Projekt - Settlers of Catan stats

36544062 Marko Pongrac, 36540794 Marko Kukolj

2024-01-26

0. MOTIVACIJA I UVOD

Kao ljubitelji društvene igre Catan te osobe koje ne vole gubiti, autori ovog rada izabrali su ovu temu s ciljem pronalaženja informacija, kako bi poboljšali svoju strategiju te šansu za pobjedu prilikom igranja s prijateljima. Zbog toga, ali i nekih iskustvenih pojava, ovaj rad pokušava dati odgovor na sljedeća pitanja:

Je li kocka zaista poštena?

Postoji li neka kombinacija resurs(a) na početnim susjednim poljima koja povećava vjerojatnost pobjede?

Postoji li neka kombinacija broj(eva) na početnim susjednim poljima koja povećava vjerojatnost pobjede?

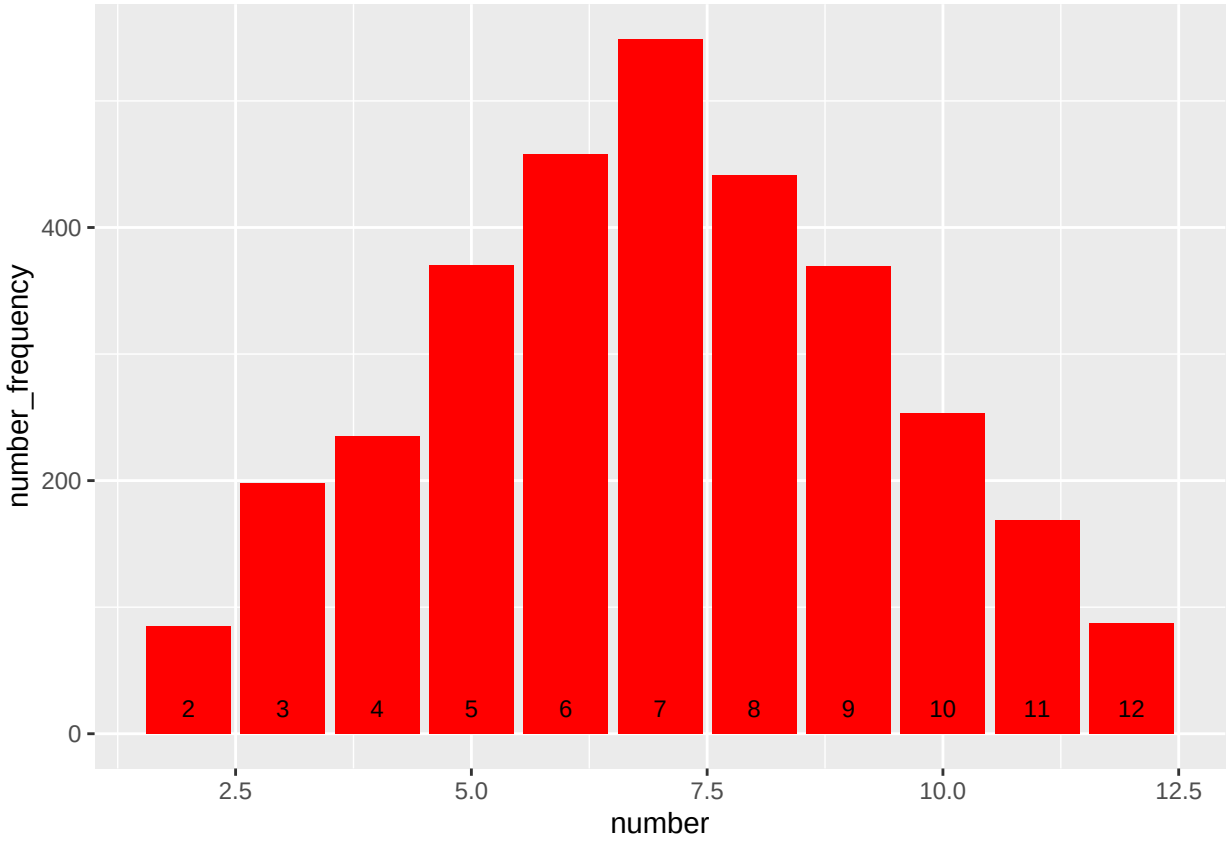
Utječe li redoslijed postavljanja naselja na pobjednika?

```
source("../R/CatanDataLoading.R")
```

