Capstone assignment

Marc A.T. Teunis

2018-01-22

Data Analysis Using R - November 2017

## Capstone definition

* A coping stone :coping
* The high point :crowning achievement

“..the capstone of her career..”

## Capstone in this course

* A showcase that you’ve learned how to use R in *every* element of the data cycle
* No statistical analysis
* Focus on read, clean and explore data
* Final product: An \*.Rmd (and corresponding rendered \*.html) file with your analysis sent to me by email **BEFORE** the deadline.

## Practical aspects

* Deadline 2 February 23:59 hrs., 2018
* Use this presentation-file “./capstone/capstone\_assignment” to make the assignment
* Questions by email, and face-to-face by appointment only (tutor-slots)
* Work in pairs, explain your code to each-other, discuss steps you made, integrate you work in one analysis document

# The data

* Toxcast
* US Airpollution
* Predict molecular properties
* Or your data from your minor-project

Choose one of the datasets below

## Toxcast data

Difficulty: \*\*

* Toxcast dataset in the folder “./capstone/data/toxcast”

## Online, open dataset 1: “US air pollution”

Difficulty: \*\*

* <https://www.kaggle.com/sogun3/uspollution>
* The data for this dataset are in “./capstone/data/uspollution”

## Online, open dataset 2: “predicting molecular properties”

* If you choose this, you can get the data by a download link: <https://www.kaggle.com/burakhmmtgl/predict-molecular-properties/downloads/predict-molecular-properties.zip> (due to the big size of the file)

Difficulty: \*\*\*

library(jsonlite)  
file\_json <- file.path(root, "capstone", "data", "predictmolprop", "pubChem\_p\_00000001\_00025000.json")  
example <- fromJSON(file\_json, flatten=TRUE)  
example$atoms[[1]]

## Your own data; prerequisites

* Quantitative data
* At least three factor (grouping) variables
* At least three different treatments
* At least 200 rows of raw data

*Sent an example to me, I will grade the difficulty and approve/disapprove the dataset.*

# The assignment

## Introduction

Write a short (max 300 words) introduction that explains some of the details of the data you are analyzing for your capstone assignment

* The type of data
* The ‘experiment’ or variables
* The source of the data
* The anticipated results or aim of the analysis: Which (research) question is being addressed

## Reading data into R

* Import the data into R with a code-chunk
* If zipped unzip the data with R before loading
* User relative paths so that your code is portable

## Inspect the integrity of the data

* Write a code chunk that checks the data, number of rows, number of variables, names of the variables, data-type of the variable
* Create a summary statistics of each variable that shows confidence intervals, max, min, mean, median values
* Create code that asserts the data-type of each variable

## Tidy data

* Write a code chunk that creates a tidy dataframe:
* Each variable in it’s own column (long format)
* Each row has it’s own observation
* Each value has it’s own cell
* Variable names are all lowercase and snake\_case and have a short descriptive name
* Reshape the data if necessary

## Select, Filter, Mutate, Summarize

* Select proper variables
* Filter values if needed
* Create new variables with dplyr::mutate() if needed
* Summarize the data by grouping variables and calculate mean values

## Exploratory data analysis

Create at least four (4) graphs that show some sort of relationship between variables present in the data

**For example:**

* Use colours to isolate groups in the data
* Use facets to show groups in the data
* Use diffrent types of graphs (e.g. geom\_point, geom\_line, geom\_bar, scale\_fille\_gradient )
* Provide a graph that indicates information on the distribution of your data

## Graph criteria

* The graph has a descriptive title
* The axes have units and decriptive *short* titles
* The graph is visually appealing and the scale is correct

## Results

* Write a short paragraph (max 500 words) on your resulting graphs.
* What does the reader have to take away from the graphs you’ve shown?

***Tell a data-story!***

## Conclusion

* Refer to the introduction and write a short (max 200 words) conclusion on you analysis - Identify gaps in you analysis
* What still needs to be done? / What is lacking?