Končno poročilo

Anže Mramor

21 5 2022

```
source("~/Faks/mag 1 letnik/MzR/Time-to-yes-contract-and-money/vizualizacija/vizualizacija_pomembni_gra
```

Uvod

Namen projekta je bil obdelati, analizirati in vizualizirati podatke o vlogah za kredite pri banki. Glavni del se je osredotočal na analizo različnih časov povezanih s krediti:

- čas od oddaje vloge do odobritve (‘time to yes‘)
- čas od oddaje vloge do podpisa pogodbe (‘time to contract‘)
- čas od oddaje vloge do prejema sredstev (‘time to money‘)

Pri projektu sem sodeloval z Novo Ljubljansko Banko, ki mi je tudi priskrbela podatke na katerih sem izvajal analize.

DISCLAIMER: ker gre za občutljive podatke, so vsi podatki anonimizirani oziroma popolnoma izmišljeni (simulirani tako, da bodo sicer odsevali realnost), kakršnekoli povezave z resničnimi podatki o kreditih/strankami ni, saj so vsi resnični podatki poslovna skrivnost banke.

V prvem delu projekta sem pridobil splošne podatke o kreditih jih analiziral in vizualiziral. Najprej sem se osredotočil na vse podatke o strankah. Sledila je analiza in vizualizacija vseh treh opazovanih časov v odvisnosti od ostalih lastnosti posamezne stranke. V zadnjem delu anlize sem s pomočjo tehnik slučajnega učenja zgradil napovedni model, s katerim sem napovedal pričakovano vrednost vseh treh časov za splošno izbiro lastnosti. za zaključek sem vse skupaj povezal in prikazal v aplikaciji zgrajeni s paketom *shiny*.
Komentar: Ker gre za projekt, ki je pretežno iz teme vizualizacije, bo poročilo zelo verjetno presegalo dolžino 10 strani, vendar samo zato, ker bo vsebovalo veliko slik (grafov). Število grafov v poročilu sem zmanjšal na tiste najnujnejše, vse generirane grafe si lahko ogledate v datoteki *porocilo_projekta.Rmd*.

Osnovna analiza podatkov

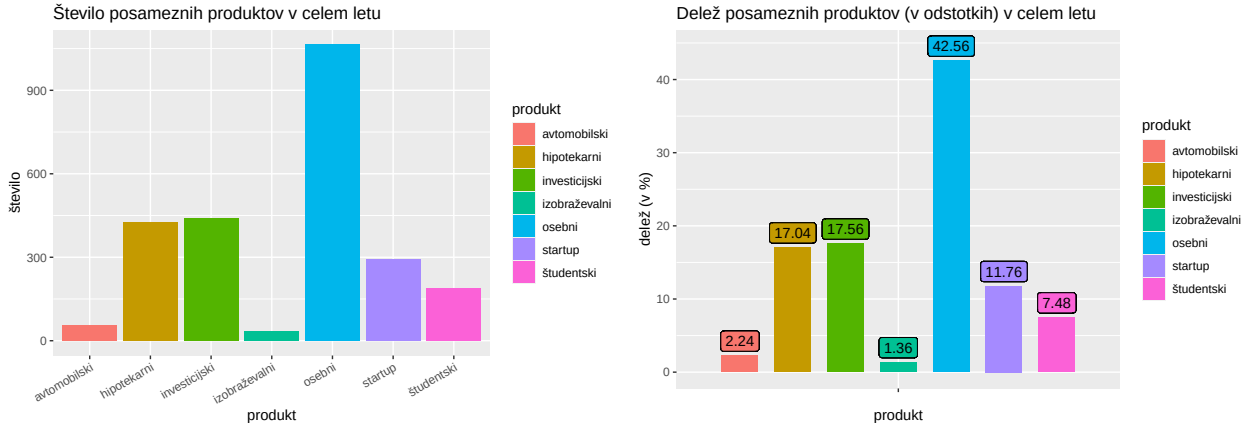
Vizualizacija in analiza podatkov je lahko ponavadi precej suhoparna, zato je zelo pomembno, da znamo iz njih izluščiti zgodba, ki nam jo podatki in grafi predstavljajo. Zato, da sem se s podatki spoznal sem najprej razčlenil za kakšne podatke pravzaprav gre. Podatki vsebujejo 10 različnih stolpcev in 2500 vrstic. Recimo, da opazujemo vloge za kredit v izmišljeni banki, ki jo poimenujemo AMBanka, v izmišljeni državi Pandamiji. Glede na stolpec `mesec` lahko privzamemo, da opzaujemo delovanje banke znotraj enega leta (saj leto ni nikjer določeno, vsi meseci pa so prisotni v stolpcu). Poleg tega lahko iz stolpcev `regija` in `poslovalnica` takoj opazimo, da banka posluje v 2 različnih regijah države - vzhodni in zahodni, ter v 7 različnih poslovalnicah. Nadalje lahko opazimo, da se 5 izmed teh poslovalnic nahaja v vzhodni, 2 pa v zahodni regiji. Če si sedaj ogledamo stolpec `produkt` lahko opazimo, da banka ponuja kredite v obliki sedmih različnih produktov - avtomobilski, hipotekarni, investicijski, izobraževalni, osebni, startup in študetnski. Vsak kredit lahko zavzame tudi drugačno obliko tipa podano v stolpcu `tip` - pri kreditu gre lahko za novo vlogo, za obnovo vloge, podaljšanje vloge ali spremembo vloge.

Prvi izmed preostalih stolpcev lastnosti vlog je stolpec `ID`, ki vsebuje 1038 različnih vrednosti, kar nam očitno pove, da v celotni državi obstaja le 1038 različnih oseb, ki je v opazovanem letu oddalo vloge za pridobitev kredita. Ker je podatkov za vizualizacijo že tako relativno malo, sem se na tej točki odločil, da bom za potrebe analize in vizualizacije atributov privzel, da so vsi krediti (tudi od istih oseb) med sabo neodvisni, ter imam tako samo 2500 podatkov o neodvisnih kreditih. S problemom več enakih `ID` se bom ukvarjal kasneje pri napovedovanju časov s strojnim učenjem, kjer neodvisnosti podatkov ne moremo tako enostavno privzeti.
Zadnji izmed stolpcev lastnosti pa je stolpec `znesek`. Med vrednostmi tega stolpca lahko hitro opazimo, da so vse vrednosti med 1 in 870. Če pobrskamo po svoji domišljiji je to logično, saj v državi Pandamiji poslujejo s posebno valuto - pandacoin, katere menjalni tečaj je 1 € = 1149,43 PC. Poleg tega vemo, da banke v opazovani državi po zakonu ne smejo posojati kreditov, katerih vrednost presega 1.000.000 € oziroma 870 pandacoinsov, ter da lahko odobravajo samo kredite v celoštevilskih vrednostih.

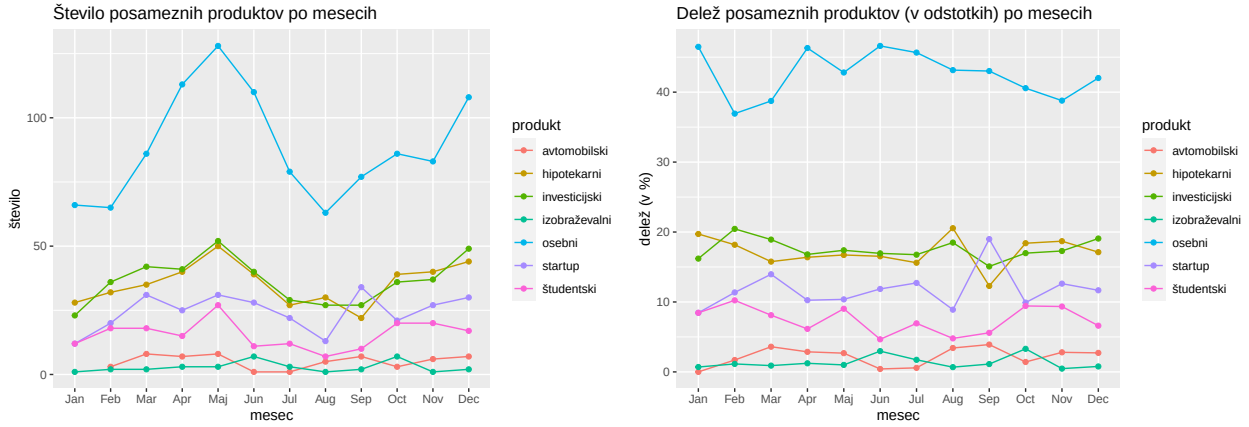
Nazadnje analizirajmo še vse tri stolpce s časi. Prva stvar, ki jo opazimo je, da so časi očitno podanu kumulaticvno - torej vsak čas se začne šteti z dnem oddane vloge, njihove končne točke pa so seveda različne. Poleg tega opazimo tudi, da vsi časi zavzamejo vrednosti manjše ali enake 100 dnem, kar bo še posebej pomembno pri napovedovanju časov s strojnim učenjem. Ob natačnejšem pregledu podatkov pa lahko opazimo tudi zanimivo lastnost, da so v nekaterih primerih časi do odobritve in časi do podpisa pogodbe lahko enaki 0 - torej je vloga odobrena in podpisana podpisana v istem dnevu kot je oddana. Ker so časi kumulativni je seveda nujno, da če do podpisa pogodbe pride v istem dnevu, je bila tudi odobrena isti dan, obratno pa seveda ni nujno res. Zanimi je opaziti še, da noben izmed časov do prejema sredstev ni enak 0, torej ne glede na hitrost birokracije na prejem denarja pri opazovani banki vedno čakamo vsaj 1 dan.

Analiza lastnosti podatkov, brez časov

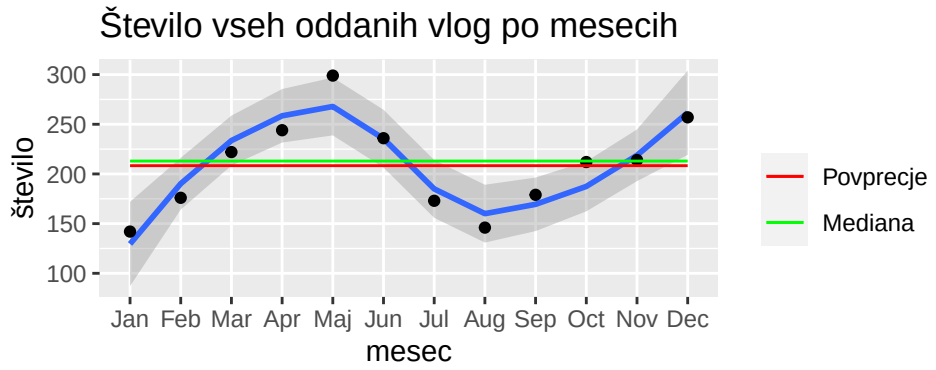
Najprej analizirajmo lastnosti vseh vlog, brez povezav s časi, zato da dobimo boljši občutek s kakšnimi podatki sploh imamo opravka.



