# Dr. Eick

COSC 6335 *“Data Mining”* Fall 2013

Project1: Exploratory Data Analysis

Group Project

**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, linear functions,…
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 task is quite challenging, as it requires background knowledge with respect to the employed data mining technique, and also practical experience.

**Due**: Thursday, October 9, 11p (electronic submission)

**Last Updated**: September 19, 11a

**Group size**: 3 (or 2)

**Project Description**:

Download Pima Indians DiabetesData Set from <http://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes>;

In this project, we center on applying exploratory data analysis to the relationships between the 4 numerical attributes, listed below and the class variable:

2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test—called first attribute in the following.

3. Diastolic blood pressure (mm Hg))—called second attribute

6. Body mass index (weight in kg/(height in m)^2)—called third attribute

7. Diabetes pedigree function—called fourth attribute

9. Class variable (0 or 1)—called class attribute

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

1. Compute the mean value and standard deviation of the 4 numerical attributes[[1]](#footnote-1). **2 points**
2. Compute the covariance matrix for the four numerical attributes you are analyzing; also compute the correlation matrix. Interpret the statistical findings! **4 points**
3. Create 6 classification scatter plots (using 2 of the 4 numerical attributes for each display)—use different colors for objects belonging to class1 and class0). Interpret the display—what does it tell you about the difficulty of predicting the class attribute based on the given attribute pairs? **9 points**
4. Create a scatter plot for the first two attributes and second scatter plot for the third and fourth attribute of the dataset. Interpret the two scatter plots! **4 points**
5. Create histograms for the third and another histogram for the fourth attribute. Give a justification how and why the reported histogram was chosen! **~~6 points~~ 4 points only** as I forgot to ask you to interpret the histograms!
6. Create 3 box plots for the second attribute and 3 box plots for the third attribute (for the whole dataset, for the instances of class 0 and for the instances of class 1). Interpret the display! Interpret the display. What does it tell you about the usefulness of the 2nd /3rd for the classification problem? **6 points**
7. Create a Star plot for the first 20 instances of class1 and the first 20 instances of class0 (based on the order in the file); interpret the 40 stat plots—star plots should be constructed for the 4 attributes of focus! **4 points**
8. Fit a linear model that predicts the class attribute (treat it as a numerical attribute that takes values 0 and 1) using the 4 attributes we focus on. Report the R2 of the linear model and the coefficients of each attribute in the obtained regression function. Do the coefficients tell you anything about the importance of the attribute in predicting the class variable; if yes, what? **6 points**
9. Are there any other interesting observations about your dataset?
10. Write a short conclusion that assesses the difficulty of predicting the class attribute using the selected 4 attributes), based on your work answering questions 1-9! **5 points for Question 9 and 10 together.**

Remark: **Extra points** might be given for exceptionally good solutions.

Summarize your answers which include all statistical results and graphics in a Word or pdf-file, including your interpretation which should be place next to graphical displays and statistical finding you analyze. Submit the file at the due date!

Other recommendations:

1. If no method is specified, use default method, if the R-function supports multiple methods
2. Some task, particularly Task8 might benefit from normalizing the dataset!
3. Note that the slide on boxplots has been updated (now different from the textbook); the information on the slide is “correct”!
4. Start early.

Remark: At least 35% of the Project1 points will be allocated to interpreting statistical findings and visualizations!

**5 Examples in the raw Pima Indians Diabetes Dataset:**

|  |  |
| --- | --- |
| 6,148,72,35,0,33.6,0.627,50,1 | |
| 1,85,66,29,0,26.6,0.351,31,0 | |
| 8,183,64,0,0,23.3,0.672,32,1 | |
| 1,89,66,23,94,28.1,0.167,21,0 | |
| 0,137,40,35,168,43.1,2.288,33,1 |  | |

Attribute information:

1. Number of times pregnant

2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test

3. Diastolic blood pressure (mm Hg)

4. Triceps skin fold thickness (mm)

5. 2-Hour serum insulin (mu U/ml)

6. Body mass index (weight in kg/(height in m)^2)

7. Diabetes pedigree function

8. Age (years)

9. Class variable (0 or 1)

**Guidelines on Submission of Reports**

1. Submit your report to [ceick@aol.com](mailto:ceick@aol.com)
2. Use e-mail header: Project1-G<groupnumber>-Report
3. A single report for each group either in Word(preferred) or pdf.
4. Discuss the tasks chronologically (task 1, 2, 3,…)
5. Put interpretations next to displays / statistical summaries that are interpreted **.** Groups that do not follow those guidelines with the guidelines

Will be penalized.

**Project1 Grading**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9/10 | Extra | Σ |
| G1 | 1 | 3 | 8 | 4 | 4 | 5 | 4 | 6 | 3 | 0 | 38 |
| G2 | 2 | 4 | 5 | 4 | 4 | 4 | 4 | 4 | 5 | 0 | 36 |
| G3 | 2 | 4 | 7 | 4 | 3 | 3 | 4 | 2 | 2 | 0 | 31 |
| G4 | 2 | 3 | 9 | 4 | 4 | 6 | 4 | 6 | 5 | 1 | 44 |
| G5 | 2 | 4 | 8 | 4 | 4 | 5 | 4 | 5 | 5 | 2 | 43 |
| G6 | 2 | 4 | 6 | 4 | 4 | 6 | 4 | 3 | 3 | 0 | 36 |
| G7 | 2 | 4 | 5 | 4 | 4 | 6 | 4 | 4 | 1 | 1 | 35 |
| G8 | 2 | 4 | 9 | 4 | 4 | 4 | 4 | 3 | 5 | 1 | 40 |
| G9 | 2 | 2 | 6 | 4 | 4 | 5 | 4 | 3 | 1 | 0 | 31 |
| G10 | 2 | 4 | 5 | 4 | 4 | 5 | 4 | 5 | 1 | 0 | 34 |
| G11 | 2 | 3 | 7 | 4 | 4 | 6 | 4 | 5 | 3 | 0 | 38 |
| max | 2 | 4 | 9 | 4 | 6 4 | 6 | 4 | 6 | 5 | — | 44+… |

Project1 Statistics:

Average: 36.9

Median: 36

SD: 4.3

1. This is more a verification of that you have the correct dataset! [↑](#footnote-ref-1)