



Adatelemzés és Kódolás

HUN-REN KRTK Közgazdaságtudományi Intézet – műhely

Reguly Ágoston

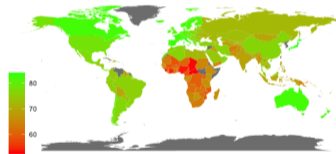
Budapesti Corvinus Egyetem & Georgia Institute of Technology



Kód tanulás

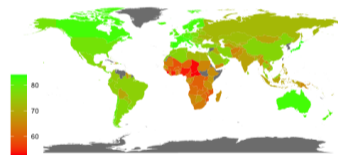
- Az elemzés ne álljon meg a tankönyvi elmélet szintjén!
- Hogy lehet megtanítani programozást korábban nem tudó embereket:
 - Regressziós módszertant vagy gépi tanulást alkalmazó algoritmusokat alkalmazni?
 - Automatizált jelentéseket készíteni pdf-ben vagy HTML formátumban?
 - Szép, olvasható és megfelelő ábrákat készíteni?
 - Nagy adatbázisokkal dolgozni és reprodukálható módon megtisztítani őket?
- Munkaerőpiacra belépőkhöz általában két típusú kérdés:
 - szakma specifikus ismeret
 - miben tud programozni? – teszt/pszeudó kód írása...

Life expectancy at birth in years (2017)



- Az elemzés ne álljon meg a tankönyvi elmélet szintjén!
- Hogy lehet megtanítani programozást korábban nem tudó embereket:
 - Regressziós módszertant vagy gépi tanulást alkalmazó algoritmusokat alkalmazni?
 - Automatizált jelentéseket készíteni pdf-ben vagy HTML formátumban?
 - Szép, olvasható és megfelelő ábrákat készíteni?
 - Nagy adatbázisokkal dolgozni és reprodukálható módon megtisztítani őket?
- Munkaerőpiacra belépőkhöz általában két típusú kérdés:
 - szakma specifikus ismeret
 - miben tud programozni? – teszt/pszeudó kód írása...

Life expectancy at birth in years (2017)



Kód tanulás

- Kód tanulást támogató kurzus anyagok
- Három nyelven:
- R (Reguly Ágoston) és Python (Duronelly Péter és Víg Ádám)
 - Könyv I-III részeit teljesen lefedi
 - 27 különálló óra
 - 1830 percnyi anyag
- Stata (Tőkés László)
 - Könyv I-II részek



STATA



- Kódolást a programozás heti begyakorlásával tanulják meg
- Megmutatni a legjobb gyakorlatokat és utána azt gyakoroltatni velük
 - Sok anyag és megközelítés van (neten), ezek közül kiválasztani a leginkább alkalmas megközelítést
 - Állandóan alakul! Alapok viszont változatlanok...
- Nem hardcore kódolási kurzus – programmozók számára egyszerűnek tűnhet
 - Fókusz az alkalmazáson van, nem a csomag/kód fejlesztésen
- R/Pythonban olyan ernyő csomagok (pl tidyverse) amit relatíve gyorsan el lehet sajátítani
- Emellett az alapabb szintű kódolási ismeretekbe is betekintünk és ott adunk kitekintést a mélyebb programozói megoldások felé.



Lecture 05: Data Exploration

Motivation

You want to know whether online and offline prices differ in your country for products that are sold in both ways. You have access to data on a sample of products with their online and offline prices. How would you use this data to establish whether prices tend to be different or the same for all products?

After collecting the data, assessing its quality, cleaning it, and structuring it, the next step is exploratory data analysis (EDA). Exploratory data analysis aims to describe variables in a dataset. EDA is important for understanding potential problems with the data and making analysts and their audiences familiar with the most important variables. The results of EDA help additional data cleaning, decisions for further steps of the analysis, and giving context to the results of the following hypothesis testing.

The lecture discusses some basic concepts such as frequencies, probabilities, distributions, and extreme values. It includes guidelines for producing informative graphs and tables for presentation and describes the most important summary statistics. Furthermore, we cover the logic and practice of testing hypotheses. We describe the steps of hypothesis testing and discuss two alternative ways to carry it out: one with the help of a test statistic and a critical value, and another one with the help of a p -value. We focus on testing hypotheses about averages, but, as we show in one of our case studies, this focus is less restrictive than it may appear.

This lecture

This lecture introduces students to data exploration. `modelsummary` is used for data descriptive tables, `ggplot2` for creating graphs, and `t.stat` for hypothesis testing. Descriptive statistics and descriptive graphs for one variable are concerned to decide on further data munging. Moreover, simple hypothesis testing is covered as well as association graphs and statistics between two variables.

Case studies connected to this lecture:

- [Chapter 03, A: Finding a good deal among hotels: data exploration](#) - emphasis on one variable descriptive analysis, different data
- [Chapter 03, D: Distributions of body height and income](#) and [Chapter 03, U1: Size distribution of Japanese cities](#) connects theoretical and empirical distributions
- [Chapter 04, A: Management quality and firm size: describing patterns of association](#) - focuses on the association between two variables, one variable descriptive is not emphasized, different data.
- [Chapter 06, A: Comparing online and offline prices: testing the difference](#) - focuses on hypothesis testing, association and one variable descriptive is not emphasized.

