Database of all business establishments in US

* Get a sense of demand (labor market)
* Get someone to really understand this
* Name, address, employee number, managers, how many in which job
* Which store closed on which day
* Which store then opened
* Automatically updated collecting available information, aggregate and use
* Min: address, name, type of business
* Uses:
  + Know competitors that laborers might work at also

List of all universities/higher education

* Downloadable format

Housing values, Zillow

* Scrape this
* Rental or owned
* Will help know where to find fitting laborers
* Demographics
* How people in an area voted

Business that exist or are dying – good info for figuring out where to open another store, etc

Ariel is doing bulk of the research, im figuring out how to make use of it

Set up a process, focus on a few data sets and understand how it can be scraped, downloaded, structured to make useful

Pick labor force size and unemployment rate

* Build process that goes to BLS website, grabs data, appends variable, writes to file
* Block of code runs every day
* Grab and append
* Think about structure
* Zip code, effective date of data
* Address is crucial, must be clean, decoded into geocodes, lat long
  + Open source geocoding - data science tool kit
  + Address cleaning and validation
    - No free open source resources
    - Future research project

Start w/ BLS

* Published clv files, api form google?
* Zillow data
* Find out which have web scraping blocks in place
* Pick one variable at a time
* Indeed.com – industry and zip codes
  + How many results come up?
  + 10 mile radius

Basic data structure:

* Postal code – value – time (date) – unemployment rate – location

Use java for scraping

Create excel or csv – in same fixed format

Code will fun and will put csv into a specific folder on server