Analiza podataka o iznajmljivanju bicikala u Seulu

Vanja Babić, IN14/2017, babicvanja11@gmail.com

I. Uvod

Seul je glavni grad Južne Koreje i broji oko 10 miliona stanovnika. Kao takav je jedan od najnaseljenijih gradova na svetu. Seul predstavlja centar međunarodnog biznisa, finansija, multinacionalnih kompanija i svetskih organizacija i grad je koji neprestano raste i napreduje.

Prema istraživanju, saobraćaj u Seulu se ne razlikuje mnogo od saobraćaja u glavnom gradu Srbije, dakle karakterišu ga gužve i veliki broj automobila na ulicama, a staze za bicikliste su počele da se pojavljuju tek pre nekoliko godina. Porast broja ljudi koji voze bicikle bi zasigurno doprineo, pre svega, zdravlju ljudi i smanjenju zagađenja životne okoline. Obzirom da su to dve glavne teme vremena u kome živimo, analizom podataka o broju iznajmljenih bicikala dolazimo do pregleda trenutne situacije, a potencijalno i do kreiranja strategije, ili poslovnog plana, kako podstaći ljude da više voze bicikle.

II. Baza podataka

Baza podataka koja je korištena za potrebe izrade ovog izveštaja naziva se *SeoulBikeData*. Ova baza sadrži 8760 uzoraka i 14 obeležja. Obeležja koja se nalaze u bazi su: broj iznajmljenih bicikala, datum, sat u toku dana, temperatura, vlažnost vazduha, brzina vetra, vidljivost, tačka rose, Sunčevo zračenje, količina kiše, količina snega, godišnje doba, da li je praznik ili ne, da li servis za iznajmljivanje bicikala radi ili ne.

Rad sa bazom nadalje je nastavljen bez obeležja datum, ali su kreirana dva nova obeležja: mesec i dan u nedelji, tako da baza nad kojom se radila analiza sadrži 15 obeležja. Kategoričkih obeležja ima 6: sat u toku dana, godišnje doba, da li je praznik ili ne, da li servis za iznajmljivanje bicikala radi ili ne, mesec i dan u nedelji. Ostala obeležja su numerička.

Jedan uzorak nam daje podatke o vrednostima svih obeležja u toku jednog sata za svaki dan. U bazi nema nedostajućih podataka. Podaci su sakupljani uglavnom tokom 2018. godine, 8016 uzoraka, ali se u bazi nalaze i 744 uzorka iz 2017. godine.

III. Analiza podataka sa posebnim osvrtom na analizu vremenskih prilika tokom godišnjih doba

Statističkom analizom dobijamo bolji uvid u vrednosti podataka i u kom opsegu se te vrednosti kreću, a ukoliko se uvide neke vrednosti koje nisu očekivane, one će biti posebno analzirane. Vrednosti za količinu kiše i količinu snega su često jednaki nuli, tačnije, oko 95% uzoraka ima vrednost nula za količinu snega, a za količinu kiše oko 94% uzoraka. Kratkim istraživanjem, došla sam do informacija da su retki dani u Seulu kada količina kiše prelazi 0.1mm, a da su zime uglavnom hladne, suve i sunčane. Obzirom da uzorci predstavljaju vrednosti u toku jednog sata, ove vrednosti su najverovatnije tačne i nećemo ih odbaciti.

Vrednosti za količinu Sunčevog zračenja su jednake nuli u oko 50% uzoraka. Ovo je parametar koji bi zahtevao istraživanje eksperata iz ove oblasti. Kao što je prethodno rečeno, obzirom da su ovo uzorci koji nam daju vrednosti obeležja za svaki sat u danu, nećemo odbaciti ovo obeležje, jer za polovinu uzoraka imamo vrednosti veće od nule.

Temperatura se kreće u opsegu od -17.8°C do 39.4°C. Najviše su temperature leti i kreću se najčešće od 22°C do 30°C. Zanimljivo je da su više temperature zabeležene u jesen nego u proleće. Tokom jeseni se temperatura najčešće kreće od 8°C do 20°C, a u proleće otpirlike od 7°C do 17°C. Zime su hladne, dakle najčešće su pojave temperatura od oko -5°C do 1°C. Vrednost temperature od 39.4°C predstavlja autlajer, dakle ovako izrazito visoke temperature su retke.

Broj uzoraka za svako godišnje doba se kreće u rasponu od 2160 do 2208. Srednja vrednost vlažnosti vazduha je slična za sva godišnja doba i kreće se od 49-64%. Najveća vrednost vlažnosti vazduha se javlja leti i u proleće i iznosi 98%, a tokom zime i jeseni je za samo 1% manja.

Srednje vrednosti brzine vetra su u takođe malom rasponu, a najveća je zimi 1.9m/s. Najveća uneta brzina vetra se javlja u proleće, 7.9m/s. Jesen je jedino godišnje doba u kom nije zabeležen ni jedan dan bez vetra.