This lecture uses [Chapter 06, A](#) as the starting point, but stresses the one variable descriptives such as in [Chapter 03, A](#) and adds the two

Minden órához kidolgozott tananyag:

- Motiváció
- Óra leírása és hogy kapcsolódik a köznyelvhez és annak fejezeteihez
- Tanulmányi kimenetek
- Adatok
- Becsült idő és kulcspontok az órához
- Házi feladatok
- További anyagok
- Mappa és fáj felépítés
 - mindenhol: nyers/órai kódok és megoldások



Használati javaslatok I

- Többnyire előre megírt kódokat tartalmaz, ahol különféle órai feladatok vannak.
- Az elején fontosnak tartjuk az órai kódolósos alkalmakat, de... az előre megírt kódok lehetővé teszik:
 - kódok nagyobb változatosságát, jó példákat mutasson a kódoláshoz
 - több parancsot és funkciót be mutatása
 - a hangsúly ne a syntaxison, hanem a probléma megoldáson
 - alapja legyen a saját elemzésüknek, ahol "csak" az adatokat kelljen megváltoztatni.
- Házi feladatokkal pedig ezeket a syntaxisokat újra gyakorolniuk kell otthon.



Használati javaslatok II

- Jól leképezi azt amilyen irányban halad a kódolás az AI hatására:
 - legyen stabil alap órai kódolásos alkalmakkal, de...
 - nem a syntaxison van a hangsúly, hanem
 - a kód ellenőrzésén és probléma megoldásán
- AI már képes komplett elemzések elvégzésére.
 - de sokszor hallucinál...
 - és tudni kell, hol van jobb megoldás!

Write an analysis in R for hotel prices using price variable as outcome and use distance as the covariate. Do simple regression, create a scatterplot with a regression line and interpret the results.

To perform an analysis of hotel prices with `price` as the outcome variable and `distance` as the covariate, you can follow these steps in R:

Steps:

1. Load the data (assuming you have a dataset with variables `price` and `distance`).
2. Fit a simple linear regression model.
3. Create a scatterplot with the regression line.
4. Interpret the results from the model summary.

Here's a full analysis in R code:

1. Load the Data

```
r Copy code
# Assuming you have a CSV file with the dataset
# data <- read.csv("hotel_data.csv")

# Alternatively, we can create a mock dataset for this example
set.seed(123)
data <- data.frame(
  distance = runif(100, 0, 20), # distances between 0 and 20 km
  price = 100 + -5 * runif(100, 0, 20) + rnorm(100, 0, 10) # prices influenced by
)
```


Example course

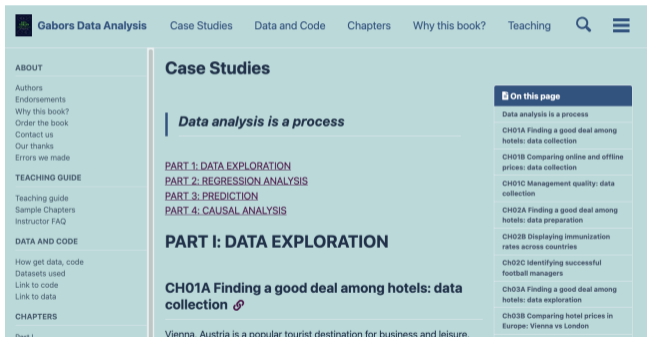
As an example for a coding course, which takes one 100-mins class per week for a semester (12 weeks), we have taught the followings:

Class	Lecture(s)	Comments
Class 01	lecture00-intro , lecture01-coding-basics	Installation of R, RStudio, and <code>tidyverse</code> package along with knitting an RMarkdown is asked to be done before the class. From coding basics some materials (e.g. numeric vs integer vs double, or indexing or lists) are left out if I run out of time.
Class 02	lecture02-data-imp-n-exp , lecture03-tibbles	Sometimes lecture03-tibbles finished on next class.
Class 03	lecture04-data-munging , start: lecture05-data-exploration	Ask about RMarkdown knitting.
		At this point, should assess students that they understand the basics of



Kapcsolat az esettanulmányokkal és adatokkal

- Amennyiben egy specifikus esettanulmányt szeretnénk átnézni, adottak, hogy melyik órákhoz tartozó anyagokat érdemes ismerni, hogy az esettanulmányt könnyen meg tudjuk érteni!
- Melyik kód melyik adatot használja!



The screenshot shows the 'Gabor's Data Analysis' website. The main navigation bar includes 'Case Studies', 'Data and Code', 'Chapters', 'Why this book?', and 'Teaching'. The 'Case Studies' section is active, displaying the title 'Data analysis is a process' and a list of parts: PART 1: DATA EXPLORATION, PART 2: REGRESSION ANALYSIS, PART 3: PREDICTION, and PART 4: CAUSAL ANALYSIS. The 'PART I: DATA EXPLORATION' section is expanded, showing 'CH01A Finding a good deal among hotels: data collection' with a link icon. Below this, the text 'Vienna, Austria is a popular tourist destination for business and leisure.' is visible. A sidebar on the left contains navigation links for 'ABOUT', 'TEACHING GUIDE', 'DATA AND CODE', and 'CHAPTERS'. A right sidebar titled 'On this page' lists various chapters and their topics, such as 'CH01A Finding a good deal among hotels: data collection' and 'CH02B Displaying immunization rates across countries'.



Kitekintés és összefoglalás

- Könyvvel együtt bevezet és felkészít az empirikus elemzésbe.
- Kódolás elkerületetlen ma a munkapiacon.
- Teljes, ingyenes kurzus! Ebben a formában unikális!
- Megadja a kapcsolódást az ismeretek további elmélyítéséhez:
 - verzió követés (git)
 - haladóbb programozás
- Ha igény van rá, fordítás és különálló honlap a könnyebb követhetőségért.

