# SDS 322E Project
## Part 1: Exploratory Data Analysis

### Overview

The research project is the culminating experience of the course: you will apply what you have learned in class to explore real data! There will be two main parts for your project:

1. Exploratory Data Analysis: explore the dataset, investigate any surprising values, recode some variables, and create appropriate visualizations and summary statistics.
2. Classification and Prediction: consider relationships with a model to classify/predict an outcome (numeric or categorical) based on some predictors.

For each part of the project, you will fill a quick check in survey on *Canvas* and, once your check-in is approved, you will work on:

- ✓ A narrative report to tell the story of your data analysis (a knitted R Markdown file),
- ✓ A short presentation to share your findings with the rest of the class (recorded video),
- ✓ 3 peer reviews to provide feedback to your peers and learn about different projects!

If you would like to explore some examples of stats/data science projects, you can refer to the Winning Projects for the USCLAP competition organized by the Consortium for the Advancement of Undergraduate Statistics Education. I highly encourage you to participate to this competition by the way, they have cash prizes! *Be aware that these are winning projects and represent the best submissions nationwide. While my expectations for this course are not for you to win, it would be pretty awesome if you do!*

### Working independently or in a team

For the project, you can work independently or in a team of 3 students maximum.
*If working in a team, you will be asked to do a few more things (see instructions below) but your team will produce one unique report/presentation. You will also assess each other's contribution to the project in the report.*

### Topic and datasets

We will focus on a few datasets accessible through the City of Austin Open Data Portal. Click on the following links to learn more about the datasets:

- Austin Animal Center Intakes and Outcomes
- Food Establishment Inspection Scores
- Imagine Austin Indicators
- Austin 311 Public Data
- Real-Time Traffic Incident Reports
- Austin MetroBike Trips
- Issued Construction Permits
- Residential Average Monthly kWh and Bills
- Water Quality Sampling Data
- Crime Reports

*If you would absolutely love to work on a different dataset, please reach out to let us know.*

Some of these datasets will be +/- **tidy**, +/- **messy**. You can also **join** some datasets together if it makes sense. For example, you can keep track of an animal going through the Austin Animal Center Intakes and Outcomes by the Animal ID. Some of these datasets are also huge so you can decide to only focus on some parts (select few variables, filter some categories, take a random sample) but make sure that you have the following in the resulting/final dataset:

- ✓ at least **50 observations** (i.e., rows)
- ✓ at least **4 variables** but *if a team project, you need at least 6 variables*.
- ✓ at least **1 outcome variable**. Note: some datasets might not have an obvious outcome variable but you can create one (e.g., in the Austin 311 Public Data, we are given the date when the service request was created and when it was closed so we calculate the number of days it took to close the request).
- ✓ at least **1 numeric variable**. Note: some datasets might not have a numeric variable but you can create one (e.g., same example with the Austin 311 Public Data, calculate the number of days it took to close the request).
- ✓ at least **1 categorical variable**. Note: most datasets have at least one categorical variable but you can still create one (e.g., split a numeric value into "small" vs "large" values).

There are no restrictions for the maximum number of datasets, observations, variables, but just be aware that R might run a little slow if you have millions!

## Report

The text of your report will provide a narrative structure around your code and outputs with *R Markdown*. Answers without supporting code will not receive credit and outputs without comments will not receive credit either: write full sentences to describe your findings. All code contained in your final project document must work correctly (knit early, knit often)! *If a team project, only submit one report per team.*

Detailed guidelines for the report:

1. **Title and Introduction.** Describe the dataset(s) you have chosen and why you are interested in this data. Cite a previous study or some article that gives you a little more background about the context of this dataset. Refer to the main variables of interests and mention what a unique row represents. What trends or relationships do you expect? What kind of research questions will your exploratory data analysis answer? *If a team project, each member suggests 1 question.*

2. **Methods.** Document any steps taken to get the dataset in shape for data analysis: describe any cleaning/wrangling you had to do (and show your code), how many rows/columns you started with, how many rows/columns you ended up with (could be the same). Quickly justify if your dataset(s) is tidy. For example, you would include in this section:
    - Steps to reshape your dataset(s) if the datasets are not tidy so that every observation has its own row and every variable its own column.
    - Steps to join several datasets and combine them in a single dataset. Make sure to report how many observations/rows were dropped/added when joining the datasets and discuss any potential issues.

