Overview

In this assignment, you will analyze criminal incident data from Seattle or San Francisco to visualize patterns and, if desired, contrast and compare patterns across the two cities.

You will produce a blog-post-style visual narrative consisting of a series of visualizations interspersed with sufficient descriptive text to make a convincing argument.

The assignment will be assessed by peer review. The rubric for assessment will include questions about the effectiveness and clarity of your argument, your use of visualization, and the completeness of your analysis. Reproducibility will also be considered, but will be evaluated subjectively -- peer reviewers will not be asked to recreate your results.

Project Ideas

You may want to consider one or more of the following types of questions when developing your submission.

- For either city, how do incidents vary by time of day? Which incidents are most common in the evening? During what periods of the day are robberies most common?
- For either city, how do incidents vary by neighborhood? Which incidents are most common in the city center? In what areas or neighborhoods are robberies or thefts most common?
- For either city, how do incidents vary month to month in the Summer 2014 dataset?
- For either city, which incident types tend to correlate with each other on a day-by-day basis?
- **Advanced** What can we infer broadly about the differences in crime patterns between Seattle and San Francisco? Does one city tend to have more crime than the other, per capita? Do the relative frequencies of types of incidents change materially between the two cities? (NOTE: The two datasets do not have the same schema, so comparisons will require some work and some assumptions. This will require extra work, but you will be working at the forefront of what is known!)
- **Advanced** For either city, do certain crimes correlate with environmental factors such as temperature? (To answer this kind of question, you will need to identify and use external data sources!)

Data

You will use real crime data from Summer 2014 one or both of two US cities: Seattle and/or San Francisco.

These reduced datasets are available on the course github repository:

https://github.com/uwescience/datasci_course_materials/tree/master/assignment6

Seattle Data

Seattle Summer 2014 dataset used in this assignment

Other Seattle data (not required; for information only):

Seattle Data Portal

Full Seattle incident dataset

San Francisco Data

San Francisco Summer 2014 dataset used in this assignment

Other San Francisco Data (not required; for information only):

San Francisco Data Portal

Full San Francisco incident dataset

All datasets are provided through their respective cities data portals, all powered by Socrata. The three portals and the links to the original datasets are

Integration You are not required to use both datasets (though inter-city comparisons of crime data is an important topic -- you will be working at the forefront of research in the area!) If you choose to use both datasets, note that these datasets do NOT agree on schema, and they do NOT agree on categories or descriptions for specific crimes. Real data is dirty! To draw comparisons and contrasts across cities, you will need to make some assumptions about correspondences. For example, LARCENY/THEFT is a category in San Francisco, but Seattle uses codes of the form THEFT-CARPROWL or THEFT-BUILDING.

So you will need to make some reasonable assumptions to compare these data, and explain and justify those assumptions!

Scale Since the goal is to focus on visualization and communication rather than algorithms and scale and many visualization tools struggle with even modestly large datasets, we have restricted all three datasets to the period Summer 2014. You are permitted to expand your analysis to the full datasets, which we also provide below.

Note that the full Chicago dataset covering 2001 to present has 5M records and is about 1.3GB as a csv file -- not a terribly large dataset, but enough to cause popular desktop tools and javascript libraries to struggle.

Review and Grading

You will be asked to review FOUR submissions.

There are ten questions in the rubric, each with a scale of 0 to 4 points. In all cases, the first option, worth 0 points, is intended to be used only when the assignment is sufficiently broken as to make it impossible to review. The highest option, worth 4 points, is intended to be used when the assignment is exceptional and/or exemplary. The fourth option, worth 3 points, is intended to be used for quality submissions when there are no major problems. Scoring a 3 on every question means you have a very good submission; you are not expected to get a score of 4 for every question!

A passing grade is 60%, which corresponds to achieving a score of 3 on four questions and a score of 2 on the remaining six. This grading is not especially strict, but do not get complacent; make sure your submission is coherent and readable, you should score well.

The peer review itself is an important part of this assignment: Critique of other people's visualization is an important skill, as discussed in the lectures.

