МОДЕЛИРАЊЕ ПОДАТАКА

Др Ђорђе М. Кадијевић, научни саветник
Институт за педагошка истраживања, Београд
Имејл: djkadijevic@ipi.ac.rs; лична веб страница: www.mi.sanu.ac.rs/~djkadij

Резиме. У оквиру ове теме, која ће се базирати како на теоријском излагању (бар један час) тако и на практичном раду на рачунарима (бар три часа), у основи ће се разматрати моделирање података коришћењем пивот табела и графикона, основних алата пословне интелигенције. Реализација те теме, која повезује садржаје из математике, информатике и статистике, заснива се на чланку

Kadijevich, Dj. M. (2016). Data modeling with dashboards: opportunities and challenges. In Engel, J. (Ed.), *Promoting understanding of statistics about society. Proceedings of the IASE Roundtable Conference, July 2016, Berlin, Germany*. Internet: http://iase-web.org/Conference_Proceedings.php .

Потребна предзнања:

- елементарна знања из математике и статистике (нпр. апсолутне и релативне фреквенце, проценти, типови графикона, просечна вредност);
- елементарна знања у вези са табелама и релационим базама података (поље, слог, реализација упита);
- основе визуелног програмирања (коришћење *drag* & *drop* приступа).

Обавезујући технолошки предуслови за реализацију ове теме:

- сваки учесник самостално ради на свом рачунару који лично обезбеђује;
- на рачунару је инсталирана новија верзија програма *Microsoft Excel*;
- учесник зна да користи тај програм (барем на почетном нивоу).

Оптимални технолошки предуслови за реализацију ове теме (поред претходно наведених и):

- рачунар на коме учесник ради има приступ интернету;
- учесник се претходно регистровао за коришћење *ZOHO* окружења за креирање *dashboards* (https://www.zoho.com/reports/dashboard.html).

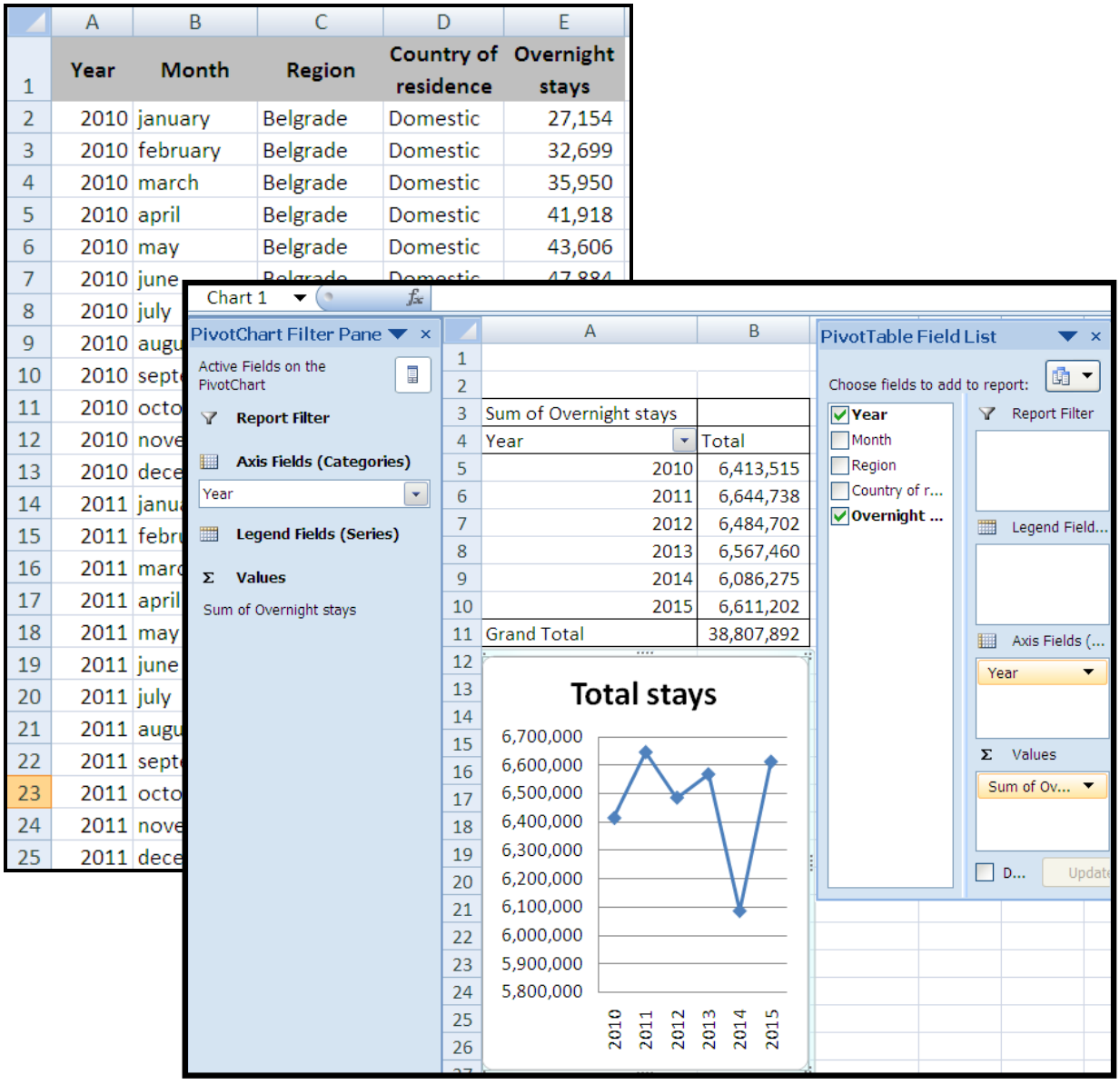
Напомена: Приступ интернету може се обезбедити месечном *pripejd* претплатом на мобилни интернат, али она није услов укључивања у реализацију ове теме. За такву претплату, коју аутор повремено успешно користи последњих неколико година, видети, рецимо, на адреси
https://www.telenor.rs/sr/privatni/ponuda/mobilni-internet/prepaid или
http://www.vipmobile.rs/privatni/internet/prepaid .

