

SVEUČILIŠTE U SPLITU
PRIRODOSLOVNO-MATEMATIČKI FAKULTET U SPLITU

SEMINAR

**Uber vožnje u NYC - analiza
Predviđanje cijena dionica u
realnom vremenu**

Matea Lukić

Mentor: *doc. dr. sc. Hrvoje Kalinić*

Split, svibanj 2021.

SADRŽAJ

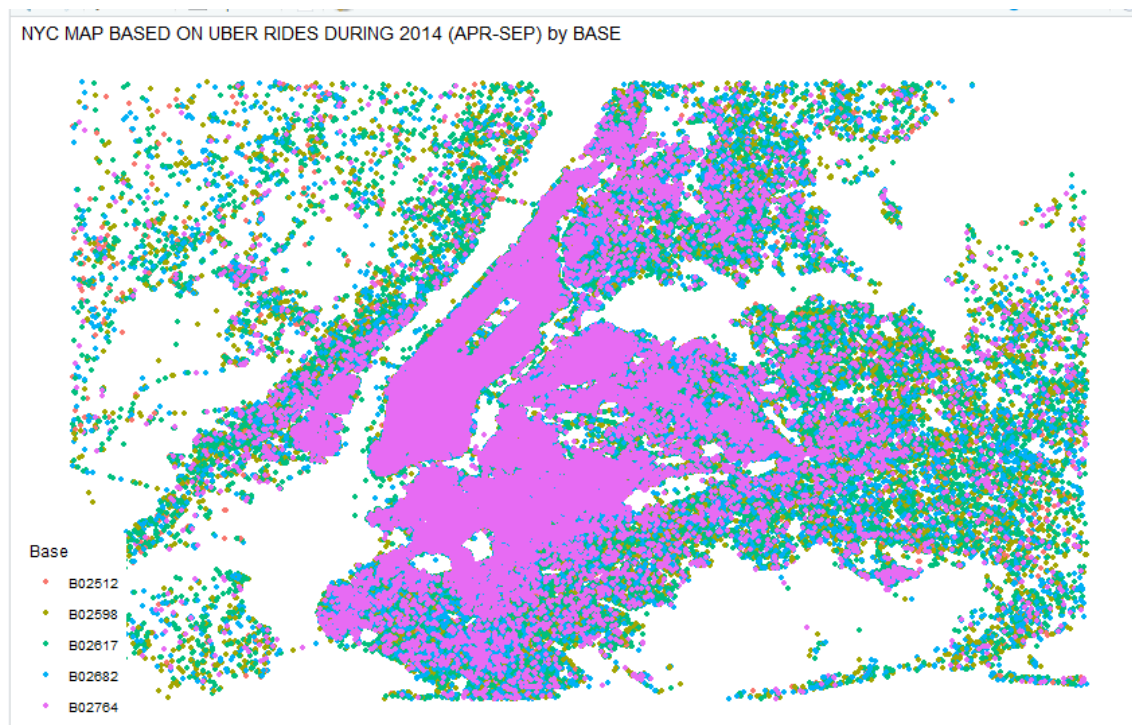
1. Uvod i motivacija	1
2. Zadatak	2
2.1. Problem	3
2.2. Metode	3
2.3. Rješenja problema - Uber vožnje	4
2.3.1. Ideje za moguće funkcionalnosti ovakve analize podataka . .	4
2.3.2. Paketi	5
2.3.3. Učitavanje podataka i podaci	6
2.3.4. Formati za vrijeme	7
2.3.5. Putovanja po satima u danu	7
2.3.6. Podaci po danima u mjesecu	9
2.3.7. Putovanja po mjesecima	11
2.3.8. Putovanja po bazama	13
2.3.9. Stvaranje Heatmap vizualizacije dana, sata i mjeseca	16
2.3.10. Stvaranje vizualizacije karte u New Yorku	21
2.4. Rješenje problema - Predviđanje cijena dionica u realnom vremenu . .	23
2.4.1. Regresijski modeli nad podacima dionica Tesle	23
2.4.2. ARIMA	25
2.4.3. Predviđanje cijena dionica Tesle u budućnosti - ARIMA . . .	27
2.4.4. Predviđanje cijena dionica u realnom vremenu	28
2.5. Usporedba rezultata Kritički osvrt	29
3. Literatura	30

1. Uvod i motivacija

Ovaj projekt će se sastojati od dva dijela tako što će prvi dio uokviriti znanja stečena u prvoj polovici slušanja kolegija Rudarenje podataka te će drugi dio, analogno, odgovarati vještinama stečenim u drugoj polovici slušanja nastave koji se tiču modela i algoritama predviđanja i strojnog učenja implementiranih u R-u. Prvi dio je vezan uz učitavanje podataka, manipulaciju okvirom podataka, tablicama i grafički prikaz. To će biti predočeno i prezentirano na podacima za Uberove vožnje u NYC-u. Nakon toga slijedi predviđanje cijene dionica u realnom vremenu. U svrhu predviđanja će biti korišteni ML modeli.

2. Zadatak

Na temelju podataka o Uberovim vožnjama u New York City-u iz 2014. godine od ožujka do rujna, potrebno je napraviti grafičku vizualizaciju podataka te provesti analizu grafova i donijeti zaključke. Krajnji je cilj dobiti mapu NYC-a s vožnjama koje su prikazane točkicama čije boje odgovaraju nekoj od Uberovih baza s legende.



Slika 2.1: Uberove vožnje po bazama

Druga tema je vezana predviđanje cijena dionica u realnom vremenu. Dakle, želimo dobiti graf za dionicu koja nam je od interesa.



Slika 2.2: Predviđanje dionica

2.1. Problem

Problem za prvi dio jeste kako podatke pripremiti da bismo dobili prikaz iz kojeg što bolje možemo pratiti koliko, kad i gdje se Uberovih vožnji dogodilo. Dakle, potrebni su nam grafovi, tablice, toplinske mape i na koncu prikaz vožnji u obliku točaka kojima su koordinate geografska širina i dužina.

Problem drugog dijela je kako predvidjeti, u realnom vremenu kako stoje dionice. Dakle, želimo dobiti neki graf funkcije za dionicu koja nam je od interesa pa da prema toj funkciji vidimo kako odabrana dionica stoji trenutno.

2.2. Metode

Metode koje će se koristiti za analizu i prikaz vožnji su metode učitavanja .csv datoteka, metode vezane uz rad s okvirom podataka i funkcije za crtanje grafova iz paketa ggplot2 s kojim smo se upoznali na predavanjima iz Rudarenja Podataka.

Metode za predviđanje dionica su iz paketa "forecast". Poslužit ćemo se istoime- nom metodom iz tog paketa. Funkcija forecast() radi s mnogo različitih vrsta ulaza. Općenito uzima vremenski niz ili model vremenskog niza kao glavni argument i prik- ladno izrađuje prognoze. Uvijek vraća predmete klase predviđanja. Forecast() je bazirana na ETS- algoritmu. ETS ili Exponential Smoothing algoritam predložen je

krajem 1950-ih (Brown, 1959; Holt, 1957; Winters, 1960), a motivirao je neke od najuspješnijih metoda predviđanja. Predviđanja proizvedena eksponencijalnim metodama smoothinga ponderirani su prosjeci prošlih promatranja, a ponderi eksponencijalno propadaju kako opažanja stare. Drugim riječima, što je novije promatranje to je veća pridružena težina. Ovaj okvir generira pouzdane prognoze brzo i za širok raspon vremenskih serija, što je velika prednost i od velike važnosti za primjenu u industriji.

2.3. Rješenja problema - Uber vožnje

2.3.1. Ideje za moguće funkcionalnosti ovakve analize podataka

Ovi podaci mogu biti korisni jer se iz njih da zaključiti u kojem godisnjem dobu, mjesecima, danima u mjesecu ili danima u tjednu ima više odnosno manje vožnji.

Možemo te podatke analizirati na razini baza Ubera pa znati koje lokacije imaju veću(manju) potrebu za prijevozom

Ovo može biti korisno Uber vozačima kako bi se znali rasporediti po lokacijama u gradu.

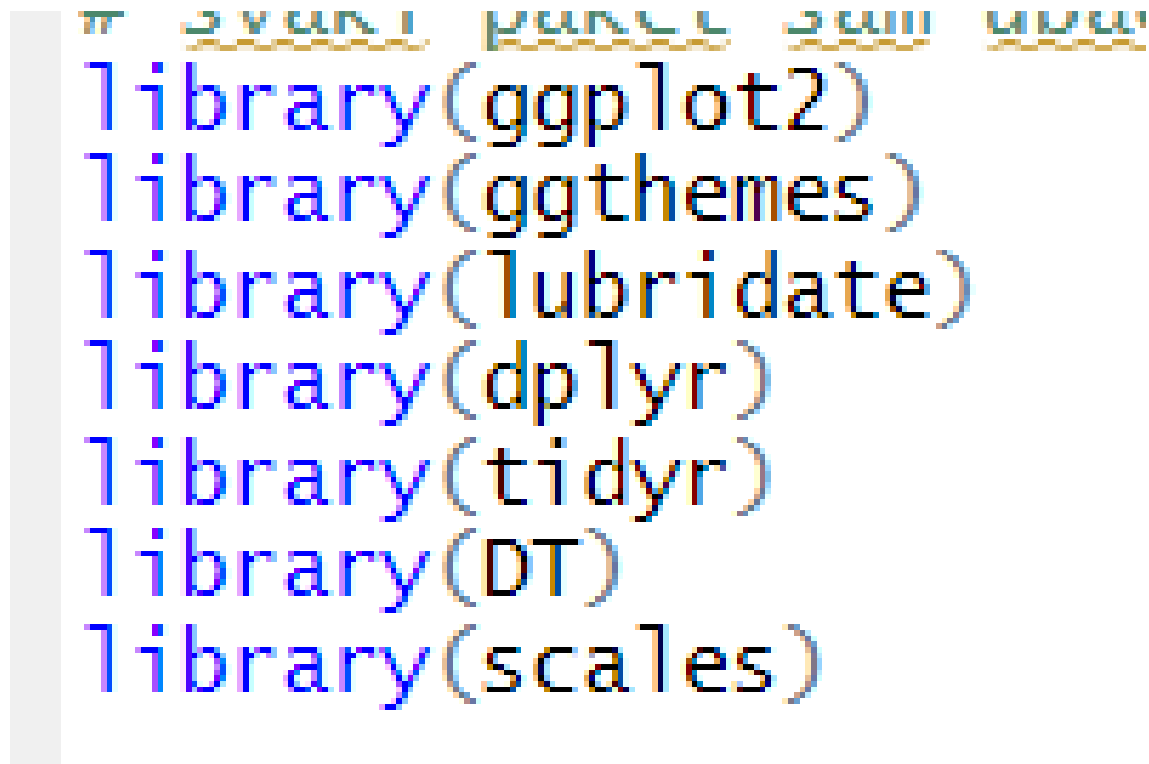
Od značaja je i kompaniji Uber jer zna kojim danima će koliko zaraditi i kada bi mogli dati popuste, a kada povećati cijene.

Također može biti korisno onima kojima taxiranje nije jedini posao pa znaju kojim danima će to raditi, a kojim ne.

