Contributions:

Thomas:

My role in this project was to create the vizualiations and to wrangle the data in order to create csv files that would give the graphs we wanted to show to people.

Once Sam gave me the unified datasets, I was able to split them up into new dataframes dependign on what was needed and then clean and fill them so as to give perfect datasets. A lot of my work was working with pandas and then working with vizualization software such as Datawrapper and Flourish.

It took a while to find the right vizulaistaion tools but I settled on these two because of their high amount of customisability.

Sam and I also worked together to make the wesbite for the project.

Samuel Woolledge:

My role in this project boiled down to handling web-scraping and some data cleaning and wrangling, in order to make Thomas's job wrangling the data into a form accepted by the visualization tools a little easier.

For the web-scraping part, I created a small tool which generated random valid patent numbers by year, so as to create a list of patents we would scrape. I then set up a program that would scrape the website (paft.uspto.gov), by constructing URLs based on the generated patent numbers and feeding those URLs to an slightly modified version of the "Patent" object from the "pypatent" library, and saved that data into .csv files by year which contained the data we planned to work with, as well as some additional data that the "Patent" object was already set up to scrape.

For cleaning, I set up a program to remove invalid records (records that contained invalid data, or where the scraping had failed entirely) and created two unified datasets - one which kept a record of how many patents a patent assignee had assigned to them, and another which was similar to the scraped datasets, but had some additional cleaning and parsing (this involved properly parsing the assignee location and figuring out the appropriate CPC category), and included a separate column to denote year.