Очекивана постигнућа биће у домену:

- припреме података за визуализацију релација у тим подацима;
- визуализације релација у анализираним подацима коришћењем једног или више графикона;
- разумевања важних статистичких појмова (величина ефекта, тренд, интеракција, конфундирајућа варијабла);
- методичког приступа једноставијем моделирању података у настави математике и информатике.

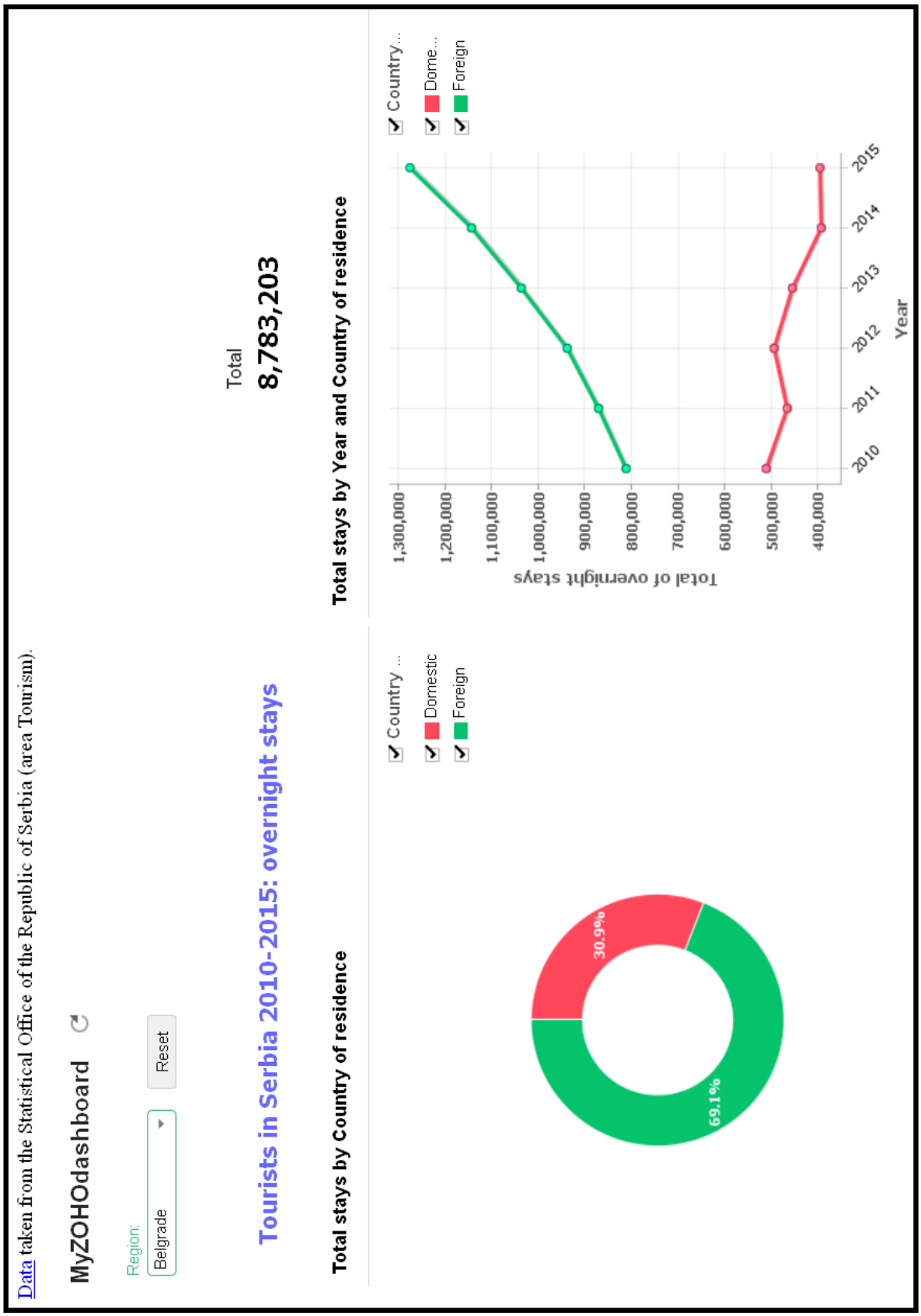
Примери моделирања података:

- *pivot chart*



Интернет: www.mi.sanu.ac.rs/~djkadij/Seminar17.xls

- *dashboard*

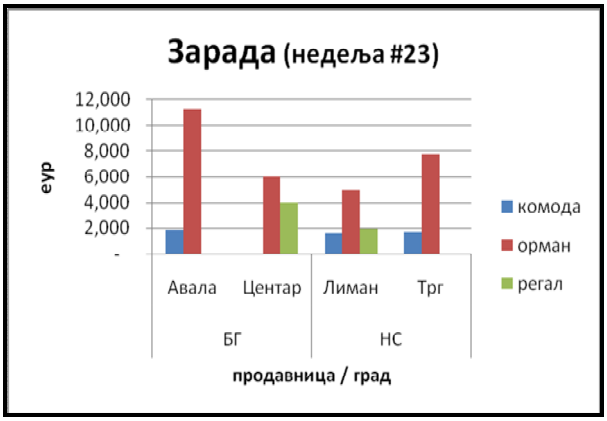


Интернет: www.mi.sanu.ac.rs/~djkadij/Dashboard.htm

Могућности коришћења пивот графикона

Пивот табеле и пивот графикони су једноставни алати пословне интелигенције, који дају сумарне извештаје у облику интерактивних табела и графикона.[#] Ти алати омогућавају да се неки сумарни показатељи (нпр. збир вредности или просечна вредност) приказују у односу на различите нивое детаља. Те детаље корисник бира у складу са својим захтевима и по потреби их сажима или даље рашчлањује.

Размотримо, на пример, следећи графикон у вези са подацима о продаји неких артикала.



Лако се уочава да се тај графикон односи на продају више артикала, да се ти артикли продају у више продавница, као и да се те продавнице налазе у два града. Стога, поред утврђивања укупне зараде (ниво 0), коришћењем пивот приступа можемо анализирати, рецимо, и следеће:

а) укупну зараду по градовима (ниво 1 тј. први ниво детаља по категоријама једне варијабле), да бисмо установили у ком граду је остварена већа зарада,

б) укупну зараду по градовима и продавницама (ниво 2), да бисмо утврдили у којој продавници у ком граду је остварена најмања зарада, као и

ц) укупну зараду по градовима, продавницама и артиклима (ниво 3), да бисмо одредили у којој продавници у ком граду је остварена најмања зарада за разматрани артикал.

Припремни задатак 1 (у оквиру самосталне припреме учесника за учешће на Семинару)
Користећи неки фајл о оствареним зарадама (рецимо www.mi.sanu.ac.rs/~djkadij/LaudonASE1.xls), генерисати графиконе које омогућавају пословне анализе попут претходно наведених. Како методички обликовати реализацију таквих садржаја да бисмо остварили успешан рад на бар два нивоа сложености (једноставнији захтеви наспрам сложенијих захтева)?

[#] Интерактивне визуализације типа *dashboard* у основи се састоје од два или више пивот графикона.

