ISA 419: Data-Driven Security

10: Visualizing Data in Python

Fadel M. Megahed, PhD

Endres Associate Professor
Farmer School of Business
Miami University

@FadelMegahed
fmegahed
fmegahed@miamioh.edu
Automated Scheduler for Office Hours

Spring 2024

Quick Refresher of Last Class

✅ Create quick visualizations using the `plot` method from pandas (with an understanding of the effect of different backends).

✅ Utilize `auto-viz` type plots to create a quick EDA of your data.

Learning Objectives for Today's Class

- Utilize standalone data viz packages to construct and tailor your graphs.
- Examine the use of network and/or spatial plots in the context of network data

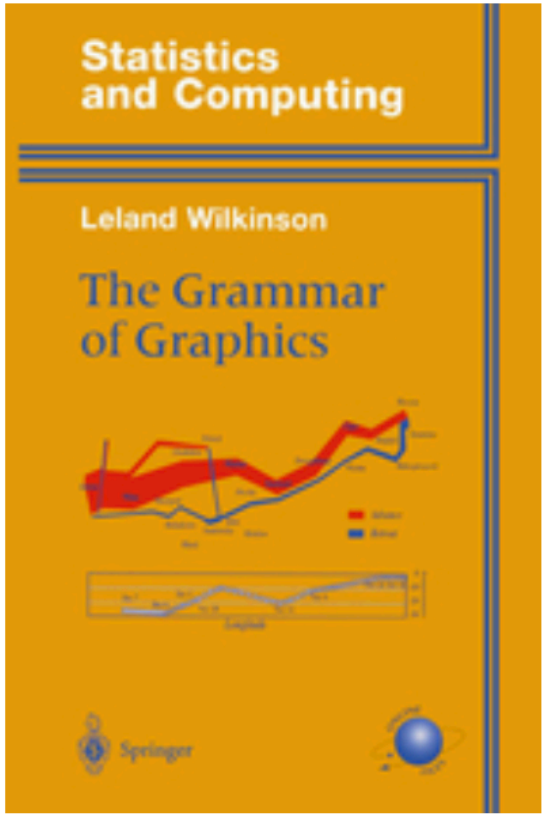
Utilize Standalone Data Viz Packages to Construct and Tailor your Graphs