Vidljivost je dobra tokom svih godišnjih doba i uglavnom iznosi oko 15km. Najniža vidljivost je u proleće, prosečno oko 12km. Najniža vrednost vidljivosti je takođe zabeležena u proleće, tek 270m.

Prosečno, najviše padavina kiše je očekivano u leto, a najveća zabeležena vrednost količine kiše se desila u proleće. Zanimljivo je da se snežne padavine javljaju i tokom jeseni, a čak je i najveća količina snežnog pokrivača pala u jesen, 8.8cm.

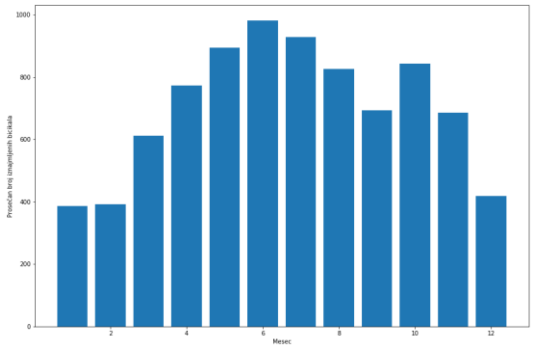
IV. Analiza obeležja broj iznajmljenih bicikala

Obeležje koje će biti izanalizirano u ovom segmentu izveštaja će biti postavljeno kao izlaz linearne regresije. Zbog toga je ova analiza od posebnog značaja.

Najveći broj iznajmljenih bicikala za jedan sat je 3556, što je za grad veličine Seula i očekivano, a najmanji broj

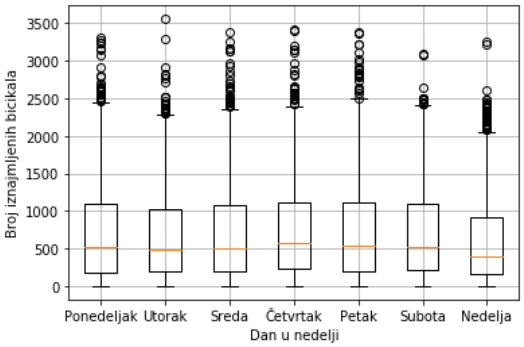
iznajmljenih bicikala je 2. Jedino se tokom dana kada servis za iznajmljivanje ne radi dešava da ni jedan bicikl ne bude iznajmljen, što je takođe očekivano. Prosečan broj iznajmljenih bicikala u toku jednog sata je 704.

Kao što je i očekivano, prosečno se najviše bicikala iznajmljuje tokom letnjih meseci, juna i jula. Potom primećujemo jedan pik (Slika 1.) koji se dešava oko meseca oktobra. Ovo bismo svakako mogli da objasnimo nešto višim temperaturama u jesen, što je već komentarisano. A svakako da veći broj iznajmljivanja bicikala očekujemo tokom lepih dana.



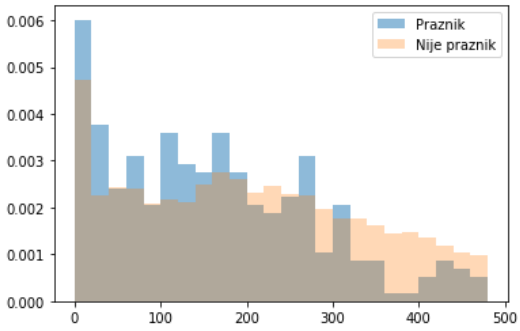
Slika 1. Prosečan broj iznajmljenih bicikala tokom meseci

Tokom dana u nedelji, broj iznajmljenih bicikala se za svaki dan kreće dosta slično, generalno je opseg od 0 pa do između 2000 i 2500. Uglavnom su to vrednosti od oko 200-300 pa do 1000-1100 iznajmljenih bicikala. Najveće odstupanje se javlja nedeljom, kada je ovaj broj nešto niži, od oko 100 do 800 iznajmljenih bicikala. Primetno je da se za svaki dan javlja veliki broj autlajera, od 2000 iznajmljenih bicikala pa naviše. Vrednosti koje prelaze 3000 su značajno ređe. Vrednost medijane je za većinu dana u nedelji 500. Izuzeci su četvrtak, kada je medijana oko 600, i nedelja, kada je taj broj oko 300.



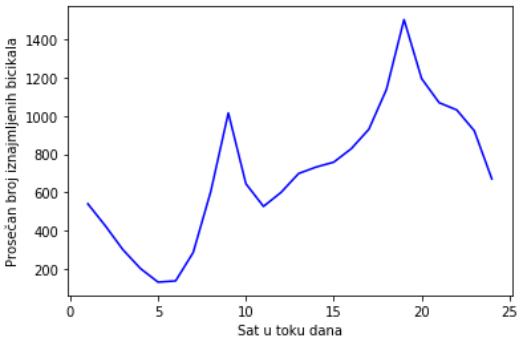
Slika 2. Uporedni prikaz boxplot-ova za za broj iznajmljenih bicikala tokom dana u nedelji

Tokom onih dana kada su praznici, primećujemo da su verovatnoće veće da će biti iznajmljen manji broj bicikala, od 0 do 200, nego onim danima kada se ne praznuje. Ovo možemo da razumemo, obzirom da tokom praznika ljudi najviše vremena provode u svojim domovima. Verovatnoća da će biti iznajmljeno više od 300 bicikala je veća onim dana kada nije praznik nego kada jeste.



