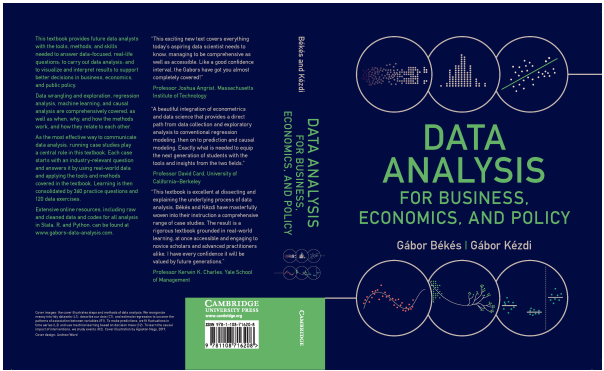
Az adatelemzés hét lépése

Békés Gábor (Central European University, HUN-REN KRTK KTI, CEPR)

HUN-REN KRTK – Közgazdaságtudományi Intézet

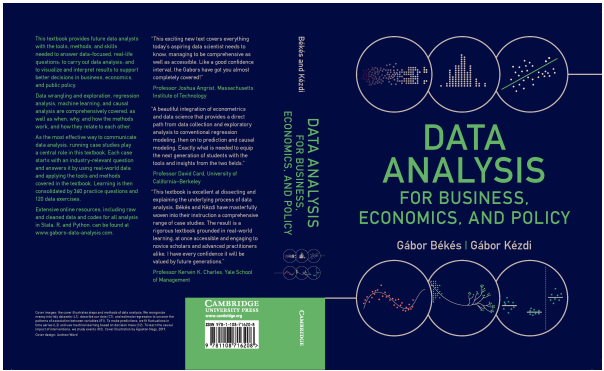
2024-10-10

Adatelemzés tankönyv



- Cambridge University Press, 2021
- cambridge.org/bekeskezdi
- gabors-data-analysis.com
- github.com/gabors-data-analysis/da_case_studies

Adatelemzés tankönyv és Kézdi Gábor



Az adatelemzés 7 lépése

Az adatelemzés folyamata

1. Kezdés egy kutatási témával és egy specifikus kutatási kérdéssel
2. Az adatgyűjtés az alapja minden empirikus munkának
3. Az adatok tisztítása és rendszerezése szükséges és időigényes feladat
4. A feltáró adatelemzés segít az adatok előkészítésében és elemzésében
5. Az analitikai munka hipotéziseket tesztel és modelleket becsül
6. Az eredmények közlése felhasználóbarát módon elengedhetetlen
7. A végén választ adunk az eredeti kérdésre és megvitatjuk az általánosíthatóságot

Az adatelemzés 7 lépése - egy folyamat

1. Kutatási kérdés
2. Adatgyűjtés
3. Az adatok tisztítása és rendszerezése
4. Feltáró adatelemzés
5. Analitikai munka
6. Eredmények közlése
7. Válasz

Erről a beszélgetésről

- Az adatelemzés 7 lépésének áttekintése
  - Ugyanazt az esettanulmányt használjuk a szemléltetésre
  - Kiemeljük a technikai kifejezéseket
- Az adatelemzéshez 7 eszközt vizsgálunk meg, egyet minden lépéshez
  - Linkeket megosztom
    https://gabors-data-analysis.com/talks-thesis#tools

1 Minden egy kérdéssel kezdődik: Egy témából x, y és z lesz

- Keress egy témát, ami érdekel vagy amelynek az eredményeire kíváncsi vagy
- Találj egy specifikus kérdést – gyakran két változó, y és x közötti kapcsolatról
  - Kezdd a mérési kérdésekkel

- Néha a munka egy új minta felismeréséről szól
- Ha lehetséges, fordítsd le egy oksági kérdésre – gondolj egy beavatkozásra
  - Találj egy y (kimeneti) és x (kezelés, oksági változó)
  - Az adat egy RCT kísérletből származik – legegyszerűbb. Véletlen hozzárendelés
  - Az elemzés egy természetes kísérleten alapul – nehéz megtalálni, könnyű végezni
  - Megfigyelési adatok – könnyen megtalálható, nehéz elemezni

- Megfigyelési adatoknál az oksági hatást a nehezebb úton kell elkülöníteni
  - Gondoljon z változókra, amelyek megakadályozhatják az oksági elemzést (pl. összemosó változók)