One question pertains to reproducibility. You are not required to provide all your code, but if you have performed convoluted transformations of the data or used complicated visualizations, you had better explain them. If you use Jupyter notebooks, review are asked to automatically give you a 4/4 for this question.

What to Turn In

You will submit a url that must resolve to your submission.

Your submission should consist of a primary finding, backed by 2-3 supporting visualizations.

Your primary finding should be prominently displayed and obvious at the top of your submission, e.g., "Violent Crime Increases at Night During Summer Months in San Francisco" or "Bike thefts are most common in the University District of Seattle"

This primary finding should be supported by evidence displayed in 2-3 visualizations. For example, your first visualization may be a timeseries of crime incidents by time of day, illustrating that crime increases at night. Your second visualization may be a histogram of specific crimes during peak hours. Your third visualization could show how the night-time pattern changes from month to month in the summer.

Each visualization should be equipped with supporting text explaining the conclusion to be drawn -- don't just "dump data" on your reviewer -- explain the important takeaways. "From the figure, we see that most armed robberies occur between 11 pm and 3 am."

Read on for details about your submission:

Format

Your submission may either be a pdf file or a rendered html page. You must not assume that your peer reviewers will have access to any other technology to read or access your submission (e.g., microsoft office).

Accessibility

It is your responsibility to ensure that the url is publicly accessible and resolves and renders properly for typical browsers. You must not require your peer reviewers to login to any system or otherwise take steps to access your submission.

Hosting

We recommend the use of github for hosting. If you create a Jupyter notebook, it will be automatically rendered by github, and you need not take any additional steps. If you create a pdf, simlpy commit the pdf to a public repository and submit the url. Make sure to submit the url to the *specific* file representing your submission, not to the overall repository. If your peer reviewers have to "guess" where your submission can be found, they may take points off. There are many materials online to help you create a github repository, including the github documentation itself (https://help.github.com/articles/create-a-repo/)..)

Tools

You are free to use whatever tool you wish.

The data may require programmatic manipulation before visualization, depending on the nature of the question you are exploring and the functionality of the tool you select.

You may consider programmatic tools such as Python with Bokeh or matplotlib, R with ggplot2, or javascript with D3 or vegalite.

You may consider desktop tools such as Tableau (students can often get a free license). (We don't recommend Excel, but you can use it.)

You may consider interactive online tools such as Google Fusion Tables.

Regardless of which tool you use, it is your responsibility to ensure that your submission is posted somewhere accessible for peer review!

Jupyter Notebooks

We encourage, but do not require, the use of Jupyter to create an online notebook as your submission. Several tutorials and videos exist online. Jupyter has several benefits for a project of this type:

1. The notebook can be hosted for free on github or nbviewer

2. Reproducibility will be automatically achieved, because all the code you use will be available as part of the notebook. Your peer reviewers will be instructed to give you full points for reproducibility.
3. Visualizations, text descriptions, and code snippets can all be used interchangeably to explain your findings.
4. Jupyter supports both Python and R, as well as other languages.

All this said, we cannot provide a custom tutorial for how to use Jupyter, so you should use whatever technology with which you are most comfortable.

Collaboration

You are encouraged to discuss the project on the discussion boards to generate ideas and solve problems, but each student must turn in their own work. We will not accept duplicate submissions.

Examples to Consider when Scoping your Submission

These examples use data from the Pronto Data Challenge.

- For illustration purposes, this project exhibits *much more* than you are expected to do for this assignment, but is otherwise a very nice example. This project answer many questions instead of just identifying one key finding, and makes extensive use of machine learning. You are only expected to explore one question and come up 2-3 supporting visualizations.
- This project on the Pronto dataset uses a literate programming style -- text inteleaved with code samples and visualizations. This is a reasonable approach to achieving reproducibility, but clarity is the priority for this assignment: make sure the code does not get in the way of communicating your primary finding. Also, this project attempts to answer many questions; for this assignment, we expect 2-3 visualizations in support of a single primary finding. And, you are not required to provide all of your code inline.