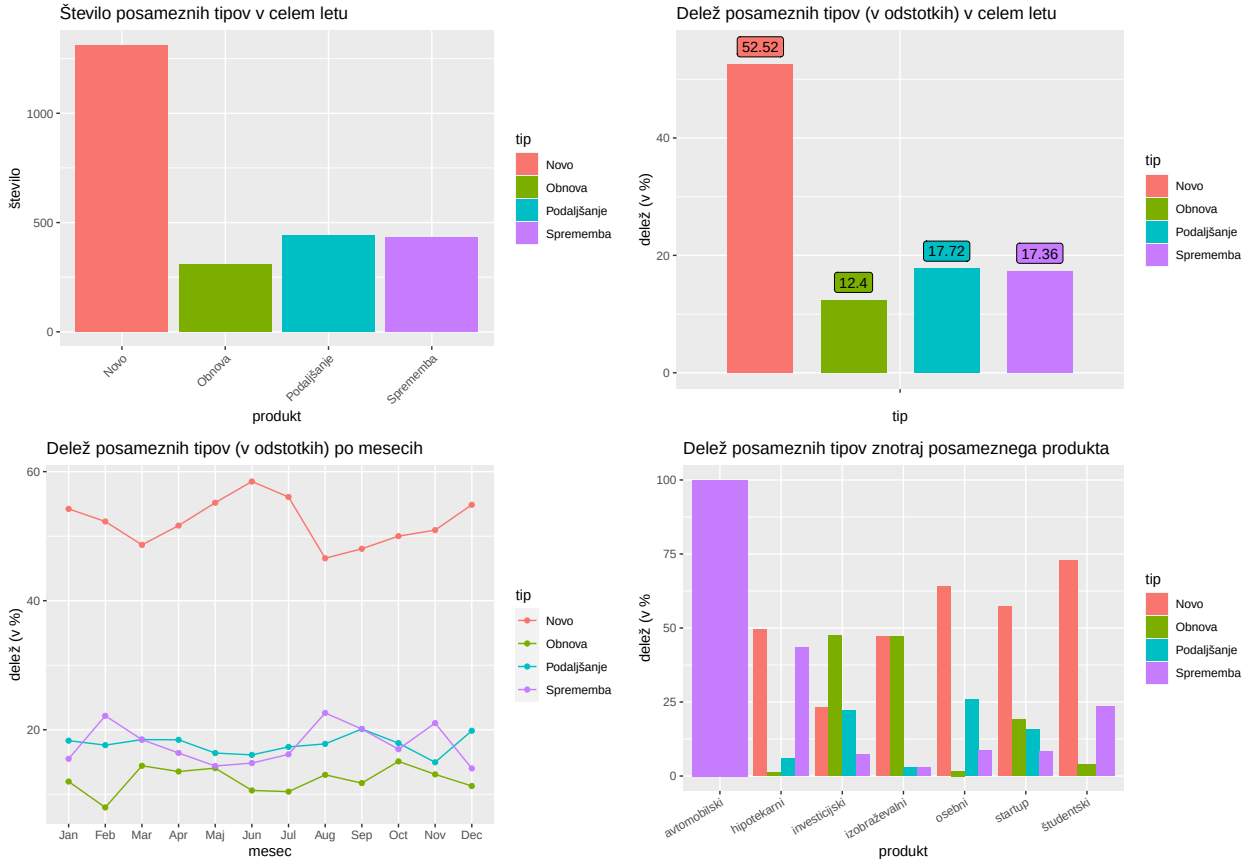
Prva dva grafa zaporedoma prikazujeta število in delež posameznih produktov v celem letu. Precej očitno prevladujejo osebni krediti, ki sestavljajo več kot 40 % vseh produktov, pri katerih lahko predvidevamo da gre za kredite manjših vrednosti, ter so zato toliko bolj pogosti. Sledita jim hipotekarni in investicijski kredit, pri katerih gre tipično za višje vrednosti, ki pa predstavljajo nujo v vsaki družbi - za razvoj podjetij ter za financiranje nakupov stanovanj oziroma hiš. Najmanjši delež kreditov v opazovani banki je avtomobilksih in izobraževalnih, kar je logično saj je Pandamija izrazito zelena država, v kateri prebivalci pogosto uporabljajo javni prevoz ali rekreativna prevozna sredstva (kolesa, skiroji, rolerji) prav tako pa zagotavljajo javno šolstvo vsem državljanom, tako da potrebe po takšnih kreditih pravzaprav ni.



Na zgornjem grafu sta prikazan še skupno število in delež posameznega produkta po mesecih za celo leto. Če opazujemo zgolj deleže se kakšnega izrazitega prevladovanja po mesecih ne da opaziti, skozi celo leto izgleda delež posameznega produkta približno konstanten (z izjemo nekaj manjših skokov). Po drugi strani pa lahko pri opazovanju števila produktov po mesecih opazimo velik porast v oddanih vlogah predvsem v osebnih, pa tudi v hipotekarnih, investicijskih in študentskih kreditih. Pri prvih in zadnjih gre verjetno predvsem za zagotavljanje zadostne količine sredstev za brezskrbno preživljanje počitnic, ter poletne prenove stanovanj, saj imajo takrat vsi največ časa za to. Število oddanih vlog doseže dno konec poletja, ko se vsi vrnejo s počitnic in jih začnejo odplačevati, nato pa proti koncu leta spet počasi naraščajo.



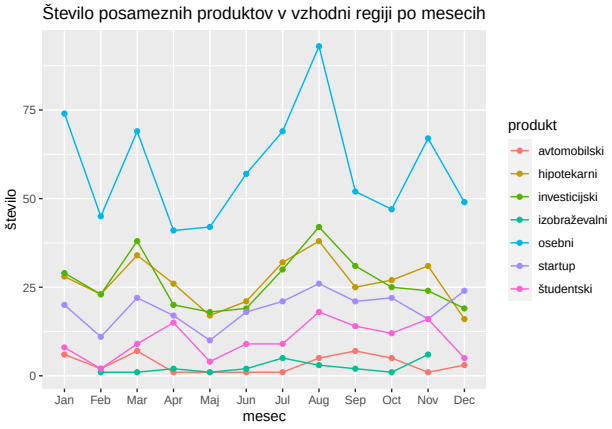
Zgornji graf prikazuje še število vseh oddanih vlog po mesecih, čez podatke je potegnjena krivulja z metodo `geom_smooth` pri 0.95 % intervalu zaupanja. Na prvi pogled bi lahko ocenili tipičen harmonični (sinusni ali kosinusni) trend na podatkih, vendar ta sam po sebi nima pravega smisla, še posebej ker bi potem pričakovali postopno zmanjšanje in ne tako izrazitega padca v januarju. Opazimo lahko tudi, da vrednosti pri večini mesecev ne odstopajo za več kot 50 vlog od povprečja ali mediane, tako da lahko privzamemo, da je število oddanih vlog precej konstantno tudi skozi celotno leto.

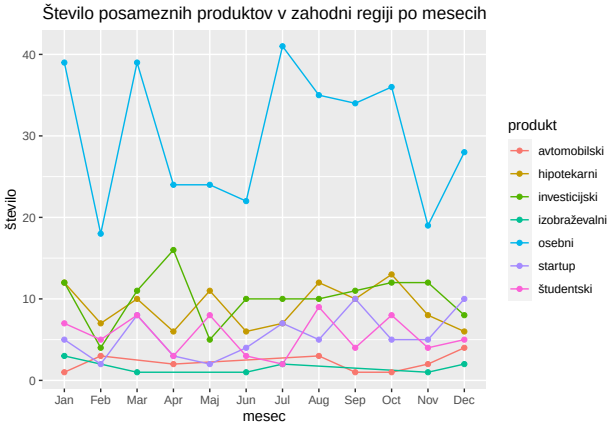


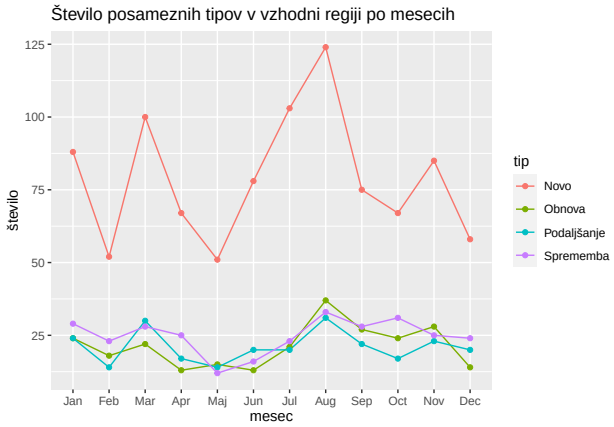
Pri obravnavi tipov hitro opazimo, da največji delež (več kot 50 %) vseh vlog predstavljajo vloge ki so oddane na novo, najmanjši pa tiste pri katerih gre za obnovo kredita. To je seveda smiselno, saj je večina kreditov običajno sklenjenih na novo, občasno pa se zaradi različnih razlogov lahko seveda pojavljajo potrebe po njihovi spremembi, podaljšanju oziroma obnovi. Ker nimamo podatkov o tem kako pomemben je banki posamezen tip, kaj več o njih ne moremo povedati, lahko pa si ogledamo še druga dva zanimiva grafa. Na prvem lahko opazimo, da so podobno kot pri produktih, tudi pri tipih deleži skozi celotno leto približno konstantni, le da je tu nekoliko bolj opazen dvig deleža novih kreditov konec pomladi in v začetku poletja. Drugi

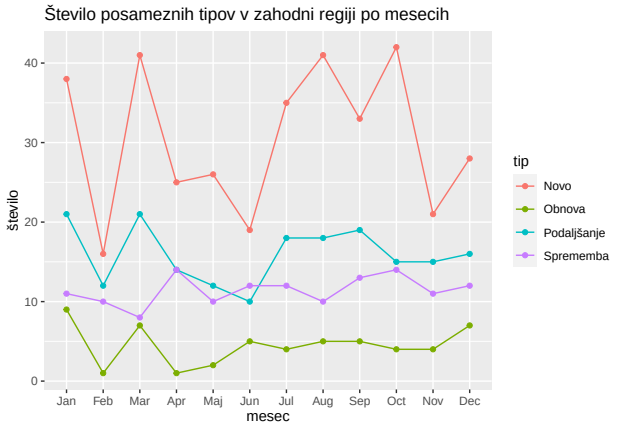
graf pa prikazuje delež posameznih tipov znotraj posameznih produktov. Pri tem ponovno ni presenetljivo, da pri skoraj vseh produktih prevladuje nov tip kredita, je pa zanimivo, da so vsi avtomobilski krediti tipa sprememba. Ker vemo, da gre za simulirane podatke, tega ne moremo pripisati drugemu kot neposrečenemu slučajnemu vzorcu. Pričakovali bi tudi, da bi podaljšanje igralo večjo vlogo, vendar ga pri večini produktov prevlada obnova. Ponovno, ker ne vemo kako v praksi za opazovano banko pomeni posamezni tip to težko točneje komentiramo.

