Assignment 4: Becoming an Independent Data Scientist

Applied Plotting, Charting & Data Representation in Python

Takashi Nishikawa

June 4, 2022

Selected region

Michigan, United States

Domain category of data

Weather phenomena

Research question

How has the number of tornados changed over the past several decades in Michigan and how does it compare to those of the other Midwestern states?

Data set

Storm Events Database, by the National Centers of Environmental Information (https://www.ncdc.noaa.gov/stormevents/). This database contains data on storms and other significant weather events in the US since 1950, including tornados.

Data collection

The names of relevant data files from https://www.ncei.noaa.gov/pub/data/swdi/stormevents/csvfiles/ were copied and pasted into a text file `file_names.txt`:

```
StormEvents_details-ftp_v1.0_d1950_c20210803.csv.gz 2021-08-05 09:53    10K
StormEvents_details-ftp_v1.0_d1951_c20210803.csv.gz 2021-08-05 09:56    12K
StormEvents_details-ftp_v1.0_d1952_c20210803.csv.gz 2021-08-05 09:56    12K
StormEvents_details-ftp_v1.0_d1953_c20210803.csv.gz 2021-08-05 09:56    21K
StormEvents_details-ftp_v1.0_d1954_c20210803.csv.gz 2021-08-05 09:56    26K
StormEvents_details-ftp_v1.0_d1955_c20210803.csv.gz 2021-08-05 09:56    52K
StormEvents_details-ftp_v1.0_d1956_c20210803.csv.gz 2021-08-05 09:56    62K
StormEvents_details-ftp_v1.0_d1957_c20210803.csv.gz 2021-08-05 09:56    80K
StormEvents_details-ftp_v1.0_d1958_c20210803.csv.gz 2021-08-05 09:56    69K
StormEvents_details-ftp_v1.0_d1959_c20210803.csv.gz 2021-08-05 09:56    66K
StormEvents_details-ftp_v1.0_d1960_c20210803.csv.gz 2021-08-05 09:56    70K
StormEvents_details-ftp_v1.0_d1961_c20210803.csv.gz 2021-08-05 09:56    81K
StormEvents_details-ftp_v1.0_d1962_c20210803.csv.gz 2021-08-05 09:56    83K
StormEvents_details-ftp_v1.0_d1963_c20210803.csv.gz 2021-08-05 09:56    70K
StormEvents_details-ftp_v1.0_d1964_c20210803.csv.gz 2021-08-05 09:56    84K
StormEvents_details-ftp_v1.0_d1965_c20210803.csv.gz 2021-08-05 09:56    102K
StormEvents_details-ftp_v1.0_d1966_c20210803.csv.gz 2021-08-05 09:56    81K
StormEvents_details-ftp_v1.0_d1967_c20210803.csv.gz 2021-08-05 09:56    95K
StormEvents_details-ftp_v1.0_d1968_c20210803.csv.gz 2021-08-05 09:56    112K
StormEvents_details-ftp_v1.0_d1969_c20210803.csv.gz 2021-08-05 09:56    100K
StormEvents_details-ftp_v1.0_d1970_c20210803.csv.gz 2021-08-05 09:56    112K
StormEvents_details-ftp_v1.0_d1971_c20210803.csv.gz 2021-08-05 09:56    123K
StormEvents_details-ftp_v1.0_d1972_c20220425.csv.gz 2022-04-25 15:06    80K
StormEvents_details-ftp_v1.0_d1973_c20220425.csv.gz 2022-04-25 15:06    157K
StormEvents_details-ftp_v1.0_d1974_c20220425.csv.gz 2022-04-25 15:06    183K
StormEvents_details-ftp_v1.0_d1975_c20220425.csv.gz 2022-04-25 15:06    172K
StormEvents_details-ftp_v1.0_d1976_c20220425.csv.gz 2022-04-25 15:06    133K
StormEvents_details-ftp_v1.0_d1977_c20220425.csv.gz 2022-04-25 15:06    137K
StormEvents_details-ftp_v1.0_d1978_c20220425.csv.gz 2022-04-25 15:06    133K
StormEvents_details-ftp_v1.0_d1979_c20220425.csv.gz 2022-04-25 15:06    151K
StormEvents_details-ftp_v1.0_d1980_c20220425.csv.gz 2022-04-25 15:06    211K
StormEvents_details-ftp_v1.0_d1981_c20220425.csv.gz 2022-04-25 15:06    159K
StormEvents_details-ftp_v1.0_d1982_c20220425.csv.gz 2022-04-25 15:06    240K
StormEvents_details-ftp_v1.0_d1983_c20220425.csv.gz 2022-04-25 15:06    270K
```

```
StormEvents_details-ftp_v1.0_d1984_c20220425.csv.gz 2022-04-25 15:06    248K