3. **Results**. Explore your data with visualizations and summary statistics. Include at least 2 visualizations representing univariate distributions reporting the appropriate statistics, and at least 2 visualizations representing relationships between 2 variables reporting the appropriate

statistics. You could do some visualizations with 3 variables if you'd like. Make sure to include labels. Write sentences interpreting each visualization with any trend/relationships and referring to appropriate statistics (remember to include units!). It might be a good idea to number your visualizations to reference them in the next section. *If a team project, each member creates 1 univariate and 1 bivariate visualization, with the appropriate statistics.*

4. **Discussion**. Putting it all together, what did you learn from your data?
   - Answer your research question(s) in context, citing important statistics and visualizations.
   - Did the data match what you expected? Anything you are curious about?
   - Consider ethical issues (from data collection to interpretation): what impacts/implications could your results have on the community?
   - If you were going to share your finding with the City of Austin (and you might actually do that!), what would be the main takeaway from your exploratory data analysis? Anything they should know about the state of the dataset (e.g., inconsistency in categories, typos, …)?

5. **Reflection, acknowledgements, and references.** Reflect on the process of conducting this project. What was challenging, what have you learned from the process itself? Include acknowledgements for any help received: who helped you with the project? Thank TAs, instructors, data owners… *If a team project, that is where you report the contribution of each member: who did what)*. Include references: dataset link(s), a citation for background context).

6. **Formatting**. Create the report using R Markdown, with headers for each section, and comments to the R code. The final report should be no more than 20 pages (the number of pages can vary greatly depending on the cleaning process though). It is extremely important that you **select pages** to corresponding sections when submitting on *Gradescope*. *If a team project, identify all team members on the submission.*

## Presentation

You will describe the main steps taken to complete your first project and share some key findings of your exploratory data analysis. *If a team project, only submit one presentation per team.*

Detailed guidelines for the presentation:

- Make slides to present and follow the structure of the report. You do not need to submit your slides. Record yourself while sharing your screen. *If a team project, you can record the video in the same room or through a Zoom meeting for example.*
- During the presentation, introduce yourself, share the title, motivations for choosing this dataset(s), and the steps taken before analysis. Discuss your research question(s) including relevant statistics and visualizations. *If a team project, each member answers their own research question.*
- The presentation should not last more than 5 minutes. *If a team project, try to split the time equally between each member.* Your video does not have to be high quality, but we should be able to hear you and see your presentation materials clearly throughout the entire video.

Detailed guidelines for the peer reviews:

- Each student will peer review 3 videos from other classmates/teams (it also means that 3 other students will watch your video). Peer reviews will be assigned the day after the presentation is due and will be due within a week the presentation was due.

- You will use the rubric (see below), evaluating the Materials, Presence, Quality for the video presentation. Write individual and constructive feedback:
  - ✓ Be specific in your comments: refer to a specific phrase, visualization, or slide.
  - ✓ Offer a solution to improve the work: mention possible revisions or links to helpful resources or examples.
  - ✓ Be kind: all your comments should be framed in a supportive way.

  Consider the following:
  - × If you leave the same comment to all 3 of your peer reviews, you won't receive full credit.
  - × If you assign a grade but don't leave any comments, you'll receive some credit.
  - × If you leave comments but don't assign a grade, you'll receive half credit.

## Grading Rubric

The first part of your project will be assessed as follows:

| Rubric Item | Description | Points |
|---|---|---|
| **Check In** | Decide if you are completing the project independently or with a team. Share your choice of dataset(s), the ID variable, and some variables of interest | **10** |
| **Report** | **Title/Introduction**: present the context of the dataset(s) and share the research question(s) your exploratory data analysis will aim to answer. | **15** |
| | **Methods**: document the process to get the dataset in shape for data analysis. | **20** |
| | **Results**: create visualizations and summary statistics to investigate your dataset. | **30** |
| | **Discussion**: answer your research question(s) in light of your results and ethical considerations. | **20** |
| | **Reflection, acknowledgements, and references**: reflect on what you have learned and refer to any external resources. | **10** |
| | **Formatting**: well organized and easy to navigate. | **5** |
| **Presentation** | **Materials**: All required elements were included. Texts and figures are easy to read, free of spelling errors. The information is presented in a logical structure. | **10** |
| | **Presence**: Presenter(s) is audible, introduces themselves, shows enthusiasm for the topic and seems prepared. *If a team project, the time is (approximately) split equally among the team members.* | **10** |
| | **Quality**: At the end of the presentation, the audience should have a good understanding of how the results can be interpreted in context. The visualizations and statistics were appropriate and conclusions are not overreaching. | **10** |
| | Complete 3 peer reviews using the rubric and leaving constructive feedback. | **10** |
| | **Total** | **150** |