Python: Data Visualization Notes

Klein
carlj.klein@gmail.com

1 Learning Objectives

What good is a data analysis if the answers or findings can't be conveyed properly? Or perhaps even more detrimental, what good is a dataset if the proper questions or hypotheses can't be formed in the first place? How can we properly tell the story the data is providing? Enter in data visualization! Both explaining the data and exploring the data can be significantly helped through the process of data visualization.

Data visualization is two-pronged:

- Exploratory Analysis: Helping to understand the data prior to analysis. Searching for relationships and insights.
- Explanatory Analysis: Helping to present analysis findings, helping to tell a story with the data. After insights were found.

And it's all a part of the entire data analysis process. We can also simplify the data analysis process into 5 steps:

1. Extract
2. Clean: Exploratory
3. Explore: Exploratory
4. Analyze: Exploratory OR Explanatory
5. Share: Explanatory

In this course, we'll be using the matplotlib, seaborn, and Pandas libraries to assist in the data analysis process.

Concepts

- Design of Visualization
- Exploration of Data

  - Univariate Exploration of Data
  - Bivariate Exploration of Data
  - Multivariate Exploration of Data
- Explanatory Visualizations
- Visualization Case Study

2 Design of Visualization

To begin our discussion of visualization, we need to cover some basic vocabulary and distinctions.
Data can be broken into two main categories, each of which can be broken down further:

- Qualitative / Categorical:
  - Nominal: No order
  - Ordinal: Intrinsic Order
- Quantitative / Numerical:
  - Interval: Absolute differences are meaningful (addition and subtraction follows logic)
  - Ratio: relative differences are meaningful (multiplication and division follows logic)

It should be noted that the quantitative data type can be also be broken down into discrete and continuous variables.

What about those 3-dimensions charts or fun backgrounds that we used to add to our science experiment plots as kids? That used to add some fun to our projects, right? While fun for the youth, it turns out there is an empirical rule when figuring out how much the additional "junk" either adds or detracts from conveying the data.

- Data-Ink Ratio = data-ink / total ink used to print the graphic. The higher the Data-Ink Ratio, the better conveyed data is.

Can visualizations be purposefully misleading, even when using the data appropriately? Absolutely!

A great example of this is trying to over-inflate the difference or change between data points during different time periods. Say a presenter is trying to make a claim there was a very large change from one year to the next. We'll say in year

1 the y-value was 100, and in year 2 the y-value was 105. The presenter changes the window of the graph to display from a y-value of 99 to a y-value of 106, and the x-values are only the two years. Obviously, this is going to look like a massive change! In reality, had the reporter shown the data at a true scale, the visual shows in actuality that the change isn't so tremendous.

This concept also has an empirical rule:

- Lie Factor = size of effect shown in graphic / size of effect shown in data

  = (change in visual / visual start) / (change in data / data start)

  As was said in the example, this can be used to purposefully distort data. In fact, a Lie Factor 1 suggests a misleading visual, and even greater than that suggest an even greater disparity from the truth.

Away from the empirical side of visuals, and more into the logical, we come across the common mistake of using too many colors! Colors can be useful when separating categories, however, they it's very easy to cause redundancy with them. Here are some tips when using color:

- Get it right in black and white (and shades of grey)
- Use less intense colors such as natural or pastel, and higher grey colors. The eye can actually concentrate longer under these conditions.
- Color facilitates communication. Use color to separate the data into groups of interest, not just to color a visual.
- Design for Color Blindness. Stay away from red / green pallets, and use blue / orange pallets.

Don't want to overdo it on the color schemes? Don't forget about other visual queues such as shape and size.

Some tips on shape, size & other tools:

- Use different types of encodings, rather than using color (square / dot vs. colors to separate groups of interest).
- Color and shape are good for categorical variables.
- Size of marker can assist in adding additional quantitative data.

3 Exploration of Data

We have a dataset on a topic or concept that has been deemed worthy for inspection! Surely, there are some insights to be gained from it. We load up the data, and then we hit a wall... Which columns are important? What questions can be answered? We can find the general statistics of the set, so what?

This section will help with the initial process of data exploration, helping to find what is actually useful and should be examined further in the data. We'll start with single variables from the data and move into visually pairing multiple data points at once.

We'll be using a few different Python libraries in this section:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sb
```

3.1 Univariate Exploration of Data

3.2 Bivariate Exploration of Data

3.3 Multivariate Exploration of Data

4 Explanatory Visualizations

5 Visualization Case Study