StormEvents_details-ftp_v1.0_d1985_c20220425.csv.gz 2022-04-25 15:06    263K
StormEvents_details-ftp_v1.0_d1986_c20220425.csv.gz 2022-04-25 15:06    291K
StormEvents_details-ftp_v1.0_d1987_c20220425.csv.gz 2022-04-25 15:06    249K
StormEvents_details-ftp_v1.0_d1988_c20220425.csv.gz 2022-04-25 15:06    250K
StormEvents_details-ftp_v1.0_d1989_c20220425.csv.gz 2022-04-25 15:06    348K
StormEvents_details-ftp_v1.0_d1990_c20220425.csv.gz 2022-04-25 15:06    377K
StormEvents_details-ftp_v1.0_d1991_c20220425.csv.gz 2022-04-25 15:06    426K
StormEvents_details-ftp_v1.0_d1992_c20220425.csv.gz 2022-04-25 15:06    468K
StormEvents_details-ftp_v1.0_d1993_c20220425.csv.gz 2022-04-25 15:06    550K
StormEvents_details-ftp_v1.0_d1994_c20220425.csv.gz 2022-04-25 15:06    1.0M
StormEvents_details-ftp_v1.0_d1995_c20220425.csv.gz 2022-04-25 15:06    1.3M
StormEvents_details-ftp_v1.0_d1996_c20220425.csv.gz 2022-04-25 15:06    6.0M
StormEvents_details-ftp_v1.0_d1997_c20220425.csv.gz 2022-04-25 15:06    6.3M
StormEvents_details-ftp_v1.0_d1998_c20220425.csv.gz 2022-04-25 15:06    9.8M
StormEvents_details-ftp_v1.0_d1999_c20220425.csv.gz 2022-04-25 15:06    9.9M
StormEvents_details-ftp_v1.0_d2000_c20220425.csv.gz 2022-04-25 15:06    7.5M
StormEvents_details-ftp_v1.0_d2001_c20220425.csv.gz 2022-04-25 15:06    6.6M
StormEvents_details-ftp_v1.0_d2002_c20220425.csv.gz 2022-04-25 15:06    6.9M
StormEvents_details-ftp_v1.0_d2003_c20220425.csv.gz 2022-04-25 15:05    6.9M
StormEvents_details-ftp_v1.0_d2004_c20220425.csv.gz 2022-04-25 15:05    7.0M
StormEvents_details-ftp_v1.0_d2005_c20220425.csv.gz 2022-04-25 15:05    7.4M
StormEvents_details-ftp_v1.0_d2006_c20220425.csv.gz 2022-04-25 15:05    7.2M
StormEvents_details-ftp_v1.0_d2007_c20220425.csv.gz 2022-04-25 15:05    9.3M
StormEvents_details-ftp_v1.0_d2008_c20220425.csv.gz 2022-04-25 15:05    12M
StormEvents_details-ftp_v1.0_d2009_c20220425.csv.gz 2022-04-25 15:05    9.8M
StormEvents_details-ftp_v1.0_d2010_c20220425.csv.gz 2022-04-25 15:05    11M
StormEvents_details-ftp_v1.0_d2011_c20220425.csv.gz 2022-04-25 15:05    15M
StormEvents_details-ftp_v1.0_d2012_c20220425.csv.gz 2022-04-25 15:05    11M
StormEvents_details-ftp_v1.0_d2013_c20220425.csv.gz 2022-04-25 15:05    11M
StormEvents_details-ftp_v1.0_d2014_c20220425.csv.gz 2022-04-25 15:05    11M
StormEvents_details-ftp_v1.0_d2015_c20220425.csv.gz 2022-04-25 15:05    9.5M
StormEvents_details-ftp_v1.0_d2016_c20220425.csv.gz 2022-04-25 15:05    8.6M
StormEvents_details-ftp_v1.0_d2017_c20220425.csv.gz 2022-04-25 15:05    9.1M
StormEvents_details-ftp_v1.0_d2018_c20220425.csv.gz 2022-04-25 15:05    10M
StormEvents_details-ftp_v1.0_d2019_c20220425.csv.gz 2022-04-25 15:05    11M
StormEvents_details-ftp_v1.0_d2020_c20220322.csv.gz 2022-03-22 14:20    9.9M
StormEvents_details-ftp_v1.0_d2021_c20220520.csv.gz 2022-05-20 09:41    10M
StormEvents_details-ftp_v1.0_d2022_c20220520.csv.gz 2022-05-20 09:42    1.6M
```

The code below reads from this file and creates a list of file names:

In [1]:
```
with open('file_names.txt', 'r') as f:
    lines = f.readlines()
file_names = []
for line in lines:
    file_names.append(line.split('\t')[0])
file_names = file_names[1:]
```

The code below downloads all the CSV files in the list to obtain data for all years (1950 - 2022) from the source URL and then saves to a CSV file `noaa_data.csv`. For some reason, `pd.read_csv()` did not work when running on the Coursera Jupyter Notebook platform, so it had to be run on a local Jupyter Notebook on a laptop computer (which took several minutes).

