# **Assignment 1: Exploratory Data Analysis**

In this assignment, you will identify a dataset of interest and perform an exploratory analysis to better understand the shape & structure of the data, investigate initial questions, and develop preliminary insights & hypotheses. Your final submission will take the form of a report consisting of captioned visualizations that convey key insights gained during your analysis.

## Step 1: Data Selection

First, you will pick a topic area of interest to you and find a dataset that can provide insights into that topic. To streamline the assignment, we've pre-selected a number of datasets included below for you to choose from.

However, if you would like to investigate a different topic and dataset, you are free to do so. If working with a self-selected dataset, please check with the course staff to ensure it is appropriate for the course. Be advised that data collection and preparation (also known as data wrangling) can be a very tedious and time-consuming process. Be sure you have sufficient time to conduct exploratory analysis, after preparing the data.

After selecting a topic and dataset – but prior to analysis – you should write down an initial set of **at least three questions** you'd like to investigate.

## Part 2: Exploratory Visual Analysis

Next, you will perform an exploratory analysis of your dataset using a visualization tool such as Tableau. You should consider two different phases of exploration.

In the first phase, you should seek to gain an overview of the shape & structure of your dataset. What variables does the dataset contain? How are they distributed? Are there any notable data quality issues? Are there any surprising relationships among the variables? Be sure to also perform "sanity checks" for patterns you expect to see!

In the second phase, you should investigate your initial questions, as well as any new questions that arise during your exploration. For each question, start by creating a visualization that might provide a useful answer. Then refine the visualization (by adding additional variables, changing sorting or axis scales, filtering or subsetting data, etc.) to develop better perspectives, explore unexpected observations, or sanity check your assumptions. You should repeat this process for each of your questions, but feel free to revise your questions or branch off to explore new questions if the data warrants.

## Final Deliverable

Your final submission should take the form of a report – similar to a slide show– that consists of **10 or more** captioned visualizations detailing your most important insights. Your "insights" can include important surprises or issues (such as data quality problems affecting your analysis) as well as responses to your analysis questions. To help you gauge the scope of this assignment, see [this example report analyzing data about motion pictures](https://drive.google.com/file/d/1oXkztG4oMiFSZ-dfbnDQr0Sp4vntMgNp/view?usp=sharing).

Each visualization image should be a screenshot exported from a visualization tool, accompanied with a title and descriptive caption (1-4 sentences long) describing the insight(s) learned from that view. Provide sufficient detail for each caption such that anyone could read through your report and understand what you've learned. You are free, but not required, to annotate your images to draw attention to specific features of the data. You may perform highlighting within the visualization tool itself, or draw annotations on the exported image. To easily export images from Tableau, use the Tablue public to save your work and share its link in your submission.

You can select any visualisation tool to perform your analysis. Another option is to download Tablue for students version by following the instructions in this [link](https://www.tableau.com/academic/students#form).

\*\*\* You need to make sure that I have access to your visualisation in public repository. Sharing the workbook is not sufficient.

## Recommended Data Sources

To get up and running quickly with this assignment, we recommend exploring one of the following provided datasets.

### Campaign Finance Data for the 2017-18 Congressional Cycle

Although Congressional elections won't be held until November 2018, political fundraising and spending is already underway. The [Federal Election Commission (FEC)](https://www.fec.gov/)maintains a database of the financial activity of Political Action Committees (PACs), including funds raised, money disbursed to committees & candidates, and operating expenditures. See [the FEC site for a full listing of available data sets](http://classic.fec.gov/finance/disclosure/ftpdet.shtml).

Here, we provide a combined data file for operating expenditures so far in the 2017-18 congressional cycle: where committees and candidates are spending their money. The provided CSV was created by joining together three tables: [the committee master file](http://classic.fec.gov/finance/disclosure/metadata/DataDictionaryCommitteeMaster.shtml), [the candidate master file](http://classic.fec.gov/finance/disclosure/metadata/DataDictionaryCandidateMaster.shtml), and [operating expenditures](http://classic.fec.gov/finance/disclosure/metadata/DataDictionaryOperatingExpenditures.shtml). Follow the per-table links for more details about the columns each table contains. In the provided file, we added a prefix to each column to indicate which table it comes from (CMTE, CAND, or EXP).

Note that this data set is large, with both many rows and many columns. It is sufficiently rich to support a variety of explorations. If you choose to explore this data set, we strongly encourage you to pick a focus area. Examples include: (1) How do presidential campaigns continue to spend money after the general election? (2) How are the primary Democratic and Republican National Committees spending funds? (3) What are Democratic Party presidential hopefuls doing? (4) What expenditures are occurring across Washington State, and by whom? (5) To what degree is money spent back in politician's home states? To what extent does it go to Beltway consultants? Focusing on any one of these topics should be sufficient for this assignment.

Data: [oppexp.csv.gz (gzipped CSV)](https://courses.cs.washington.edu/courses/cse442/17au/data/oppexp.csv.gz)

### Chicago Crimes, 2001-Present

This dataset reflects reported incidents of crime (with the exception of murders where data exists for each victim) that occurred in the City of Chicago from 2001 to present, minus the most recent seven days. Data is extracted from the Chicago Police Department's CLEAR (Citizen Law Enforcement Analysis and Reporting) system.

