

**Data Mining, CSCI 347, Fall 2019**  
**Data Sets Presentation and Paper, Sept. 13**

Chosen topics:

- Government website, Data.gov, with hundreds of datasets, Carson Fiechtner
- Kaggle, datasets and competitions, Keelie Lawler
- Wine quality dataset from the UCI Machine Learning Repository, Dalton Caron
- Million Song dataset, Kaleb Bausch
- Github “Awesome Public Datasets”, Xaavan Dolence
- Airline safety, Joseph Weeks

Links to public sources of data are given under the Resource tab of the class website. They are repeated here.

- kaggle, world's largest data mining and machine learning community <https://www.kaggle.com/datasets>
- Aviation Weather Center <https://www.aviationweather.gov/metar>
- University of California, Irvine Machine Learning Repository <https://archive.ics.uci.edu/ml>
- World Bank Open Data <https://data.worldbank.org/>
- Awesome Public Datasets on github <https://github.com/awesomedata/awesome-public-datasets>
- Robert Wood Johnson Foundation County Health Rankings & Roadmaps <http://www.countyhealthrankings.org/explore-health-rankings/rankings-data-documentation>
- 25+ websites to find datasets for data science projects (Analytics Vidhya) <https://https://www.analyticsvidhya.com/blog/2016/11/25-websites-to-find-datasets-for-data-science-projects/>

Explore these sites and others and choose a site to present to the class. Prepare a 5 minute presentation, along with a 3-6 page write-up, on a source of data. Describe and show the type of data that is available, who has collected the data, how the data was collected, whether data sets in general appear complete, balanced (for datasets with a class value, the number of instances is approximately equivalent for each of the class values), consistent (for datasets with a class value, the same evidence generally gives the same prediction). If available, give examples of how some of the data has been used, and any other things of interest about the data.

When you select a site, let me know. I will post what you have chosen so others do not also choose it.

Your write-up should be in your own words, with quotes and descriptions referenced.

Turn in the paper and visual aids (~4 slides to a page, if using slides) before beginning your presentation. Your presentation will be graded using the following form:

**Oral Assessment Form**  
Form updated: 4/21/2014

**Course Number:** CSCI 347

**Semester:** Fall 2019

**Date:** Sept. 13

**Presenter:**

**Evaluator:**

**Topic:**

Audience types: (1) General (2) Peers

**Content** 1 = Poor, 2 = Needs Improvement, 3 = Good, 4 = Excellent

Material is relevant to topic	1 2 3 4
Topic is explored in depth Description of the data Who collected the data How data was collected Characteristics of the data – consistent, complete, well-formatted, balanced How has data been used	1 2 3 4
Student is competent and knowledgeable about topic	1 2 3 4
Reliable sources were used	1 2 3 4

**Organization**

Clear overview / objectives are given	1 2 3 4
Major issues are covered (see above)	1 2 3 4
Presentation has a logical sequence with smooth transitions	1 2 3 4
Clear conclusion or summary is given	1 2 3 4

**Delivery**

Student projects confidence and mastery	1 2 3 4
Student is relaxed, natural and spontaneous*	1 2 3 4
Student speaks clearly*	1 2 3 4
Appropriate vocabulary and grammar are used	1 2 3 4
Infrequent use of notes are made	1 2 3 4
Students maintains eye contact with audience	1 2 3 4
Effective use is made of visual / technical aids*	1 2 3 4
Presentation is within the time limitations*	1 2 3 4
Answers questions effectively (demonstrates knowledge of topic)	1 2 3 4

\* See notes on the back.

**Notes**

**Delivery**

- Student is relaxed, natural and spontaneous – the presentation was not rushed
- Student speaks clearly – No um’s, ah’s, etc.
- Effective use is made of visual / technical aids - Avoid distracting special effects
- Presentation is within the time limitations - The length of the presentation is within the specified times, neither too long nor too short

**Comments:**

The point breakdown will be:

<b>Presentation (70 points)</b>	
Completeness of presentation content Description of the data Who collected the data How data was collected Characteristics of the data – consistent, complete, well-formatted, balanced How has data been used	1 2 3 4
Quality of presentation content	1 2 3 4
Organization of presentation	1 2 3 4
Delivery or presentation	1 2 3 4
<b>Paper (20 points)</b>	
Completeness of paper content	1 2 3 4
Quality of paper content	1 2 3 4
Organization of paper (Title and subheadings are used, paragraphs are cohesive, introductory and concluding paragraph)	1 2 3 4
Mechanics (grammar, spelling, tone and length are appropriate, references are in APA format)	1 2 3 4
<b>Slides (10 points)</b>	
Printed 4 to a page.	