1 Esettanulmány: Egy témából x, y és z lesz

- Miért jobb néhány vállalatnál a vállalatvezetés minősége?
- Alapító/családi tulajdon és a vállalatvezetési minőség

1 Esettanulmány: Egy témából x, y és z lesz

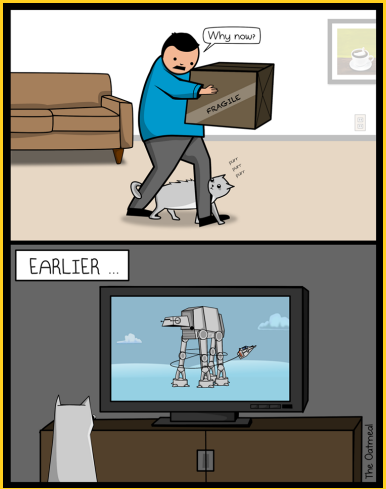
- Miért jobb néhány vállalatnál a vállalatvezetés minősége?
- Alapító/családi tulajdon és a vállalatvezetési minőség

- Jobb lesz-e a vállalatvezetés minősége annak hatására, ha az alapítók tulajdonosok maradnak?
  - Gondolatkísérlet: vegyünk alapító tulajdonban lévő vállalatokat, adjunk el véletlenszerűen részesedéseket, és nézzük meg, mi történik később

- y (kimeneti változó) a vállalatvezetés minősége, és x (oksági változó) a tulajdon
- Összemosó változók z: intézmények...
- Az adatokat egy, a vállalatvezetési minőséget mérő felmérésből gyűjtjük

1 Kulcspont: Legyen egy érdekes kérdés, és mérd meg

- Egy érdekes kérdés kell
- Amíg nem tudod, mi az y és mi az x, és nem tudod, hogyan lehet ezeket mérni, addig nincs projekted



2 Az adatgyűjtés az empirikus munka alapja

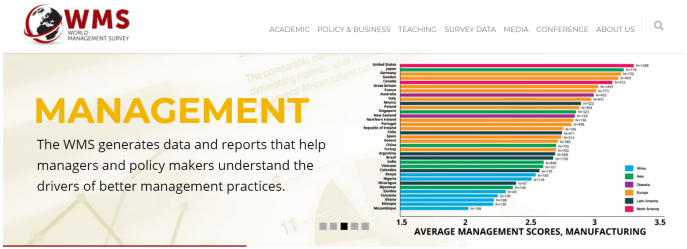
- Két mód arra, hogyan gondolkozzunk a kutatási kérdésről és az adatgyűjtésről

- A: Kérdés megfogalmazása és megfelelő adatok gyűjtése a válaszhoz
- B: Annak felmérése, hogy a rendelkezésre álló adatok segíthetnek-e a válaszadásban.

- Az adatgyűjtés sokféle formája létezik
  - Adminisztratív adatok – nagy, de nehéz hozzáférni
  - Online adatok 1: Letöltés/API – bizonyos esetekben remek, de nem mindig elérhető
    - Számos kiváló forrás: World Bank, FRED, EBRD, US Census, Kaggle stb.
    - Lásd gyűjteményemet is gabors-data-analysis.com/data-source-ideas
  - Online adatok 2: Webszkréping – nagyszerű, a tisztítás fárasztó, némi programozási készség kell hozzá
  - Felmérés – fókuszált, időigényes, nehéz előre megmondani, hogy működik-e