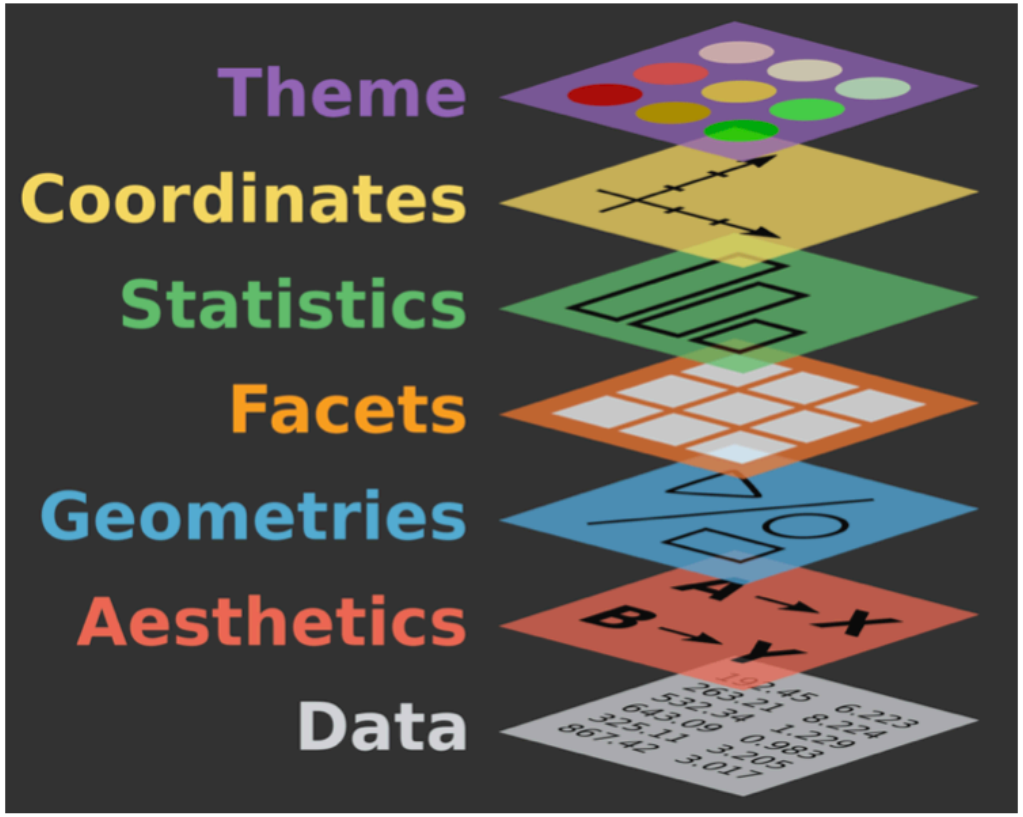
Our Data

- We will use the `merged_ips` data set from a previous class to demonstrate how to plot data in pandas.

```
import pandas as pd
import numpy as np

np.random.seed(2024) # so we get the same random sample

toxic_ips = pd.read_csv(
  "https://raw.githubusercontent.com/fmegahed/isa419/main/data/listed_ip_90_all.csv",
  header = None, names = ['ip', 'frequency', 'lastseen']
)

geolocation = pd.read_csv(
  'https://raw.githubusercontent.com/fmegahed/isa419/main/data/ip_geolocation.csv',
  names = ['ip', 'country', 'city', 'latitude', 'longitude']
)

merged_ips = (
  toxic_ips
  .merge(right = geolocation, how = 'left', on ='ip')
  .dropna()
  .assign( lastseen = lambda df: df['lastseen'].astype('datetime64[ns]') )
  .sample(1000) #<< a random sample of 1000 rows for faster plotting
  .query('frequency < 100')
)
```

Grammar of Graphics





The Grammar of Graphics with `plotnine`

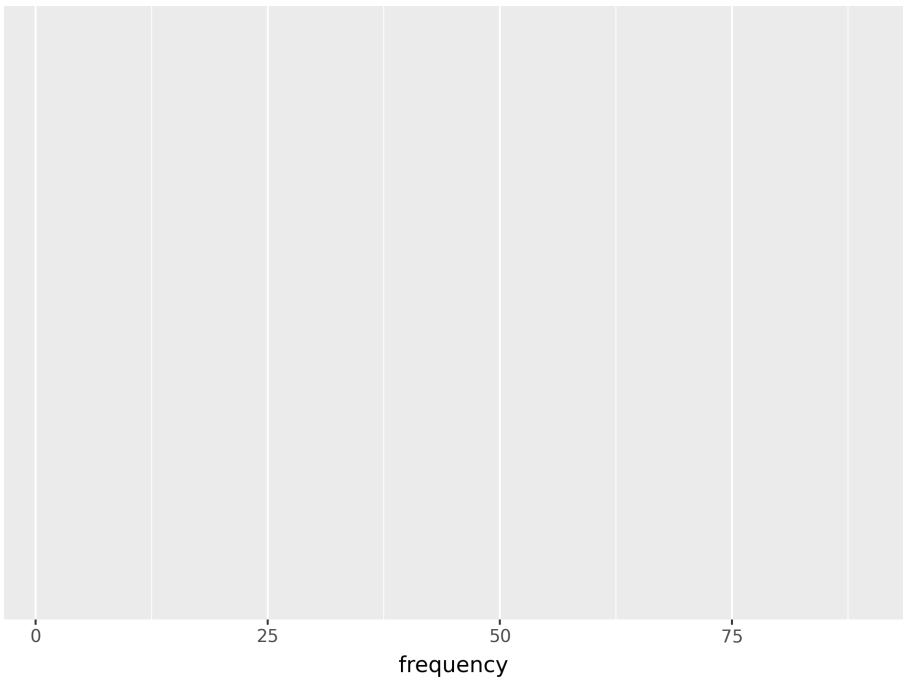
```
from plotnine import ggplot, aes, geom_histogram, la

(
  ggplot(merged_ips) # data
)
```

```
## <Figure Size: (640 x 480)>
```

