# Dr. Eick

COSC 6335 *“Data Mining”* Fall 2012

Draft Project1: Exploratory Data Analysis

Individual or Group Project[[1]](#footnote-1)



**Learning Objectives**:

1. Learn how to manage and preprocess datasets and how to compute basic statistics and to create basic data visualizations in R
2. Learn how to interpret popular displays, such as histograms, scatter plots, box plots,…
3. Get some practical experience in exploratory data analysis
4. Learn how to create background knowledge for a dataset
5. Learn to distinguish expected from unexpected results in data analysis and data mining—in general, this is quite difficult, as it requires background knowledge with respect to the employed data mining technique, and also practical experience.

**Due**: Wednesday, September 19, 11p (electronic Submission)

**Last Updated**: September 5, 11a

**Project Description**:

Download Win Qualify Data Set dataset from <http://archive.ics.uci.edu/ml/datasets/Wine+Quality> ; we will be limiting ourselves to red wines and to analyzing subsets of the dataset; use all 1599 red wine examples to create your subset:

If your last name starts with A-K or your group number is odd, you analyze the following attributes:

1 - fixed acidity

10 - sulphates

11 - alcohol

12 - quality (score between 0 and 10)

If your last name starts with L-Z or your group number is even, you limit your analysis to the following attributes:

4 - residual sugar

9 - pH

11 - alcohol

12 - quality (score between 0 and 10)

Apply the following exploratory data analysis techniques **using R** to your dataset:

1. Compute the covariance matrix for the three numerical attributes you are analyzing, excluding the output attribute; also compute the correlation for each of the three pairs of numerical attributes. Interpret the statistical findings!
2. Create 3 classification scatter plots (using 2 of the 3 numerical attributes for each display)---use different colors for visualizing different wine qualities (e.g. yellow=3, orange=4, red=5, blue=6, black=7). Interpret the display—what does it tell you about the difficulty of predicting wine quality based on the given attributes?
3. Create a scatter plot for the first two attributes in your dataset. Interpret the scatter plot!
4. Create histograms for the first and second numerical attributes. Interpret the obtained displays!
5. Create box plots for the second and third attribute for the instances with wine quality 5, for instances of wine quality 6 and for instances of wine quality 7. Interpret and compare the 2x3=6 box plots!
6. Use one more display of your own liking that sheds some light on the predicting wine quality using the three numerical attributes in your subset! Interpret the display!
7. Are there any other interesting observations about your dataset?
8. Write a short summary that assesses the difficulty of predicting the wine quality (in which the class variable has to be predicted just using the selected 3 attributes), based on your work answering questions 1-7! What technique(s) you believe has (have) the best potential to solve the classification problem and why?

Remark: About 35-40% of the Project1 points will be allocated to interpreting statistical findings and visualizations! A few extra points will be allocated for really good answers to the questions in green!

**Background Information Red Wine Quality Dataset**

5 Examples in the raw Red Wine Quality Dataset

|  |  |
| --- | --- |
| 7.3;0.48;0.32;2.1;0.062;31;54;0.99728;3.3;0.65;10;7 | |
| 6.7;1.04;0.08;2.3;0.067;19;32;0.99648;3.52;0.57;11;4 | |
| 7.3;0.48;0.32;2.1;0.062;31;54;0.99728;3.3;0.65;10;7 | |
| 7.3;0.98;0.05;2.1;0.061;20;49;0.99705;3.31;0.55;9.7;3 | |
| 15.6;0.685;0.76;3.7;0.1;6;43;1.0032;2.95;0.68;11.2;7 |  | |

Attribute information:

For more information, read [Cortez et al., 2009].

Input variables (based on physicochemical tests):

1 - fixed acidity

2 - volatile acidity

3 - citric acid

4 - residual sugar

5 - chlorides

6 - free sulfur dioxide

7 - total sulfur dioxide

8 - density

9 - pH

10 - sulphates

11 - alcohol

Output variable:

12 - quality (score between 0 and 10)

[Cortez et al., 2009] P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In *Decision Support Systems*, Elsevier, 47(4):547-553, 2009.

Links:

<http://www.vinhoverde.pt/en/>

1. Unless the enrollment of the course is below 20, groups of 2 students will work on Project1 jointly; otherwise, this project will be an individual project. **Groups will be set up by Zechun/Dr. and announce during the lab on Sept. 6, 2012.** [↑](#footnote-ref-1)