ΔΠΜΣ ΕΠΙΣΤΗΜΗ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ

ΜΑΘΗΜΑ: *Προγραμματιστικά Εργαλεία και Τεχνολογίες για Επιστήμη Δεδομένων*

ΔΙΔΑΣΚΩΝ: *Δημήτρης Φουσκάκης*

ΑΚΑΔΗΜΑΙΚΟ ΕΤΟΣ: *2020-2021*

Εργαστήριο στην R

Title: Data Tables and Visualizations

Make sure before coming to the lab that you have installed the latest version of R in your laptop, as well as, all the packages that we used in the slides during class. Mostly in this lab we will use the packages `data.table` and `ggplot2`.

(1) Using the function `fread` from the package `data.table` read the data "flights13.csv" from the website of the course. The data are similar to the flights data we used in class; giving information for flights departing from New York airports for the year 2013. We will use the `data.table` package to make several calculations.

(a) Remove the first column that is just a flight index number.
(b) Using the function `na.omit` remove rows with missing values.
(c) What percentage of trips have had total delay less than zero?
(d) Calculate the percentage of trips corresponding to each departure airport. Named appropriately your new resulting variable.
(e) Calculate the percentage of trips corresponding to each departure airport and for each month. Named appropriately your new resulting variable. Order your object by all grouping variables.
(f) In your previous result sort the output by the percentage of trips in decreasing order.
(g) Calculate the average total delay time for each carrier. Named appropriately your new resulting variable and sort your answers in decreasing order according to this new variable.
(h) Calculate the average arrival and departure delay for each origin and destination pair, for each month for carrier "AS". Named appropriately your new variables. Sort your results by month.
(i) For the carrier "UA" only compute the mean arrival and departure delays for every origin and month. Sort your results in increasing order by arrival and departure delay.
(j) On the 24th of December you want to travel from any New York airport to Los Angeles (LAX). Money is not an issue! You wish to be in LAX (schedule arrival) before 9pm. Therefore you have to depart (schedule departure) before 6pm from New York. Which specific flight you would choose?

(2) It is time now to create some nice visualizations on our data, using the R package `ggplot2`. To start, we focus only on flights departing from New York on the 24th of December to LAX, before 6pm.

(a) Produce a scatter plot of arrival versus departure delay by carrier. Add separate linear regression lines without standard errors.
(b) Using now `facets` create multiple scatter plots as in above by carrier.
(c) Produce a new variable that takes the value "<0" if the total delay is negative and ">=0" if it is not. Create percentage stacked bar-plots using this new variable and carrier. Also create frequencies grouped bar-plots using this new variable and carrier.
(d) Produce again the scatter plot from (a) without adding regression lines. Add for each carrier the flight number corresponding to the flight with the lowest total delay.
(e) Produce again the scatter plot from (a) without adding regression lines. Change the size of your points proportionally to total delay.
(f) For the whole dataset, with destination the Los Angeles airport (LAX), produce a box-plot of the total delay by carrier. Adjust the width of the boxes to be proportional to the number of observations it contains.
(g) Produce the same boxplots as above, for different origins as well.
(h) Produce a pie-chart of carriers that flight to Los Angeles airport (LAX), for the whole dataset.
(i) For the whole dataset, with destination the Los Angeles airport (LAX), produce a time series plot of the mean total delay.
(j) Produce the same plot as above but also for different carriers.