Understanding Complex Results

Nick DeRobertis

November 18, 2020

1 Introduction to Visualization

- Visualization is a key modeling concept as often we have many different outputs to understand, but humans are terrible at getting understanding by looking at lots of numbers

- Thoughtfully creating appropriate visualizations will allow someone to glance at your model and gain immediate understanding at a much richer level

- Tables are a more primitive form of visualization which lay out the numbers in a better format, while charts/graphs can summarize a lot of numbers in one picture

- For the most part, visualization in Excel is straightforward: insert chart and follow the prompts. Your numbers should already be in tables.

- Python, being open-source and developed by the community, has a dizzying array of options for visualization. There is far more than you can do in Excel, including interactive plots, but it is generally a bit more complicated to work with

- In this course, we will focus on Pandas (powered by matplotlib) to produce graphs simply

2 Visualization in Excel Example

- Recommended Charts is a nice way to scan through a few possibilities which probably work well for your data, but take a look at All Charts if nothing seems right

- Make sure that you have an appropriate title and axis titles for your chart so the reader immediately knows what it is about.

3 Introduction to Pandas

- We will be using pandas to produce tables and graphs in this course though the custom DataFrame type

- You will also find these DataFrames useful for general problem-solving purposes. Many use them as a primary way to store and work with data in their models

- Pandas does far more than we will cover in the course. It is the top Python package for manipulating and analyzing data. I use it extensively on a daily basis.

- In this course, with Pandas we will focus on loading and exporting data, doing math, other basic operations and summarizations, and presenting data in a tabular format

4 Styling Pandas DataFrames

- Just as it is important to format tables in Excel to increase readability, we should do the same with any Pandas DataFrames we display to the reader of the model

- There is a philosophical difference in how styling is done in Excel versus Pandas. In Excel, you directly format the table which stores your data. In Pandas, you create a styled object immediately before displaying which is separate from the original data, the data itself does not get formatted

- Because of this difference in philosophy, the way I recommend working with Pandas styling is to create a styler function that accepts a DataFrame and returns the styled object. This way you can just call it on your DataFrame as you display it. This has a couple advantages: your data logic is completely separate from formatting code, and you can apply consistent formatting to multiple different DataFrames easily.

5 Introduction to Graphs in Python with Pandas

- All the main graph types that you would expect are available in Pandas

- See the official Pandas visualization guide on how to adjust any plots to your liking, but the defaults are already pretty good

6 Visualization in Python Example

- If you have structured your model well, it should be easy to add visualization at the various stages of your model

- Visualizations are especially helpful in Python as you don't automatically see the tabular representation of the data like you do in Excel

7 Lab Exercises

- Complete all the exercises in the Pandas and Visualization Labs Jupyter notebook