# Dataset + Project Proposal

### BA775

**Due: Nov 7 @ 9:30 AM**

Work with your team to find a dataset that interests all team members. Use any data source below to propose a suitable project. The initial proposed ideas by teammates should be prioritized in your search, but it is acceptable to propose a brand-new project. You may discover well-defined projects and datasets on Kaggle, but I suggest searching outside Kaggle. A cleaned dataset doesn’t allow you to get full credit for the data-cleaning steps of the team assignment. Your team assignment and team project are based on this dataset. You should define a broad project goal. This goal can be finalized after this deadline, but having the first iteration will help you set your project’s tone. You may select one or several datasets as you see fit for the project.

**You must submit your project proposal to me in the team’s Slack channel.**

What to submit:

* **Project title** (<15 words)
* **Problem definition** (<100 words): The project goal should roughly be a brief statement or bullet points.
* **Data sourc**e (the actual source and the link - note that, for instance, Kaggle is not a source itself).
* **Motivation** (<50 words): why are you interested in this problem?

**Requirements:**

* All the team members should agree with the choice of dataset and problem definition
* This cannot be a dataset previously used in BA775 or BA780
* It should have at least 10,000 rows and be smaller than 20 GB
* It should have at least eight columns and a mixture of data types: string/categorical/numeric
* Your dataset should have at least 2 tables so you can practice building relations between the tables and practice joins
* Proposals should be submitted as plain text on the team’s Slack and not as a document. They should follow the same format as the example below

**Example of proposal submission:**

Boston's Income-Restricted Housing Analysis

**Problem Definition:** Evaluate the distribution, accessibility, and effectiveness of income-restricted housing units across Boston. We will address the following:

* Identify areas with the highest concentration of income-restricted housing units.
* Examine the impact of various funding sources on the availability of these units.
* Investigate any potential disparities in the distribution of these units concerning neighborhood demographics.

**Data Source:** The Mayor’s Office of Housing (MOH) - Income-Restricted Housing Inventory. Direct access link [here](https://data.boston.gov/dataset/income-restricted-housing).

**Motivation:** With rising housing costs and affordability issues becoming more prominent, understanding the distribution and impact of income-restricted housing in Boston is essential to ensure equitable housing opportunities for all citizens.

Platforms for Public Data Sources:

* [Google Dataset Search](https://toolbox.google.com/datasetsearch) - Dataset search engine
* [Data.gov](https://data.gov/) - The US government’s open data portal
* [European Union Open Data Portal](https://data.europa.eu/en)
* [World Bank Open Data](https://data.worldbank.org/)
* [Open Data on AWS](https://registry.opendata.aws/) - The Registry of Open Data on AWS
* [BigQuery Datasets](https://console.cloud.google.com/marketplace/browse?filter=solution-type:dataset) - Datasets publicly available on BigQuery and managed by Google
* [BigQuery Datasets](https://www.reddit.com/r/bigquery/wiki/datasets) - Datasets publicly available on BigQuery
* [Google Trends](https://trends.google.com/trends/?geo=US)
* [Harvard Dataverse](https://dataverse.harvard.edu/)
* [UC Irvine Datasets](https://archive.ics.uci.edu/ml/datasets.php)
* [Awesome Public Datasets on GitHub](https://github.com/awesomedata/awesome-public-datasets)
* [538 datasets](https://github.com/fivethirtyeight/data)
* [Buzzfeed open datasets](https://github.com/BuzzFeedNews/everything)
* [Kaggle](https://www.kaggle.com/datasets)