The Grammar of Graphics with `plotnine`

```
from plotnine import ggplot, aes, geom_histogram, la

(
  ggplot(merged_ips, aes(x = 'frequency') ) # data
)
```

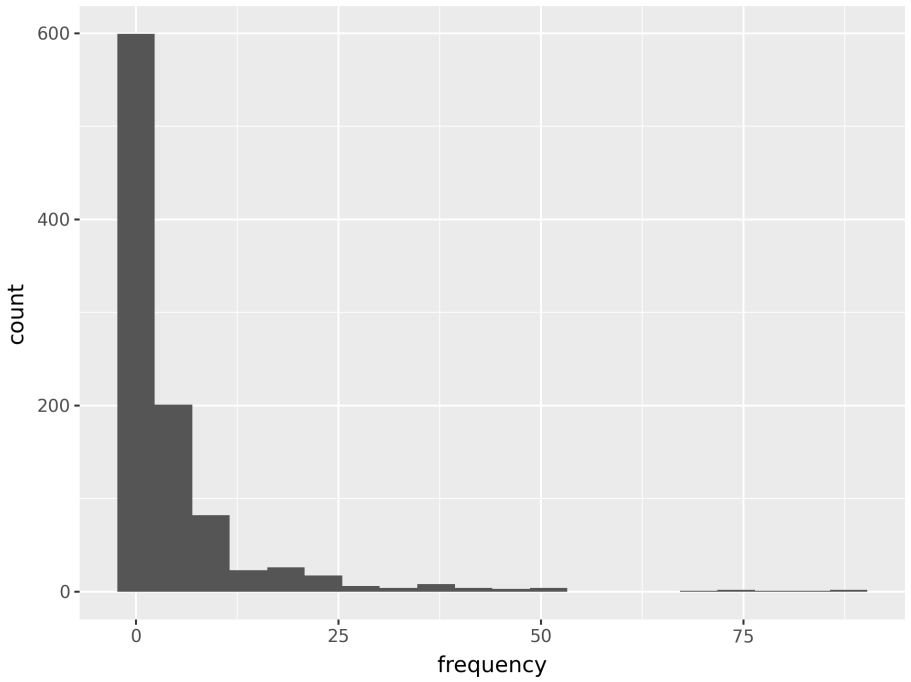
```
## <Figure Size: (640 x 480)>
```



The Grammar of Graphics with `plotnine`

```
from plotnine import ggplot, aes, geom_histogram, la

(
  ggplot(merged_ips, aes(x = 'frequency') ) + # data
  geom_histogram(bins = 20) # geometry
)
```