1. Je li kocka poštena?

Svima se sigurno dogodila partija u kojoj su naselja lukavo postavili na poziciju susjednu poljima s brojevima koji bi teoretski trebali češće padati, međutim magično su se brojevi 3 i 12 pojavljivali češće nego 8 i 9. Budući da se i njima samima dogodila takva partija, autori rada odlučili su prvo istražiti je li kocka poštena. Odnosno, pojavljuje li se zaista broj 7 najčešće, malo rjeđe brojevi 6 i 8, i tako do bojeva 2 i 12, koji bi se najrеđe trebali pojavljivati. Budući da imamo prilično velik uzorak bacanja, distribucija zbrojeva bi prema centralnom graničnom teoremu trebala biti slična normalnoj.

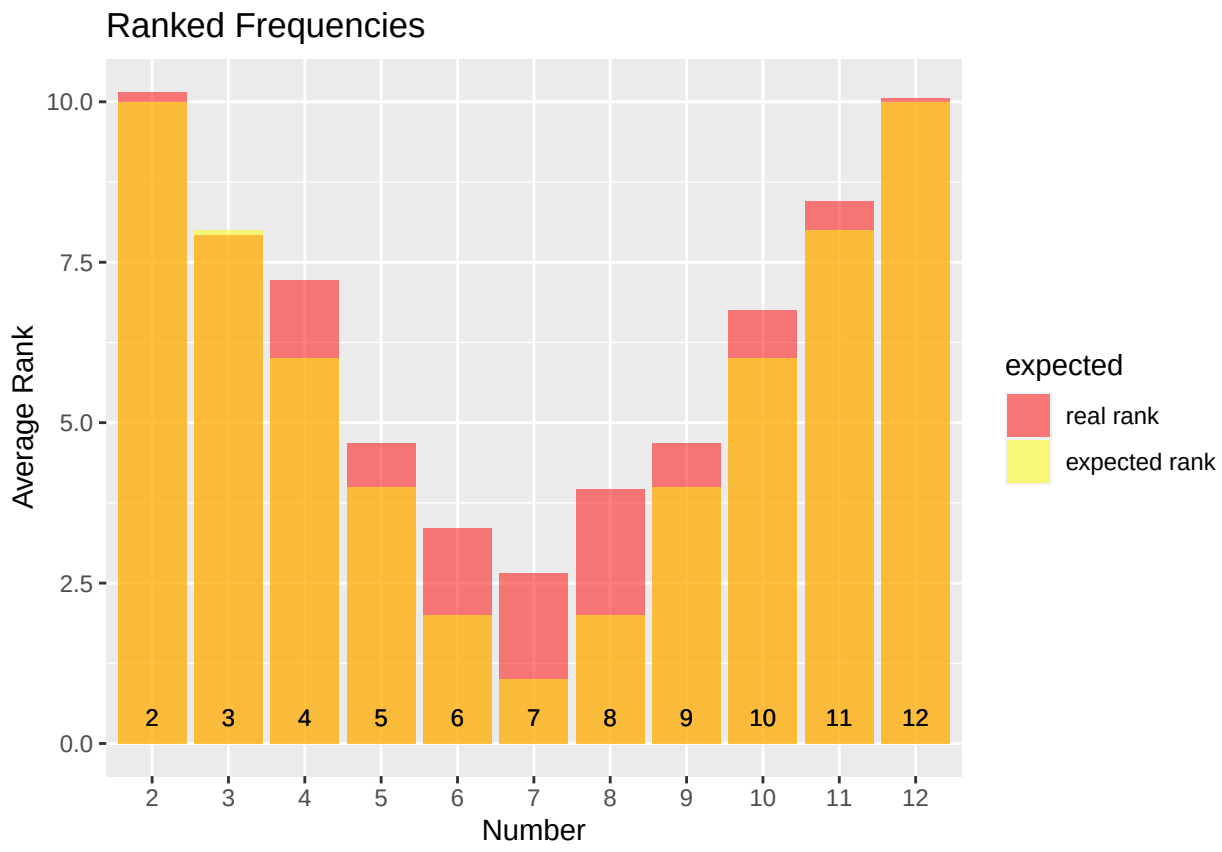
```
source("../R/CatanAnalyseMatchData.R")

dice_frequency_plot <- ggplot(frequency_df, aes(x = number, y = number_frequency)) +
  geom_bar(fill = "red", stat = "identity") +
  geom_text(aes(x = number, y = 0, label = number), stat = "identity", vjust = -1, size = 3)
dice_frequency_plot %>% print()
```