Финални задатак (у оквиру петочасовног рада на Семинару)

У оквиру самосталног рада учесника, планирано је да свако од њих, у складу са темом коју изабере, реализује моделирање података помоћу интерактивних графикона. (Ако се користи *ZOHO* окружење, корисна упутства за моделирање помоћу *dashboards* дата су у фајлу на адреси https://www.youtube.com/watch?feature=player_embedded&v=aPLg4dp-f28 .)

То моделирање би требало базирати на реалним подацима које обезбеђује нека релевантна институција, рецимо Републички завод за статистику, *World Bank,* или *Eurostat*.

Посебну пажњу би требало посветити методичком обликовању реализације таквог моделирања (на конкретном нивоу образовања; нпр. час математике у осмом разреду), која омогућава успешан рад ученика на бар два нивоа сложености (док ће неки ученици за моделирање података имати једноставније захтеве, други ће кроз визуализацију релација у подацима одговарати на сложеније захтеве).

Актуелност теме и њене образовне вредности

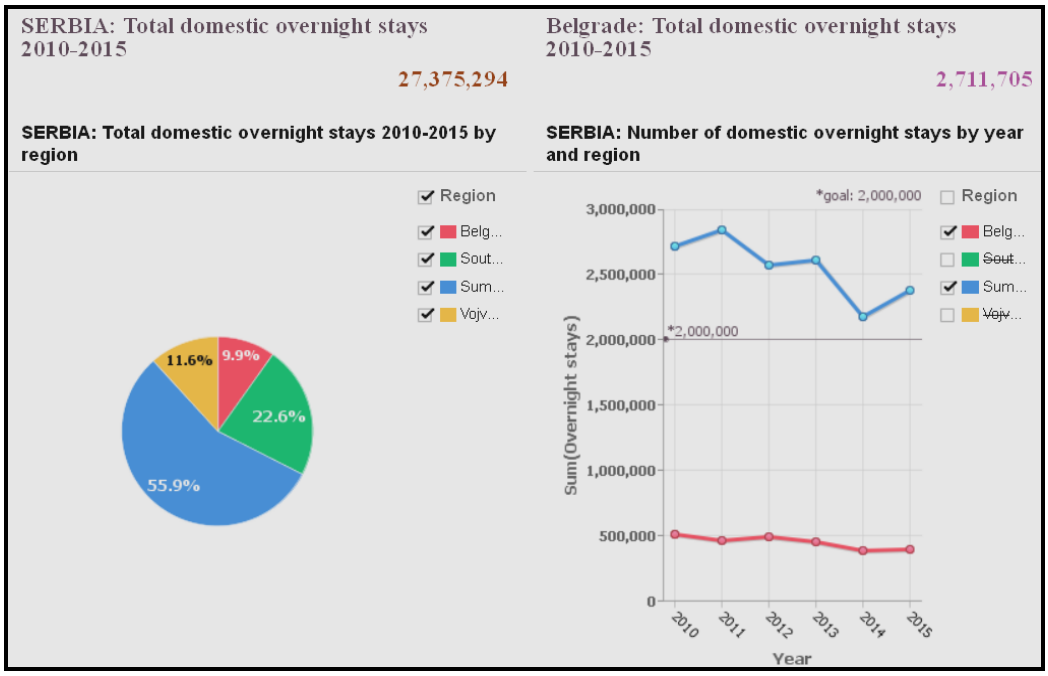
Наука о подацима (енгл. *data science*) — која је на неки начин већ више од пола века присутна у рачунарству (подсетимо се да је термин *datalogy* користио дански научник Питер Наур (1928–2016) пре скоро шездесет година; извор https://en.wikipedia.org/wiki/Data_science) — поново је актуелна последњих година. Имајући у виду све присутније захтеве да се огромне количине података моделирају у пословне, научне или друге сврхе (да би се из њих добиле потенцијално корисне информације које могу довести и до новог знања), не изненађује чињеница да је посао научника који ради са подацима (енгл. *data scientist*) све траженији. Иако статистичари обично кажу да је наука о подацима у ствари статистичка анализа података (тј. *data science* „=“ *data analysis*), то ипак није тачно, јер наука о подацима захтева не само знања и вештине из рачунарства (нпр. из програмирања и основа база података) и математике (примена разноврсних математичких и статистичких модела), већ, између осталог, и висок степен креативности и вештина комуникације. Стога не изненађује да такав посао доноси и високе зараде (и преко 100,000 $ на годишњем нивоу). С обзиром на то да ће некакво моделирање података вероватно бити пристуно у будућем професионалном раду већине ученика, све више је присутан захтев да основно моделирање података буде заступљено и на ранијим нивоима образовања.[##]

