Dr. Eick

COSC 3337 *“Data Science I”* Fall 2021

Problem Set1

Third Draft

Last Updated: September 23, 7p

Task1: Exploratory Data Analysis for an Abalone Dataset



Task1 Due: Monday, Sept. 27, 11:59p (electronic Submission)

**Learning Objectives**:

1. Learn how to manage and preprocess datasets and how to compute basic statistics and to create basic data visualizations (using R or other tools)
2. Learn how to interpret popular displays, such as histograms, scatter plots, box plots, density 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 task is quite challenging, as it requires background knowledge with respect to the employed data mining technique, and also practical experience.

**Abalone** is a shellfish considered a delicacy in many parts of the world. The abalone shell and the meat is of value. The goal of this project is to perform exploratory data analysis for the Y*Abalone dataset* which is a modification of the Abalone Dataset (<http://archive.ics.uci.edu/ml/datasets/Abalone>). The original Abalone dataset is a 9D dataset and YAbolone is a 10D dataset with an ordinal Age attribute added; YAbalone has the the following attributes:

Sex / nominal / -- / M, F, and I (infant)

Length / continuous / mm / Longest shell measurement

Diameter / continuous / mm / perpendicular to length

Height / continuous / mm / with meat in shell

Whole weight / continuous / grams / whole abalone

Shucked weight / continuous / grams / weight of meat

Viscera weight / continuous / grams / gut weight (after bleeding)

Shell weight / continuous / grams / after being dried

Rings / integer / -- / +1.5 gives the age in years

Age / ordinal/ ---/ Y, M, and O (age classes ‘young’, ‘medium’ and ‘old’; derived from Rings attribute; see below)

4 Examples in the YAbalone Dataset:

M,0.455,0.365,0.095,0.514,0.2245,0.101,0.15,15,O

M,0.35,0.265,0.09,0.2255,0.0995,0.0485,0.07,7,Y

F,0.53,0.42,0.135,0.677,0.2565,0.1415,0.21,9,M

M,0.44,0.365,0.125,0.516,0.2155,0.114,0.155,10,M

The values of the Age attribute have been computed from the Rings attribute as follows: 0-8🡪Y, 9-11🡪M, 12-29🡪O. In general, we are interested to predict Attributes 9 and 10 using the other attributes; that is, we like to predict the age of abalones based on their physical properties described by attributes 2 through 8. Other things we are interested in are finding relationships between the continous attributes in the dataset, and to understand differences between male and female abalones.

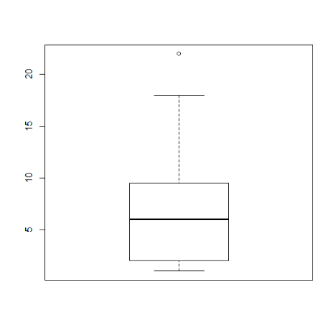
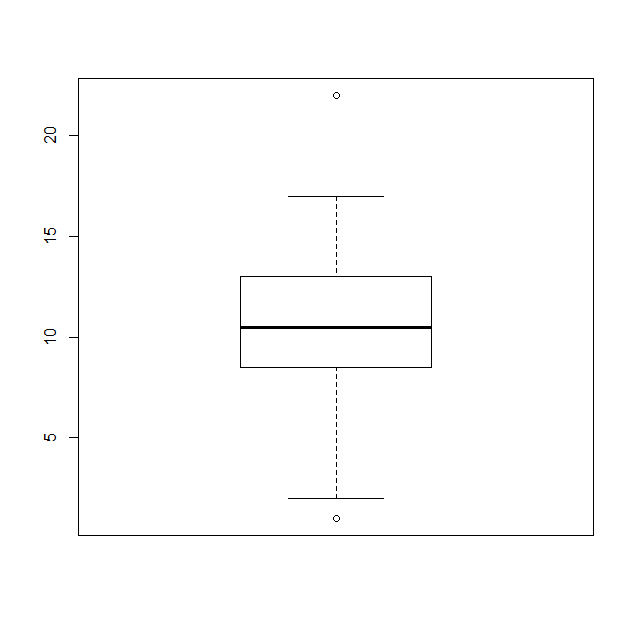
Assignment1 Tasks:

Apply the following exploratory data analysis techniques **using R** or other tools of your liking to your dataset:

1. Create the YAbalone dataset or use the one created by the TA!
2. Compute the mean value and standard deviation of the 7 numerical attributes**[[1]](#footnote-1). 1 point**
3. Compute the covariance matrix for each pair of the following attributes: Length, Diameter, Shucked Weight, and Rings (treat this attribute as a continuous attribute); next, compute the correlations for each of the 6 pairs of the 4 attributes. Interpret the statistical findings! **3 points**
4. Create a scatter plot for the Shell weight and Length of your dataset. Interpret the scatter plot**! 3 points**
5. Create histograms for Diameter, Whole Weight, and Rings attributes for both the male and the female abalones; interpret the obtained 6 histograms. **6 points**
6. Create box plots for the Whole Weight attribute for the instances of each age class—one for Y, M and O— and a fourth box plot for all instances in the dataset. Interpret and compare the 4 box plots for each attribute! **4 points**
7. Create supervised scatter plots/supervised density plots for the following 3 pairs of attributes using the Age attribute as a class variable: Diameter&Viscera Weight, Diameter&Shell Weight and Viscera Weight& Shell Weight. Use different colors for the class variable. Interpret the obtained plots; in particular address what can be said about the difficulty in predicting the correct age class and the distribution of the instances of the three classes. **6 points**
8. Create 3 density plots for the instances of the 3 age classes in the Diameter/Shell Weight space. Compare the 3 density plots! **6 points**
9. Create a new dataset ZAbalone from the YAbalone dataset by transforming the 7 continuous attributes into z-scores. Fit a linear model that predicts the Rings attribute using the 7 z-scored, continuous attributes as the independent variables. Report the R2 of the linear model and the coefficients of each attribute in the obtained regression function. What do the obtained coefficients tell you about the importance of each attribute for predicting the number of rings of an abolone? **6 points**
10. Create 3 decision tree models with 20 or less nodes for the dataset (leaf nodes count; do not submit models with more than 20 nodes!); use the age attribute as the class variable, and do not use the first and 9th attribute when building the decision tree model; that is, the model uses attributes 2-8 to predict attribute 10! Explain how the 3 decision tree models were obtained! Report the training accuracy and the testing accuracy of the submitted decision trees. Interpret the learnt decision tree. What does it tell you about the importance of the 7 attributes for the classification problem? **9 points**
11. Write a conclusion (at most 13 sentences!) summarizing the most important findings of this task; in particular address the findings obtained related to predicting the age of an abalone (the values of attributes 9 and 10) using attributes 1-8. **6 points (and up to 4 extra points)**

Remark: About 30-40% of the Task1 points will be allocated to interpreting statistical findings and visualizations!

**Submission Guidelines Task1**: When you submit your task 1 for problem set 1, you should submit a compressed (zipped) folder that contains a word file that displays your graphs and your interpretations. Each interpretation should use complete sentences to describe your findings. Also in the folder, you should include all files used to complete your tasks, such as your R or python files. If you have doubts what to submit send Mathew an e-mail.

Task 2: An Intelligent Tool Which Compares Boxplots 

Task2 Due: Monday, October 4, 11p (electronic Submission)

**Learning Objectives**:

1. Comparing box plots
2. Converting numbers into natural language summaries
3. Tools for data storytelling

Develop a tool that compares 2 box plots and reports its findings as a story. The story summarizes the similarity of two input box plots[[2]](#footnote-2) with respect to median, box location, IQR, maximum and minimum values that are no outlier (and maybe[[3]](#footnote-3) also compares the reported outliers) and skewness; moreover, the tool summarizes the overall agreement / disagreement of the two boxplots. Be prepared to demo your tool; moreover, a benchmark of 4 example pairs of boxplots to compare will be posted by September 16.

Basically, the input of the tool are the following values of each boxplot:

Median1, median2:= median of each boxplot

75perc1,75perc2:= 75% percentile value of each boxplot

25perc1,25perc2:= 25% percentile value of each boxplot

max1, max2:= maximum value in each dataset from which the box plot was created which is not an outlier

min1, min2:= minimum value in each dataset from which the box plot was created which is not an outlier

These values can be obtained for a boxplot b in R using: b.$stats

Other values that are derived from those input parameters:

IQR1, IQR2 := IQRs of the respective boxplots

If your tool additionally analyzes outliers, the following inputs are additionally considered:

Out1:= set of outlier values boxplot1; could be empty!

Out2:= set of outlier values boxplot2; could be empty!

For a boxplot b outliers can be obtained in R using: b$out

Based on those input values the tool creates a “summary story” concerning the similarity of the two input boxplots.

Three Test cases for Problemset1 Task2:

**Case 1:**

L1 = [18, 58, 22, 50, 44, 64, 68, 10, 58, 6, 82, 42, 39, 26, 18, 44, 80, 26, 59, 35, 20, 81, 23, 77, 18, 72, 24, 3, 30, 81, 189, 149, 109, 139]

L2 = [58, 98, 90, 93, 55, 14, 15, 94, 33, 84, 26, 29, 98, 22, 24, 47, 52, 51, 71, 48, 83, 15, 0, 50, 99, 20, 21, 85, 57, 49]

**Case 2:**

class1 = [18, 58, 22, 50, 44, 64, 68, 10, 58, 6, 82, 42, 39, 26, 18, 44, 80, 26, 59, 35]

class2 = [58, 98, 90, 93, 55, 14, 15, 94, 33, 84, 26, 29, 98, 22, 24, 47, 52, 51, 71, 48

**Case 3:**

Dallas = [18, 58, 22, 50, 44, 64, 14, 15, 58, 6, 26, 29, 26, 18, 44, 178, 170]

Houston = [149, 180, 165, 166, 147, 171, 154, 125, 136, 162, 141, 139, 131, 132, 143, 11, 34, 18]

1. This is more a verification of that you have the correct dataset! [↑](#footnote-ref-1)
2. Moreover, we assume that boxplots are generated using the default parameter setting of the respective boxplot function of the boxplot tool you use. [↑](#footnote-ref-2)
3. A comparison of outliers is not required, but a small amount of extra credit will be given to students whose storytelling tool compares the presence and location of outliers in the two box plots. [↑](#footnote-ref-3)