Može biti korisno kompanijama da rasporede zaposlenike kojim danima rade od kuće, a kojim iz ureda što doprinosi uređenju prometa i smanjenju gužvi.

Naravno, korisno je imati sliku o prijevozima u gradu općenito za kontroliranje prometa.

2.3.2. Paketi

A screenshot of an R console window showing the installation of several packages. The code is as follows:

```
# SVAKI PAKET SAM UBACU  
library(ggplot2)  
library(ggthemes)  
library(lubridate)  
library(dplyr)  
library(tidyr)  
library(DT)  
library(scales)
```

The text is displayed in a monospaced font with syntax highlighting: comments are green, function names like 'library' are blue, and package names are black.

Slika 2.3: Paketi

- ggplot2 - Najpopularnija biblioteka za vizualizaciju podataka koja se najviše koristi za stvaranje parca estetske vizualizacije.
- ggthemes - Dodatak za glavnu ggplot2 biblioteku. Ovim možemo učiniti bolje stvaranje dodatnih tema i ljestvica pomoću mainstream ggplot2 paketa.
- lubridata - Skup podataka uključuje različite vremenske okvire. Da bismo razumjeli podatke u odvojenim vremenskim kategorijama, koristimo paket lubridata.
- tidyr - Osnovno načelo uređenja je sređivanje stupaca u kojima je svaka varijabla prisutna u stupcu, svako opažanje predstavljeno je retkom, a svaka vrijednost prikazuje ćeliju.
- DT - Povezivanje s JavaScript bibliotekom pod nazivom - Datatables.
- scales - Pomocu grafičkih ljestvica podatke možemo automatski preslikati na ispravne ljestvice dobro postavljenim osima i legendama.

Svaki paket sam ubacila na ovaj način : `install.packages('imepaketa')`

2.3.3. Učitavanje podataka i podaci

Pročitat ćemo nekoliko CSV datoteka koje sadrže podatke od travnja 2014. do rujna 2014. Pohranit ćemo ih u odgovarajuće okvire podataka kao što su `apr_data`, `may_data` itd. Nakon što pročitam datoteke, kombinirat ću sve te podatke u jedan okvir podataka nazvan `'data_2014'`. Zatim ću u sljedećem koraku izvršiti odgovarajuće oblikovanje stupca `Date.Time`. Zatim ću nastaviti stvarati čimbenike vremenskih objekata poput dana, mjeseca, godine itd.

```
apr_data <- read.csv("uber-raw-data-apr14.csv")
may_data <- read.csv("uber-raw-data-may14.csv")
jun_data <- read.csv("uber-raw-data-jun14.csv")
jul_data <- read.csv("uber-raw-data-jul14.csv")
aug_data <- read.csv("uber-raw-data-aug14.csv")
sep_data <- read.csv("uber-raw-data-sep14.csv")

data_2014 <- rbind(apr_data,may_data, jun_data, jul_data, aug_data, sep_data)
```

Slika 2.4: Učitavanje podataka

`rbind()` - Kombiniranje R objekata po redovima

Uzmimamo niz argumenata vektora, matrice ili okvira podataka i kombiniramo po redovima u ovom slučaju.

```
head(data_2014)
# Date.Time Lat Lon Base Time day month year dayofweek hour minute second
# 1 2014-04-01 00:11:00 40.7690 -73.9549 B02512 00:11:00 1 Apr 2014 Tue 0 11 0
# 2 2014-04-01 00:17:00 40.7267 -74.0345 B02512 00:17:00 1 Apr 2014 Tue 0 17 0
# 3 2014-04-01 00:21:00 40.7316 -73.9873 B02512 00:21:00 1 Apr 2014 Tue 0 21 0
# 4 2014-04-01 00:28:00 40.7588 -73.9776 B02512 00:28:00 1 Apr 2014 Tue 0 28 0
# 5 2014-04-01 00:33:00 40.7594 -73.9722 B02512 00:33:00 1 Apr 2014 Tue 0 33 0
# 6 2014-04-01 00:33:00 40.7383 -74.0403 B02512 00:33:00 1 Apr 2014 Tue 0 33 0

tail(data_2014)
# Date.Time Lat Lon Base Time day month year dayofweek hour minute second
# 4534322 2014-09-30 22:57:00 40.7300 -73.9565 B02764 22:57:00 30 Sep 2014 Tue 22 57 0
# 4534323 2014-09-30 22:57:00 40.7668 -73.9845 B02764 22:57:00 30 Sep 2014 Tue 22 57 0
# 4534324 2014-09-30 22:57:00 40.6911 -74.1773 B02764 22:57:00 30 Sep 2014 Tue 22 57 0
# 4534325 2014-09-30 22:58:00 40.8519 -73.9319 B02764 22:58:00 30 Sep 2014 Tue 22 58 0
# 4534326 2014-09-30 22:58:00 40.7081 -74.0066 B02764 22:58:00 30 Sep 2014 Tue 22 58 0
# 4534327 2014-09-30 22:58:00 40.7140 -73.9496 B02764 22:58:00 30 Sep 2014 Tue 22 58 0
```

Slika 2.5: Pregled podataka

Dakle, radi se o okviru podataka koji sadrži:

- Lat - Zemljopisna sirina Uber-ovih preuzimanja (latitude)
- Lon - Zemljopisna dužina preuzimanja Ubera (longitude)
- Base - Baza tj. TLC osnovni kod tvrtke povezan s Uberovim preuzimanjem
- Ostalo je ocito vrijeme, dan, mjesec, godina...

2.3.4. Formati za vrijeme

```
data_2014$Date.Time <- as.POSIXct(data_2014$Date.Time, format = "%m/%d/%Y %H:%M:%S")
# as.POSIXct je funkcija za datum-vrijeme konverziju --> služi za manipulaciju objekata koji predstavljaju datum(vrijeme)

data_2014$Time <- format(as.POSIXct(data_2014$Date.Time, format = "%m/%d/%Y %H:%M:%S"), format="%H:%M:%S")
# format() je funkcija koju koristimo da za formatiranje R objekta koji nam je zgodan za ispis

data_2014$Date.Time <- ymd_hms(data_2014$Date.Time)

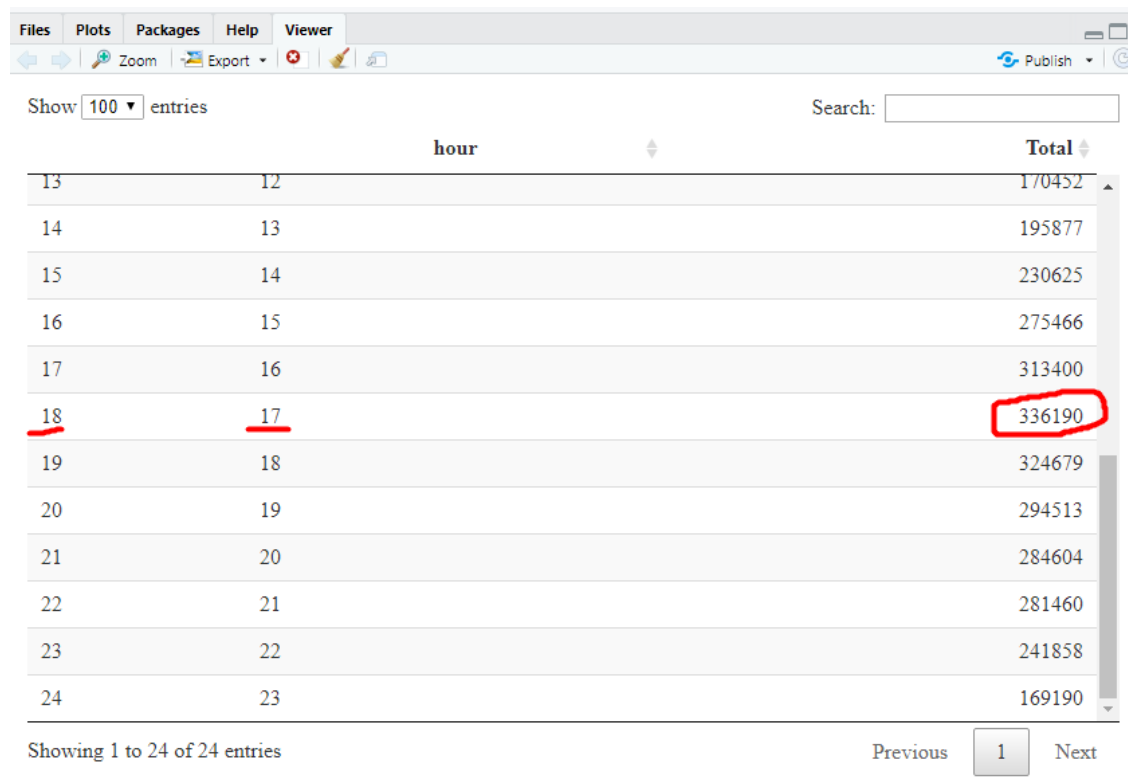
data_2014$day <- factor(day(data_2014$Date.Time))
data_2014$month <- factor(month(data_2014$Date.Time, label = TRUE))
data_2014$year <- factor(year(data_2014$Date.Time))
data_2014$dayofweek <- factor(wday(data_2014$Date.Time, label = TRUE))
#factor() se koristi za enkodiranje vektora u faktor ('category' i 'enumerated type' se također koriste kao faktori)
# funkcije day(), month(), year() su za izoliranje odgovarajućih datumskih segmenata iz standardnog datumskog formata
data_2014$hour <- factor(hour(hms(data_2014$Time)))
data_2014$minute <- factor(minute(hms(data_2014$Time)))
data_2014$second <- factor(second(hms(data_2014$Time)))

# hms() je funkcija koja parsira sate,minute,sekunde
```

Slika 2.6: Formati za datum i vrijeme

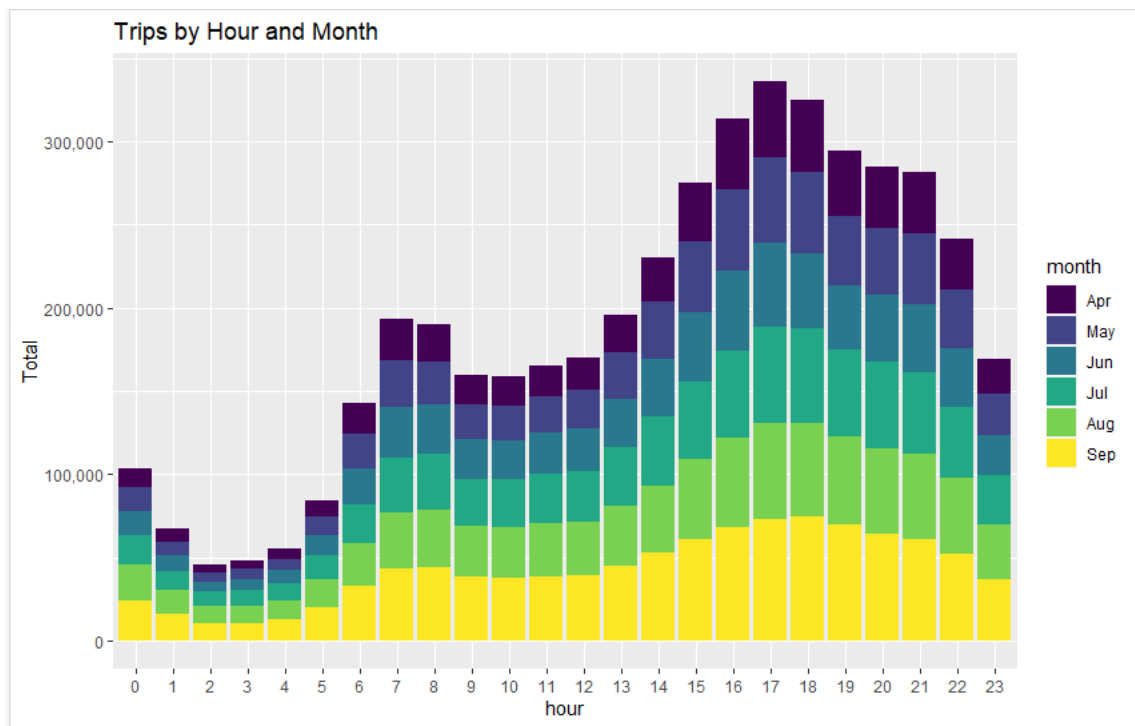
2.3.5. Putovanja po satima u danu

U sljedećem koraku koristila sam funkciju ggplot za crtanje broja putovanja koja su putnici obavili u danu. Također je upotrebljen dplyr za prikupljanje podataka. U vizualizacijama se vidi kako se broj putnika kreće tijekom dana. Primjećujemo da je broj putovanja veći navečer oko 17:00 i 18:00.