Međutim, i dalje je moguće da postoje nepravilnosti u broju pojavljivanja određenog zbroja na kockama. Zbog toga uvodimo novu metriku "average rank". Ona u svakoj partiji rangira broj pojavljivanja određenog zbroja(zbroj koji se najčešće pojavio ima rang 1 u partiji, sljedeći 2 itd.) te ćemo odrediti koji je prosječni rang svakog zbroja na kockama. Taj broj ćemo zatim usporediti s teoretskim prosječnim rangom.

```
ranked_frequencies_plot <- ggplot(ranked_df, aes(x = factor(number), y = average_rank, fill = expected))
  geom_col(position = "identity", alpha = 0.5) +
  geom_text(aes(x = number, y = 0, label = number), stat = "identity", vjust = -1, size = 3) +
  scale_fill_manual(values = c("FALSE" = "red", "TRUE" = "yellow"), labels = c("real rank", "expected ra
  labs(title = "Ranked Frequencies", x = "Number", y = "Average Rank")
```

```
ranked_frequencies_plot %>% print()
```



Primijetimo da oko sredine, odnosno najčešće dobivenih brojeva postoje određene razlike u očekivanom i stvarnom rangu. To je očekivano, budući da je taj dio najosjetljiviji na promjene. I iako stvarni poredak broja pojavljivanja odgovara teoretskom, oko između kombinacija brojeva 5, 6, 8, i 9 postoje izrazito male razlike.

ZAKLJUČAK

Možemo zaključiti da kocka zaista je poštena, te da ako se igračima čini da je nepoštena, ne igraju dovoljno često Catan da bi primijetili da je poštena ili jednostavno ne mogu prihvatiti da su loši u igri.

2. Utjecaj početka igre na pobjedu

Najbitnija stvar u igri Catan je kao i u svakoj društvenoj igri pobjediti, pa ćemo nakon prethodne, očito pogrešne, analize da je kocka poštena (svatko tko je igrao bilo kakvu društvenu igru zna da kocka nikako nije poštena) pokušati otkriti kako pobijediti. Prvo ćemo razmatrati kako početak igre utječe na izglede pobjede. Pokušati ćemo odgovoriti na dva pitanja u ovom dijelu:

Utječu li, i ako da kako, početni resursi na vjerojatnost pobjede?

Utječu li, i ako da kako, brojevi na poljima oko našeg početnog settlementa na vjerojatnost pobjede?

2.1. Utjecaj početnih resursa na pobjedu

Prije početka legenda resursa:

L = lumber (hrv. drvo)

C = clay (hrv. cigla)

S = sheep (hrv. ovca)

W = wheat (hrv. žito)

O = ore (hrv. kamen)

3G = 3:1 general port (hrv. luka)

2(X) = 2:1 port for resource X (hrv. luka za resurs X)

