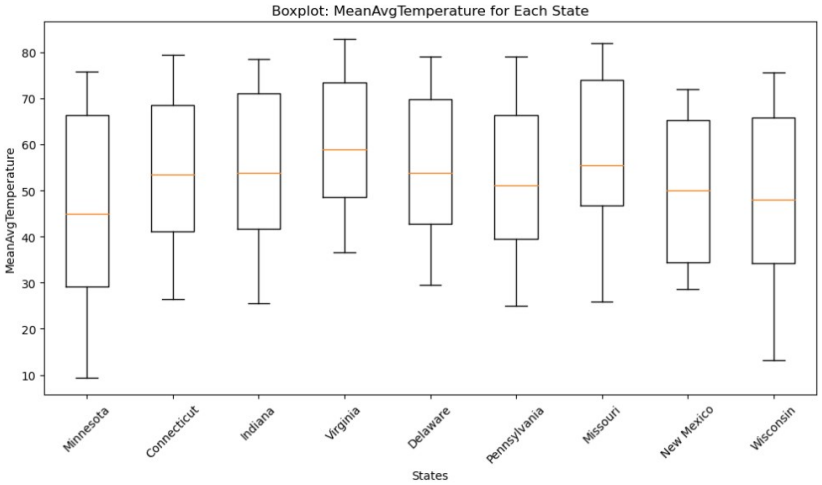
As the weather fluctuates with the seasons, I attempted to represent this change using a line graph. Figure A displays the temperature cycle over 36 months in Minnesota. This matches with my expectation, as temperatures reach their highest points around June and drop to their lowest around December to January.



Next, I aim to illustrate the differences in temperature across all states on a single graph to observe both the similarities and distinctions. Looking at the line graph below, we observe that the temperature patterns follow a similar cycle in all states, yet there are variations in the temperature ranges among them.



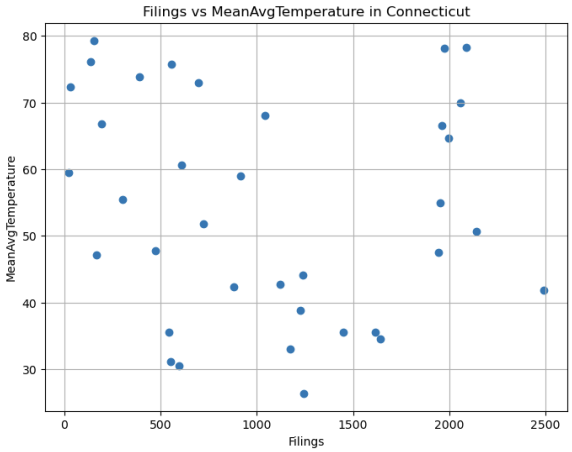
Because the lines on the graph are too closely packed, making it difficult to discern specific differences in the ranges, I also generated box plots for each state. This will provide a clearer and more detailed visualization of the data distribution for better comparison and understanding.



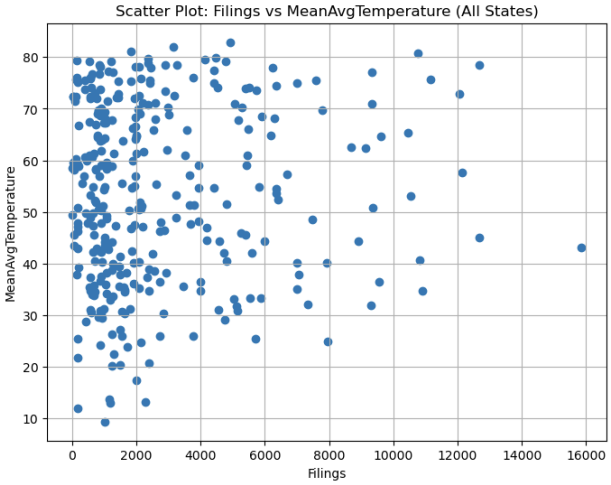
From the boxplot above, we can clearly see that Minnesota and Wisconsin are two of the coldest states, while Virginia and Missouri are two of the hottest states in the sample size we selected. Virginia lowest recored temperature is only a little bit higher than to Minnesota's mean temperature. This boxplot bettershown the comparison these state's weather to each other. Besides demonstrate critical information fiven by weather dataset, this task highlights how different types of data visualization can offer distinct insights into the dataset.

Association of Temperature and Filings Count

First, I show the correlation of Filings and Mean Temperature through scatter plot. From the scatterplot of state Connecticut, it's hard to observe a clear connection between temperature and the number of conviction filings in just one state, Connecticut. I reckon that having data from only one state across three years (36 months) is too limited to identify any correlation accurately.



Hence, I attempted to create a scatterplot depicting the relationship between Filings and Mean Temperature across all states over 36 months. Here, we observe that there isn't a prominent linear trend evident. Nevertheless, it's noticeable that areas experiencing extremely cold temperatures (30 degrees Fahrenheit or below) tend to have significantly fewer conviction filings.



Although these scatterplot already shows the insignificant correlations between Conviction Filings Count and Mean Temperature of a Place, I wanted to see the numerical result of the this association. So, I also did a Chi-square algorithm on the same Data Set.

```
# Discretize MeanAvgTemperature into bins
allstate_df['MeanTempBins'] = pd.cut(allstate_df['MeanAvgTemperature'], bins=7)

# Create contingency table
contingency_table = pd.crosstab(allstate_df['filings'], allstate_df['MeanTempBins'])

# Chi-square test on contigency table created
chi2, p_val = chisquare(contingency_table)
print(f"P-value: {p_val}")
print(f"Chi-square value: {chi2}")
```

```
P-value: [0.56904695 0.80853594 0.98311701 0.99575108 0.98518872 0.99367361
 0.9685701 ]
Chi-square value: [309.         292.         263.16129032 252.         262.
 255.         269.05263158]
```

From the result returned, we can see that the P-value is no where near 0.05. This alligns with the result above shown from the scatterplot.

Overall, I enjoyed working on this project and gained valuable insights into what to anticipate when undertaking a data analytics project. I learned the significance of comprehending the dataset at hand and the importance of adapting to various datasets throughout the process. As we did not look closely at Weather API data, we did not notice they do not store historical data, which makes us replace API task to Data Wrangling Task.

Something I'd love to add if I have more time are additional factors that might have influenced the number of conviction filings, such as population size, income levels, and so on. While working on this project, I've come to realize that focusing solely on one variable like temperature might not provide a comprehensive view when trying to establish associations. Each state likely has other unique characteristics that could contribute to conviction filings. Moreover, by incorporating more variables, we could employ techniques like K-clustering to potentially uncover different patterns and gain a more comprehensive understanding of conviction filings.