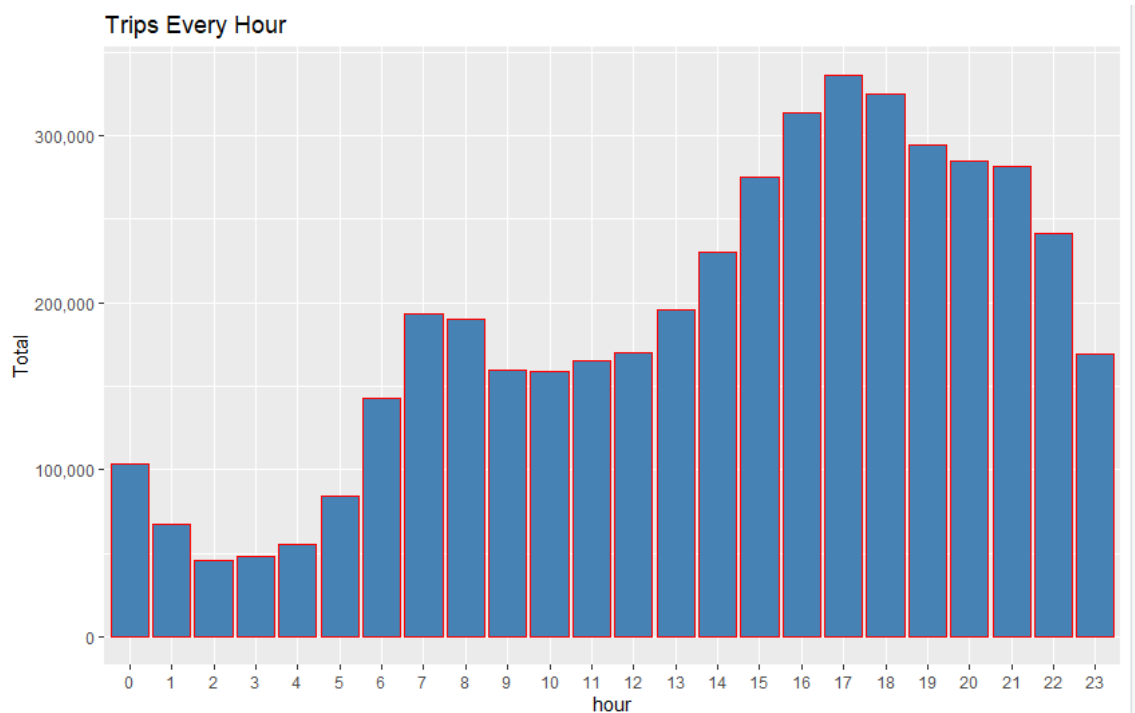
	hour	Total
13	12	170452
14	13	195877
15	14	230625
16	15	275466
17	16	313400
18	17	336190
19	18	324679
20	19	294513
21	20	284604
22	21	281460
23	22	241858
24	23	169190

Slika 2.7: Tablica putovanja po satima u danu



Slika 2.8: Putovanja po satima i mjesecima

Graf na x-osi ima sate, a na y-osi putovanja. Boje odgovaraju mjesecima. Odavdje zaključujemo da se s proljeća na ljeto povećava broj vožnji



Slika 2.9: Putovanja po satima u danu

2.3.6. Podaci po danima u mjesecu

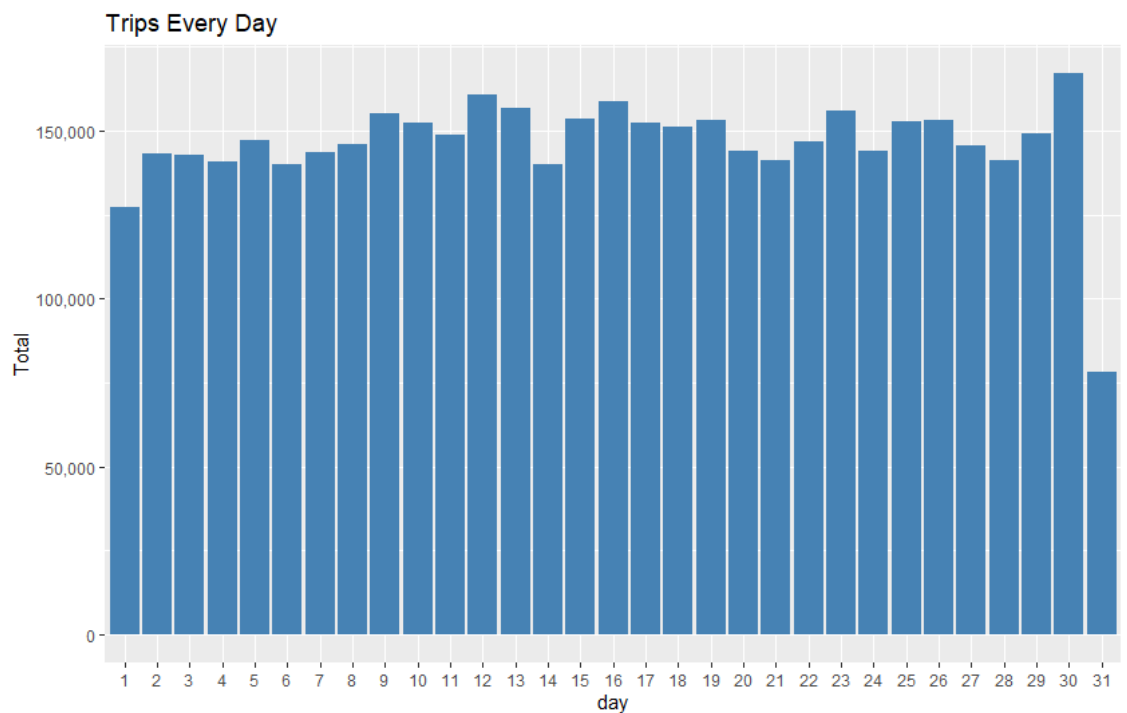
Sastavljanje podataka na temelju svakog dana u mjesecu. Iz vizualizacije koja proizlazi se da uociti da je 30. u mjesecu imao najviše putovanja u godini, čemu je najviše pridonio mjesec travanj.

Show entries Search:

	day	Total
20	20	144179
21	21	141112
22	22	146952
23	23	156032
24	24	144169
25	25	152667
26	26	153405
27	27	145652
28	28	141157
29	29	149086
30	30	167160
31	31	78073

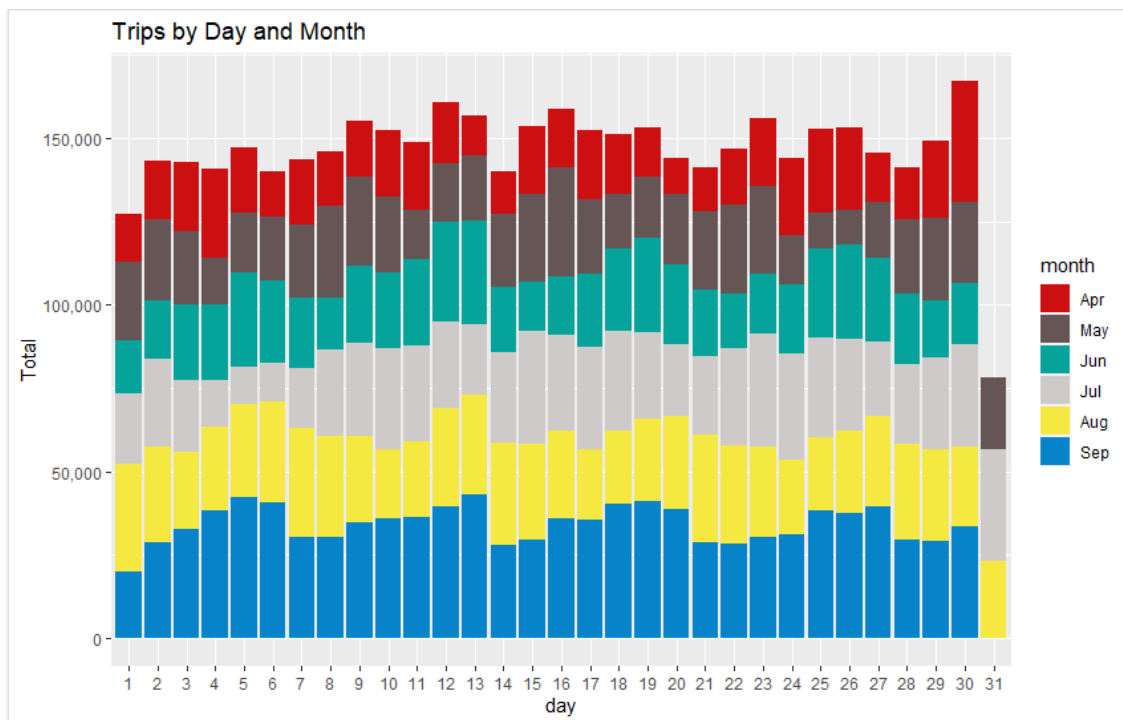
Showing 1 to 31 of 31 entries Previous 1 Next

Slika 2.10: Tablica putovanja po danima u mjesecu



Slika 2.11: Histogram putovanja po danima u mjesecu

Koristim ggplot() za crtanje grafa. Na x-osi su dani u mjesecu, a na y-osi broj vožnji. Generalno, najmanje vožnji je 31. u mjesecu, najviše 30., ostali dani imaju podjednak broj vožnji. Neki mjeseci niti nemaju 31 dan. Očekujemo da će za 31. dan stupac biti upola manji.