In [ ]:
```
import pandas as pd

url = 'https://www.ncei.noaa.gov/pub/data/swdi/stormevents/csvfiles/'
df_list = []
for fn in file_names:
    print(fn)
    df = pd.read_csv(url + fn)
    df = df[['YEAR', 'STATE', 'TOR_F_SCALE']]
    df_list.append(df)
pd.concat(df_list).to_csv('noaa_data.csv')
```

Data cleaning

The code bellow loads the file `noaa_data.csv`, cleans the data frame, and use `groupby()` to count the number of tornado occurrences for the 12 Midwestern state (according to the Census Bureau's definition; source: https://en.wikipedia.org/wiki/Midwestern_United_States). The row for the current year 2022 was dropped since only partial data was

available. (Note: `noaa_data.csv` ended up being larger than 30MB and thus could not be uploaded to the Coursera Jupyter Notebook platform.)

In [2]:
```
import pandas as pd

df = pd.read_csv('noaa_data.csv')
df = df.drop(['Unnamed: 0'], axis=1)
g = df.groupby(['YEAR', 'STATE'])
df = g.count().unstack().droplevel(0, axis=1)
df.index.name = 'Year'
df.columns = [s.title() for s in df.columns]
df = df.fillna(0)
df.drop(2022, axis=0, inplace=True)
midwestern_states = [
    'Illinois', 'Indiana', 'Iowa', 'Kansas', 'Michigan',
    'Minnesota', 'Missouri', 'Nebraska', 'North Dakota',
    'Ohio', 'South Dakota', 'Wisconsin'
]
df = df[midwestern_states]
df
```

Out[2]:

Year	Illinois	Indiana	Iowa	Kansas	Michigan	Minnesota	Missouri	Nebraska	North Dakota	Ohio	South Dakota	Wisconsin
1950	11.0	2.0	4.0	32.0	0.0	1.0	6.0	6.0	2.0	3.0	1.0	5.0
1951	5.0	5.0	5.0	47.0	8.0	4.0	6.0	9.0	1.0	3.0	2.0	7.0
1952	4.0	2.0	6.0	18.0	0.0	11.0	11.0	11.0	9.0	2.0	4.0	1.0
1953	5.0	9.0	32.0	28.0	19.0	10.0	6.0	37.0	8.0	8.0	12.0	9.0
1954	9.0	31.0	19.0	56.0	12.0	7.0	40.0	18.0	3.0	12.0	16.0	12.0
...	...	...	...	...	...	...	...	...	...	...	...	...
2017	59.0	37.0	68.0	63.0	10.0	71.0	91.0	35.0	33.0	44.0	15.0	28.0
2018	60.0	16.0	76.0	48.0	16.0	57.0	50.0	33.0	29.0	19.0	19.0	35.0
2019	51.0	32.0	59.0	102.0	8.0	56.0	81.0	36.0	15.0	58.0	26.0	32.0
2020	74.0	17.0	31.0	17.0	3.0	64.0	21.0	23.0	22.0	24.0	23.0	21.0
2021	82.0	21.0	146.0	39.0	18.0	68.0	61.0	56.0	15.0	33.0	21.0	46.0

72 rows × 12 columns

The code below computes the 10-year rolling average for each state:

In [3]:
```
df = df.rolling(10).mean().dropna()
df
```

Out[3]:

Year	Illinois	Indiana	Iowa	Kansas	Michigan	Minnesota	Missouri	Nebraska	North Dakota	Ohio	South Dakota	Wisconsin
1959	20.4	16.0	12.9	45.4	10.6	9.3	25.2	25.4	5.6	6.9	9.0	10.5
1960	23.2	16.7	15.4	46.9	11.1	9.9	28.5	28.8	6.3	7.3	9.8	10.6
1961	27.3	19.5	16.5	45.7	10.8	10.3	32.2	28.7	6.2	8.8	10.9	10.7
1962	28.0	20.7	17.0	48.7	11.3	10.7	33.3	31.2	6.0	8.9	14.1	11.4
1963	29.0	23.1	15.3	48.8	9.9	10.4	34.3	29.0	5.7	9.7	13.7	11.4
...	...	...	...	...	...	...	...	...	...	...	...	...
2017	60.5	37.8	57.9	101.4	14.0	47.8	62.3	48.3	32.2	26.9	23.2	28.2
2018	61.0	35.7	54.0	84.4	14.0	48.6	56.5	45.0	31.6	27.3	22.4	27.2
2019	60.0	37.8	57.3	83.9	14.5	51.7	59.6	44.5	30.0	31.8	21.6	28.7
2020	62.1	36.4	57.0	76.8	12.1	45.5	54.7	43.0	25.6	29.1	19.7	25.5
2021	62.5	30.0	66.2	73.3	12.1	48.5	51.5	43.0	20.9	28.4	19.9	25.0