2 Esettanulmány: Vállalatvezetési minőség adatgyűjtése

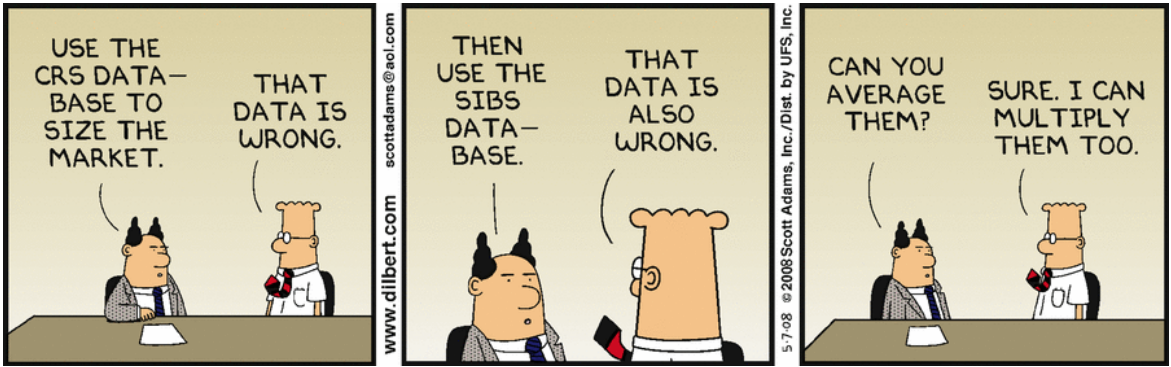
- World Management Survey (WMS) – központosított kérdések, globális `www.worldmanagementsurvey.org` – felmérés vállalatokról és menedzsmentről.
- Pontozólap 18 monitoring, célkitűzési és ösztönzési gyakorlat, mint pl. lean menedzsment
- Vállalatvezetési minőség = pontszám (átlag)
- Standardizált. Pilot tanulmány. Nyilvános.



2 Kulcspont: Az egyedi adatbázis hatalmas előny

- Egy egyedi adatbázis, ha megfelelő minőségű, hatalmas előny
- Webszkréping, felmérés, különböző források adatainak összekapcsolása

Kitérő: A rossz adatokkal való munka kellemetlen lehet



3 Az adatok tisztítása és rendszerezése szükséges és időigényes feladat

Adattisztítás az a folyamat, amely a nyers adatokat adattáblákká alakítja át, amelyek különféle későbbi célokra, például elemzésre használhatók. Tele van döntésekkel.

Megértés és tárolás

- kezdés a nyers adatokból
- az adatszerkezet és tartalom megértése
- kapcsolatok megértése a táblák között
- nagy adatok - mérnöki megoldások

Adattisztítás

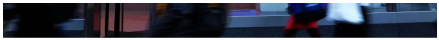
- ismérvek, változótípusok megértése
- duplikátumok szűrése
- hiányzó megfigyelések keresése és kezelése
- korlátok megértése

3 Esettanulmány: A WMS adatok előkészítése

- Hibák és furcsa értékek ellenőrzése.
  - Iskolázottság évei, numerikus változó, 999 jelentése hiányzó adat
- Hiányzó értékek törlése vagy imputálása
  - A megfigyeléseket töröltük, amikor a kulcsváltozók hiányoztak (14%).
- Szűrés a cél érdekében
  - Néhány céget kizártunk, ahol kevesebb, mint 50 vagy több, mint 5000 alkalmazott volt (3%).

- Néhány döntés szükséges az elemzéshez.
- Néhány döntés önkényes.

Kitérő: Változók tárolása: Példa a Washington Post-tól (2016)



(Jewel Samad/AFP/Getty Images)

By **Christopher Ingraham**
August 26, 2016

A surprisingly high number of scientific papers in the field of genetics contain errors introduced by Microsoft Excel, according to an analysis recently published in the journal Genome Biology.

A team of Australian researchers analyzed nearly 3,600 genetics papers published in a number of leading scientific journals — like Nature, Science and PLoS One. As is common practice in the field, these papers all came with supplementary files containing lists of genes used in the research.

The Australian researchers found that roughly 1 in 5 of these papers included errors in their gene lists that were due to Excel automatically converting gene names to things like calendar dates or random numbers.

[This new model for training scientists could create a conflict of interest]

You see, genes are often referred to in scientific literature by symbols — essentially shortened versions of full gene names. The gene "Septin 2" is typically shortened as SEPT2. "Membrane-Associated Ring Finger (C3HC4) 1, E3 Ubiquitin Protein Ligase" gets mercifully shortened to MARCH1.

What you type	What you see	How Excel stores it
MARCH1	1-MAR	42430
SEPT2	2-SEP	42615

https://www.washingtonpost.com/news/wonk/wp/2016/08/26/an-alarming-number-of-scientific-pa

3 Kulcspont: A reprodukálható adattisztítás elengedhetetlen, időigényes

- Az elemzési projektek idejének körülbelül 80%-a adattisztítással és adatfeldolgozás telik.
- "Adat és kód nélkül nem történt meg".