Slika 2.12: Histogram putovanja po danima u mjesecu i mjesecima

Graf na x-osi ima dane u mjesecu, a na y-osi putovanja. Boje odgovaraju mjesecima. Oдавдје zaključujemo da se s proljeca na ljeto povećava broj vožnji, a na kraju mjeseca je najveći broj vožnji.

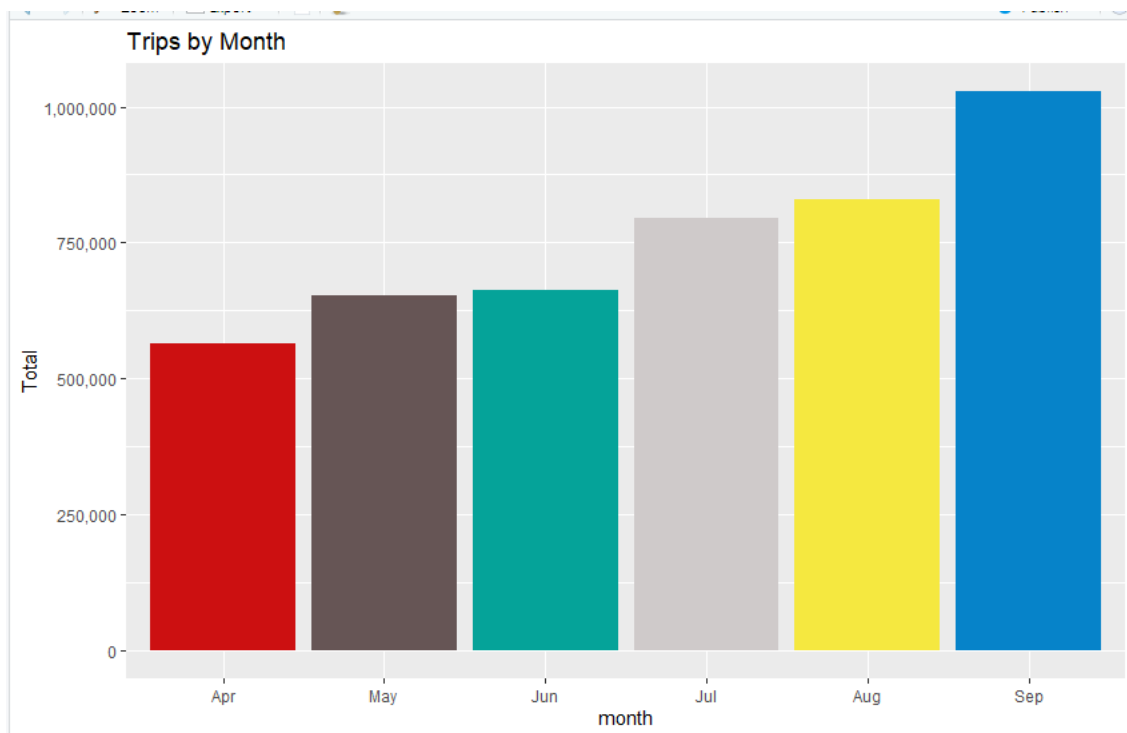
2.3.7. Putovanja po mjesecima

U vizualizaciji se da uočiti da je većina putovanja obavljena tijekom mjeseca rujna. Nadalje, dobivamo i vizualna izvješća o broju putovanja koja su obavljena svakog dana u tjednu.

Show entries Search:

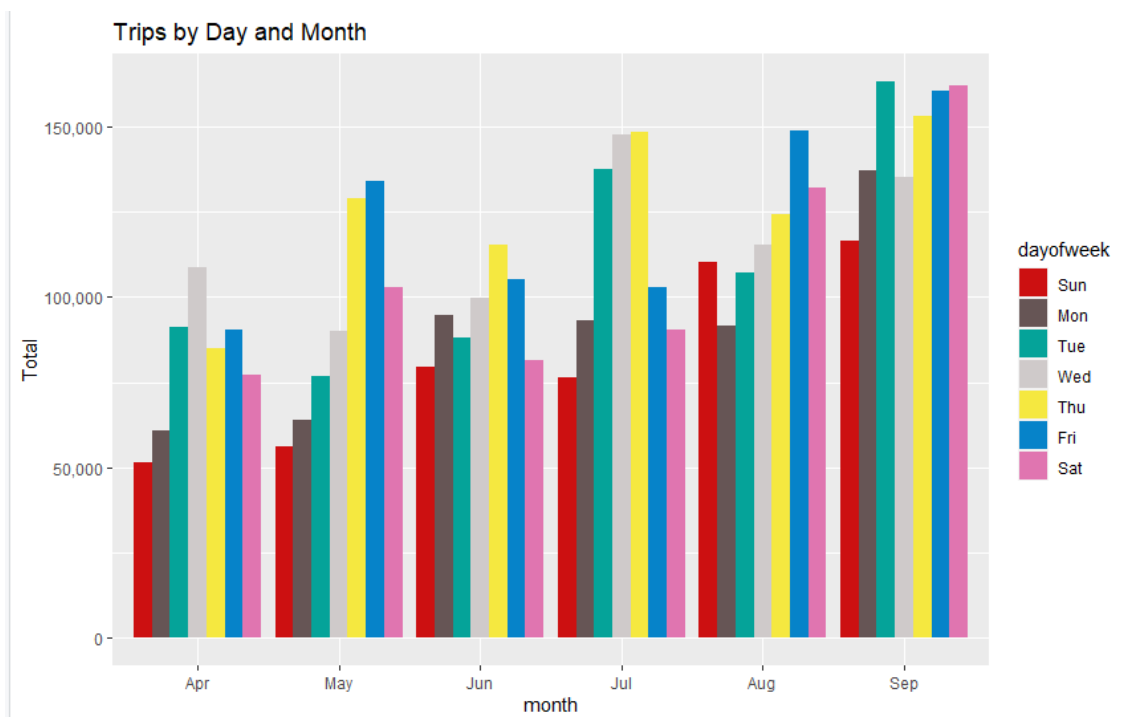
	month	Total
1	Apr	564516
2	May	652435
3	Jun	663844
4	Jul	796121
5	Aug	829275
6	Sep	1028136

Slika 2.13: Tablica putovanja po mjesecima



Slika 2.14: Histogram putovanja po mjesecima

Graf koji na x-osi ima mjesec od travnja do rujna, a na y-osi broj vožnji. Najviše vožnji je u rujnu, a najmanje u travnju s tim da je broj vožnji od travnja do rujna rastao. Možemo to interpretirati na način da se promet povećava kako dolazi ljeto. No, Uber je osnovan 2009. u San Franciscu, tako da u travnju 2014. u NYC-u još bio u razvoju. Prema ovim podacima "raširio" se u pola godine. Dodala bih i da je Uberov konkurent Lyft osnovan 2012. godine te da zbog Lyfta, Uber u 2014. nije imao nagli porast i velike razlike među susjednom mjesecima.



Slika 2.15: Histogram putovanja po danu u tjednu i mjesecu

Na x-osi su prikazani mjeseci [travanj, rujan] , a na y-osi broj vožnji. Stupci odgovaraju danima u tjednu, svaki dan ima svoju boju. Nedjeljom i ponedjeljkom, u pravilu uvijek ima manje vožnji u odnosu na ostale dane. U svibnju su Njujoržani najčešće petkom birali Uber za prijevoz, dok su u ostalim mjesecima se nešto češće vozili četvrtkom.

2.3.8. Putovanja po bazama

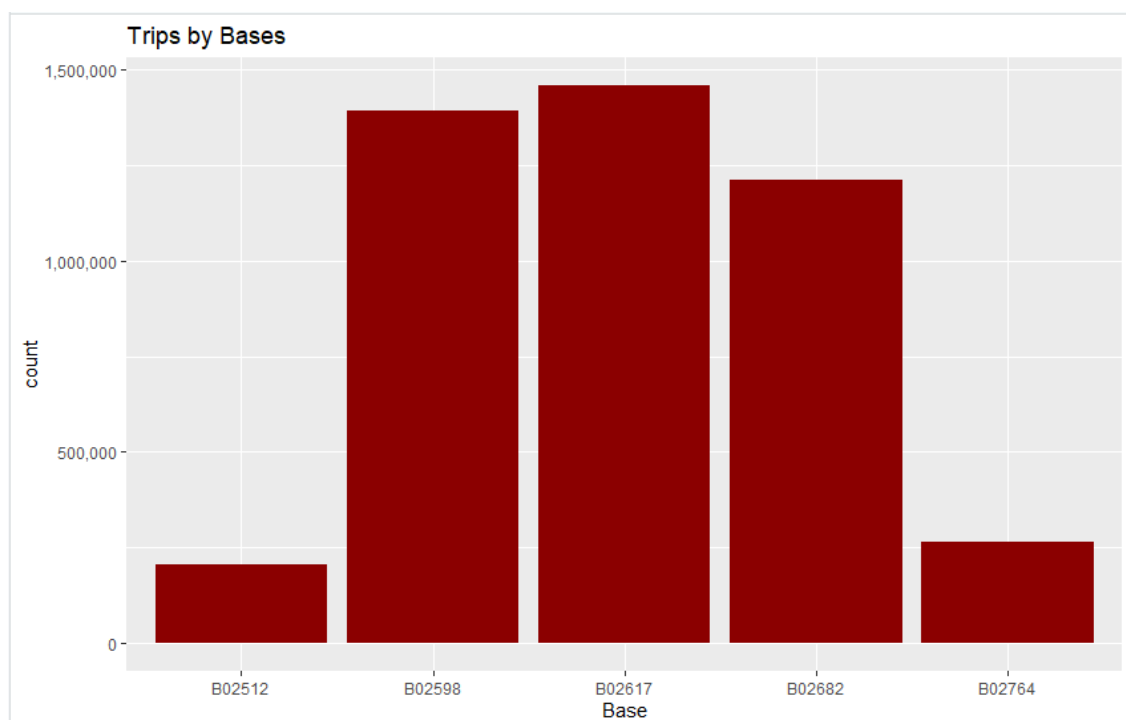
U sljedećoj vizualizaciji ucrtavam broj putovanja koja su putnici obavili iz svake baze. Svega je pet baza, primijetili smo da je B02617 imao najveći broj putovanja. Nadalje, ova baza imala je najveći broj putovanja u mjesecu B02617. U četvrtak je zabilježeno najviše putovanja u tri baze - B02598, B02617, B02682.

