

Phase 3 Project - Choosing a Dataset



<https://github.com/learn-co-curriculum/dsc-phase-3-choosing-a-dataset>



<https://github.com/learn-co-curriculum/dsc-phase-3-choosing-a-dataset/issues/new>

You have the option to either choose a dataset from a curated list or propose your own dataset not on the list. The goal is to choose a dataset appropriate to the type of business problem and/or classification methods that most interests you. **It is up to you to define a stakeholder and business problem appropriate to the dataset you choose.**

If you choose a dataset from the curated list, inform your instructor which dataset you chose and jump right into the project. If you would like to propose your own dataset, run the dataset and business problem by your instructor for approval before starting your project.

Your Get Hired 'Game Plan'



Help set yourself up for success by being strategic about your project/dataset choices.


Already know what your job search focus will be? Consider choosing a dataset that relates to the companies/industries you are interested in and the types of business problems/data they navigate day to day. Doing so demonstrates your subject matter knowledge in their area, significantly elevating your relevance and value as a candidate -- we've seen this strategy WOW companies time and time again!

Still exploring what type of role you would like to get once you graduate? That's okay! Try to focus on a topic or problem that you are interested in and passionate about. Doing so will help you produce a better project overall that you enjoy creating and that you can speak about confidently and naturally.

Coming out of Flatiron School your projects will be listed on your resume and will showcase your specific subject matter knowledge and interest/passions once you're job seeking. Help yourself put your best foot forward and make the strongest first impression possible.

Here are two grads who successfully did just this...

This student was interested in working with government and public sector data, and focused specifically on traffic data and safety. They utilized the Chicago Car Crashes dataset in [one project](https://github.com/jmarkowi/Chicago-Crashes)  then later created a bike lane image dataset from multiple sources for their [capstone project](https://github.com/jmarkowi/NYC_bike_lanes)  https://github.com/jmarkowi/NYC_bike_lanes *. Based on their combination of technical skills and subject-matter expertise, this student landed a government consulting role at **ASR Analytics** where they work to prevent identity theft in tax fraud.

This student ([GitHub link here](https://github.com/kbaranko/NYC-Building-Energy-Intensity/blob/master/README.md)  <https://github.com/kbaranko/NYC-Building-Energy-Intensity/blob/master/README.md> *) focused on working in the clean energy sector, and created

their project *NYC Building Energy Density* using data from the 2016 Energy and Water Data Disclosure for New York City Local Law 84. The student landed a role at **Kevala**, a clean energy software company, in under two months of job seeking.

**Keep in mind that the Flatiron School Data Science program has changed over time, so these projects may or may not reflect the current project requirements. They are intended as inspiration for your dataset/project choice.*

Curated List of Datasets

You may select any of the datasets below - we provide brief descriptions of each. Follow the links to learn more about the dataset and business problems before making a final decision.

If you are feeling overwhelmed or behind, we recommend you choose dataset #1.

1) [SyriaTel Customer Churn](https://www.kaggle.com/becksddef/churn-in-telecoms-dataset)

[\(https://www.kaggle.com/becksddef/churn-in-telecoms-dataset\)](https://www.kaggle.com/becksddef/churn-in-telecoms-dataset)

Build a classifier to predict whether a customer will ("soon") stop doing business with SyriaTel, a telecommunications company. This is a **binary** classification problem.

Most naturally, your audience here would be the telecom business itself, interested in reducing how much money is lost because of customers who don't stick around very long. The question you can ask is: are there any predictable patterns here?

2) [Tanzanian Water Wells](https://www.drivendata.org/competitions/7/pump-it-up-data-mining-the-water-table/page/23/)

[\(https://www.drivendata.org/competitions/7/pump-it-up-data-mining-the-water-table/page/23/\)](https://www.drivendata.org/competitions/7/pump-it-up-data-mining-the-water-table/page/23/)

This dataset is part of an active competition until April 7, 2023!

Tanzania, as a developing country, struggles with providing clean water to its population of over 57,000,000. There are many water points already established in the country, but some are in need of repair while others have failed altogether.

Build a classifier to predict the condition of a water well, using information about the sort of pump, when it was installed, etc. Your audience could be an NGO focused on locating wells needing repair, or the Government of Tanzania looking to find patterns in non-functional wells to influence how new wells are built. Note that this is a **ternary** classification problem by default, but can be engineered to be binary.

3) [H1N1 and Seasonal Flu Vaccines](https://www.drivendata.org/competitions/66/flu-shot-)

[\(https://www.drivendata.org/competitions/66/flu-shot-\)](https://www.drivendata.org/competitions/66/flu-shot-)

learning/).

This dataset is part of an active competition until March 31, 2022!

As the world struggles to vaccinate the global population against COVID-19, an understanding of how people's backgrounds, opinions, and health behaviors are related to their personal vaccination patterns can provide guidance for future public health efforts. Your audience could be someone guiding those public health efforts.

This challenge: can you predict whether people got H1N1 and seasonal flu vaccines using data collected in the National 2009 H1N1 Flu Survey? This is a **binary** classification problem, but there are two potential targets: whether the survey respondent received the seasonal flu vaccine, or whether the respondent received the H1N1 flu vaccine. Please choose just one of these potential targets for your minimum viable project.

4) Chicago Car Crashes ➞

(<https://data.cityofchicago.org/Transportation/Traffic-Crashes-Crashes/85ca-t3if>)

Note this links also to Vehicle Data ➞ (<https://data.cityofchicago.org/Transportation/Traffic-Crashes-Vehicles/68nd-jvt3>) and to Driver/Passenger Data ➞ (<https://data.cityofchicago.org/Transportation/Traffic-Crashes-People/u6pd-qa9d>)

Build a classifier to predict the primary contributory cause of a car accident, given information about the car, the people in the car, the road conditions etc. You might imagine your audience as a Vehicle Safety Board who's interested in reducing traffic accidents, or as the City of Chicago who's interested in becoming aware of any interesting patterns.