Na zgornjem grafu lahko opazujemo število posameznih produktov v posamezni regiji. Kot pričakovano, je v vzhodni regiji precej več vlog za kredite, saj je očitno večja oziroma je v njej več poslovalnic. Drugih presenečenj na grafu ni, vse kar o številu posameznih produktov velja za celotno državo, velja tudi za vsako njeno posamezno regijo.

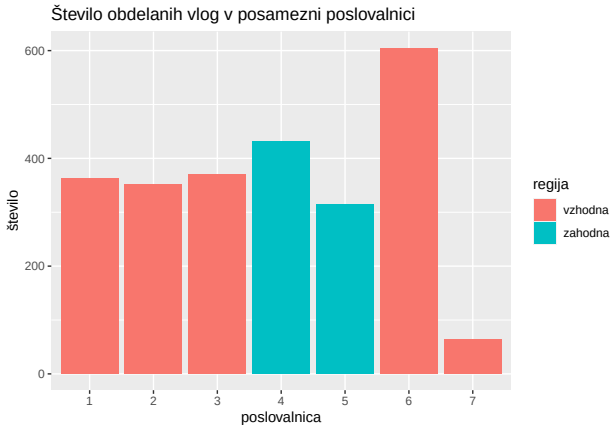


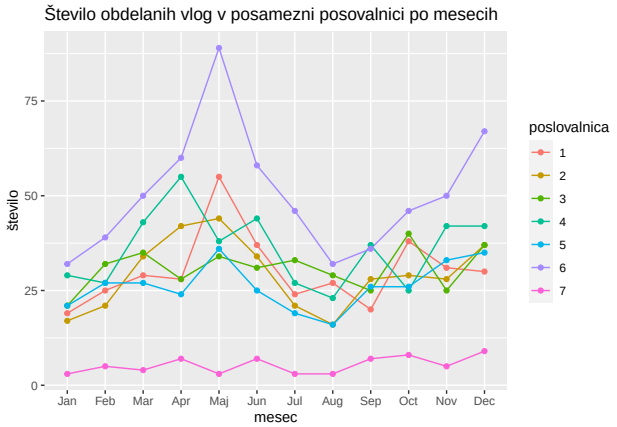




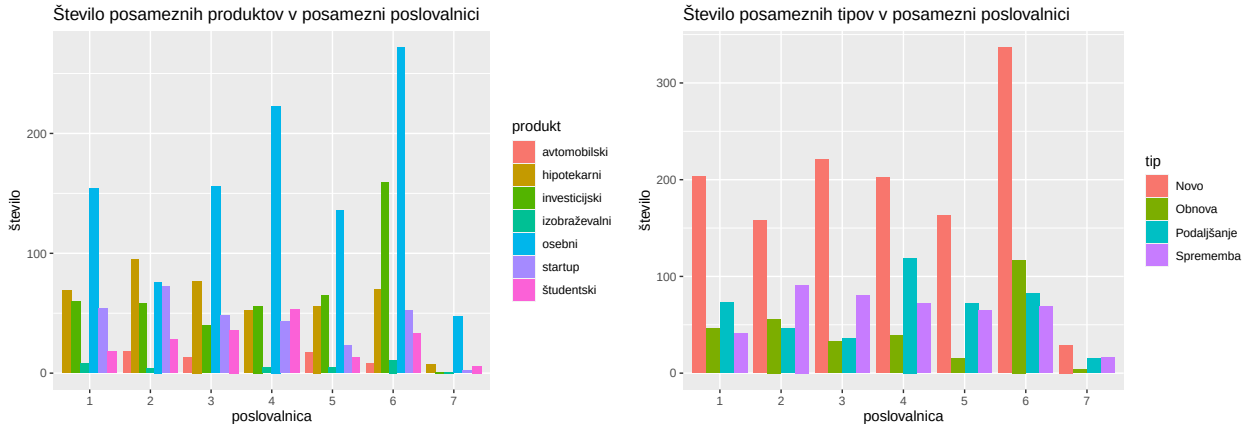


Nekoliko bolj zanimivi so grafi zgornjih primerjav vzhodne in zahodne regije. Najprej ponovno opazimo, da je očitno glede na skalo veliko več obdelanih vlog v vzhodni regiji kot v zahodni. Ponovno pri produktih prevladujejo osebni krediti, v obeh regija, pri tipih je največ seveda novih. Pri osebnih produktih in novih tipih je opazno tudi največje nihanje skozi celotno opazovano leto, medtem ko so vrednosti za ostale produkte in tipe približno enake (precej bolj konstantne). Neke očitne sezonske komponentne na danih podatkih ni mogoče opaziti, za to bi seveda potrebovali tudi podatke za več let, da bi potrdili ali so nihanja števila podobna v istih mesecih vsako leto. Tako pa lahko v našem primeru ta velika nihanja pripišemo zgolj neposrečenemu generiranju podatkov in jim ne pripisujemo pomembnejše vloge.



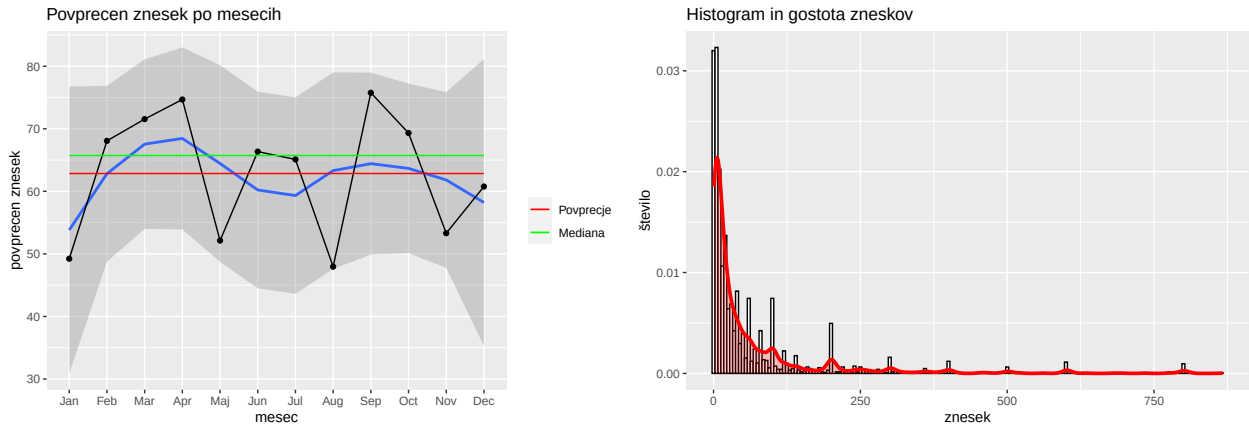


Če razdelimo regije dodatno še po poslovalnicah in jih primerjamo med sabo kar lahko opazujemo na zgornjih dveh grafih, se hitro izkaže, da sta obe regiji približno enako učinkoviti - obe v večini poslovalnic obdelata približno enako število vlog. Daleč največ vlog obdelajo v poslovalnici s številko 6, najmanj pa v poslovalnici s številko 7. Ker ne poznamo geografske razdelitve poslovalnic po državi, ne moremo zagotovo reči zakaj do tega prihaja, najbolj verjetna razlaga je da se poslovalnica 6 nahaja v večjem (verjetno glavnem) mestu države, medtem ko se poslovalnica 7 nahaja v kakšnem manjšem kraju, v katerem ni toliko ljudi, ki bi potrebovali kredite. Če bi poznali kakšne dodatne geografske dejavnike bi lahko bolj natančno izmerili učinkovitost poslovalnic - recimo izračunali razmerje med številom vlog in številom prebivalcev v kraju/občini, kjer stoji poslovalnica ali pa primerjali dve poslovalnici med sabo, če bi vedeli da se nahajata v istem kraju.
Graf na desni nam prikazuje število obdelanih vlog v posamezni poslovalnici po mesecih. Opazimo lahko izrazit skok v mesecu maju, kar smo opazili tudi že prej. Logično so te spremembe najbolj izrazite pri poslovalnicah, ki povprečno obdelajo največ vlog, saj so verjetno postavljene v večjih mestih in se v njih lahko v določenem času, ko je to potrebno, lahko za kredit prijavi več ljudi kot recimo v poslovlanicah, kjer število obdelanih vlog tudi povprečno manjše (7), kjer takšnih izrazitih skokov ni.



Zgoraj lahko opazujemo še dva grafa povezana s poslovalnicami, ki prikazujeta število posameznih produktov in posameznih tipov v posamezni poslovalnici. Ni presenetljivo, da so zopet najbolj pogosti osebni produkti in novi tipi, kot smo do zdaj že večkrat opazili. Zanimivo pa je na tej točki opaziti, da je na drugi najbolj pogost produkt v poslovalnici 6 investicijski produkt, medtem ko je v večini drugih drugi najbolj pogost hipotekarni. To nam potrjuje prejšnjo hipotezo, da je poslovalnica 6 zelo verjetno v glavnem mestu države, kjer je največ podjetij, ki oddajo največ vlog v (verjetno) glavno poslovanico banke, medtem ko v drugih poslovalnicah večino vlog oddajajo fizične osebe. Še en dejavnik, ki nam potrjuje to hipotezo je, da je na drugem mestu v poslovalnici 6 obnova kreditov, kar pomeni da je večino strank (podjetij) rednih in se samo vračajo (mesečno), da kredite obnovijo.