Najpopularnije Uber- baze u NYC (prvih 5 prikazujem na grafu) :

- B02512: Unter
- B02598: Hinter
- B02617: Weiter
- B02682: Schmecken
- B02764: Danach-NY

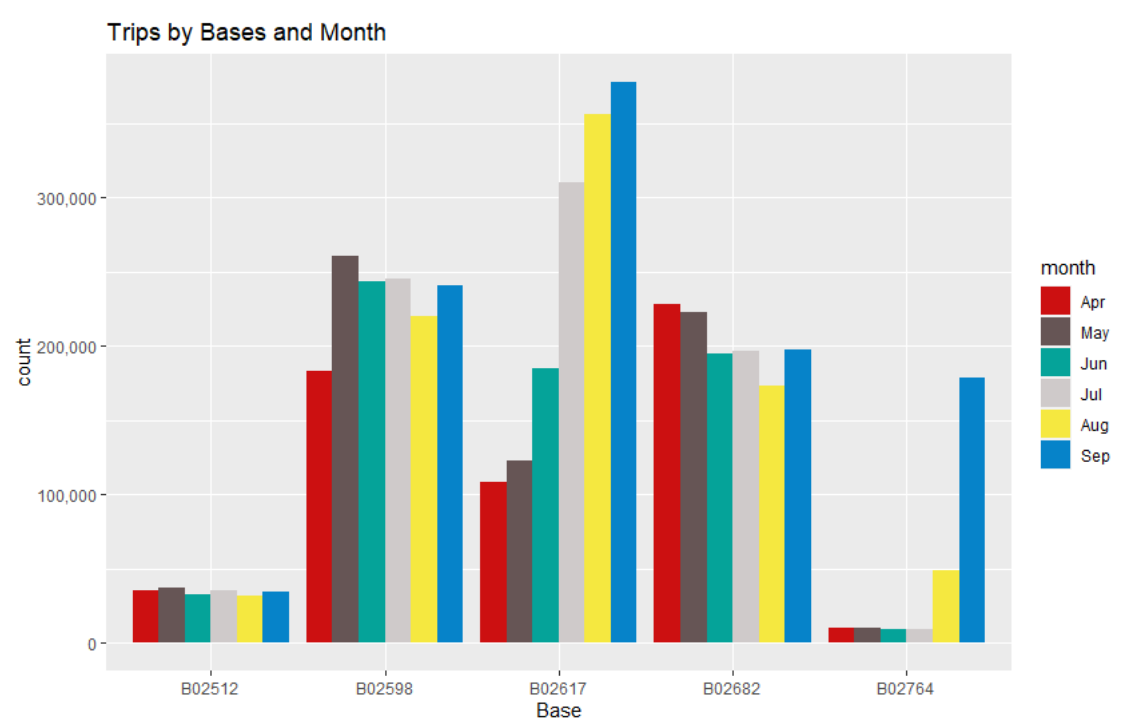
- B02765: Grun
- B02835: Dreist
- B02836: Drinnen

Vidimo da baze označavaju prijedloge na njemačkom jeziku pa možemo zaključiti koja baza odgovara kojem dijelu grada. (Unter= ispod; Hinter= iza; Weiter= sljedeći ili nastavak itd, Danach= nakon toga)



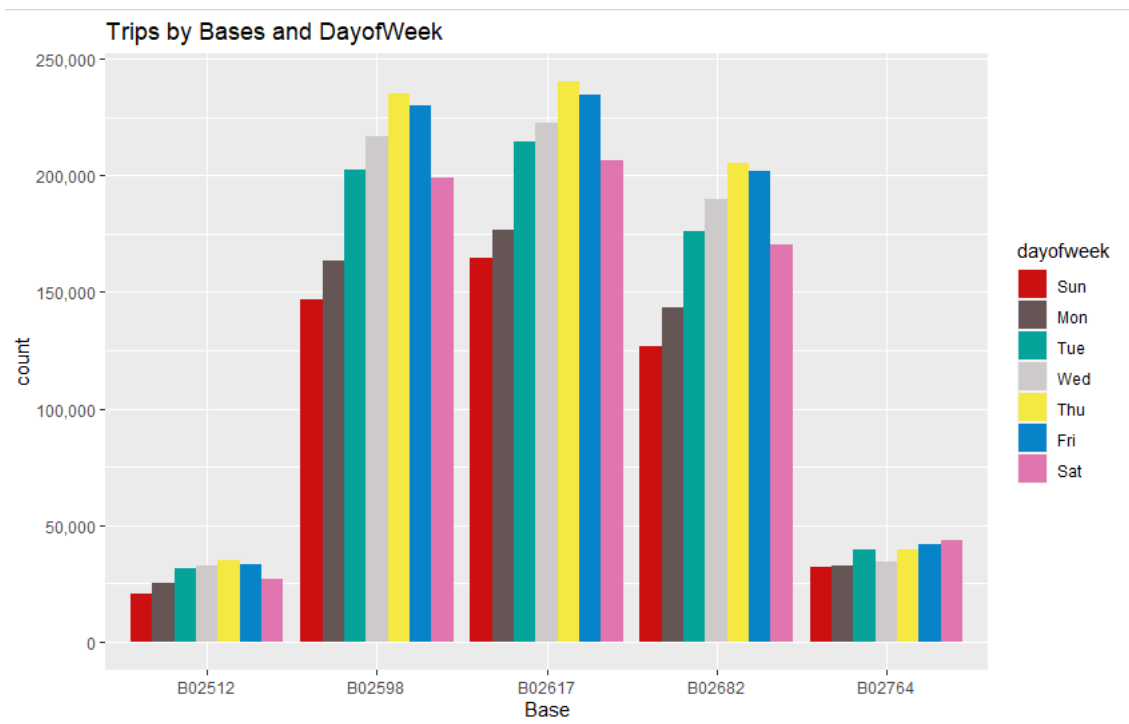
Slika 2.16: Histogram putovanja po bazama

Na x-osi su prikazane baze, a na y-osi broj vožnji. Iz grafa vidimo da je B02617 (Weiter) najpopularnija baza s gotovo 1.5M vožnji, a slijede ju Hinter (1.4M) i Unter (1.2M)



Slika 2.17: Histogram putovanja po bazama i mjesecima

X-os su baze, y-os broj vožnji, stupci predstavljaju mjesece, a svaki mjesec je označen drugom bojom. Što se tiče baze Unter, u svim mjesecima je imala podjednak broj vožnji, kao i Hinter i Schmecken.



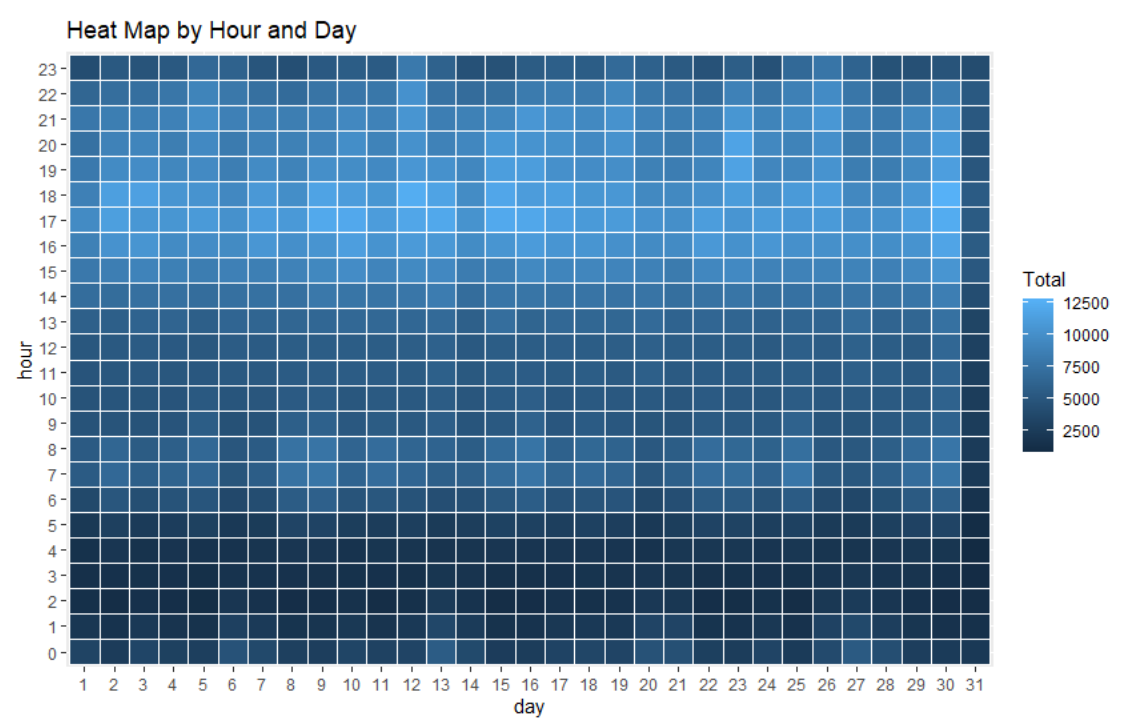
Slika 2.18: Histogram putovanja po bazama i danu u tjednu

X-os su baze, y-os broj vožnji, stupci predstavljaju dane tjedna, a svaki dan je označen drugom bojom. Što se tiče baze Unter, u svim danima je imala podjednak broj vožnji, kao i Danach. Ostale tri baze su imale manje prometa nedjeljom i ponedjeljkom, više ostalim danima, a najviše četvrtkom i petkom. Bilo bi zanimljivo provjeriti kakvo je stanje sada četvrtkom i petkom s obzirom na to da je u mnogim firmama u NYC petak dan za rad od kuće i to se popularno zove "Petak u papučama".

2.3.9. Stvaranje Heatmap vizualizacije dana, sata i mjeseca

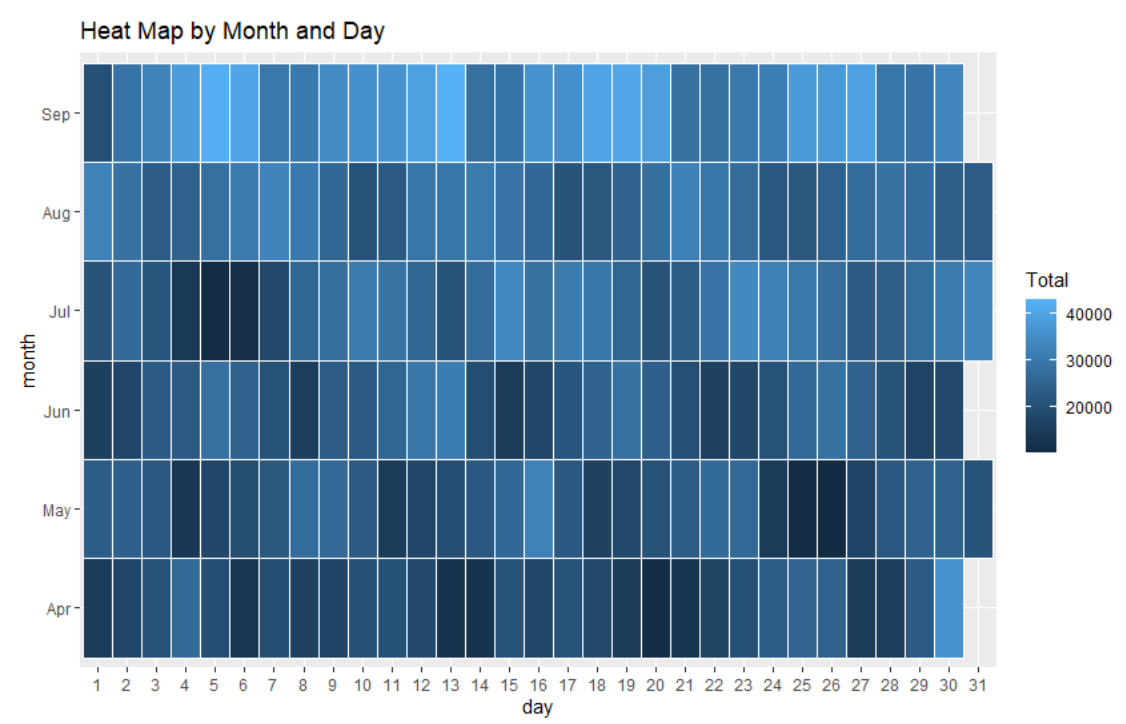
Ucrtat cu pet parcela toplotne karte.

