CSCI 114
Final Project

CSCI 114 Final Project

- Project Goals:
  - Use C++ to read and analyze a data set
  - Use one or more data analysis techniques (e.g. k-means, knn, linear regression, etc) to explore a hypothesis related to the data
- Project Data Sources
  - NASA: https://www.earthdata.nasa.gov/ or https://firms.modaps.eosdis.nasa.gov/
  - UCI: https://archive.ics.uci.edu/dataset/186/wine+quality
  - USDA: https://www.ers.usda.gov/data-products/food-environment-atlas/data-access-and-documentation-downloads/
  - Atmospheric data: https://weather.uwyo.edu/upperair/sounding.html
  - Kaggle: https://www.kaggle.com/datasets/
- Data Source Goals
  - Find a non-trivial dataset that interests you
  - Can be from one of the sources above or from another source
  - If you're doing research, you may use data from that project

CSCI 114 Final Project Steps

- For your chosen data source, identify a hypothesis (question) that the data might help answer
  - e.g does the moon phase affect the number of bugs caught in the bug traps? Can we cluster patient data to predict heart disease? Can we build classifiers that correctly identify the Iris species?
- Explore/visualize your data
  - Create at least 5 plots that examine and explore your data as it relates to your hypothesis
- Analyze your data
  - Using data analysis techniques (e.g linear regression, linear separators/SVMs, knn, k-means) and C++ code, analyze your data set to answer your hypothesis
  - You don't have to limit yourself to the techniques we looked at in class. For example, dlib contains many data analysis algorithms that may be useful
  - You should graph or plot your results to show how your data analysis technique helps answer your hypothesis
    - These plots count towards your required 5 plots

Final Project Report

- Write report on your project
  - Include a brief introduction of your dataset and why you found it interesting.
  - Describe your hypothesis and how the dataset can help resolve your hypothesis
  - Describe your data analysis technique and how it will use the dataset to resolve your hypothesis
  - Describe your plots and results
  - Give a conclusion, does the dataset resolve the hypothesis?
  - Describe step-by-step your code: what it does and how it works. Document the data structures (e.g classes or structs) you created to help process your data
  - Include your code, formatted nicely.
  - Include instructions for how to download your data and compile and run your code.

Final Project Deliverables

- E-mail Professor Goodney a project proposal by Friday December 6th at 9am
  - Briefly describe the dataset you're going to use
  - Briefly describe the hypothesis you plan on testing
  - Briefly describe the data processing technique(s) you plan on using
  - Briefly describe the plots/visualizations you plan on generating
- A 'project' sub-directory in your Github Classroom "assignments" repo will be created.
  - Develop your project in this directory, including the report (final delivery format PDF)
  - git commit and git push before **Tuesday December 17th at 6pm**