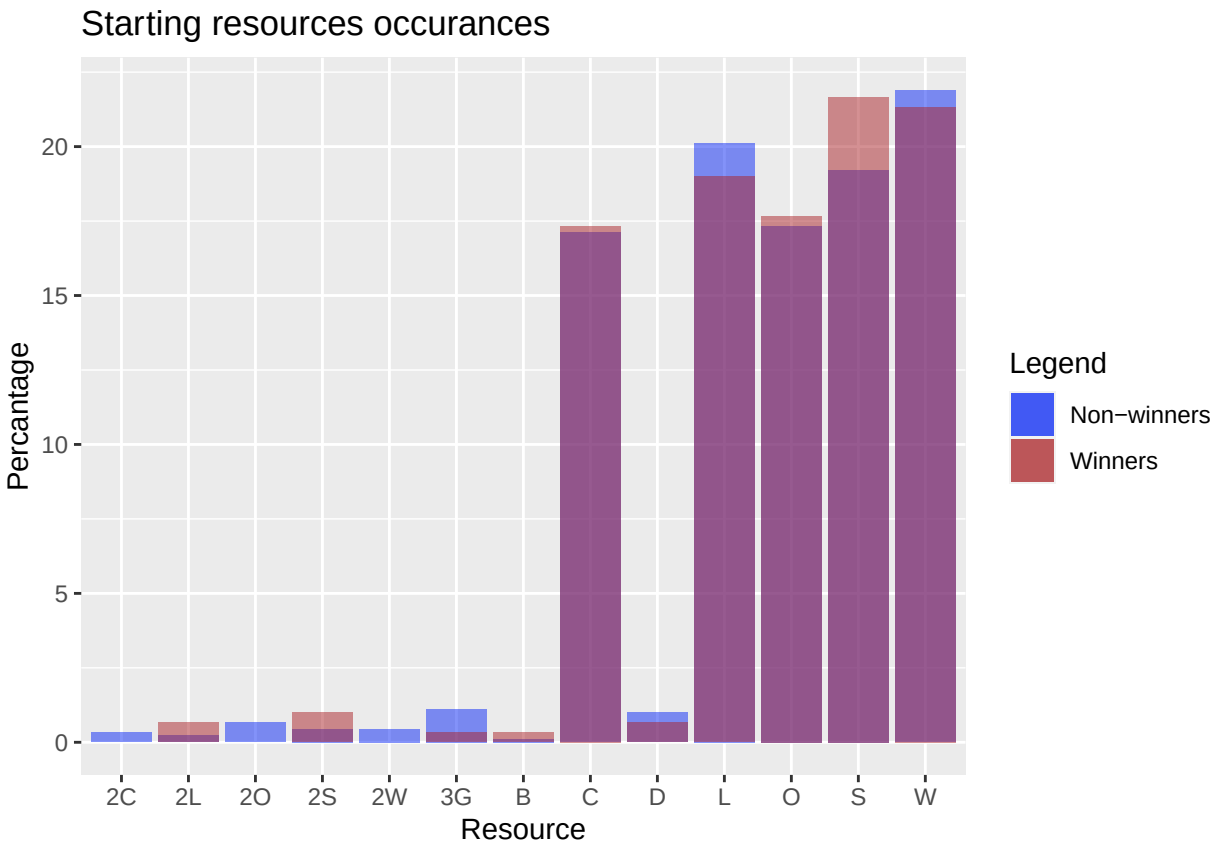
D = desert (hrv. pustinja)

B = blank (moguće je postaviti settlement da graniči s 2 polja, umjesto 3)

*Translator note: nazivi nisu doslovno prevedeni te vjerojatno svi imaju načine kako zovu polja u svojim igrama. Korišten je prijevod koji autori koriste kad igraju ovu igru.

```
source("../R/AnalyseStartingResources.R")
ggplot() +
  geom_col(aes(x = names(starting_resources_table),
               y = starting_resources_table,
               fill = "Non-winners"),
           alpha = 0.5, position = "dodge") +
  geom_col(aes(x = names(starting_resources_winners_table),
               y = starting_resources_winners_table,
               fill = "Winners"),
           alpha = 0.5, position = "dodge") +
  scale_fill_manual(name = "Legend",
                    breaks = c("Non-winners", "Winners"),
                    values = c("Non-winners" = "#0827F5", "Winners" = "#AB2328")) +
  labs(title = "Starting resources occurances", x = "Resource", y = "Percantage")
```

```
## Don't know how to automatically pick scale for object of type <table>.
## Defaulting to continuous.
```



Iz grafa je evidentno da svi igrači imaju prilično sličnu viziju kako igrati Catan. Pobjednici cijene ovce(S na grafu) kao resurs više od ostalih igrača, no osim toga čini se da igrači stavljaju vrlo slične vrijednosti na resurse.