- Prvo cu nacrtati toplinsku kartu po satima i danima.
- Drugo, sastavit cu toplinsku kartu po mjesecima i danima.
- Treće, toplotna karta po mjesecima i danima u tjednu.
- Četvrto, toplinska karta koja ocrtava mjesec i baze.
- Na kraju cu ucrtati toplotnu kartu prema bazama i danima u tjednu.



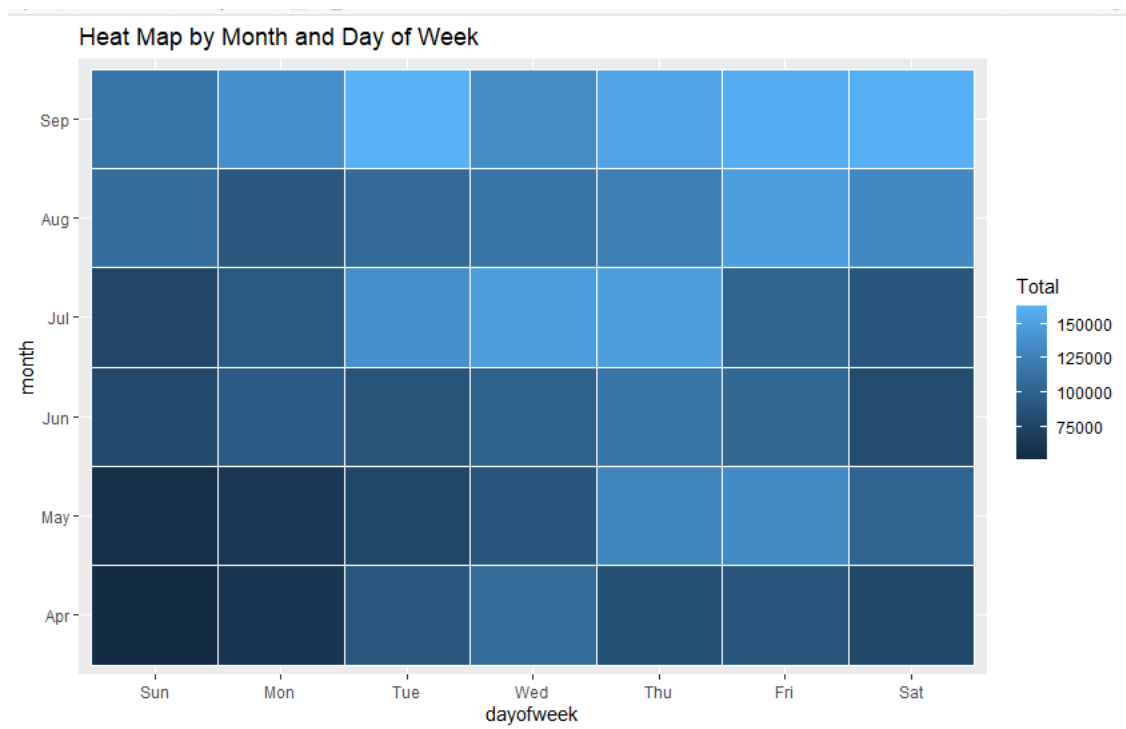
Slika 2.19: Toplinska mapa po satu i danu u mjesecu

Toplotna mapa kojoj su na x-osi dani, na y-osi sati, a jačina boje odgovara broju vožnji. Iz toplotne karte vidimo da je najviše vožnji između 17 i 20h, a najmanje u intervalu od 1 do 5h.



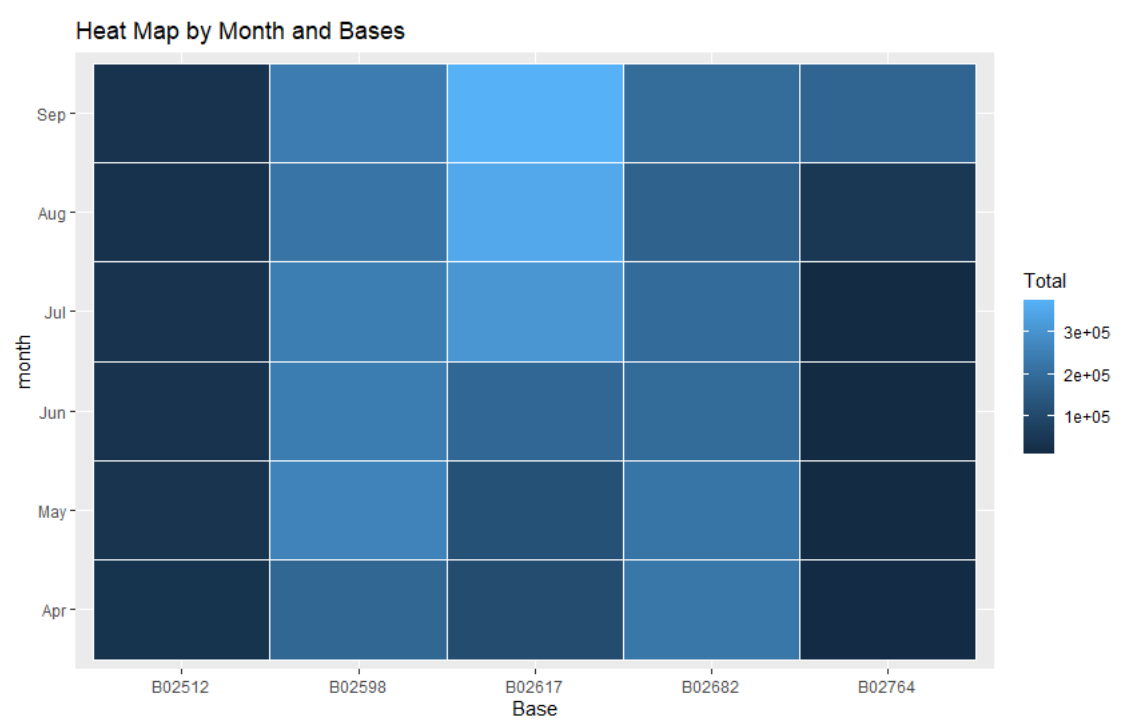
Slika 2.20: Toplinska mapa po mjesecu i danu u mjesecu

Toplotna mapa gdje su dani u mjesecu na x-osi, a mjeseci na y-osi...dok jačina boje odgovara broju vožnji. Vidimo da je nekako najviše vožnji bilo u rujnu u razmacima od 7 dana, a najmanje u travnju. Opcenito se da zaključiti da preko proljeća bude manje vožnji u odnosu na ljetne mjesece!



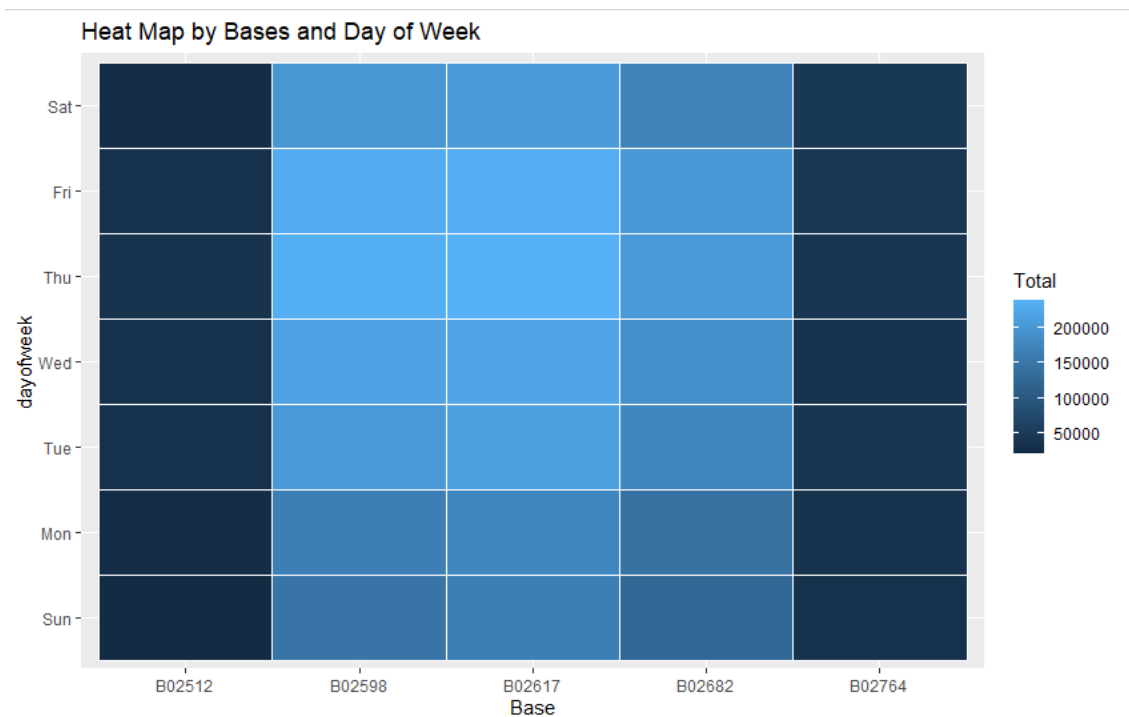
Slika 2.21: Toplinska mapa po mjesecu i danu u tjednu

Toplotna mapa sto na x-osi ima dane u tjednu, a na y-osi mjesece...jačina boje označava broj vožnji. Najmanje vožnji je bilo nedjeljama u travnju i svibnju, a najviše petkom i subotom u kolovozu i rujnu. Možemo zaključiti da bude manje vožnji početkom tjedna u proljetnim mjesecima, pa se to povećava kako ide ljeto, radni tjedan i vikend.



Slika 2.22: Toplinska mapa po mjesecu i bazi

Na x-osi je 5 Uber baza, a na y-osi mjeseci od travnja do rujna. Jačina boje označava broj vožnji. Baze B02512 i B02764 imaju manje vožnji tokom mjeseci koje promatramo u odnosu na ostale tri baze. Baza B02617 ima znatno više vožnji tijekom ljetnih mjeseci u odnosu na druge baze.



