Data Description

We provide publicly available datasets on US patents granted between 1976 and 2022. The data comes from the PatentsView (PV) and Patent Assignment (PA) databases:

- g_patent
- g_assignee_disambiguated
- g_location_disambiguated
- g_us_patent_citation_1
- g_us_patent_citation_2
- g_cpc_current
- pa

To reduce the file size, we removed some variables that you do not need for the task. The identifier of a patent is the *patent_id*. PatentsView uses a disambiguation algorithm to create an identifier for patent assignees (i.e., the owner of the patent): *assignee_id*.

Programs

Complete the following tasks using Stata and/or Python.

Deliverables

- A report containing your answers to the tasks set out below, any major assumptions you made along the way, and any suggestions to improve the analysis you think could be constructive.
- Your code files to allow us to replicate your work.

Tasks

We are interested in the role of US startups in innovation between 1985 and 2015. Assume "innovative startups" are *founded at the grant date of their first patent*. The grant_date is labeled patent_date in the g_patent dataset. For example, *GoPro, Inc* is a startup founded in the year 2004, as their first patent (6,799,702) was granted on the date 2004-10-05.

1. What is the number of startups founded per year? Visualize the number of startups founded per year.
2. What is the share of startups' first patents among all patents? Visualize the share of startups' first patents per year.
3. Which are the US states with the largest number of startups?
4. What are the technology fields (cpc_class) with the largest number of startups?

Next, we want to examine *the quantity and quality of startups' innovative output*. You can measure the quantity by the number of patents granted to a startup within the five years following the grant date of its first patent. You can use patent citations to measure the quality of innovation. The quality of a patent is higher if it receives more citations from patents granted by the patent office within the subsequent five years. You can use the citations to a startup's first patent as measure of its innovative quality.

5. What are the average quantity and quality of startups' innovative output? Visualize the average quantity and quality of startups' innovative output by founding year.
6. Compare the quantity and quality of startup innovation by startups founded between 1995 and 1999 and by startups founded between 2003 and 2007. Are the differences statistically significant?

Most startups are not successful and die sooner or later. A measure of the success of a startup is an "exit" that allows founders, employees, and early investors to sell their shares. Exit events can be Initial Public Offerings (IPOs) or acquisitions by other companies. The Patent Assignment (PA) database allows us to observe patent transactions. For example, *superDimension's* first patent (no. 7,233,820) was reassigned to *Covidien Group* in the year 2012. Assume a startup is acquired by another company if its first patent is subsequently reassigned in a patent transaction classified as "assignment" or "merger" within the ten years following its grant date. For this analysis, consider only startups founded between 1985 and 2010.

7. What is the share of startups that are acquired? Visualize the share of acquired startups by founding year.
8. Estimate linear regression models that explain whether or not a startup is acquired based on the startup's innovation quantity, innovation quality, technology field, US state, and founding year. Explain your findings and interpret the most interesting coefficients in a way that can be understood by a reader who has not read this task-assignment document (i.e., make your explanation self-contained).