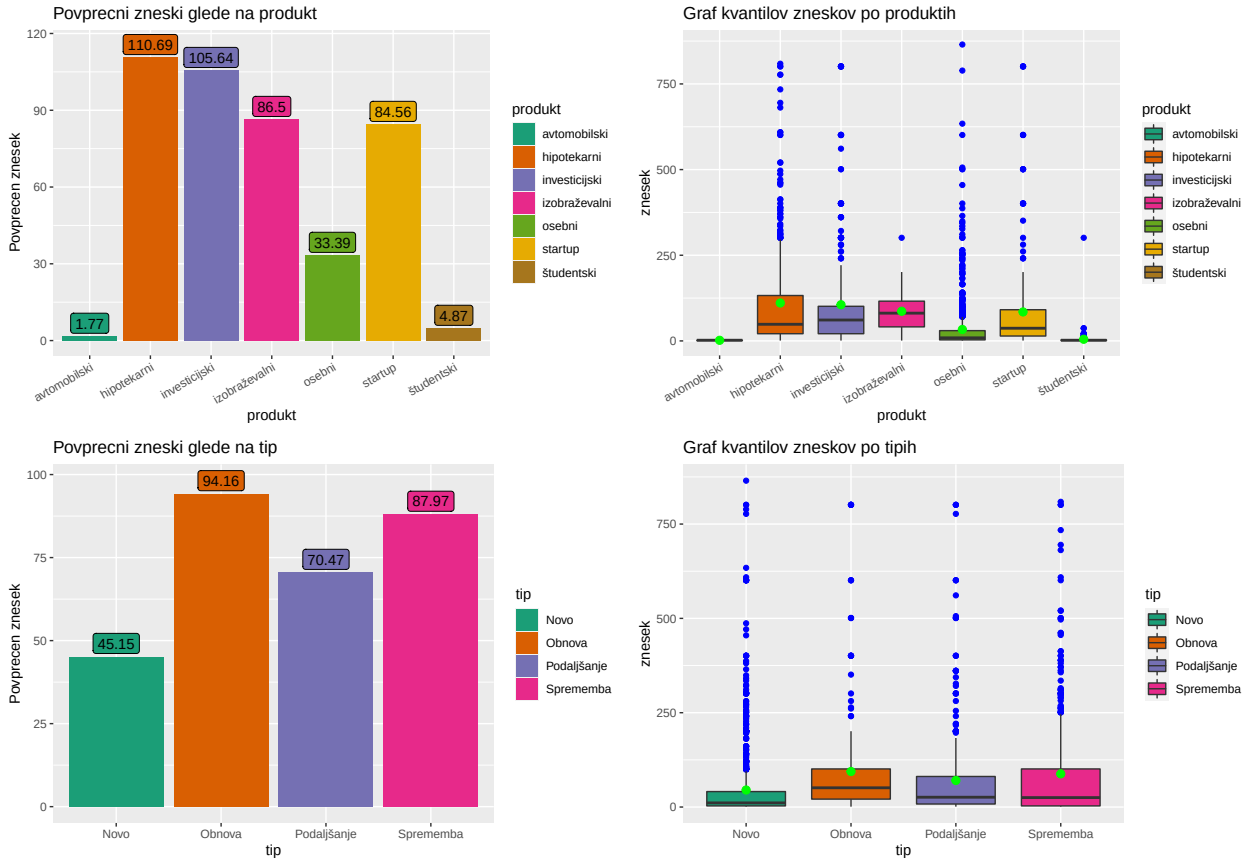
Oglejmo si sedaj še najbolj zanimiv del lastnosti - zneski. Na spodnjih grafih sta prikazana povprečen znesek po mesecih povezan s funkcijo `geom_smooth` ter 0.95 % intervalom zaupanja, čemer sta dodana še povprečen letni znesk in mediana, ter histogram vseh zneskov z gostoto narisano s funkcijo `density`. Na prvem grafu lahko izrazito opazimo, da zneski nihajo in da krivulja čez njih nima pravega smisla, ter je verjetno posledica le generiranja simuliranih podatkov, saj se večina zneskov nahaja v okolici 10 pandacoinsov od povprečja oziroma mediane. Tako lahko privzamemo, da so zneski skozi celotno leto pri opazovani banki konstantni. Na drugem grafu lahko opazimo ogromno zneskov malih vrednosti, zelo hitro, lahko bi ocenili da eksponentno ali pa vsaj primerjalno s funkcijo $\frac{1}{x}$ padajo. Imajo še nekaj vrhov pri res visokih vrednostih - to lahko razlagamo kot to, da banka v večini posoja kredite fizičnim osebam in manjšim pravnim osebam v manjših vsotah, hkrati pa posluje tudi z nekaj večjimi podjetij, ki so odgovorni za ogromne zneske na koncu.



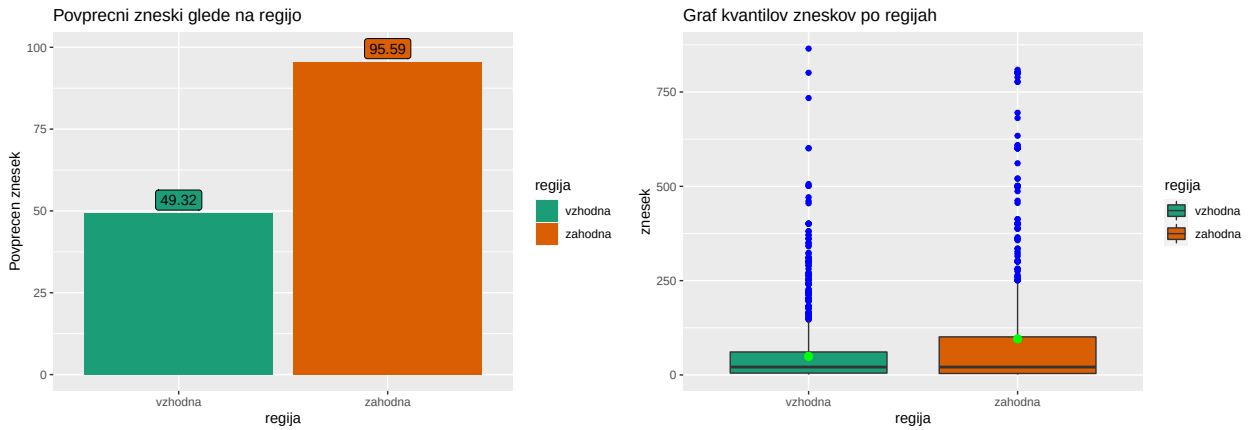
Analizirajmo zneske sedaj še glede na vsak drug atribut posebej.

Spodaj si lahko ogledamo grafa povprečnih zneskov in kvantila zneskov glede na produkt. Na grafu kvantilov zneskov so prikazane povprečne vrednosti iz prejšnjega grafa s svetlo zeleno točko na vsaki škatli, zato da se lažje primerja povprečje in mediano. Očitno je, da je povprečje v splošnem višje od mediane, kar je logično,

saj kot smo videli obstaja majhno število visokih zneskov, ki višajo povprečje, mediana pa se nahaja nekje med nižjimi zneski, saj jih je količinsko precej več. Kot smo pravilno predvidevali je iz grafov očitno tudi, da so hipotekarni in investicijski krediti v povprečju bistveno višji kot osebni, lahko pa opazimo tudi precej velike osamelce pri kvantilu osebnih produktov, torej obstajajo posamezniki, ki si izposojajo velike vsote kot osebni produkt.

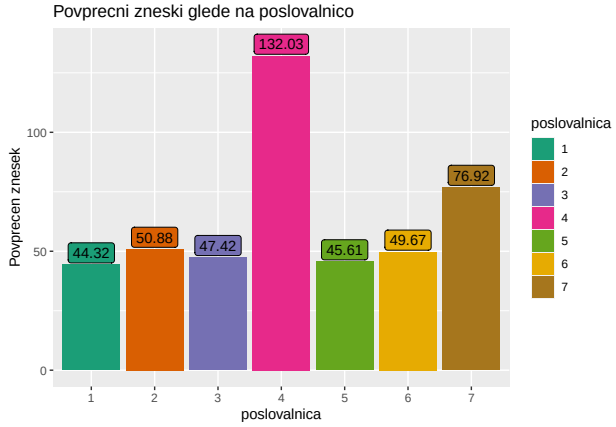


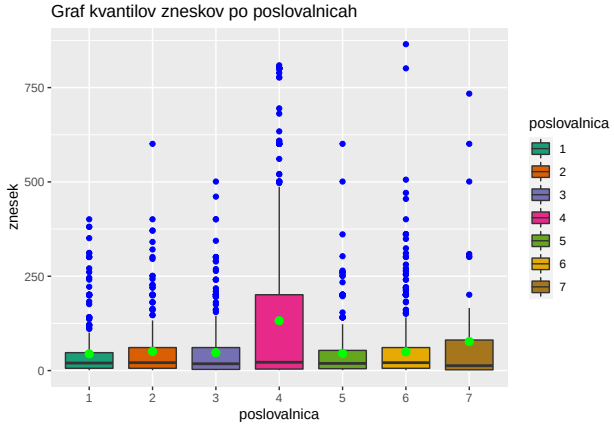
Pri primerjavi tipov ni presenetljivo, da je povprečni znesek tipa novo najnižji, saj je takšnih kreditov največ, ter tako gotovo veliko v manjših vrednostih. Je pa razlika v povprečju v primerjavi z ostalimi tipi, precej velika, kar je zanimivo kljub temu, da imamo pri tipu novo veliko število osamelcev, povprečna vrednost pa leži že pravzaprav izven škatle. Ponovno tu v igro pride dejstvo, da je takšnih kreditov več kot polovica vseh.



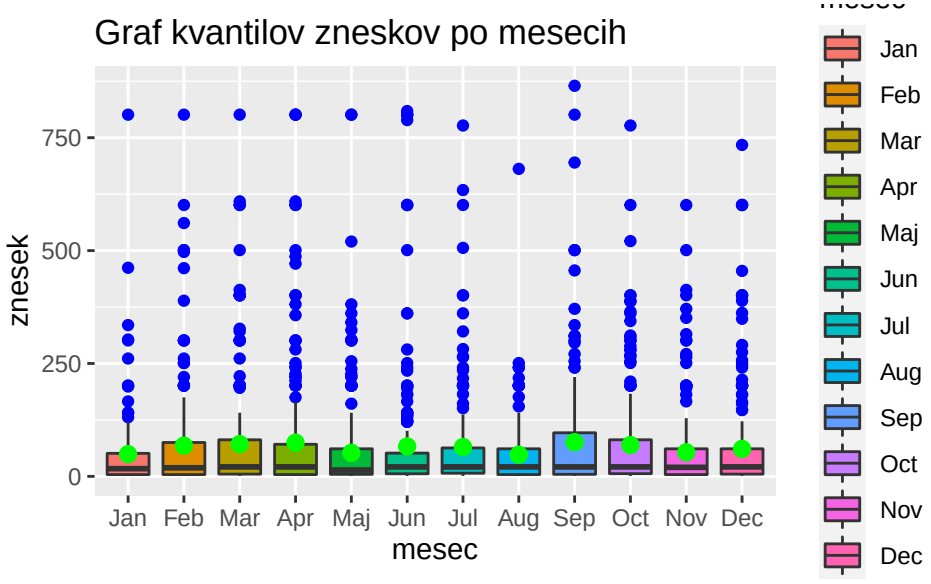
Primerjajmo sedaj še regije in poslovalnice. Pri regijah je izjemno zanimiva razlika med povprečnim zneskom

vzhodne in povprečnim zneskom zahodne - zahodna ima višji povprečni znesek za skoaj 50 pandacoinsov, kar je morda rahlo presenetljivo. Po drugi strani pa vemo, da sta v zahodni regiji samo dve poslovalnici, kar nam lahko podobno kot prej napeljuje na dejstvo, da večino manjših kreditov obdelajo v poslovalnicah vzhodne regije.