4 A feltáró adatelemzés segíti az előkészítést és az elemzést

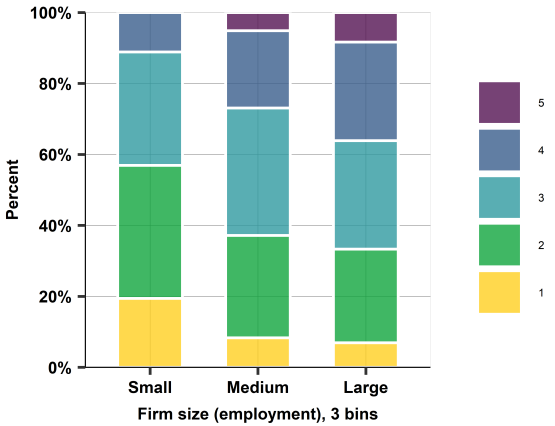
- Feltáró Adatelemzés (EDA)
- Kapcsolódik az adatok előkészítéséhez
  - Kontextust ad a végső eredményekhez.
  - Segít eldönteni az alkalmazandó elemzési módszer részleteit.
- Első fontos (leíró) eredményeket hoz létre.
- Mélyebb kutatást irányít.
- Feltételes átlagok és eloszlások összehasonlítása.
  - Táblázatok, grafikonok.

4 Esettanulmány: feltáró adatelemzés

- Pilot tanulmány: a minta tervezése
  - Értsük meg az eloszlásokat, mérjük a minőséget
  - Csoportosítsuk alágazatok szerint: iparág, ország
  - Csoportosítsuk tulajdonformák szerint – döntsük el, mit tartsunk meg és mit ne
    - Folyamat: esetleg térjünk vissza az oksági gondolkodáshoz és a tisztításhoz.
- Írjuk le a mintákat
  - Mutassuk a menedzsment minőség és a tulajdonformák közötti korrelációkat
  - Mutassunk korrelációkat néhány z változóval
    - Folyamat: az eredményektől függően: térjünk át az elemzésre vagy vissza az oksági gondolkodáshoz.

4 Esettanulmány – Vállalatvezetési minőség és vállalatméret

- Lean menedzsment pontszám 1–5
- Méret: kicsi, közepes, nagy
- Feltételes valószínűség:
  - a kicsi vállalatokra feltételes pontszám=1 aránya körülbelül 20%
  - a nagy vállalatokra feltételes pontszám=5 aránya körülbelül 10%
- Kapcsolati mintázatot mutat



Megjegyzés: *Forrás: A vállalatvezetési minőség 18 változó átlagolt pontszáma. A vállalatméret az alkalmazottak száma.* `wms-management-survey` adatok. Mexikói minta, n=300.

4 Kulcspont: sokat dob egy jó leíró táblázat vagy grafikon

- Gyakran egy jó leíró táblázat vagy egy regressziós egyenessel ellátott pontdiagram meggyőzi az olvasókat, hogy valami érdekes történik.
- Mindenképpen informatív

5 Az elemzés modelleket tesztel és becsül

- Cél gyakran a okság megközelítése

- Keresztmetszeti adatok OLS – gondosan átgondolni a okságot
- Különbség a különbségekben – lehetséges, hogy a változás más tényezők miatt következik be?
- Panel fix hatások és esemény tanulmányok
  - amikor a beavatkozás időben változik vagy gyakran fordul elő
  - gyakran ez a legtöbb, amit megtehetünk megfigyelési adatokkal
- Párosítás (matching) – nagyszerű módszer a közös támogatás biztosítására
- Regressziós diszkontinuitás – remek, ha sikerül találni egyet
- Instrumentális változók – ritkán működik meggyőzően.
  - Hacsak nincs randomizálás a háttérben.

Kitérő: magyar fordítási munka

- Nagyon fontos de nehéz a magyar fordítás
- Első 12 fejezet: gabors-data-analysis.com/dictionary-hun
- Fontos folytatni...
  - Igények, források: email Horn Dánielnek.

5 Esettanulmány: OLS és párosítás

- Keresztmetszeti adatok – OLS, párosítás
  - párosítás a legközelebbi szomszédra: kezelt megfigyelések csoportja számára hasonló jellemzőkkel rendelkező kezeletleneket talál
  - Itt: iparág, ország, vállalat kora, technológia típusa szerint csoportosítva
    - Algoritmus
- Nagyon hasonló eredmények – a párosítás azt sugallja, hogy bizonyos típusú cégeket, amelyek csak családi vagy csak állami tulajdonúak, kizárunk.
- A párosítás fő előnye, hogy felismerjük, vannak olyan cégtípusok, amelyeknek nincs hasonló megfelelőjük.