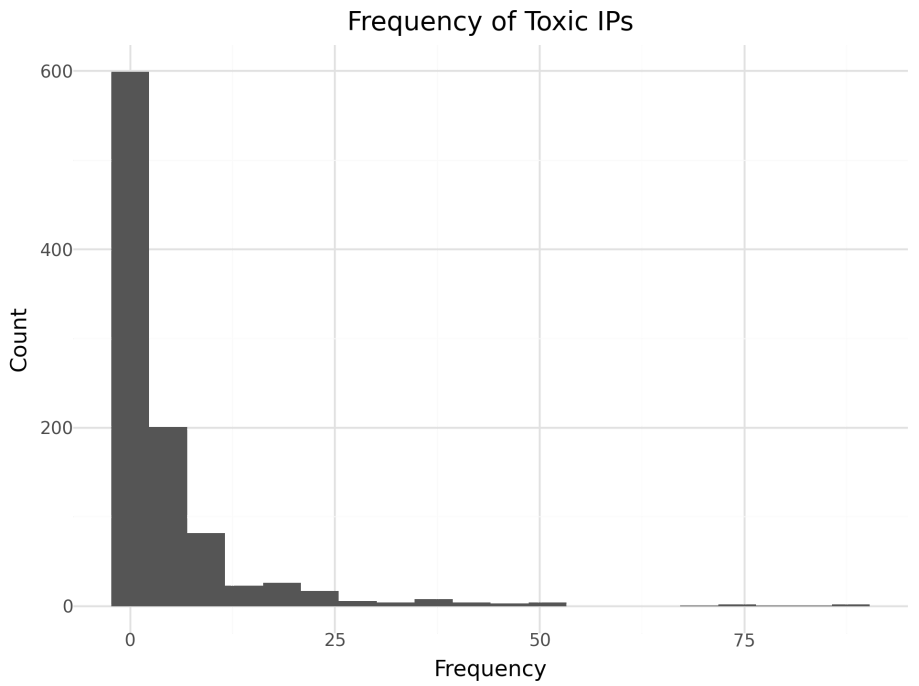
```
## <Figure Size: (640 x 480)>
```



The Grammar of Graphics with `plotnine`

```
from plotnine import ggplot, aes, geom_histogram, la

(
  ggplot(merged_ips, aes(x = 'frequency') ) + # data
  geom_histogram(bins = 20) + # geometry
  labs(title = 'Frequency of Toxic IPs', x = 'Freque
  theme_minimal() # theme
)
```

```
## <Figure Size: (640 x 480)>
```



Class Activity

15:00

Task | Solution

- After going through the plotnine Reference and Gallery tabs, create a scatter plot of the `latitude` and `longitude` variables from the `merged_ips` data set using `plotnine`.
- Add a title to the plot and label the x and y axes.
- Use the `theme_minimal` theme.
- Annotate the plot with the `country` variable.

Class Activity

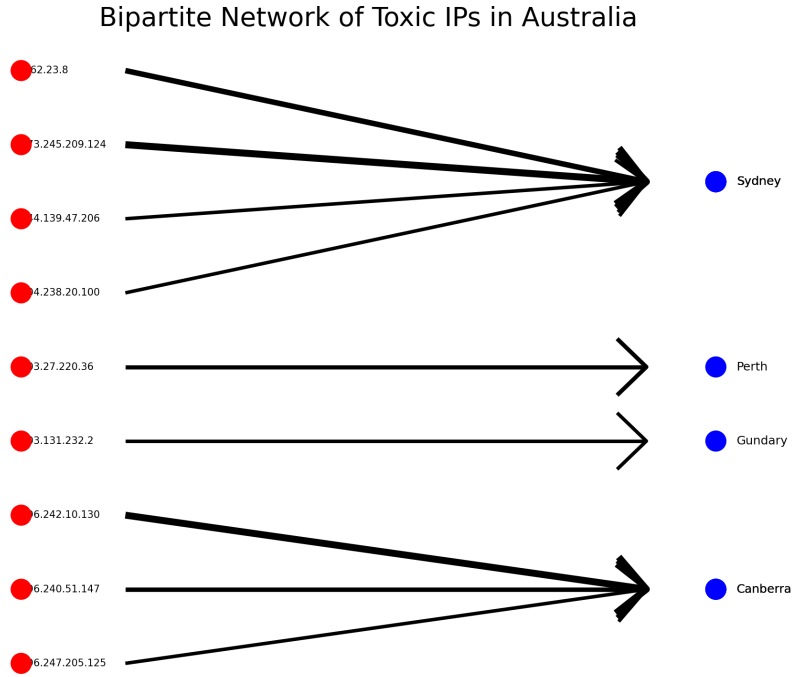
15:00

Task | Solution

Examine the Use of Network Plots in the Context of Network Data

Network Data

Let us create a simple network data set to demonstrate how to plot network data in Python. We will focus on querying the `merged_ips` data set to only include toxic IPs from Australia. We will create a simple bipartite network with the `ip` and `city` variables.



Recap

Summary of Main Points

By now, you should be able to do the following:

- Utilize standalone data viz packages to construct and tailor your graphs.
- Examine the use of network and/or spatial plots in the context of network data.

 Review and Clarification 

- **Class Notes**: Take some time to revisit your class notes for key insights and concepts.
- **Zoom Recording**: The recording of today's class will be made available on Canvas approximately 3-4 hours after the end of class.
- **Questions**: Please don't hesitate to ask for clarification on any topics discussed in class. It's crucial not to let questions accumulate.