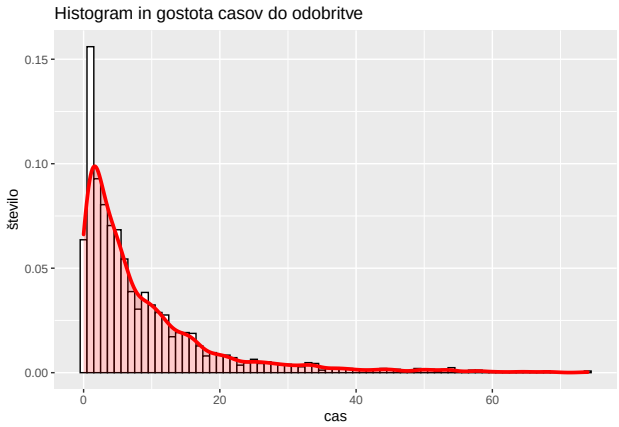
Če pa si ogledamo še grafa povprečnih zneskov in kvantilov zneskov poslovalnic, postane jasno zakaj ima zahodna tako višji povprečni znesek. Poslovalnica 4 ima daleč najvišje povprečne zneske, torej je bila naša predpostavka od prej (vsaj delno) napačna - poslovalnica 6 res obdela največ vlog v celotni državi, vendar se večina velikih poslov zgodi v poslovalnici 4 v zahodni regiji. Tako bi lahko predpostavili, da se poslovalnica 6 nahaja v mestu z veliko prebivalstva, medtem ko se poslovalnica 4 nahaja v kakšnem bolj poslovno orientiranem okolju, kjer delajo večja podjetja in bogatejši ljudje, ki posledično potrebujejo višje kredite. Dejstvo, da je poslovalnica 7 druga po povprečnem znesku lahko zavržemo s tem, da imajo vseeno občutno manj obdelanih vlog kot ostale poslovalnice, torej jim že majhno število velikih zneskov občutno zviša povprečje - če bi to normalizirali, glede na recimo število prebivalcev kraja, v katerem se nahajajo posamezne poslovalnice, bi bil njihov povprečni znesek v primerjavi z ostalimi gotovo nižji, na kar kaže tudi to da imajo najnižjo mediano izmed vseh poslovalnic.

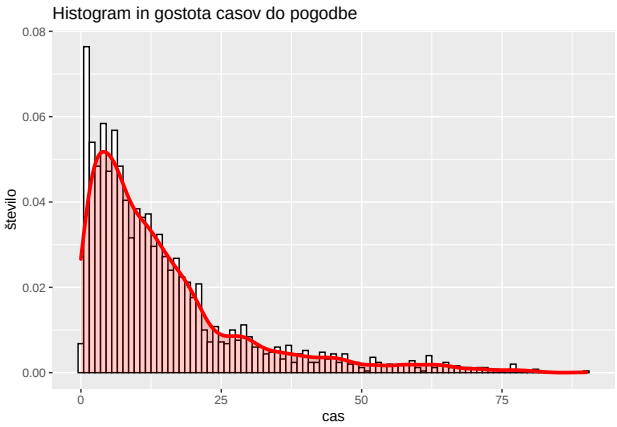


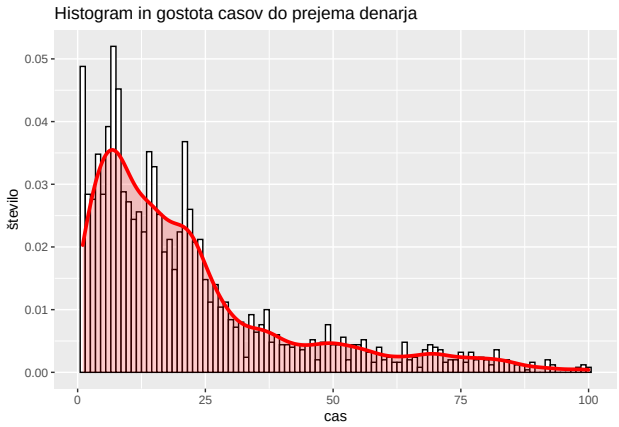
Na zadnjem grafu prvega dela zgoraj si lahko ogledamo kvantile zneskov po mesecih. Meseci so med seboj precej primerljivi, ni nekih občutnih razlik, kar smo videli tudi že na začetku analize zneskov. Zanimivo je morda opaziti le, da čeprav smo opazili velik porast števila vlog konec pomaldi in v začetku poletja, so kvantili zneskov v teh mesecih med najnižjimi.

Analiza časov

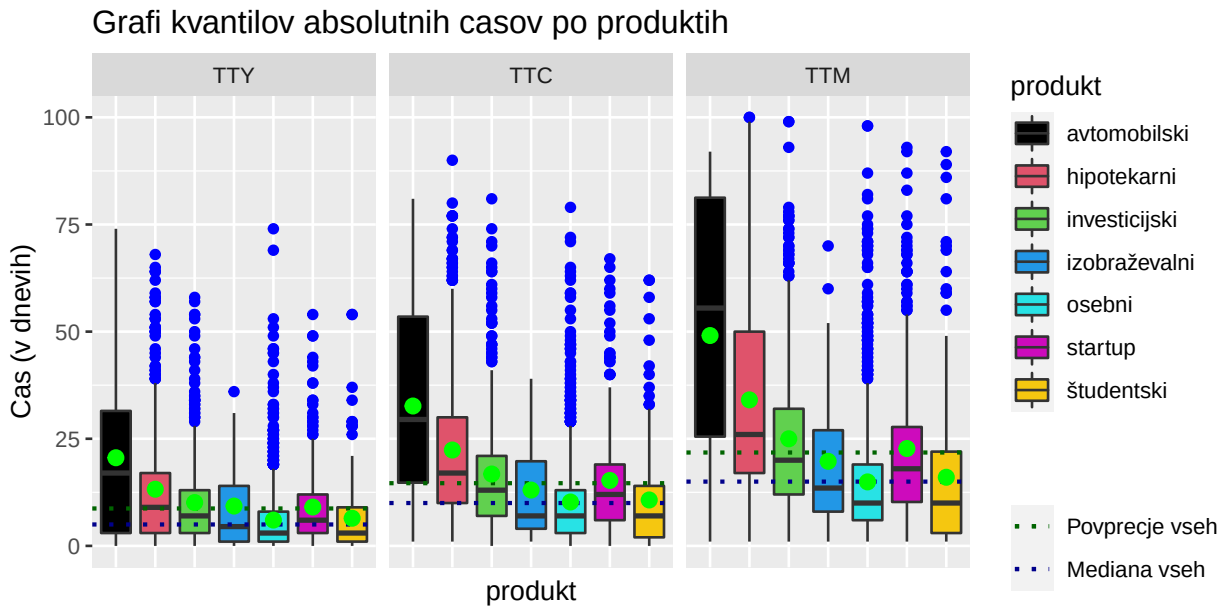
analizo časov začnimo z ogledom histogramov vseh časov. Podobno kot pri zneskih, se vsi časi zgodijo najpogosteje pri majhnih vrednostih (v 5 ali 6 dneh) in potem število vrednosti za posamezen čas ponovno počasi pada. Predvsem čas do odobritve bi se morda dalo opisati s porazdelitvijo *Gama* z neko β manjšo od 1. Pri rugih dveh časih bi imeli s tem že več težav, saj oblika ni tako lepa kar je logično, saj se časi za čas do podpisa pogodbe in prejema sredstev podaljšujejo in se vedno pogosteje zgodi, da moramo čakati tudi 50 in več dni, zato se dodatni lokalni maksimumi lahko pojavljajo tudi kasneje.