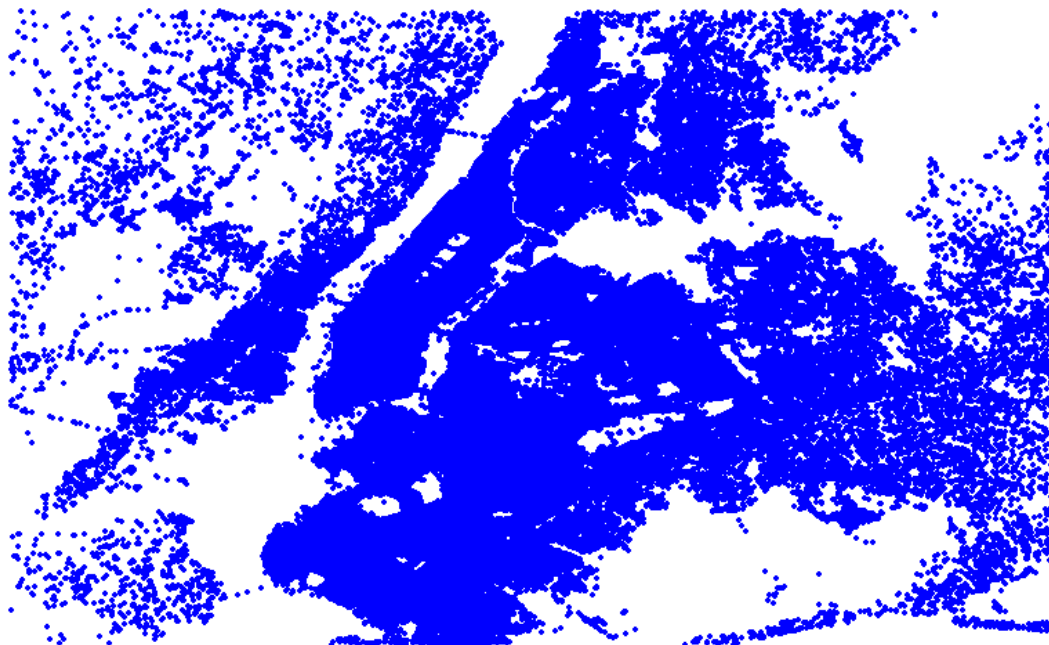
Slika 2.23: Toplinska mapa po bazi i danu u tjednu

Na x-osi je 5 Uber baza, a na y-osi dani u tjednu. Jačina boje označava broj vožnji. Baze B02512 i B02764 imaju manje vožnji tokom cijelog tjedna u odnosu na ostale tri baze. Baze B02617 i B02598 imaju jako puno vožnji četvrtkom i petkom.

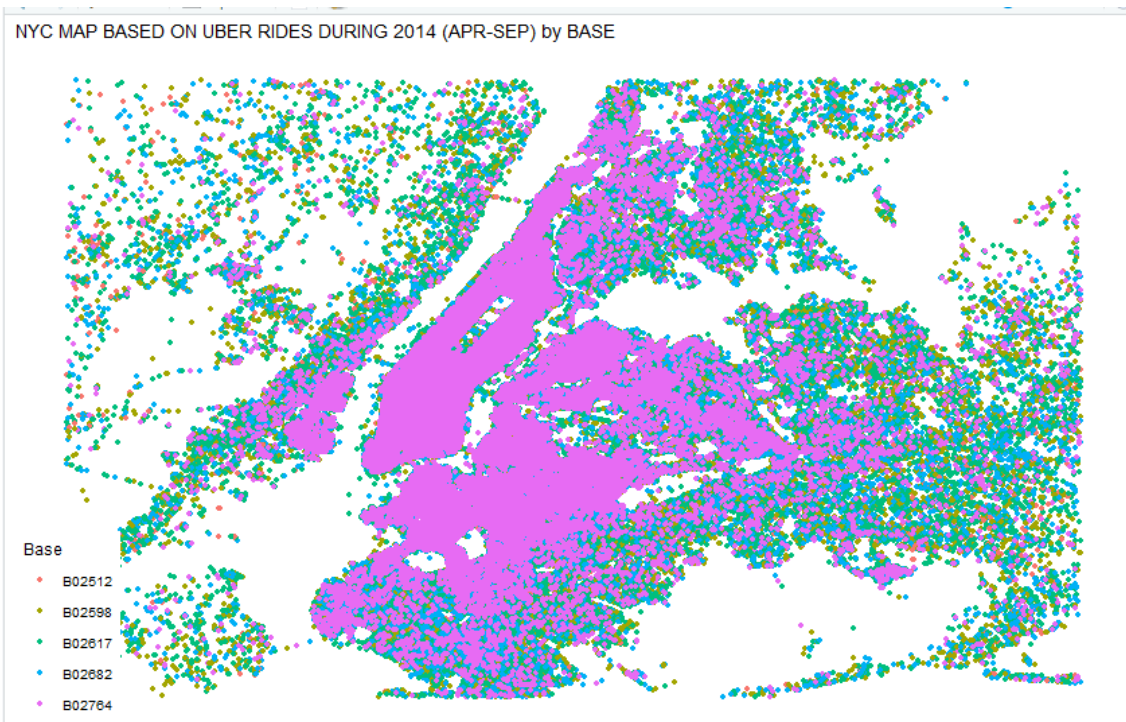
2.3.10. Stvaranje vizualizacije karte u New Yorku

U posljednjem ćemo dijelu vizualizirati vožnje u New Yorku stvaranjem geo-parcele koja će nam pomoći da vizualiziramo vožnje tijekom 2014. (travanj - rujna) i baza u istom razdoblju.

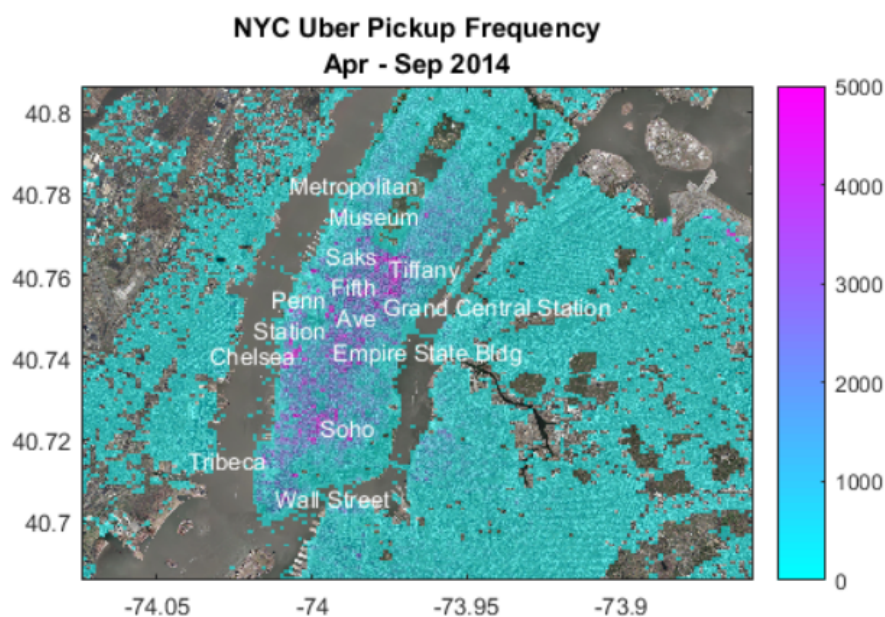
NYC MAP BASED ON UBER RIDES DURING 2014 (APR-SEP)



Slika 2.24: Vožnje na karti NYC



Slika 2.25: Vožnje na karti NYC s bazama prema legendi



Slika 2.26: Frekvencija vožnji po bazama na karti NYC

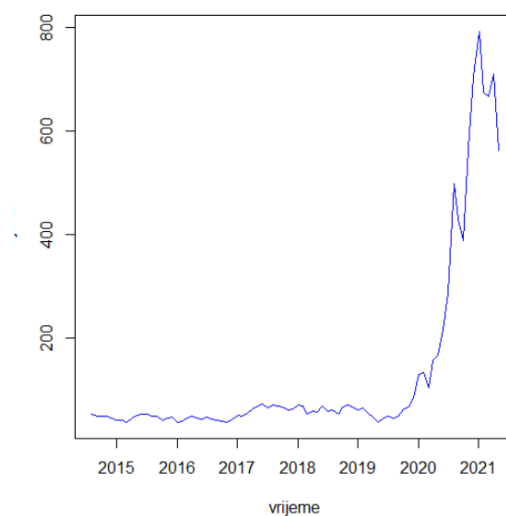
Ovime bismo mogli zaključiti kako je vrijeme utjecalo na putovanja kupaca. Ko-
načno, imamo geo-parcelu New York Cityja koja nam je pružila detalje o tome kako
su različiti korisnici putovali iz različitih baza.

2.4. Rješenje problema - Predviđanje cijena dionica u realnom vremenu

2.4.1. Regresijski modeli nad podacima dionica Tesle

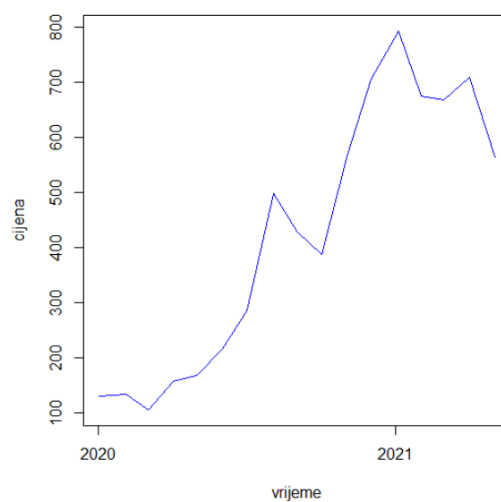
Vidjeli smo kako se kreću podaci za Uber, a sada je na redu Tesla. Za početak ću
prikazati kretanje cijena dionica Tesle.

Na sljedećem grafu možemo vidjeti kako su se kretale cijene dionica Tesle od
2014.-2021. godine.



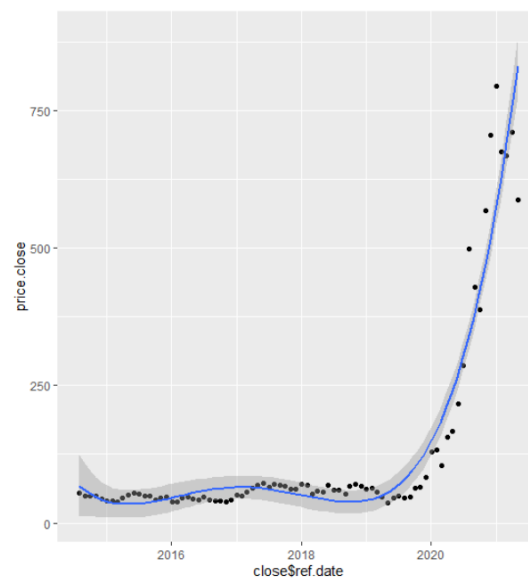
Slika 2.27: Kretanje cijena Teslinih dionica

Nadalje, približimo dio od 2020.-2021. da bismo pogledali svježije podatke iz bliske prošlosti.



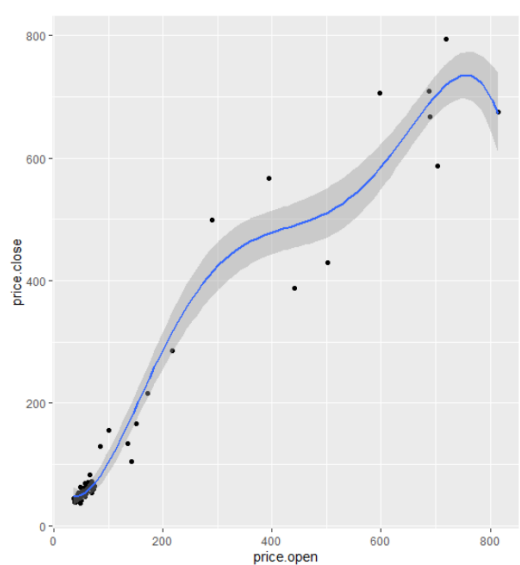
Slika 2.28: Kretanje cijena Teslinih dionica od 2020.-2021.