This is a **multi-class** classification problem. You will almost certainly want to bin, trim or otherwise limit the number of target categories on which you ultimately predict. Note that some primary contributory causes have very few samples, for example.

5) Terry Traffic Stops ➞ (<https://data.seattle.gov/Public-Safety/Terry-Stops/28ny-9ts8>)

In Terry v. Ohio ➞ (<https://www.oyez.org/cases/1967/67>), a landmark Supreme Court case in 1967-8, the court found that a police officer was not in violation of the "unreasonable search and seizure" clause of the Fourth Amendment, even though he stopped and frisked a couple of suspects only because their behavior was suspicious. Thus was born the notion of "reasonable suspicion", according to which an agent of the police may e.g. temporarily detain a person, even in the absence of clearer evidence that would be required for full-blown arrests etc. Terry Stops are stops made of suspicious drivers.

Build a classifier to predict whether an arrest was made after a Terry Stop, given information about the presence of weapons, the time of day of the call, etc. This is a binary classification problem.

Note that this dataset also includes information about gender and race. You may use this data as well. You could conceivably pitch your project as an inquiry into whether race (of officer or of subject) plays a role in whether or not an arrest is made.

If you do elect to make use of race or gender data, be aware that this can make your project a highly sensitive one; your discretion will be important, as well as your transparency about how you use the data and the ethical issues surrounding it.

Proposing Your Own Dataset

Sourcing new data is a valuable skill for data scientists, but it requires a great deal of care. An inappropriate dataset or an unclear business problem can lead you to spend a lot of time on a project that delivers underwhelming results. The guidelines below will help you complete a project that demonstrates your ability to engage in the full data science process.

Once you've sourced your own dataset and identified the business problem you want to solve with it, **you must run them by your instructor for approval.**

Data Guidelines

Your dataset must be:

1. **Appropriate for classification.** It should have a categorical outcome or the data needed to engineer one.
2. **Usable to solve a specific business problem.** This solution must rely on your classification model.
3. **Somewhat complex.** It should contain a minimum of 1000 rows and 10 features.
4. **Unfamiliar.** It can't be one we've already worked with during the course or that is commonly used for demonstration purposes (e.g. Titanic).
5. **Manageable.** Stick to datasets that you can model using the techniques introduced in Phase 3.

Problem First, or Data First?

There are two ways that you can source your own dataset: **Problem First** or **Data First**. The less time you have to complete the project, the more strongly we recommend a Data First approach to this project.

Problem First: Start with a problem that you are interested in that you could potentially solve with a classification model. Then look for data that you could use to solve that problem. This approach is high-risk, high-reward: Very rewarding if you are able to solve a problem you are invested in, but frustrating if you end up sinking lots of time in without finding appropriate data. To mitigate the risk,

set a firm limit for the amount of time you will allow yourself to look for data before moving on to the Data First approach.

Data First: Take a look at some of the most popular internet repositories of cool data sets we've listed below. If you find a data set that's particularly interesting for you, then it's totally okay to build your problem around that data set.

Potential Data Sources

There are plenty of amazing places that you can get your data from. We recommend you start looking at data sets in some of these resources first:

- [UCI Machine Learning Datasets Repository](https://archive.ics.uci.edu/ml/datasets.php) ➞ [_](https://archive.ics.uci.edu/ml/datasets.php)(https://archive.ics.uci.edu/ml/datasets.php)
- [Kaggle Datasets](https://www.kaggle.com/datasets) ➞ [_](https://www.kaggle.com/datasets)(https://www.kaggle.com/datasets)
- [Awesome Datasets Repo on Github](https://github.com/awesomedata/awesome-public-datasets) ➞ [_](https://github.com/awesomedata/awesome-public-datasets)(https://github.com/awesomedata/awesome-public-datasets)
- Local data portals for state and local government resources
 - Examples: [NYC](https://opendata.cityofnewyork.us/) ➞ [_](https://opendata.cityofnewyork.us/)(https://opendata.cityofnewyork.us/), [Houston](http://data.houstontx.gov/) ➞ [_](http://data.houstontx.gov/)(http://data.houstontx.gov/), [Seattle](https://data.seattle.gov/) ➞ [_](https://data.seattle.gov/)(https://data.seattle.gov/), [California](https://data.ca.gov/) ➞ [_](https://data.ca.gov/)(https://data.ca.gov/)
- [Inside AirBNB](http://insideairbnb.com/) ➞ [_](http://insideairbnb.com/)(http://insideairbnb.com/)
- [FiveThirtyEight's data portal](https://data.fivethirtyeight.com/) ➞ [_](https://data.fivethirtyeight.com/)(https://data.fivethirtyeight.com/)
- [Data is Plural's Archive Spreadsheet](https://docs.google.com/spreadsheets/d/1wZhPLMCHKJvwOkP4juchhjEgqIY8fQFMemwKL2c64vk/edit#gid=0) ➞ [_](https://docs.google.com/spreadsheets/d/1wZhPLMCHKJvwOkP4juchhjEgqIY8fQFMemwKL2c64vk/edit#gid=0)(https://docs.google.com/spreadsheets/d/1wZhPLMCHKJvwOkP4juchhjEgqIY8fQFMemwKL2c64vk/edit#gid=0)
- [Datasets Subreddit](https://www.reddit.com/r/datasets/) ➞ [_](https://www.reddit.com/r/datasets/)(https://www.reddit.com/r/datasets/)