5 Kulcspont: A oksághoz való közeledés kemény munka

- Néha találhatunk egy okos trükköt, mint az RDD.
- Gyakran szembe kell néznünk azzal, milyen messze vagyunk az oksági értelmezéstől.

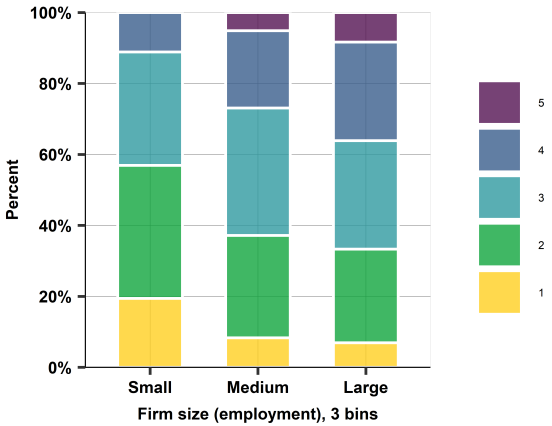
6 Az eredmények felhasználóbarát módon történő közlése

- Az eredmények értelmezése és hatékony bemutatása
- Az adatvizualizáció összegzi az eredményeket / üzeneteket közvetít.

- Vannak szabályok és elérhető sok segítség a jó táblázatok és grafikonok készítéséhez
  - Segíts, hogy a felhasználó megértse, szabd a közönséghez
  - Biztosítsd, hogy a ábrához az állványozás is ott legyen.

6 Esettanulmány – A grafikon tervezése

- Grafikon felépítése: három csoport a vállalatméret alapján
- Dönts a grafikon típusáról: Összevont oszlopdiagram a relatív gyakoriságok megjelenítésére
- Válassz egy színsémát (viridis)
- Adj hozzá megjegyzést, amely tartalmazza a kulcsfontosságú információkat, mint például alhalmaz, N, a változó definíciója



6 Kulcspont: Fejleszd az ábrázolási készségeidet

- A jó grafikonok készítése gyakorolható és fejleszthető.
- Nagyon hasznos készség a való életben.

7 Válaszolj az eredeti kérdésre és vitasd meg az általánosíthatóságot

- Válaszold meg a kérdést
  - Pontosan. A legjobbnak tartott modelledből
  - Általánosabban
- Állást kell foglalni és megvitatni, hogyan értelmezed az eredményeket. Megbízható? Oksági?
- Általánosítás arra az adatkészletre, ami érdekel
  - Statisztikai következtetés: SE, CI, p-értékek a populációban
  - Külső érvényesség: az adathalmazon és a populáción túl
- A statisztikai következtetés és a külső érvényesség egyaránt fontos.
  - Néha kompromisszumokat kell kötni.

7 Esettanulmány: eredmény és értelmezés

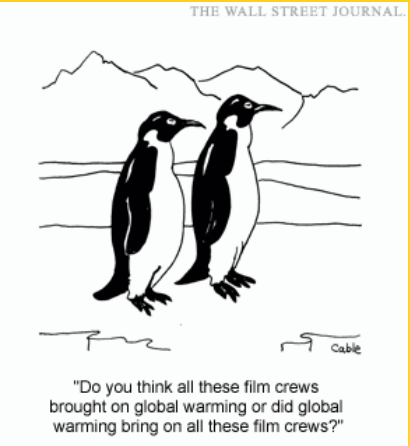
- A vállalatvezetés minősége átlagosan körülbelül 30%-kal alacsonyabb az alapító/családi tulajdonban lévő cégeknél, mint más cégeknél
  - ugyanabban az országban, iparágban, méretben, korban, azonos arányú felsőfokú végzettséggel rendelkező munkavállalókkal és hasonló számú versenytárssal.
- Az állami tulajdon szorosan kapcsolódik a vállalatvezetés minőségéhez, valószínűleg oksági kapcsolat van.
- Magas nem megfigyelt szórodás – nem lehetünk biztosak.

