Introduction to Visualization

This lesson will cover the importance of visualizing data, and why visualization is required in data science as well as common tools.

WE'LL COVER THE FOLLOWING

The importance of data visualization

So far, we have looked at understanding data via descriptive statistics and tables. Another useful tool is visualization.

Visualizations of data can provide the following benefits:

- A better understanding of the data
- A more compelling story when explaining the data
- An easier to comprehend medium

Data visualization is a core skill necessary for any analyst or data scientist. Being able to use great visualizations to help tell a data story often significantly adds to the comprehension of others. That added understanding can be the difference in driving a project forward.

Data visualization tools in Python

In Python, two of the most popular tools for visualizing data are **Matplotlib** and **Seaborn**. These are the tools we will focus on, but there are many others including **Bokeh**, **ggpy**, and **D3**.

Matplotlib is sort of the base plotting library in Python. Think of it as a low-level library that allows you to do all sorts of things, but this flexibility can sometimes make it hard to work with. Matplotlib has been around for a while and sometimes can look a bit dated in style.

Seaborn was created to help deal with some of these issues. It is built on top of Matplotlib and in its own words, “provides a high-level interface for drawing attractive statistical graphics.” For the most part, when possible, I lean towards using Seaborn. When I need more low-level control, I pull in Matplotlib. Since Seaborn is built on top of Matplotlib, it is pretty easy to mix the two.

Our dataset

Before we get started with visualization, I will show how to load in the dataset we will be using:

```
from sklearn.datasets import load_boston
import pandas as pd

# Load the boston dataset from sklearn.datasets
boston_data = load_boston()
# Enter the boston data into a dataframe
boston_df = pd.DataFrame(boston_data.data, columns=boston_data.feature_names)

# Print the first 5 rows to confirm ran correctly
print(boston_df.head())
```

You will notice that we load in the Boston dataset from sklean to use as data to visualize and convert it to a Pandas dataframe in the code above.

Now! Let’s start visualizing some data! We will cover the following types of plots:

- Scatter
- Bar
- Distribution
- Line
- Heat map
- Data aware

- Data aware