Zaključak

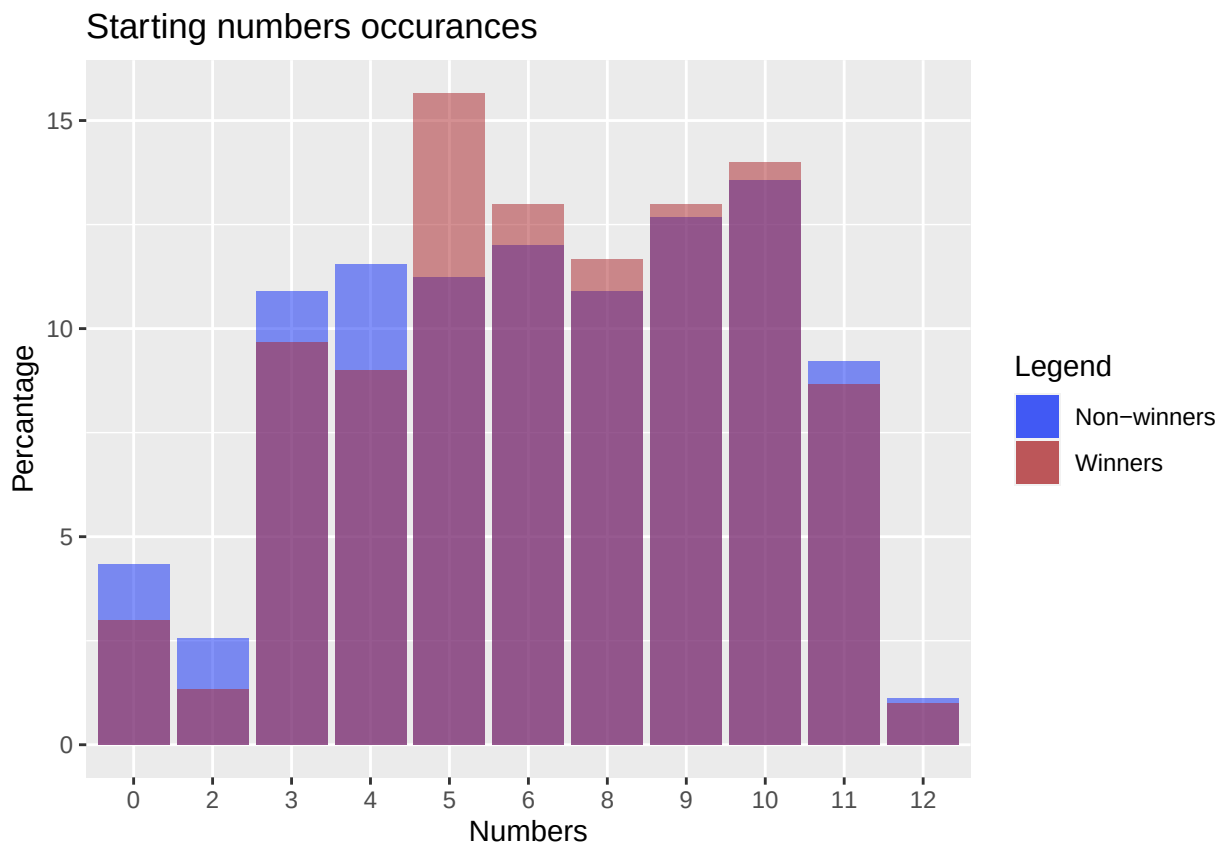
Čini se da pšenica i ovcu imaju najveću vrijednost pri postavljanju početna 2 naselja, nakon čega slijedi drvo te veliku vrijednost, no manju od prethodna 3 resursa, imaju i kamen i cigla. Luke imaju izuzetno malu vrijednost, te pobjednici jedino stavljaju bitno veći prioritet na ovce od ostalih igrača (čak nevjerojatnih 2.5% više pobjednika u prva dva naselja osigura pristup ovcama u odnosu na ostale igrače), tako da je možda pametno osigurati pristup ovcama na početku igre.