7 Kulcspont: Mutasd be az eredményt és vitasd meg a problémákat

- Okság és a belső érvényesség: őszinteség az eredményekkel kapcsolatban.
- Külső érvényesség: mire számíthatsz, ha a modellt az adatokon kívül alkalmazod.
- Ha össze tudod foglalni az eredményeket néhány tweetben, akkor van egy cikked.

7 Kulcspont: Mutasd be az eredményt és vitasd meg a problémákat

- Okság és a belső érvényesség: őszinteség az eredményekkel kapcsolatban.
- Külső érvényesség: mire számíthatsz, ha a modellt az adatokon kívül alkalmazod.
- Ha össze tudod foglalni az eredményeket néhány tweetben, akkor van egy cikked.

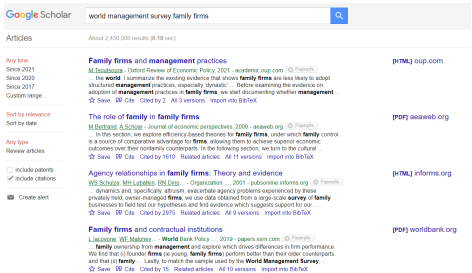


Eszközök, amelyek segítik a folyamatot

- Rengeteg technológia és eszköz áll rendelkezésre az adatelemzési folyamat segítésére.
- Tekintsünk át néhányat mind a hét lépéshez!

1 Ismerd meg a témád, és kezeld a forrásokat

- Cikkek gyűjtése és kezelése
- Számos eszköz áll rendelkezésre a bibliográfia és a források kezelésére, mint például: Paperpile, vagy Zotero.
- repec, Google Scholar továbbra is a leghasznosabb



2. Online felmérések készítése

- Adatgyűjtés felméréssel a legrégebbi adatgyűjtési módszer
- Számos online platform segíthet
- Az adatgyűjtés nehéz, mert
  - a kérdések megírása és tesztelése nehéz
  - alacsony válaszadási arány

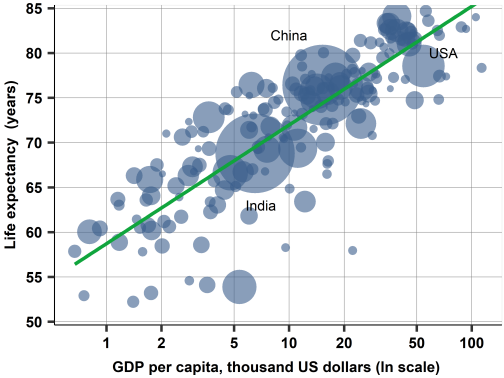
- surveymonkey.com
- docs.google.com/forms
- Qualtrics

3. Kódkörnyezetek adatkezeléshez és elemzéshez

- Kódolás reprodukálható kutatás érdekében
- Stata, R, Python (+Matlab, Gretl, SPSS, SAS, Julia)
  - Stata: akadémia, NGO, kormányzati szektor gazdag országokban
  - R: akadémia, kormányzat, statisztika, tanácsadás, újságírás
  - Python: számítástechnika, pénzügy, akadémia
- Kódkörnyezetek nagyban segítenek
  - Rstudio az R-hez tervezve
  - Jupyter notebook a Pythonhoz tervezve
  - Quarto (Posit) notebook mindkettőhöz – egyszerű pdf, prezentáció, weboldal
  - VScode bármilyen nyelvhez

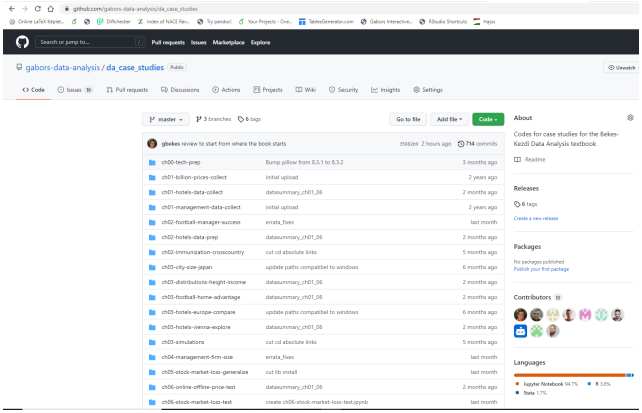
4. Adatfeltárás és vizualizáció GGplot-tal (R) és plotnine-nel (Python)