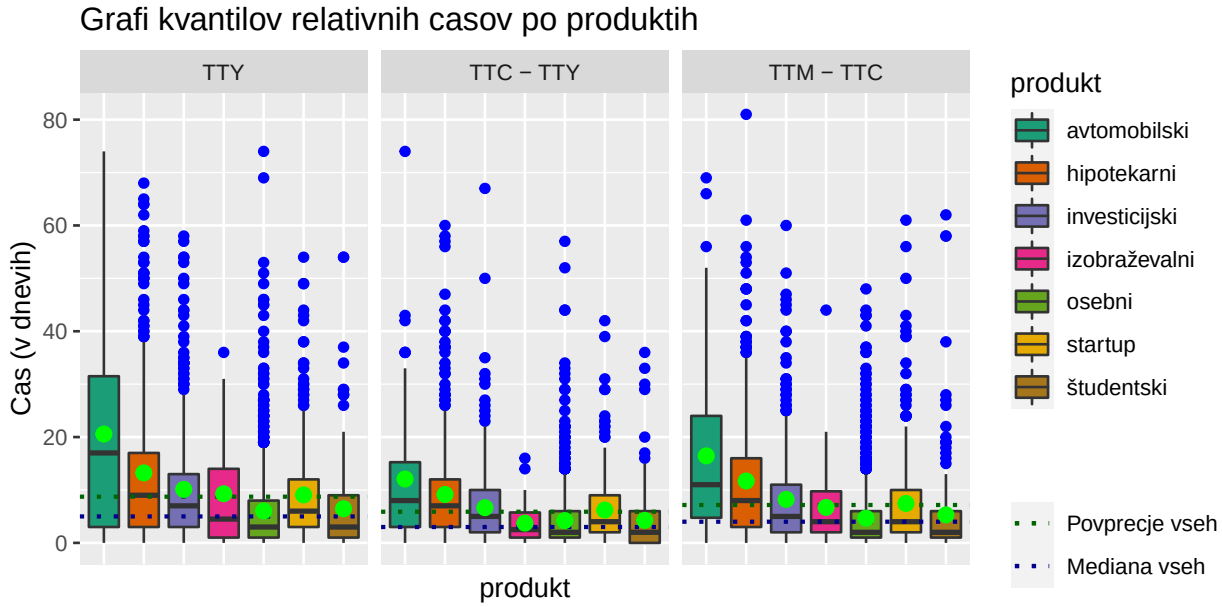




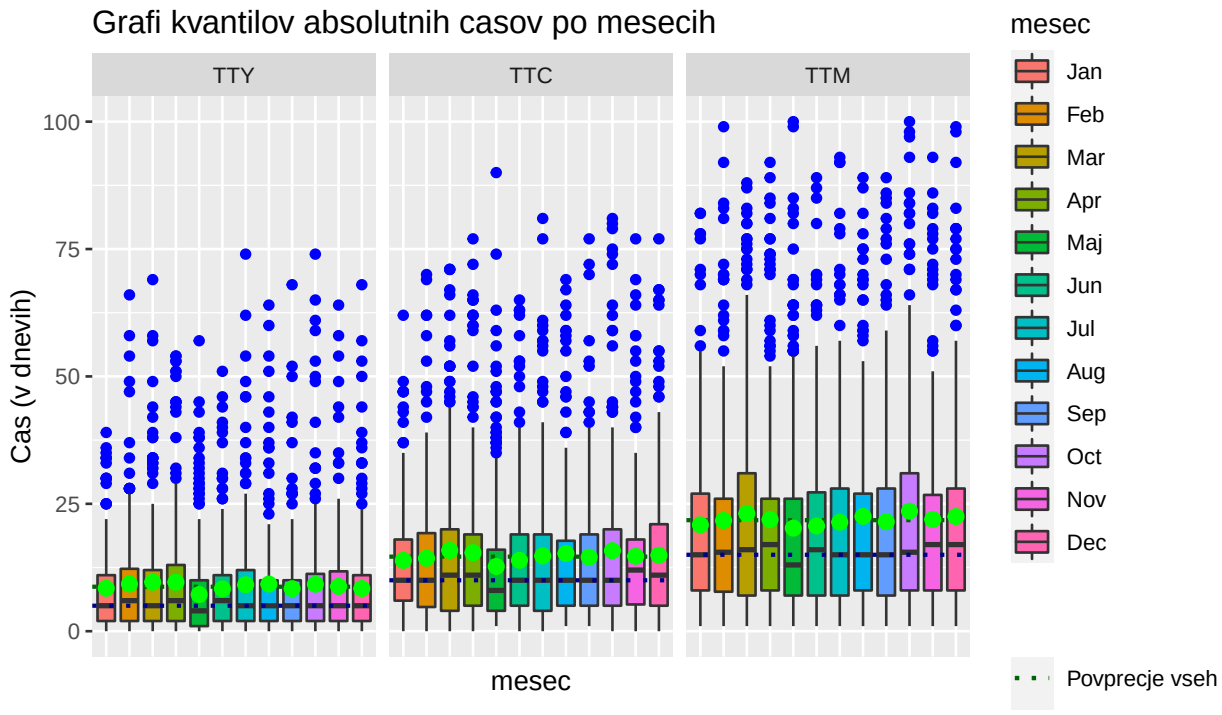


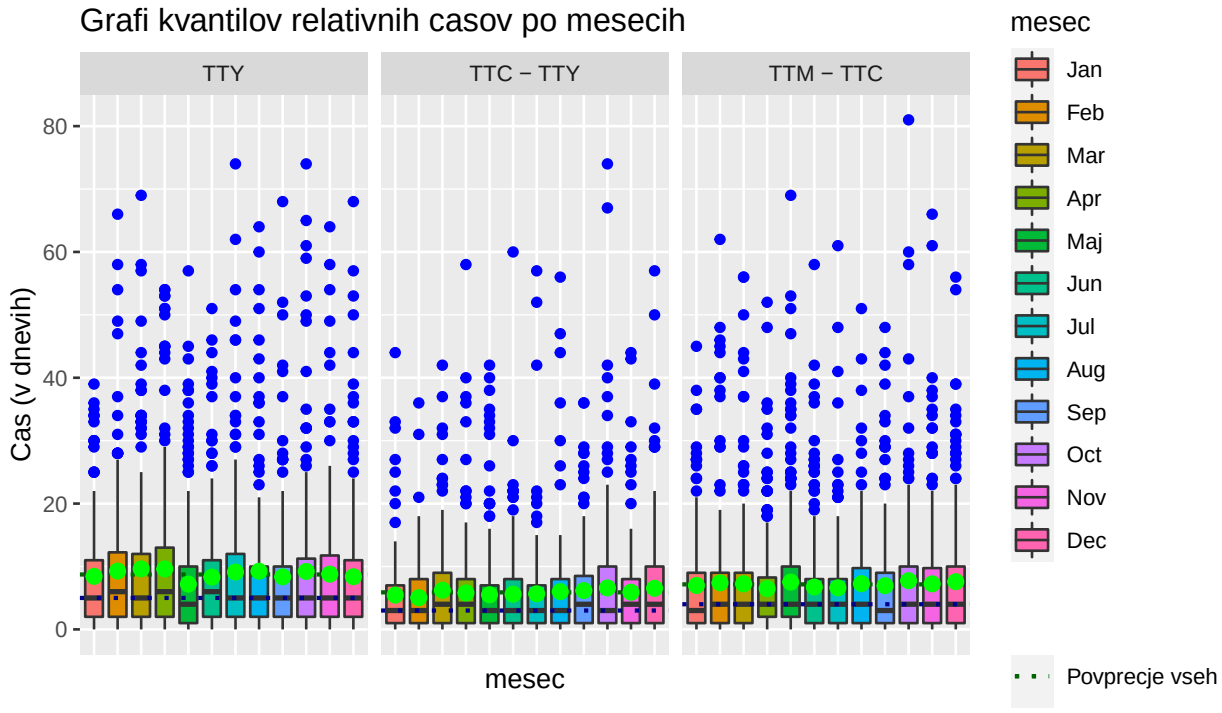
V nadaljevanju si bomo ogledali grafe kvantilov vseh časov v odvisnosti od ostalih lastnosti kreditov. Za vsako lastnost sem narisal dva grafa - enega z absolutnimi časi in enega, kjer so časi relativni. Relativni časi v tem primeru pomenijo, da se vsi začnejo meriti od 0, ko se zgodi prejšnji dogodek - za čas do odobritve je to oddaja vloge, za čas do podpisa pogodbe je to odobritev kredita in za čas do prejema sredstev je to podpis pogodbe. Za primerjavo sem na vsak graf dorisal še povprečje in mediano vseh podatkov za posamezen čas