Slika 3. Verovatnoća broja iznajmljenih bicikala u zavisnosti od toga da li je praznik ili ne

Analiza broja iznajmljenih bicikala u odnosu da sate u toku dana daje interesantne podatke. Od 05:00h do 10:00h vidimo rast prosečnog broja iznajmljenih bicikala. Potom prosečan broj iznajmljenih bicikala naglo pada do 12:00h, a zatim raste do 20:00h kada dostiže svoj maksimum. Dakle, u toku dana prosečan broj iznajmljenih bicikala je najveći oko 10:00h i 20:00h.



Slika 4. Prikaz prosečnog broja iznajmljenih bicikala tokom sati u danu

V. Analiza korelacije

Ukoliko želimo da utvrdimo da li postoji zavisnost između dva ili više obeležja, tada govorimo o utvrđivanju postojanja korelacije između tih obeležja. Kada je koeficijent korelacije pozitivan, to znači da ukoliko rastu vrednosti jednog obeležja, tada rastu i vrednosti drugog obeležja. Kada je koeficijent negativan, tada rast vrednosti jednog ukazuje na pad vrednosti drugog obeležja.

Obeležja u ovoj bazi podataka su generalno slabo korelisana. Najveća pozitivna korelacija postoji izmedju obeležja temperatura i tačka rose, i iznosi 0.91. Nešto veća negativna korelacija od -0.54 se javlja između obeležja vidiljivost i vlažnost vazduha, što nam govori da, što je vidljivost veća, vrednosti za vlažnost vazduha su manje. Takođe, obeležje vlažnost vazduha je relativno korelisano sa obeležjem brzina vetra. Korelacija je negativna, pa što je veća brzina vetra, vlažnost vazduha je manja. Obeležje broj iznajmljenih bicikala ima najveću pozitivnu korelaciju sa vrednostima temperature, što je i očekivano, dakle kao vrednosti temperature rastu, tako se povećava i broj

iznajmljenih bicikala. Takođe postoji pozitivna korelacija između temperature i godišnjih doba od oko 0.6.

VI. LINEARNA REGRESIJA

Linearna regresija predstavlja metodu koja se koristi za predviđanje kontunualne izlazne promenljive y, uz pretpostavku da je tu vrednost moguće dobiti kao linearnu kombinaciju ulaznih obeležja x. Izlazna promenljiva za model koji će se obrađivati je obeležje broj iznajmljenih bicikala.

Prvi model (M1) koje je isproban jeste osnovni oblik linearne regresije sa hipotezom y =b0+b1x1+b2x2+...+bnxn. Ovaj model je dao loše rezultate.

Tabela 1: Mere uspešnosti modela M1

Srednja kvadratna greška	167874
Srednja apsolutna greška	310
Koren srednje kvadratne greške	409
R2 skor	0.5681
Prilagođeni R2 skor	0.5676

Nakon ovoga je izvršena standardizacija i selekcija obeležja, međutim rezultati nad početnim modelom nisu napredovali.

Sledeća isprobana hipoteza podrazumeva interakcije između obeležja i njihove stepene. Za prvi model sa ovom hipotezom su isprobane kombinacije interakcija obeležja i njihovih stepeni, do drugog (M2). Ovaj model je dao nešto bolje rezultate. Nakon toga isproban je i model koji uključuje treći stepen obeležja (M3), koji je do sada dao najbolje rezultate. Srednja kvadratna greška je gotovo duplo manja nego kod početnog modela. Model koji uključuje i obeležja četvrtog stepena (M4) je dao najgore rezultate od svih do sada isprobanih modela.

Tabela 2: Mere uspešnosti modela M3

Srednja kvadratna greška	84092
Srednja apsolutna greška	213
Koren srednje kvadratne greške	289
R2 skor	0.7836
Prilagođeni R2 skor	0.7772

Lasso regularizacijom su dobijeni nešto gori rezultati od modela M3. Ridge model je dao za nijansu gore rezultate u odnosu na M3, ali bolje u odnosu na Lasso. Međutim, koeficijenti su značajno manji kod Ridge modela i kreću se u opsegu od -400 do 800, dok su te vrednosti kod modela M3 između -12500 i 5000. Iz ovog razloga, kao konačan model je odabran Ridge model.

Tabela 3: Mere uspešnosti konačnog modela

Srednja kvadratna greška	87187
Srednja apsolutna greška	218
Koren srednje kvadratne greške	295
R2 skor	0.7757
Prilagođeni R2 skor	0.7690