- Mindenki megtanulhatja a jó grafikon készítésének alapjait...
  - Rengeteg online segítség
- R: ggplot, Python: plotnine (azonos szintaxis), seaborn
  - sokoldalú, megéri befektetni
- Grafikon itt (ggplot): github.com/gabors-data-analysis/da_case_studies and ch08-life-expectancy-income



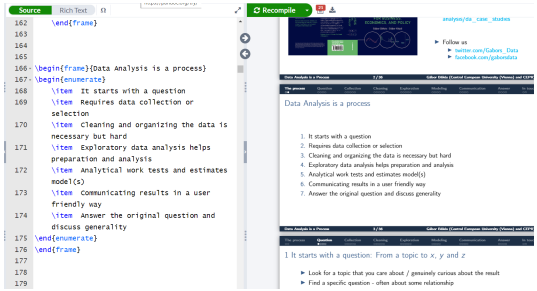
5. Reprodukálható kutatás Git és Github használatával

- Reprodukálható kutatás
- Git egy verziókezelő rendszer
- Github egy felhőalapú kódtároló rendszer, amely a git-en alapul
- A tankönyvünk összes kódja megtalálható a Githubon: github.com/gabors-data-analysis/da_case_studies



6. Szakdolgozat írása és prezentálása

- Tex/latex egy dokumentum előkészítő rendszer (mint az MS Word).
  - A felhasználó teljes ellenőrzése alatt áll.
- Overleaf egy felhő alapú megoldás
  - Egyszerű latex használathoz és együttműködéshez
- Ez a prezentáció latexben készült, Overleaf-ben szerkesztve
  - a tankönyv is
- Markdown is nagyszerű



7 Mesterséges intelligencia

Kódolás

- Github Copilot AI kódolási asszisztens
  - A legtöbb fejlesztőkörnyezetbe beépítve, mint VScode, Rstudio
  - Olvassa a kódot, javaslatokat tesz, befejez kódot
- Claude.ai vagy chatgpt kód megírásához. Hibakeresésre van szükség. Legjobb gyakori problémák esetén.

7 Mesterséges intelligencia

Kódolás

- Github Copilot AI kódolási asszisztens
  - A legtöbb fejlesztőkörnyezetbe beépítve, mint VScode, Rstudio
  - Olvassa a kódot, javaslatokat tesz, befejez kódot
- Claude.ai vagy chatgpt kód megírásához. Hibakeresésre van szükség. Legjobb gyakori problémák esetén.

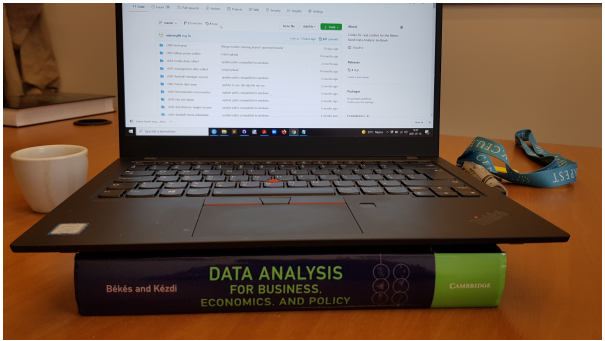
Kutatástámogatás

- ChatGPT: források PDF-ből BibTeX-be való átalakításához
- scite.ai és consensus.app források keresésére
- Számos alkalmazás van, amelyek összefoglalják a cikkeket (de én még kellek)
- Forditás: prezentációra jó alap, szakszövegre közepes
  - Biztos van jobb megoldás már

Eszközök áttekintése

- Keresés Google Scholar, és forráskezélés Paperpile-lal
- Online felmérések készítése a Google Forms és a SurveyMonkey használatával
- Kódkörnyezetek reprodukálható kutatáshoz: R/Rstudio és Python/Jupyter
- Adatfeltárás és vizualizáció a ggplot (R) és a Plotnine (Python) segítségével
- Reprodukálható kutatás a Git és a Github használatával
- Szakdolgozat írása és prezentálása Latexben és a Overleaf segítségével
- Copilots mindenhol

Itt elértek:



- X.com/GaborBekes
- gaborbekes.bsky.social
- LinkedIn/bekesgabor

- facebook.com/gaborsdata
- x.com/Gabors_Data

- Kézdiről: kezdigabor.life/