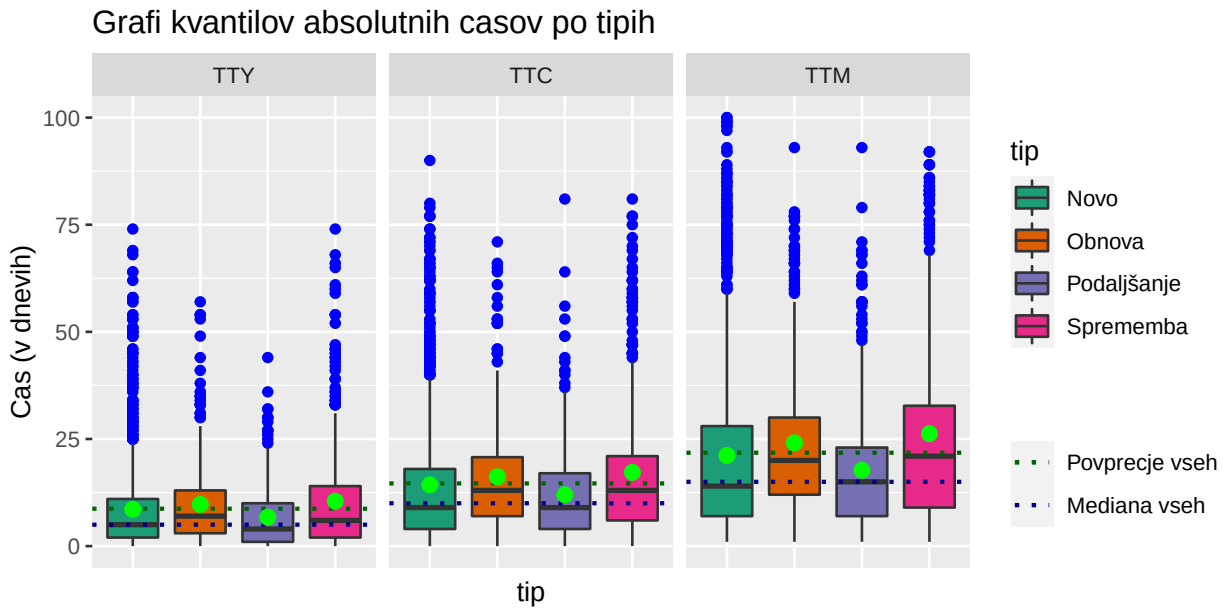


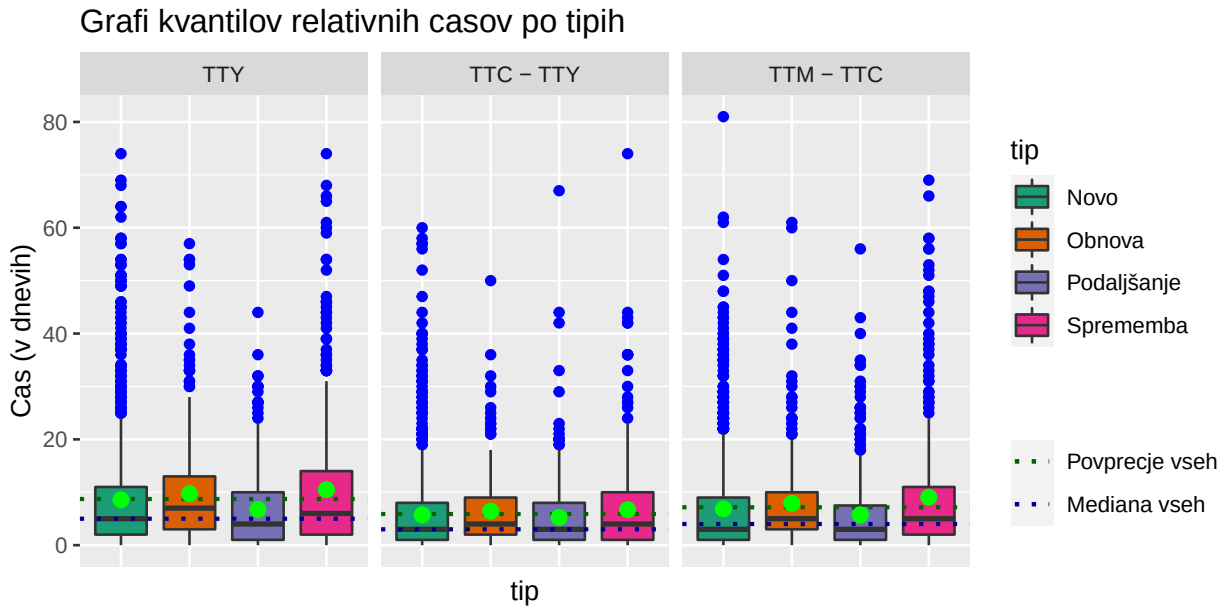
Oba grafa produktov nam razjasnita precej neodgovorjenih vprašanj od prej. Najprej postane jasno, zakaj se prebivalci Pandamije ne odločajo za avtomobilske kredite pri AMBanki - absolutno najdlje časa v povprečju traja, da banka odobri kredit, prav tako se potem na podpis pogodbe in prejem sredstev čaka še dodatno izjemno dolgo. Morda torej ni razlog v tem, da so prebivalci ekološko ozaveščeni in ne vzemajo avtomobilskih kreditov, temveč v tem da vedo, da jih lahko dobijo pri kakšni drugi banki precej hitreje. Zanimivo pa je, da po najdaljših časih sledijo hipotekarni in investicijski krediti, vendar je potreba po njih dovolj velika, da so stranke vseeno bolj pripravljene počakati na njih, kot na avtomobilske. Poleg tega postane jasno tudi zakaj so osebni produkti tako pogosti pri opazovani banki - odobreni so v povprečju najhitreje - poleg študentskih edini ki so odobreni pod povprečnim časom vseh podatkov. Zanimivo je videti tudi, da so izobraževalni krediti najbolj konstantni - pri njih lahko opazimo zelo malo osamelcev. Za konec lahko opazimo še splošno stvar, da so čas od odobritve do podpisa pogodbe v povprečju traja najmanj, najdlje pa traja čas od oddaje vloge, do njene odobritve



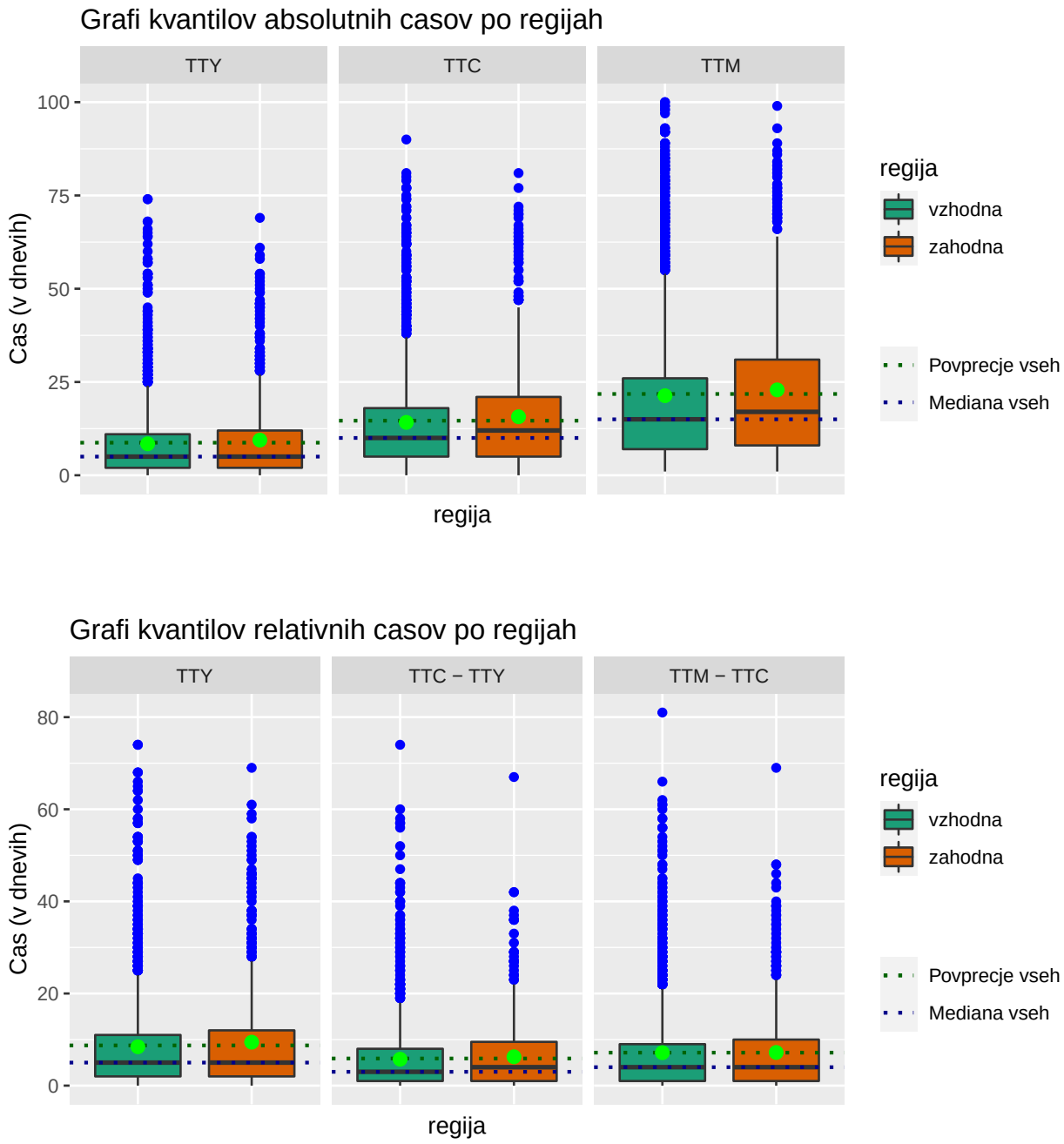


Analiza časov po mesecih je v tem primeru dokaj dolgočasna - kot pri zneskih se tudi tu pokaže, da mesec v katerem je bila oddana vloga za kredit ne igra bistvene vloge pri časih do odbritve, podpisa pogodbe ali prejema sredstev. Povprečja in mediane posameznih mesecev se pri večini časov ujemajo s povprečjem in mediano vseh podatkov.

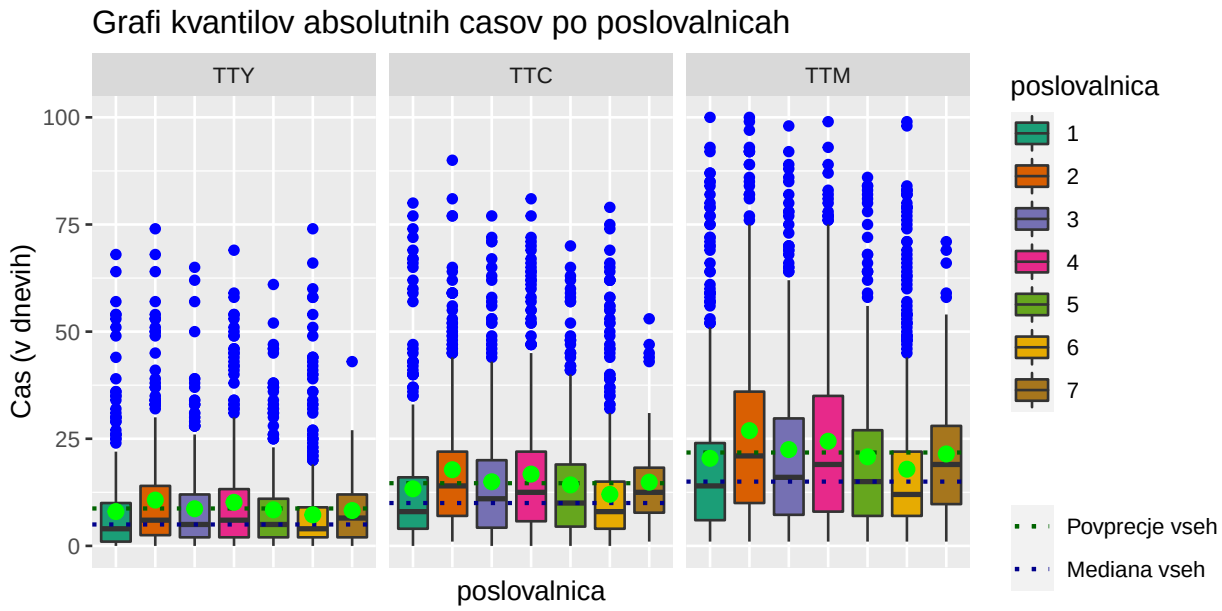


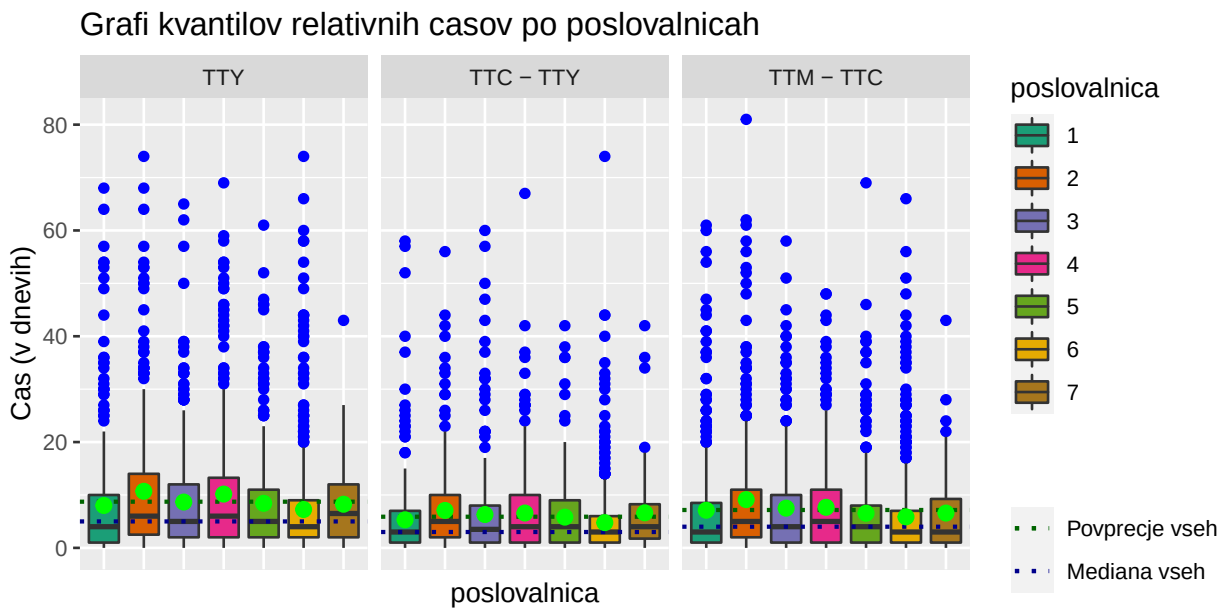


Najkrajši čas pri tipih kreditov imajo podaljšanja, najdaljše čase pa imajo spremembe. Ponovno tu ni popolnoma jasno kaj bi bil razlog za krajše ai daljše čase, saj nimamo podatka o tem kako banka obravnava posamezen tip. V povprečju tudi ni nekih odstopanj - tako povprečja kot mediane posameznih kvantilov se pri vseh časih ujemata s povprečjem in mediano vseh podatkov.