[##] Рад са подацима (енгл. *Data*) је једна од области провере знања ученика основне школе из математике, која се, у оквиру четворогодишњих циклуса, реализује у оквиру међународне студије TIMSS (http://timssandpirls.bc.edu/). За разлику од ранијих циклуса тог истраживања у којима су ученици из Србије имали резултат испод и на нивоу међународног просека од 500 поена (2003: 456 поена – VIII разред; 2007: 456 – VIII; 2011: 503 – IV), у последњем тестирању, у 2015. години, ученици IV разреда постигли су изнадпросечан резултат (515 поена). Задаци из области *Data* у IV разреду углавном се односе на табеларно и графичко приказивање података, као и извођење закључака на основу таквих приказа података, а то су важни аспекти елементарног моделирања података.

Образовне вредности моделирања података помоћу интерактивних графикона (а уз примену елементарних модела из математике и статистике) су последњих десетак година у жижи проучавања једног броја истраживача који се баве унапређивањем наставе статистике. Према нашем недавном увиду у литературу у тој области, међу таквим истраживачима предњачи група која ради при центру *Smart* (на Durham University, UK, https://www.dur.ac.uk/smart.centre/). Начин на који ови истраживачи приступају моделирању података приказан је, на пример, у чланку *Visualise then Conceptualise* (који је доступан на адреси https://core.ac.uk/display/23138591). Њихова истраживања указују да таква врста моделирања, и при неформалном приступу, омогућава, између осталог, разумевање важних статистичких појмова, као што су величина ефекта, тренд, интеракција и конфундирајућа варијабла.

Припремни задатак 2 (у оквиру самосталне припреме учесника за учешће на Семинару)
Користећи визуализацију *Cardiovascular Disease Risk Factors* наведеној на интернет адреси https://www.dur.ac.uk/smart.centre/freeware/, сумирати неколико налаза који се односе на неке, по вама битне, релације у анализираним подацима. Експериментишући са том визуализацијом, покушајте да илуструјте неке од горе поменутих статистичких појмова (у том циљу користити подсетник дат на крају овог материјала).

Имајући на уму да се интерактивни објекти типа *dashboard*, између осталог, састоје од два или више пивот графикона, ти објекти омогућавају квалитетније моделирање података него појединачни пивот графикони. То је због тога јер је, између осталог, сада могуће поредити стања на више графикона, као што приказује доња слика.



Истраживачи у настави статистике тек почињу да проучавају начине како да успешно користе интерактивне објекте типа *dashboard* у настави. Упркос томе, а имајући у

виду образовне потенцијале таквог моделирања података, ученицима би требало помагати при креирању и коришћењу таквих објеката као скупова пивот графикона чија се структурна комплексност постепено повећава. При томе би требало имати у виду бројне изазове који се тичу 1) података које треба анализирати, 2) *dashboard* објеката које у том циљу треба креирати и 3) процеса моделирања података које треба реализовати. Ти изазови, које су истраживања у настави статистике до сада углавном занемаривала, детаљније су разматрани у раду који је наведен на првој страници приказа ове теме.

Мали статистички подсетник

- Величина ефекта нам указује која варијабла је више повезана са циљним обележјем. На пример, на основу графикона, уочава се да пушење увећава ризик од инфаркта око два пута, а недовољна физичка активност чак 3–4 пута.
- Тренд нам описује како се вредности једне варијабле мењају током времена. На основу графичког приказа можемо установити да ли те вредности са протоком времена расту или опадају, као и да ли је уочени тренд линеаран или не.
- Интеракција независних варијабли постоји када се ефект једне од њих на зависну варијаблу разликује за различите вредности. Нека се у истраживању, на пример, утврдило да се за разлику од машинског факултета, младићи и девојке који студирају саобраћајни факултет разликују у погледу просечне дужине студирања. Тада постоји интеракција између варијабли пол и факултет. Интеракција се лако уочава при графичком приказивању резултата, јер се растојање између просечних вредности зависне варијабле за вредности једне независне варијабле (нпр. пол) мења променом вредности друге независне варијабле (нпр. факултет).
- Конфундирајућа варијабла је обележје које би могло утицати на релације између разматраних варијабли. На пример, установљена разлика у просечној тежини студената природних и друштвених наука би могла бити резултат старости студената, јер се упоређиване групе студената разликују по просечној старости.