

Homework 3 — Idea Generation and Feasibility Study

Quantitative Reasoning II (INF 1220)

Due Monday, October 19, 11:59pm, on Populi

This homework assignment is the first step of the research project. Read the project brief before you start.

Task: Pick two topic areas you are interested in. Brainstorm several research questions in each topic area. Investigate whether data exists that can answer at least one of your questions. I will provide feedback on questions and data sources so that you can refine your project before you begin analysis.

1. Deliverable

One-page worksheet (provided on the website), as an edited `.docx` or `.pdf`.

It's okay if one of your chosen topic areas turns into a dead-end. Document it anyway! If both areas are dead-ends, book an appointment with me next week so we can find you a question that works.

2. Guidelines for questions

- **Don't go too narrow:** Your question shouldn't be answerable with a single statistic that requires no judgment (e.g. "how many people moved to Austin last year?")
- **Don't go too broad:** "What causes poor sleep?" is too broad. "What is the relationship between screen time and sleep quality?" is better.
- **You care about the answer:** You'll spend a lot of time on your project, and the next three weeks will be way more fun if you find your topic interesting!

3. Where to find data

Leverage any and all resources to find data for your project! Here are some places to start. If you get stuck, please reach out to me for help. Think about using multiple datasets to answer a question (you might want to combine census data with education data, for example).

Where	What is there
<i>Many topics</i>	
data.gov	the federal government's open data catalog
data.census.gov	American Community Survey (ACS): income, education, employment, and demographics.
data.austintexas.gov	Austin's own open data portal: permits, crime, code complaints, transit, city finances.
ourworldindata.org	Country-level series over long periods, with sources documented.
<i>Education</i>	

Where	What is there
Texas Academic Performance Reports	K-12 school performance data: test results, spending per pupil, teacher experience, student demographics. Pick your categories and it builds a CSV. Choose narrowly, because the full download runs past 2,000 columns, and read the separate data dictionary to find out what the column names mean. Change the year in the url for earlier files.
IPEDS	Every US college: enrollment, admissions, tuition, finances, completion. Covers Texas institutions, which the TAPR files do not.
College Scorecard	Earnings and loan repayment by institution and field of study.
<i>Health</i>	
County Health Rankings	One row per US county, with health outcomes and their usual predictors in the same file, plus a data dictionary. The easiest starting point on this list.
CDC BRFSS	Adult survey on health behavior and conditions, by state and year. CDC's YRBS is the same idea for high school students.
HRSA data warehouse	Where the country is short of doctors, dentists, and mental health providers, by county and tract, as CSV and zip files.
<i>Politics</i>	
MIT Election Lab	Election returns by county and precinct, cleaned and consistent across years.
FEC	Who gave money to which federal campaign, and what the campaign spent it on.
<i>Arts and culture</i>	
Playbill grosses	Weekly Broadway box office by show: gross, potential gross, seats sold, seats in theatre, average and top ticket price, percent of capacity. There is no file to download, but <code>pd.read_html</code> reads the table off the page, and adding <code>?week=2025-12-28</code> to the address returns that week, so you can loop over weeks and build a panel.
IMDb datasets	Titles, ratings, runtimes, genres, and credits, as downloadable files.
Spotify Web API	Track and album features. Needs a free developer account, so budget time for the setup.
<i>Flights</i>	
Bureau of Transportation Statistics	Every domestic flight by a reporting carrier since 1987, one row per flight: scheduled and actual times, delay in minutes split by cause, cancellations, diversions, distance. The zip files here are one month each, named by year and month. This is the source of the Austin file from Homework 1, trimmed from 110 columns to 20.
<i>Housing and local</i>	
Zillow research data	Rent and home value indices by ZIP code, city, and metro, monthly.

Where	What is there
Inside Airbnb	Every short-term rental listing in a city, Austin included, with price and location.
Eviction Lab	Eviction filing and judgment rates by county and census tract.

Guidelines for data sources: you can download the data without a paywall or an untenable rate limit, there is documentation of what the variables are and how the data is collected, and a named organization or individual is accountable for the data. Well compiled datasets on GitHub might be okay, but check the documentation carefully.

4. Feasibility study

Once you've defined your areas of interest and brainstormed some questions, explore some datasets. Actually download and open the data! The best way to open and explore the data is in a python notebook with the pandas package. You are not turning the file or any code in, but you will document your exploration in the worksheet.

Data is often not ready to go straight out of the box. Sometimes you have to clean up the file before you can load it into a dataframe, or you have to import several files and combine them. This is part of the exercise and the learning process! Again, let me know if you find yourself spinning your wheels.