## ISTA 350 Final Project: Rubric and Instructions

Your work product will consist of a series of documents with different due dates, a presentation, and your Bitbucket repository. All deadlines are hard deadlines – no points will be given for late documents. Document and presentation descriptions along with point values and due dates follow. The essence of the project is using a web-based dataset (one or more) of your choice that you scrape using Python to generate three visualizations, also using Python. So this is a repeat of the 131 final project with an additional web scraping component. **Even if you are working in a group, you will need to create three images of your own. About `git`, use unique filenames! If everyone has a file on their development branch called `image1.py`, disaster will ensue!**

Friday, 2/19, by 11:59 pm (5 pts): Upload a plain text file call `readme.txt` to your Final Project Assignments folder on D2L. This file must contain a list of your group members. All group members must upload their own copy. You may work alone if you don't want to be in a group, in which case yours will be the only name in the file. Deductions:

- -5 for late (no points).
- -1 incorrect filename.
- -5 not a plain text file (no `.docx`, `.xlsx`, or anything other than `.txt`).
- **+2.5 extra credit if you are in a group.**

Friday, 3/5, by 11:59 pm (5 pts): Upload a plain text file called `dataset.txt` to your Final Project Assignments folder on D2L. **All group members must upload their own copy.** This file should contain the web address of the dataset(s) you plan to analyze in your project and a brief explanation of why you chose it and what you think you might do with it. This file may be as short as one paragraph. Your final product may be significantly different than what you describe in this document and that's ok. You may abandon this idea and finish with something completely different. Or you may decide that this isn't meaty enough and add more datasets. Your final project may, for example, consist of interesting perspectives on three unrelated datasets. Deductions:

- -5 for late (no points).
- -1 incorrect filename.
- -5 not a plain text file (no `.docx`, `.xlsx`, or anything other than `.txt`).

Monday, 4/26, 11:59 pm (15 pts): Upload a pdf or Word document containing the images you were responsible for (see below). Include the filenames of the scripts used to generate the images that are on Bitbucket. There must be at least 3 images to get full credit. Deductions:

- -5 each missing image.
- -5 each image that isn't generated from data scraped from the web using Python.

Tuesday, 4/27 – Tuesday, 5/4 (75 pts): You and your group will do a 10-minute presentation (assuming a 3-person group – 3 minutes per person) to either me or your SL or both via Zoom. There may be an audience with questions after each presentation. You are responsible for creating at least 3 visualizations to be shown in your presentation using Python. **Instead of presenting the visualizations in a PowerPoint or PDF, you will run the scripts that generate them and they need to scrape the web live!** Each visualization is worth 20 points. Each visualization must have a corresponding Python script in your repo (or turned into the final assignments folder, if you're not doing `git`) that lists you as head developer for the script in its documentation. You may get help from your team and elsewhere. You

should note any help in your documentation. Visualizations will be graded according to the following rubric:

- +2 each adequate horizontal and vertical axis titles (must include units) and labels.
- +2 adequate chart title (must be informative).
- +6 visual appeal and readability (make your text big enough).
- +10 content.
- -10 no corresponding script in your repo or a script that doesn't reproduce your image.
- -20 no web scraping.

The remaining 15 points will be my subjective evaluation of your performance, distributed as follows:

- +5 demeanor. You will at least occasionally have to speak to groups of people throughout your career. You will be rewarded with respect and advancement for poise and composure, rewarded with pity (from your friends and nicer colleagues) and contempt (from your detractors and the more douchy of your colleagues) for visible anxiety and signs of panic. Don't worry, it is normal to be nervous – it's what you do with it that counts. I recommend practice in advance and positive self-talk immediately before the presentation. Get excited about the material! Your enthusiasm will show. Focus on how cool your project is instead of how much you hate public speaking. It will work wonders. Of course, the more work you put in on your project, the easier it will be to generate enthusiasm.
- +10 effort. This will be a totally subjective assessment on my part of how hard it seems to me you worked on the project.

At least one visualization must be a scatter plot with a regression line. Deductions:

- -15 no scatter plot.
- -5 scatter plot but no regression line.

**Git, due the same day and time as the images file (10 extra credit pts): I repeat, use unique filenames! If everyone has a file on their development branch called `image1.py`, disaster will ensue!** You must not do any work on the master branch after it's created. You must create and check out a development branch and do your work on that, then merge when you think you have a stable product. You must have at least three merges (4 pts each) and one issue (3 pts) per group member. Deductions:

- -10 did work on the master branch. If you have problems at the beginning and manage to do some work on the master unintentionally, come see as soon as you realize you have worked on the master and made a commit. You won't be penalized.

**Audience (2.5 extra credit pts)**: Arrange with me or the SL's to watch one of the presentations and ask a question of one of the presenters afterward. Deductions:

- -2.5 didn't ask a question.

A sample of the many, many, many available data sets:
https://www.kaggle.com/datasets
https://www.kaggle.com/norc/general-social-survey
http://www.cdc.gov/nchs/nsfg/
https://earthdata.nasa.gov/
http://nsidc.org/
http://www.spc.noaa.gov/wcm/#data
https://www.data.gov/
http://census.gov/
https://nces.ed.gov/

http://www.nsf.gov/statistics/
http://grouplens.org/
https://www.ssa.gov/oact/babynames/limits.html
http://www.fec.gov/disclosure.shtml
https://www.treasury.gov/resource-center/financial-education/Pages/fdd.aspx
http://data.worldbank.org/
http://www.imf.org/external/datamapper/
http://murderdata.org/
https://www.cdc.gov/nchs/nhanes/
https://www.cdc.gov/nchs/
https://www.quandl.com/
https://www.ncdc.noaa.gov/
https://www.ncdc.noaa.gov/data-access/quick-links
https://ourworldindata.org/
http://aa.usno.navy.mil/index.php
https://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml
https://archive.ics.uci.edu/ml/index.php
https://archive.ics.uci.edu/ml/datasets.html
https://blog.bigml.com/list-of-public-data-sources-fit-for-machine-learning/
http://www.mldata.org/
http://www.mldata.org/repository/data/
http://www.argo.ucsd.edu/Argo_data_and.html
https://data.detroitmi.gov/
http://gis-michigan.opendata.arcgis.com/
http://portal.datadrivendetroit.org/
http://detroitsoundconservancy.org/
https://detroitography.com/
https://wdet.org/series/detroit-numbers/

You can use one of these or some other one that you find.