Communicate Data Findings Project

Dataset Options

To complete this project, select one of the datasets in the table, or you can select your own dataset. For guidelines on choosing your own dataset, see below the table.

Dataset	Overview and Notes	Example Topics/Questions
Ford GoBike System Data	This data set includes information about individual rides made in a bike-sharing system covering the greater San Francisco Bay area. Note that this dataset *will* require some data wrangling in order to make it tidy for analysis. There are multiple cities covered by the linked system, and multiple data files will need to be joined together if a full year's coverage is desired. If you're feeling adventurous, try adding in analysis from other cities, following links from [this page].	When are most trips taken in terms of time of day, day of the week, or month of the year? How long does the average trip take? Does the above depend on if a user is a subscriber or customer?
Flights	This dataset reports flights in the United States, including carriers, arrival and departure delays, and reasons for delays, from 1987 to 2008. You may want to try downloading multiple years worth of data and joining them, to do a year-by-year comparison. Not all features will be interesting for performing your exploration. Note that the linked page points towards another page with more detailed variable descriptions in the original, full data.	Are there certain destination or arrival cities that are home to more delays or cancellations? What are the preferred times for flights to occur? Are there any changes over multiple years?

Loan Data from Prosper	This data set contains 113,937 loans with 81 variables on each loan, including loan amount, borrower rate (or interest rate), current loan status, borrower income, and many others. This data dictionary explains the variables in the data set. You are not expected to explore all of the variables in the dataset! Focus your exploration on about 10-15 of them.	What factors affect a loan’s outcome status? What affects the borrower’s APR or interest rate? Are there differences between loans depending on how large the original loan amount was?
PISA Data Note: The unzipped PISA Data csv file is 2.75 GB.	PISA is a survey of students' skills and knowledge as they approach the end of compulsory education. It is not a conventional school test. Rather than examining how well students have learned the school curriculum, it looks at how well prepared they are for life beyond school. Around 510,000 students in 65 economies took part in the PISA 2012 assessment of reading, mathematics and science representing about 28 million 15-year-olds globally. Of those economies, 44 took part in an assessment of creative problem solving and 18 in an assessment of financial literacy. **PISA Data Dictionary** The data and topics of investigation come from the PISA Data Visualization Competition. If you want to know more about the survey design, the details can be found in the technical report here.	How does the choice of school play into academic performance? Are there differences in achievement based on gender, location, or student attitudes? Are there differences in achievement based on teacher practices and attitudes? Does there exist inequality in academic achievement?
Or select your own dataset!	See below for guidelines on whether or not a dataset will be appropriate for use in this project. **Remember that finding**	

	and cleaning your own data set could take significant time and effort!	

If you’re finding your own dataset…

Your dataset should:

include at least 600 observations. (This is the number of rows after tidying your data - see the bullet point below about tidy data.)

include at least eight variables.

include at least one qualitative / categorical variable. (This can also be engineered / created.)

include at least one numeric variable.

be in a tidy format. In a nutshell, tidy data has each row as a single observation and each column reporting a single variable. You can read more about tidy data in Hadley Wickham’s paper [here]. You may need to do some cleaning and reshaping to tidy your dataset, before you actually get started with your exploration.

be in a common data format. This includes .csv, .tsv, .txt, and .xls. Basically, there should be a reasonable pandas.read_*() function to open up your data in a tidy format as a pandas DataFrame.

Here are some resources to help you find a dataset:

http://www.data.gov/

http://databank.worldbank.org/data/home.aspx

https://github.com/awesomedata/awesome-public-datasets

http://www.pewglobal.org/category/datasets/ (requires account registration)

https://www.kaggle.com/datasets (requires account registration)

https://toolbox.google.com/datasetsearch

https://www.quora.com/Where-can-I-find-large-datasets-open-to-the-public

https://dreamtolearn.com/ryan/1001_datasets

https://mpopov.com/blog/advice-for-grads-entering-industry-datasci - Also includes tips for breaking into the data scene!