Instructions for Data Science Project

Timeline

April 5, 8pm:	Project proposals due (15% of the project grade);
April 5, midnight:	Project approval;
April 6, 9am:	Start working on the project;
April 13, noon:	Progress Report I (10% of the project grade);
April 20, noon:	Progress Report II (15% of the project grade);
April 23 and 27, in class/lab:	Lightning talk (20% of the project grade);
May 4, noon:	Final reports due (40% of the project grade);

Project Theme and Teams

Teams. Each project could have one or two students in a team. The expected amount of work for 2-person projects is close to double the work for 1-person projects. The expected amount of work for the whole project per student is 30 hours.

Topic: You are free to choose your topic. Start from the links provided at the end of this document and also look back at other labs and propose a topic. The work should go well beyond what you did in labs – you are expected to put an extra effort in some or all of: data collection, data processing, exploratory data analysis, predictive analytics, data visualization, or storytelling. Consider creating an interesting dashboard or a web page or an app that serves a customer. In case of doubt, you should feel free to consult with the professor or your teaching assistant. Some projects will be heavy on data collection and processing, others will be heavy on running experiments, while some might be heavy on implementation, but the total work should equal 30-hour effort.

Project proposal (submit as pdf file)

Proposal should be 1.5 - 2 pages in length with 11pt font size and single spacing. It should contain

- **Project title and student name(s)**
- **Introduction Section**: give motivation; describe the problem; mention related work and results
- **Proposed Work Section**: explain the idea; explain proposed approaches and methodology
- **Timeline**: propose the timeline for the project (state what will be done after week 1, 2, and for the final report)
- **References**: provide at least 2 related references (could be a web link)

Progress Report I and II (submit as pdf file)

A 2-page report summarizing the project status – what has been done, what has not been done, what will be done during the following week, if needed describe unforeseen problems and propose modifications to the original project goals.

Final Report

Project deliverables should include a project report and relevant programs and result listings. Project report should consist of: Title, Abstract, Introduction, Methodology, Results, and References sections. It should be made clear on what aspects of the project you spent your 30 hours of effort. Programs and results listings to be delivered will depend on the specific project. The report should be 5 pages in length with 11pt font size and single spacing. It should contain

- **Project title and student name**
- **Introduction Section**: give motivation; describe the problem
- **Approach**: explain the idea; explain proposed approaches and methodology
- **Results**: explain what was done, show results in form of tables and figures, discuss results
- **Conclusion**: summarize the whole project and its outcome
- **Acknowledgements**: clearly acknowledge people that helped you finish the project and the web resources you used (please exclude the professor and the TAs)

Other Deliverables

- **Programs** could be submitted as ipynb file(s)

- **Additional figures and tables** – if you really like your results but cannot fit them in the 5-page report, place them in a separate file called the appendix. Tables and figures should be clearly labeled and with captions describing what we are looking at.
- **Pointers** to any artifacts such as project web pages.

Useful Links:

- https://www.kaggle.com/datasets
- https://registry.opendata.aws/
- Open Data Philadelphia: https://www.opendataphilly.org/dataset
- Some links: https://towardsdatascience.com/top-sources-for-machine-learning-datasets-bb6d0dc3378b
- APIs with OAuth to extract specific data from Twitter, Linkedin, Facebook, Google
- If you have some particular source of data in mind, you might be able to see how other people analyzed it. For example, try 'ipynb boston housing' or 'python scraping indeed'
- Storytelling is an important part of data science. Some nice examples:
    - http://www.nytimes.com/interactive/2012/05/17/business/dealbook/how-the-facebook-offering-compares.html?_r=0
    - https://www.maptive.com/17-impressive-data-visualization-examples-need-see/
    - https://econsultancy.com/blog/67465-data-visualization-14-jaw-dropping-examples/