Data: [Chicago Access Page (click Export to download a CSV file)](https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present-Dashboard/5cd6-ry5g)

### Daily Weather in the U.S., 2017

This dataset contains daily U.S. weather measurements in 2017, provided by the [NOAA Daily Global Historical Climatology Network](ftp://ftp.ncdc.noaa.gov/pub/data/ghcn/daily/). This data has been transformed: some weather stations with only sparse measurements have been filtered out. See the accompanying [weather.txt for descriptions of each column](https://courses.cs.washington.edu/courses/cse442/17au/data/weather.txt).

Data: [weather.csv.gz (gzipped CSV)](https://courses.cs.washington.edu/courses/cse442/17au/data/weather.csv.gz)

### Yelp Open Dataset

This dataset provides information about businesses, user reviews, and more from Yelp's database. The data is split into separate files (business, checkin, photos, review, tip, and user), and is available in either JSON or SQL format. You might use this to investigate the distributions of scores on yelp, look at how many reviews users typically leave, or look for regional trends about restaurants. Note that this is a large, structured dataset and you don't need to look at all of the data to answer interesting questions.

In order to download the data you will need to enter your email and agree to [Yelp's Dataset License](https://s3-media2.fl.yelpcdn.com/assets/srv0/engineering_pages/e926cc12796d/assets/vendor/yelp-dataset-license.pdf).

Data: [Yelp Access Page (data available in JSON & SQL formats)](https://www.yelp.com/dataset)

## Additional Data Sources

If you want to investigate datasets other than those recommended above, here are some possible sources to consider. You are also free to use data from a source different from those included here. If you have any questions on whether your dataset is appropriate, please ask the course staff ASAP!

* [data.seattle.gov](https://data.seattle.gov/) - City of Seattle Open Data
* [data.wa.gov](https://data.wa.gov/) - State of Washington Open Data
* [nwdata.org](https://nwdata.org/) - Open Data & Civic Tech Resources for the Pacific Northwest
* [data.gov](https://www.data.gov/) - U.S. Government Open Datasets
* [U.S. Census Bureau](https://www.census.gov/data.html) - Census Datasets
* [IPUMS.org](https://www.ipums.org/) - Integrated Census & Survey Data from around the World
* [Federal Elections Commission](http://www.fec.gov/finance/disclosure/ftpdet.shtml) - Campaign Finance & Expenditures
* [Federal Aviation Administration](https://www.faa.gov/data_research/) - FAA Data & Research
* [fivethirtyeight.com](https://github.com/fivethirtyeight/data/) - Data and Code behind the Stories and Interactives
* [Buzzfeed News](https://github.com/BuzzFeedNews)
* [Socrata Open Data](https://opendata.socrata.com/)
* [17 places to find datasets for data science projects](https://www.dataquest.io/blog/free-datasets-for-projects/)

## Visualization Tools

You are free to use one or more visualization tools in this assignment. However, in the interest of time and for a friendlier learning curve, **we strongly encourage you to use Tableau**. Tableau provides a graphical interface focused on the task of visual data exploration. You will (with rare exceptions) be able to complete an initial data exploration more quickly and comprehensively than with a programming-based tool.

* [Tableau - Desktop visual analysis software](https://www.tableau.com/academic/students). Available for both Windows and MacOS; register for a free student license.
* [Voyager - Research prototype from the UW Interactive Data Lab](https://vega.github.io/voyager2/). Voyager combines a Tableau-style interface with visualization recommendations. Use at your own risk!
* [R](https://www.r-project.org/), using the [ggplot2](http://ggplot2.org/) library or with R's built-in plotting functions.
* [Jupyter Notebooks (Python)](http://jupyter.org/), using libraries such as [Altair](https://altair-viz.github.io/) or [Matplotlib](http://matplotlib.org/).

## Data Wrangling Tools

The data you choose may require reformatting, transformation or cleaning prior to visualization. Here are tools you can use for data preparation. We recommend first trying to import and process your data in the same tool you intend to use for visualization. If that fails, pick the most appropriate option among the tools below. Contact the course staff if you are unsure what might be the best option for your data!

#### Graphical Tools

* [Tableau](https://www.tableau.com/academic/students) - Tableau provides basic facilities for data import, transformation & blending.
* [Trifacta Wrangler](https://www.trifacta.com/start-wrangling/) - Interactive tool for data transformation & visual profiling.
* [OpenRefine](http://openrefine.org/) - A free, open source tool for working with messy data.

#### Programming Tools

* [JavaScript data utilities](https://bocoup.com/work/learn-js-data) and/or the [Datalib JS library](https://github.com/vega/datalib).
* [Pandas](http://pandas.pydata.org/) - Data table and manipulation utilites for Python.
* [dplyr](https://cran.r-project.org/web/packages/dplyr/vignettes/introduction.html) - A library for data manipulation in R.
* Or, the programming language and tools of your choice...

## Grading Criteria

Each submission will be graded based on both the analysis process and included visualizations. Here are our grading criteria:

* Poses clear questions applicable to the chosen dataset.
* Appropriate data quality assessment and transformation.
* Sufficient breadth of analysis, exploring multiple questions.
* Sufficient depth of analysis, with appropriate follow-up questions.
* Expressive & effective visualizations crafted to investigate analysis questions.
* Clearly written, understandable captions that communicate primary insights.

## Submission Details

This is an individual assignment. You may **not** work in groups.