Grafi o časih v posamezni regiji nam ne povedo ničesar novega, razen tega da poslovalnice v zahodni regiji morda potrebujejo malodlje časa za obravnavo kreditov kot v vzhodni. Hkrati pa je očitno da tako povprečja kot mediane obeh kvantilov pri vseh časih ležijo na povprečju in mediani vseh podatkov. Tako lahko zaključimo, da so postopki znotraj vsake regije dovolj standardizirani, da izbira regije v splošnem ne bi smela preveč vplivati na to, kako dolgo čaka stranka na kredit.





Zanimivo pa si je v tem primeru pogledati še čase posamezenih poslovalnic in jih razčleniti. Takoj postane jasno, zakaj je ravno poslovalnica 6 tista najbolj obiskana v vsej državi - saj imajo v povprečju najkrajše vse tri čase. Zanimivo v nasprotju, da izmed dveh poslovalnic v zahodni regiji tista z višjimi povprečnimi zneski ni tudi hitrejša - to je verjetno tudi logično, saj višji zneski pomenijo bolj tvegane posle, ki jih je pametno natančneje preveriti in pregledati. Še bolj kot to pa je izjemno zanimivo, da poslovalnica z najdaljšimi časi ni nobena izmed najbolj obiskanih - poslovalnica 2. V tem primeru bi se splačalo predlagati vodstvu naj v kratkem tja pošlje kakšno nadzorno ekipo, ki bi preverila zakaj točno prihaja do teh daljših časov. Morda je v poslovalnici 2 zaposlena mlajša ekipa, ki še nima toliko izkušenj in se radi vzamejo več časa preden odobrijo posle in podpišejo pogodbe, morda je zaposlena starejša ekipa, ki ob vse moderni tehnologiji potrebuje dalj časa da uredi vso birokracijo. Različnih razlogov za to je lahko veliko, vendar iz podatkov, ki so nam dani ne moremo razbrati kateri je pravi. Vsekakor pa je odstopanje od povprečja dovolj majhno da ta podaljšanj čas ni preveč zaskrbljujoč.

Napovedovanje časov

V zadnjem delu projekta sem se lotič napovedovanja časov z metodami strojnega učenja. Najprej je bilo seveda treba narediti preureditev podatkov - precej jasno je bilo, da bomo napovedovali številske vrednosti, torej v podatkih ne moremo imeti nikakršnih besednih opisov spremenljivk. Zato sem lotil spreminjanja vrednosti lastnosti.`ID`, `Znesek` in `poslovalnica` so seveda ostali enaki. Spreminjanja stolpca `mesec` sem se lotil na način, da sem podatke za 12 mesecev zapisal kot funkciji `sin` in `cos`. Kot vemo nam kombinacija teh funkcij daje enotsko krožnico in če pravilno izberemo vrednosti, lahko zakodiramo mesece na način, da sta decemeber in januar zelo blizu skupaj, in hkrati najdlje od junija in julija kot je možno. Uporabil sem formulo $sin((mesec-1)*(2*\pi/12))$ in $cos((mesec-1)*(2*\pi/12))$, kjer sem vsakemu mesecu najprej pripisal številsko vrednost (1-12). Po takšni transformaciji sem dobil dve vrednosti za mesec, ki enolično določata njegovo mesto na enotski krožnici. Nadalje sem stolpec `regija` pretransformiral v binomski zapis, saj sem imel samo dve različni vrednosti. Tako je `vzhodna` postala vrednost 1, `zahodna` pa 0. Podatke o produktu in tipu sem zakodiral kot `dummy variables`, kar pomeni, da se za vsako različno vrednost stolpca produkt (in tip) ustvari nov stolpec, katerega vrednost je 1, če je kredit tistega produkta (tipa) in 0 sicer. Ker nisem vedel ali bo samo transformacija dovolj, sem ustvaril še dodatne stolpce kot so skupna vrednost kreditov (saj smo imeli več `ID`, ki so podali vloge za več kreditov), povprečna vrednost kreditov, ter skupno število kreditov za posamezen `ID` v opazovanem obdobju. Na koncu sem ustvaril še ločeno tabelo takih podatkov, ki sem jo še standardiziral.

Sledilo je ukvarjanje s problemom več istih `ID` v podatkih. V splošnem je pri strojnem učenju to lahko hud problem, saj se nam pogsto zgodi (recimo pri kliničnih testih), da so atributi odvisni od osebe. V našem primeru bi lahko argumentirali, da ostali atributi nimajo velike povezave z osebo samo, kaj šele z časom, ki ga poskušamo napovedati. Edina povezava, ki bi lahko vplivala je to, da velika večina strank, ki je oddala več vlog, vedno odda vlogo v isti poslovalnici (torej tudi v isti regiji) - ljudje poslujejo v poslovalnicah, ki so jim verjetno najbližje, ter jim zaupajo, ker jih že poznajo. V vsakem primeru sem se tega problema lotil na dva načina. V prvem primeru sem samo pazil, da sem, ko konstruiral učno in testno množico za model, shranil vse kredite z istim `ID` bodisi v učno bodisi v testno, torej nikoli se ni zgodilo, da bi bila oseba z istim `ID` hkrati v obeh množicah. V drugem primeru sem za vsako množico iz nabora vseh kreditov za posamezen `ID` naključno izbral enega in ga določil učni oziroma testni množici. Od tam naprej sem obravnaval kredite znotraj množice kot neodvisne med seboj.

Sledila je konstrukcija funkcije, ki bo izvajal prečno preverjanje na podatkih in tako pomagala določiti optimalni model. Ker je bilo jasno, da bo šlo za regresijski problem, sem se odločil izbirati med linearnim, log-linearnim, posplošenim linearnim, posplošenim log-linearnim in lmer modelom. Prečno preverjanje sem izvedel tako, da sm celotno množico podatkov naključno razporedil v 10 skupin (in pri tem upošteval zgornja pravila o `ID`). Nato sem preko zanke vsako izmed 10 množic enkrat nastavil za testno, vse ostale pa za učno množico, na njih naučil model za vsakega od potencialnih kandidatov, si izračunal napake (izbiral sem na podlagi povprečne kvadratične napake - MSE) in si shranil rezultate. Na koncu sem rezultate primerjal glede na različne izbire atributov (nisem vedno izbral vseh dodatnih spremenljivk). Rezultati so bili sledeči:

- za napovedovanje časa TTY je bil najboljši splošni log-linearni model z vključenim atributom povprečni znesek
- za napovedovanje časa TTC je bil najboljši splošni log-linearni model brez vključitve dodatnih atributov
- za napovedovanje časa TTM je bil najboljši splošni log-linearni model brez vključitve dodatnih atributov

Napake za algoritme so bile nekaj več kot 100, torej okvirno 10 dni, kar se mi zdi precej dober približek, glede na to da podatki niso najboljši za napovedovanje, da jih je malo in da se jih morda sploh ne da učinkovito napovedovati. Končno sem še enkrat naučil vse tri modele, tokrat na vseh podatkih in jih shranil. To so modeli, ki jih uporablja tudi *shiny* alikacija za izračun vseh treh časov.

Shiny aplikacija

Po končanih vseh analizah sem se lotil oblikovanja uporabniškega vmesnika s paketom *shiny.* Naredil sem spletno stran banke, na kateri si lahko vsak ogleda najpomembnejše izmed vseh grafov. Razdelki so ločeni glede na splošne atribute, brez povezave s časi, ter na grafe, ki so povezani s časi. Poleg tega je na aplikaciji dostopna tudi tabela osnovnih podatkov, da si jo lahko uporabnik sam podrobneje ogleda. Zadnji zavihek pa je interaktiven izračun napovedi vseh treh časov za splošnega uporabnika. Vsak lahko vnese željeni znesek (med 1 in 870) ter izbere vse lastnosti kredita od produkta, tipa, do izbire regije in poslovalnice, ter meseca v katerem bo vlogo oddal. Nato program glede na izbrane modele izpiše, koliko časa ocenjuje, da bo uporabnik čakal na kredit. Če je kakšen čas slučajno daljši od 100 dni, časa ne računa in uporabnika obvesti, da bo čakal dlje kot 100 dni.

Zaključek

Zdi se mi da je bil projekt izveden precej uspešno. Naredil sem vse, kar sem si zadal na začetku - podrobno analizo in vizualizacijo podatkov, tako časov kot ostalih, napoved časov s strojnim učenjem in interaktivno aplikacijo. Seveda so možne izboljšave - če bi imeli več podatkov (tako po številu kreditov, kot po številu atributov) ali natančnejše informacije o njih, bi jih lahko analizirali bolj natančno in ugotovili oziroma potrdili še kakšne dodatne lastnosti ali hipoteze, ki so se nam porodile tekom dela. Tudi strojno učenje bi lahko izvedel še bolj strokovno, uporabil še več različnih modelov, jih med sabo pokombiniral ter tako morda dobil še boljše približke za čase (ali pa tudi ne). Tudi aplikacija bi lahko bila še bolj estetska, lahko bi ponujala več funkcij in ima na splošno še možnosti za izboljšanje. V splošnem pa je bilo bistvo projekta doseženo, tudi pri NLB so projekt pohvalili in komentirali njegov uporabnost za nadalje - vsekakor pa bodo potrebne modifikacije, predvsem pri napovednih modelih, saj so dejanske tabele podatkov nekoliko kompleksnejše kot je bila moja.