63 rows × 12 columns

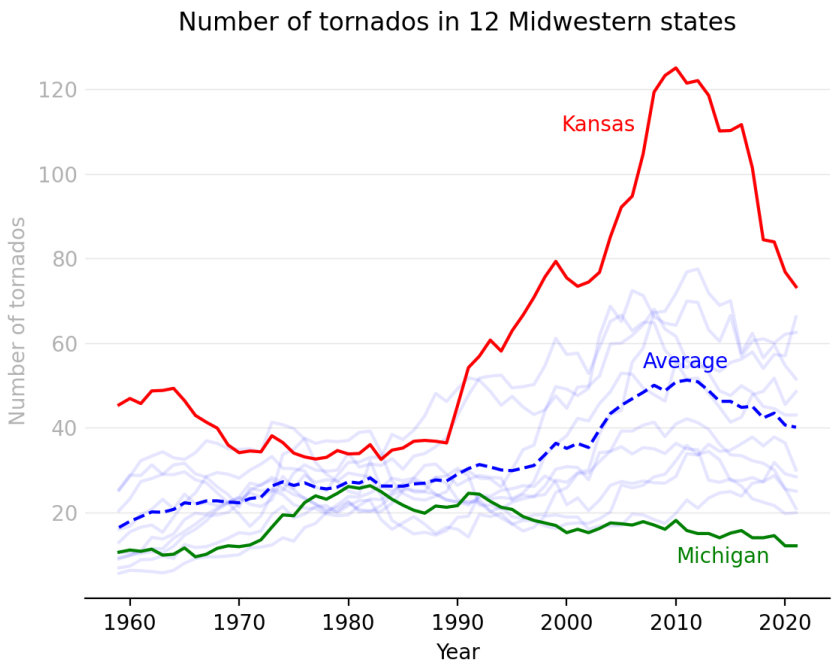
Data visualization

The code below creates the visualization of the cleaned data:

In [4]:

```
%matplotlib notebook
import matplotlib.pyplot as plt

fig = plt.figure()
yticks = range(20, 121, 20)
for y in yticks:
    plt.axhline(y, linewidth=0.5, color='k', alpha=0.1)
plt.plot(df.index, df, color='b', alpha=0.1)
plt.plot(df.index, df['Michigan'], label='Michigan', color='g')
plt.plot(df.index, df['Kansas'], label='Kansas', color='r')
plt.plot(df.index, df.mean(axis=1), 'b--', label='Midwestern states average', alpha=1)
ax = plt.gca()
ax.annotate('Kansas', (1999.5, 110), color='r')
ax.annotate('Michigan', (2010, 8), color='g')
ax.annotate('Average', (2007, 54), color='b')
plt.xlabel('Year')
plt.ylabel('Number of tornados', alpha=0.3)
for text in ax.get_yticklabels():
    text.set_alpha(0.3)
plt.title('Number of tornados in 12 Midwestern states')
ax.set_yticks(yticks)
ax.tick_params(axis='y', length=0)
for x in ['top', 'left', 'right']:
    ax.spines[x].set_visible(False)
plt.show()
```



The green curve shows that the number of tornados in Michigan has stayed low over the past six decades and slightly decreased in the last four decades. This is in contrast with the steady increase in the average number of tornados for the 12 Midwestern states (dashed blue curve) through 2010. The contrast is even larger with the number of tornados in Kansas (red curve), which show much more drastic increase from 1990 to 2010, followed by a sharp drop. The light blue curves in the background correspond to the rest of the Midwestern states.

Here is how the visualization incorporates Cairo's principles:

- **Truthfulness.** The choice of using 10-year rolling averages simultaneously allowed for an accurate representation of the trends over several decades and a reduction of noise in the data (i.e., large year-to-year fluctuations).
- **Beauty.** The format and color choices for the visual were made to keep it simple and beautiful.
- **Functionality.** The design is free of any decorative items and focused on the minimum that is need to tell the story.

- **Insightfulness.** The use of color and transparency to highlight the curves for Michigan, Kansas, and the average helps convey the story of how the Michigan number has changed over the last six decades compared to the numbers for the other Midwestern states.

Finally, the code below saves the visualization as a PNG image file:

In [5]:
```
plt.savefig('Assignment_4_plot.png', dpi=300)
```