Na idućem grafu možemo vidjeti kako se graf polinoma petog stupnja prilagodio podacima korištenjem polinomijalne regresije i funkcije `lm()`. Promatramo cijene u odnosu na vrijeme.



Slika 2.29: Regresijski model za predviđanje cijena dionica u odnosu na vrijeme

Graf ispod prikazuje regresijski model pomoću kojeg predviđamo kolika će biti cijena zatvaranja ako znamo cijenu otvaranja.



Slika 2.30: Predviđanje cijena zatvaranja na temelju cijena otvaranja

2.4.2. ARIMA

ARIMA modeli pružaju jedan pristup predviđanju vremenskih nizova. Eksponencijalno izgladivanje (Exponential smoothing) i ARIMA modeli dva su najčešće korištena pristupa predviđanju vremenskih serija i pružaju komplementarne pristupe pro-

blemu. Iako se eksponencijalni modeli zaglađivanja temelje na opisu trenda i sezonalnosti u podacima, ARIMA modeli imaju za cilj opisati autokorelacije u podacima.

Nesezonski (Non-sesional) modeli

Nesezonski ARIMA modeli Ako kombiniramo diferenciranje s autoregresijom i modelom pokretnog prosjeka, dobivamo nesezonski ARIMA model. ARIMA je skraćenica od AutoRegressive integrated moving average (integrirani pomični prosjek). Cjeloviti model možemo zapisati kao

$$y'_t = c + \phi_1 y'_{t-1} + \dots + \phi_p y'_{t-p} + \theta_1 \varepsilon_{t-1} + \dots + \theta_q \varepsilon_{t-q} + \varepsilon_t,$$

Slika 2.31: ARIMA model - matematički izraz

Gdje je y'_t serija diferencija (možda je diferencirana više puta). "Prediktori" s desne strane uključuju obje zaostale vrijednosti y_t i zaostale pogreške. To nazivamo ARIMA (p,d,q) model, gdje :

- p = redoslijed autoregresivnog dijela;
- d = stupanj prve razlike koja je uključena;
- g = redoslijed dijela pokretnog prosjeka.

Specijalni slučajevi ARIMA modela :

White noise	ARIMA(0,0,0)
Random walk	ARIMA(0,1,0) with no constant
Random walk with drift	ARIMA(0,1,0) with a constant
Autoregression	ARIMA(p,0,0)
Moving average	ARIMA(0,0,q)

Slika 2.32: Specijalni slučajevi ARIMA modela

ARIMA modeliranje u R-u

Funkcija `auto.arima()` u R-u koristi varijaciju Hyndman-Khandakar algoritma (Hyndman Khandakar, 2008), koja kombinira jedinstvene testove korijena, minimiziranje AICc i MLE kako bi se dobio ARIMA model. Argumenti za `auto.arima()` pružaju mnoge varijacije algoritma. Ovdje je opisano zadano ponašanje.

Hyndman-Khandakar algoritam za automatsko ARIMA modeliranje

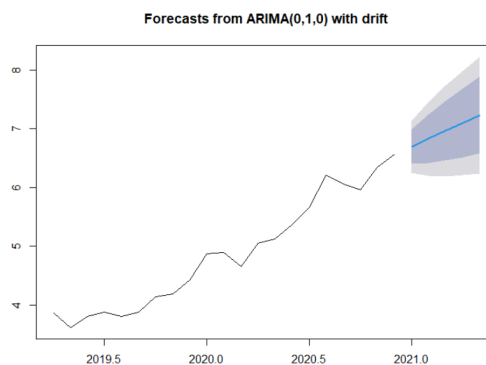
1. Broj razlika $0 < d < 2$ je obilježen korištenjem ponavljajućih KPSS testova.
 2. Vrijednosti p i q su izabrane minimiziranjem AICc nakon diferenciranja podataka d puta.
Algoritam ne provjerava sve kombinacije za p i q već koristi stepwise pretragu.
- a) Četiri inicijalna modela su prilagođena:
- ARIMA(0,d,0),
 - ARIMA(2,d,2),
 - ARIMA(1,d,0)
 - ARIMA(0,d,1)
- Konstanta je uključena osim ako je $d=2$. Ako je $d < 1$, dodatni model je također prilagođen:
- ARIMA(0,d,0) bez konstante.
- b) Najbolji model (s najmanjom AICc vrijednošću) prilagođen u koraku a) se postavlja kao "trenutni model".
- c) Varijacije na trenutni model se razmatraju:
- p i/ili q iz trenutnog modela (+/-1)
 - uključivanje/isključivanje konstante iz trenutnog modela
- Nakon toga se određuje trenutni model (ili onaj koji je već bio određen ili neka varijacija)
- d) Ponavljanje koraka 2.c) dok se ne može naći niži AICc.

Slika 2.33: Hyndman-Khandakar algoritam za automatsko ARIMA modeliranje

- *MLE - U statistici je procjena maksimalne vjerojatnosti (MLE) metoda procjene parametara raspodjele vjerojatnosti maksimiziranjem funkcije vjerojatnosti, tako da su promatrani podaci prema pretpostavljenom statističkom modelu najvjerojatniji. Točka u prostoru parametara koja maksimizira funkciju vjerojatnosti naziva se procjenom najveće vjerojatnosti.
- *KPSS test - statistički test za provjeru stacionarnosti niza oko determinističkog trenda
- *Stacionarne vremenske serije su one serije čije su statističke značajke poput srednje vrijednosti, varijance i autokorelacije, konstantne kroz vrijeme
- * Kada imamo neko razumno fizičko objašnjenje trenda, to želimo iskoristiti pa objasnimo trend deterministički. Npr, deterministički rastući trend može biti posljedica porasta populacije, a neke promijene koje se ciklički ponavljaju mogu nastati zbog određene sezonske komponente. Deterministički trendovi i sezonske varijacije mogu se modelirati pomoću regresije.
- *Diferenciranje - jedan od načina da se nestacionarne vremenske serije učine stacionarnima - računanje razlika između uzastopnih promatranja
- *AICc – AIC s ispravkom za male veličine uzorka
- *AIC - Akaikeov informacijski kriterij (AIC) procjenjuje pogrešku predviđanja, a time i relativnu kvalitetu statističkih modela za zadani skup podataka. S obzirom na zbirku modela podataka, AIC procjenjuje kvalitetu svakog modela u odnosu na svaki drugi model. Dakle, AIC pruža sredstva za odabir modela.

2.4.3. Predviđanje cijena dionica Tesle u budućnosti - ARIMA

Grafički prikaz odgovara cijenama TESlinih dionica (crna linija) i cijenama koje su dobivene korištenjem funkcije predict() iz paketa ARIMA (plava linija).



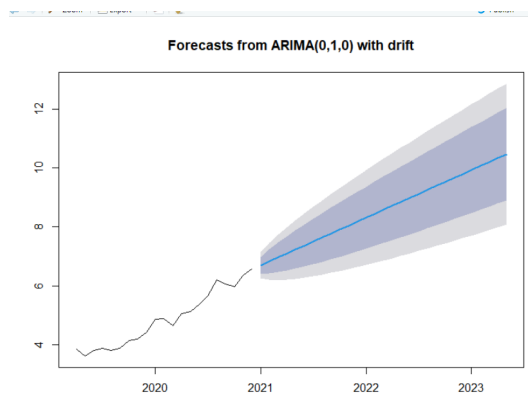
Slika 2.34: Predviđanje cijena - ARIMA

Možemo usporediti predviđene cijene s podacima za testiranje. Vidimo da su predviđene cijene OK. Razlikuju se za 100, što je prihvatljivo za potrebe ovog seminara.

```
> okvir
Prava cijena Predvidjena cijena
1      793.53      807.3996
2      675.50      923.7945
3      667.93     1056.9690
4      709.44     1209.3420
```

Slika 2.35: Usporedba predviđenih cijena i podataka za testiranje

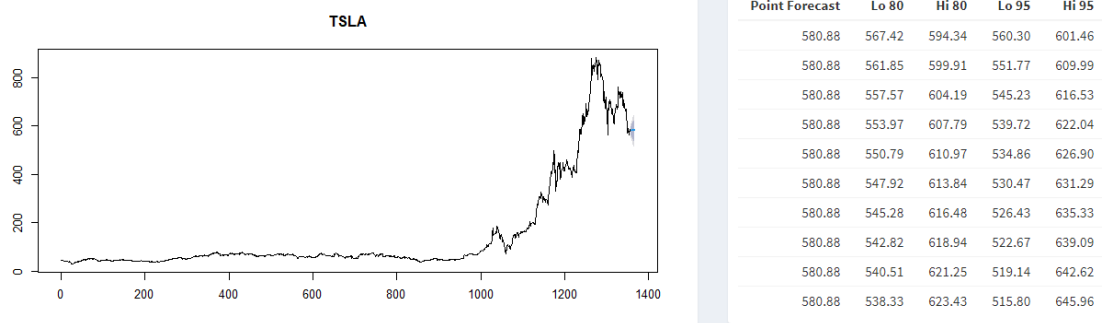
Predviđanje cijena u budućnosti (do 2023. godine) je prikazano na grafu ispod.



Slika 2.36: Predviđanje cijena u budućnosti

2.4.4. Predviđanje cijena dionica u realnom vremenu

Ovako izgleda primjena ARIMA modela u aplikaciji napravljenoj pomoću Shiny-a i R-a.



Slika 2.37: Predviđanje cijena dionica Tesle u realnom vremenu

2.5. Usporedba rezultata Kritički osvrt

Što se tiče mog prvog problema, tu sam pokušavala koristiti razne metode za vizualizaciju, ali je ggplot paket pruža daleko najbolje metode prikaz i vizualizaciju Uber vožnji zato što omogućuje stvaranje histograma, toplotnih karata i najzanimljivije, prikaz podataka prema geografskoj karti.

Predvidjeti cijene dionica sam prvo pokušala modelom linearne regresije te sam odmah odustala jer pravac ne oponaša dobro kretanje podataka koji osciliraju kroz vrijeme. Zatim sam pokušala raditi s polinomijalnom regresijom te sam dobila dobre rezultate što se tiče podataka koje sam već imala. Graf polinoma kojeg sam dobila je dobro opisivao kretanje cijena dionica i to bi se moglo iskoristiti ako nemamo podatke za neki datum, a trebaju nam. Na primjer, ako želim znati kolika je bila cijena zatvaranja 23.6.2018., ako znam cijenu otvaranja. Ali, ono što se događalo u prošlosti često nije od interesa. Zanimalo me kako mogu predvidjeti kako će se cijene kretati u budućnosti. Zbog toga sam koristila ARIMA model. Konkretno, ARIMA(1,0,1) model s konstantom koji se zove "Random walk with draft". Prvo sam uradila model s predviđanjem i usporedila s podacima za testiranje. Rezultati su bili zadovoljavajući. Zatim sam modelirala predviđanje cijena dvije godine unaprijed te sam konačno dobila ono što mi je bio inicijalni interes.

3. Literatura

[1] <https://otexts.com/fpp2/arma-r.html>. ARIMA modelling in R.