2.2 Utjecaj početnih brojeva na pobjedu

Nakon razočaravajućeg rezultata da početni resursi ne čine bitno razliku u određivanju pobjednika okrećemo se brojevima u našem nastojanju da maksimiziramo svoje šanse pobjede već na početku igre.

```
source("../R/AnalyseStartingNumbers.R")
ggplot() +
  geom_col(aes(x = fct_reorder(names(starting_numbers_table),
                               as.numeric(names(starting_numbers_table))),
               y = starting_numbers_table,
               fill = "Non-winners"),
           alpha = 0.5, position = "dodge") +
  geom_col(aes(x = fct_reorder(names(starting_numbers_winners_table),
                               as.numeric(names(starting_numbers_winners_table))),
               y = starting_numbers_winners_table,
               fill = "Winners"),
           alpha = 0.5, position = "dodge") +
  scale_fill_manual(name = "Legend",
                    breaks = c("Non-winners", "Winners"),
                    values = c("Non-winners" = "#0827F5", "Winners" = "#AB2328")) +
labs(title = "Starting numbers occurances", x = "Numbers", y = "Percantage")
```

```
## Don't know how to automatically pick scale for object of type <table>.
## Defaulting to continuous.
```



Za razliku od prethodne analize vidimo da postoji bitnija razlika u biranju brojeva na početnim poljima između pobjednika i ostalih igrača. Pobjednici češće biraju brojeve u intervalu [5, 10], dok ostali igrači češće biraju preostale brojeve.

Zaključak

Čini se da je puno bitnije odabrati polja koja imaju "dobre" brojeve (brojeve koji imaju veću vjerojatnost da padnu) na početku igre, nego fokusirati se na resurse, tako da sljedeći put kad igrate dobro razmislite kako postaviti svoja naselja tako da maksimizirate vjerojatnost da dobijete nešto u svakom bacanju. Ovakav rezultat ide u prilog rezultatu da je kocka poštena, iako mislim da se svi iskustveno možemo složiti da je kocka sve osim poštena.

3. Tradeanje (naslov WIP)

4. Redoslijed postavljanja naselja (naslov WIP)

5. Predikcija

U svrhu pobjeđivanje pokušati ćemo zaključiti imaju li dobitci i gubljenje resursa značajnu vezu s pobjeđivanje. Iz tog razloga stvaramo linearan model koji će predviđati broj bodova na temelju svih mogućih dobitaka i gubitaka resursa.

```
source("../R/Model.R")
summary(linMod)
```

```
##
## Call:
## lm(formula = points ~ production + tradeGain + robberCardsGain +
##     tradeLoss + robberCardsLoss + tribute, data = catan_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.1578 -0.9776 -0.0627  1.0611  3.1781
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1.24580    0.37428   3.329 0.001046 **
## production       0.14984    0.01181  12.689  < 2e-16 ***
## tradeGain        0.18339    0.04625   3.965 0.000103 ***
## robberCardsGain  0.12626    0.02271   5.560 8.91e-08 ***
## tradeLoss       -0.14729    0.02533  -5.816 2.47e-08 ***
## robberCardsLoss -0.14936    0.03347  -4.462 1.38e-05 ***
## tribute         -0.14688    0.02159  -6.804 1.25e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.376 on 193 degrees of freedom
## Multiple R-squared:  0.6046, Adjusted R-squared:  0.5923
## F-statistic: 49.18 on 6 and 193 DF,  p-value: < 2.2e-16
```

Zaključak

Rezultati koje nam daje linearan model su vrlo loši, predviđa bodove s vrijednosti +-3, s tim da 50% predikcija ima grešku +-1. S obzirom da igrač pobjeđuje u Catanu kada skupi 10 ili više bodova, a sve u igri pridonosi 1 ili 2 boda, ovakva greška je vrlo velika i zaključujemo da linearan model nije dobar za predviđanje bodova u igri Catan. To je konzistentno s rezultatima na kaggleu na ovom datasetu gdje linearni modeli daju loše rezultate, no ljudi su imali više uspjeha s logističkom regresijom gdje binarno klasificiraju rezultat kao pobjedu ili gubitak te su neki postigli preciznost ~80%.

Zaključak

TODO lorem ipsum