2104529 Computational Methods in Industrial Engineering

Workshop 4: Data Exploration and Visualization

Question 1

Consider '`iris`' dataset (one of the most famous data set in Data Mining) and explore the basis of '`data.frame`' package (the most basic and popular data template in R)

(a) access 'Sepal.Length' column of '`iris`' '`data.frame`' as matrix(·) and list(·)

(b) explore top/bottom/random sampling

(c) view and check for dimension and duplication of the data.frame

(d) use '`summarytools`' package to explore the data.frame

Question 2

Re-Consider '`iris`' dataset do the following simple data analysis and handling with R with '`base`', '`dplyr`', and '`data.table`' package

(a) summarize the data using basic descriptive statistic (mean, median, sd kurtosis, sd, IQR, CV)

(b) find record that '`Species` = `versicolor`' and '`Petal.Width`' is between 1.0 and 1.5

(c) summarize 'Sepal.Width' and 'Sepal.Length' by its Species

(d) find 'Species' that contain texts '*color*' and '*vir*'

(e) Arrange data by '`Species`, '`Petal.Length`' and '`Petal.Width`' respectively

(f) Count number of data in each Species and summarize using the following criteria

type	width of petal	length of petal
low	$[0.00, 0.75)$	$[0.0, 2.5)$
medium	$[0.75, 1.75)$	$[2.5, 5.0)$
high	$[1.75, \infty)$	$[5.0, \infty)$

(g) [0 points (bonus)] Use you knowledge to query and mutate iris in the following step

- label by its 'Petal.Length' into equal groups, (i.e., 'PL.H','PL.M', 'PL.L')
- label by its 'Sapal.Length' into equal groups, (i.e., 'PL.H','PL.M', 'PL.L')
- compute median and sd of column 'Petal.Width' of groups
- compute mode and CV of column 'Sepal.Width' of groups
- represent result as labels of 'Petal.Length' and 'Sapal.Length'

(h) [0 points (bonus)] convert dataset into long format (see below)and convert back

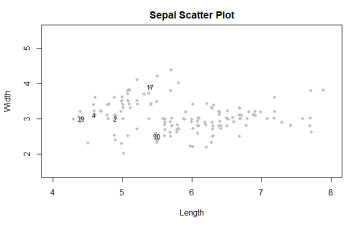
id	attribute	value	Species
1	Sepal.Length	5.1	setosa
2	Sepal.Length	4.9	setosa
⋮	⋮	⋮	⋮
300	Petal.Width	1.8	virginica

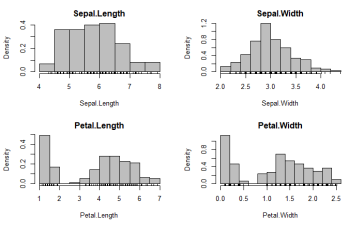
Question 3

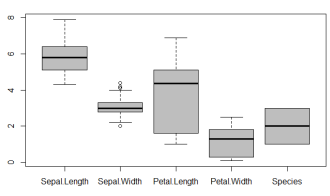
Reconsider '`iris`' dataset (again) and explore the following basis visualization

(a) explore the following ASCII plots stem(·) , sunflower(·)

(b) recreate the following plots







(c) compare standard boxplot(·) with '`lattice::bwplot()`'

(d) [0 points (bonus)] the standard boxplot(·) be use to detect and eliminate outlier. detect and report the outlier of each attributes

(e) [0 points (bonus)] Apply R command that you know to prepare '`iris`' dataset in the following step.

- check for duplication/ missing value / incorrect /irregular values
- consider remove or impute such data points

Question 4

Visual `iris` dataset using `ggplot2` package, a popular visualization package that interfaced seamlessly with `data.frame` or `data.table` package

(a) use geom_histogram(·) , geom_density(·) , geom_column(·) , and geom_violin(·) to visualize a single attribute

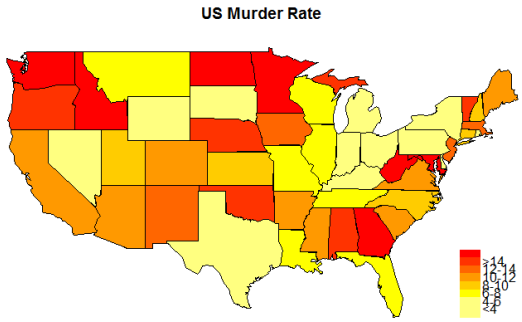
(b) use geom_scatter(·) , geom_density_2d(·) , geom_point(·) , and geom_lineplot(·) to visualize multiple attribute

Question 5

After marked 30 questions, an instructor notice a possible cheating of the following 10 students. The questions are TRUE-FALSE question, and instructor has marked '1' for correct answer and '0' for incorrect answer. Can you detect cheaters (source and copier)?

Question 6

Consider the following visualization example of number of murders in US from '`USArrests`' in package '`datasets`' by state with the thermal map (**hint:** heat.colors (·), legend (·), map (·) in package '`maps`')



(a) Represent other three types of arrest, i.e. '`Assult`', '`UrbanPop`', and '`Rape`' with the similar manner with function plotThermalMap (type=1,quantLv=c(0.1,0.25,0.5,0.75,0.9))

(b) Automatically generate the thermal map and export as files

Question 7

Re-consider again '`iris`' dataset and classify this dataset using the following tasks

(a) based on the data exploration so far, list useful insights

(b) separate data into training set (120 data points) and testing set (30 data points)

(c) apply the classification technique the training dataset and discuss the predicted result with testing data (**hint** '`class::knn()`' '`base::glm()`' , '`party::ctree()`'). why you choose this technique?

note: This question serves as an introduction to machine learning and clustering technique

last updated: August 4, 2022