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, i ako da kako, tradeanje (trgovanje) na ishod igre?

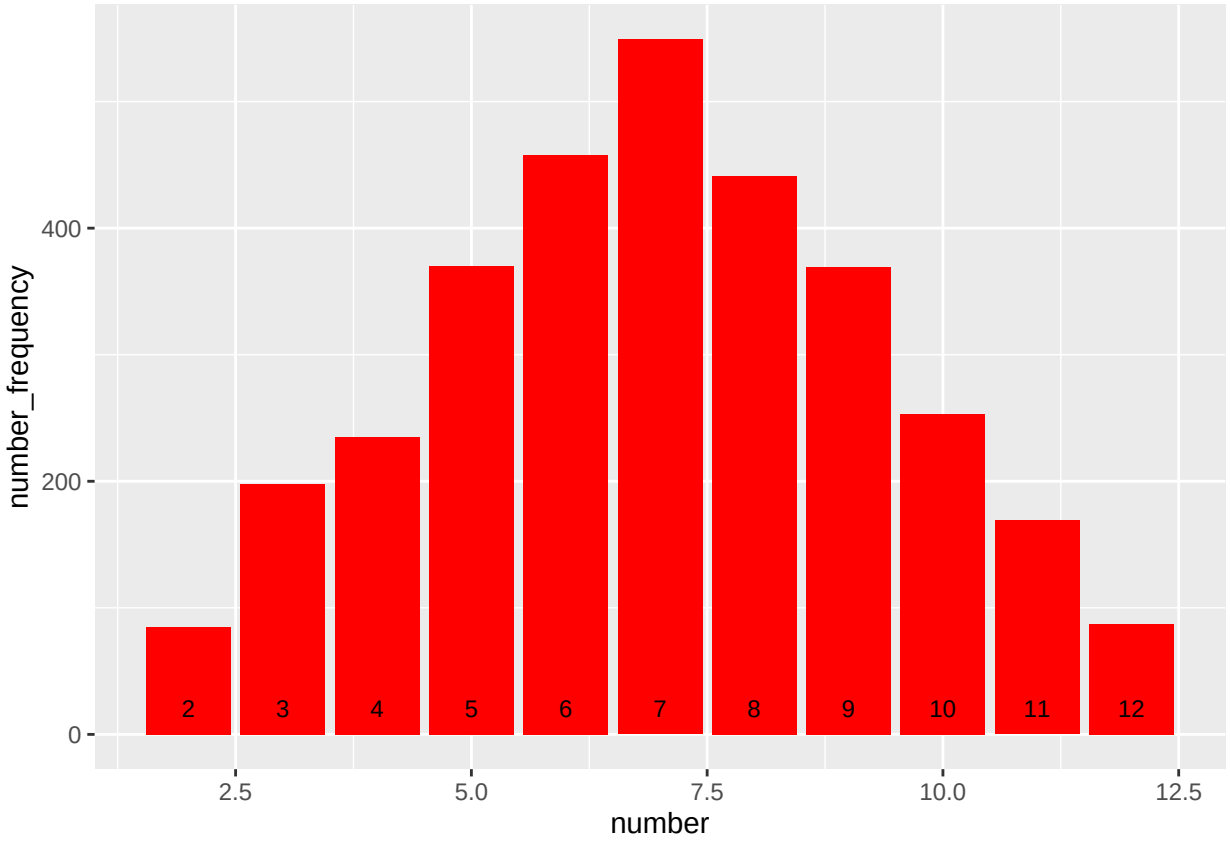
Utječe li redoslijed postavljanja naselja na pobjednika?

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

1. Je li kocka poštena?

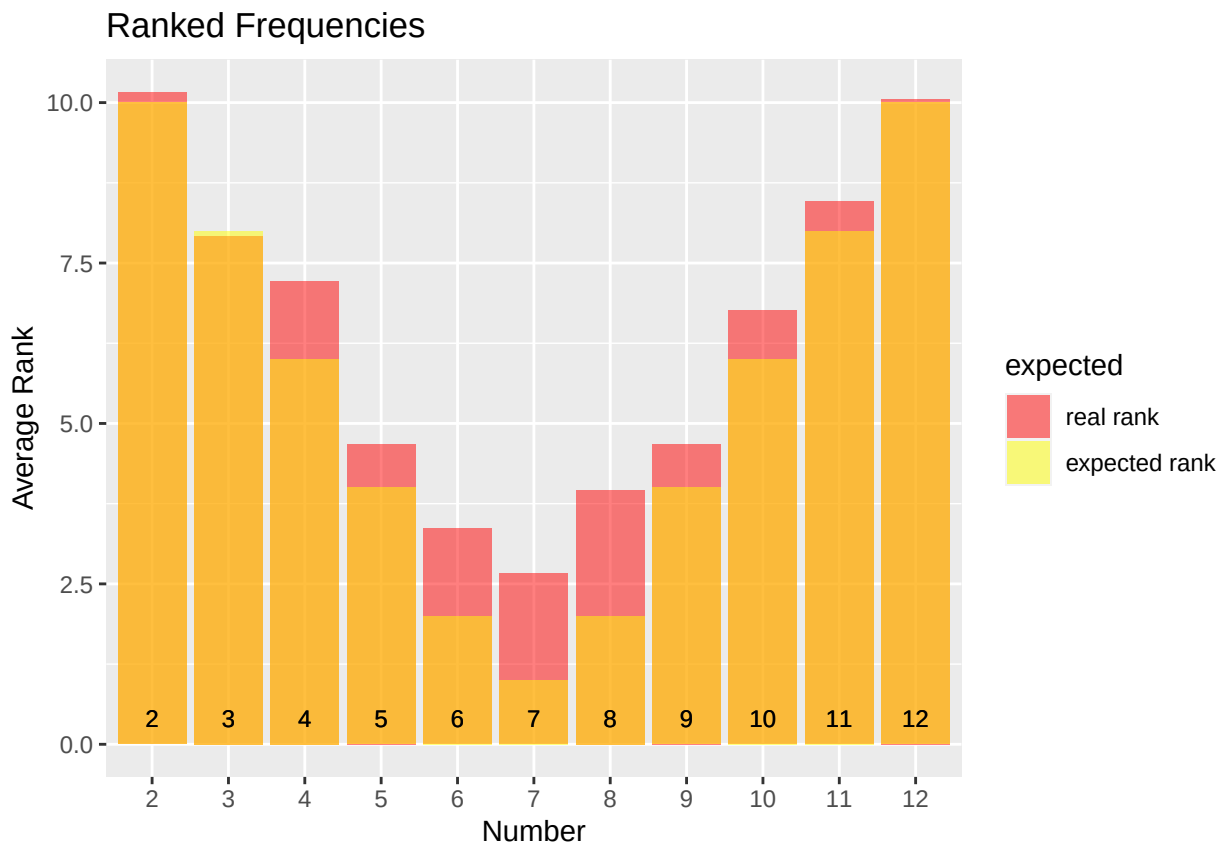
Svima se sigurno dogodila partija u kojoj su naselja lukavo postavili na poziciju susjednu poljima s brojevima koji bi teoretski trebali češće padati, međutim magično su se brojevi 3 i 12 pojavljivali češće nego 8 i 9. Budući da se i njima samima dogodila takva partija, autori rada odlučili su prvo istražiti 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

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

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)

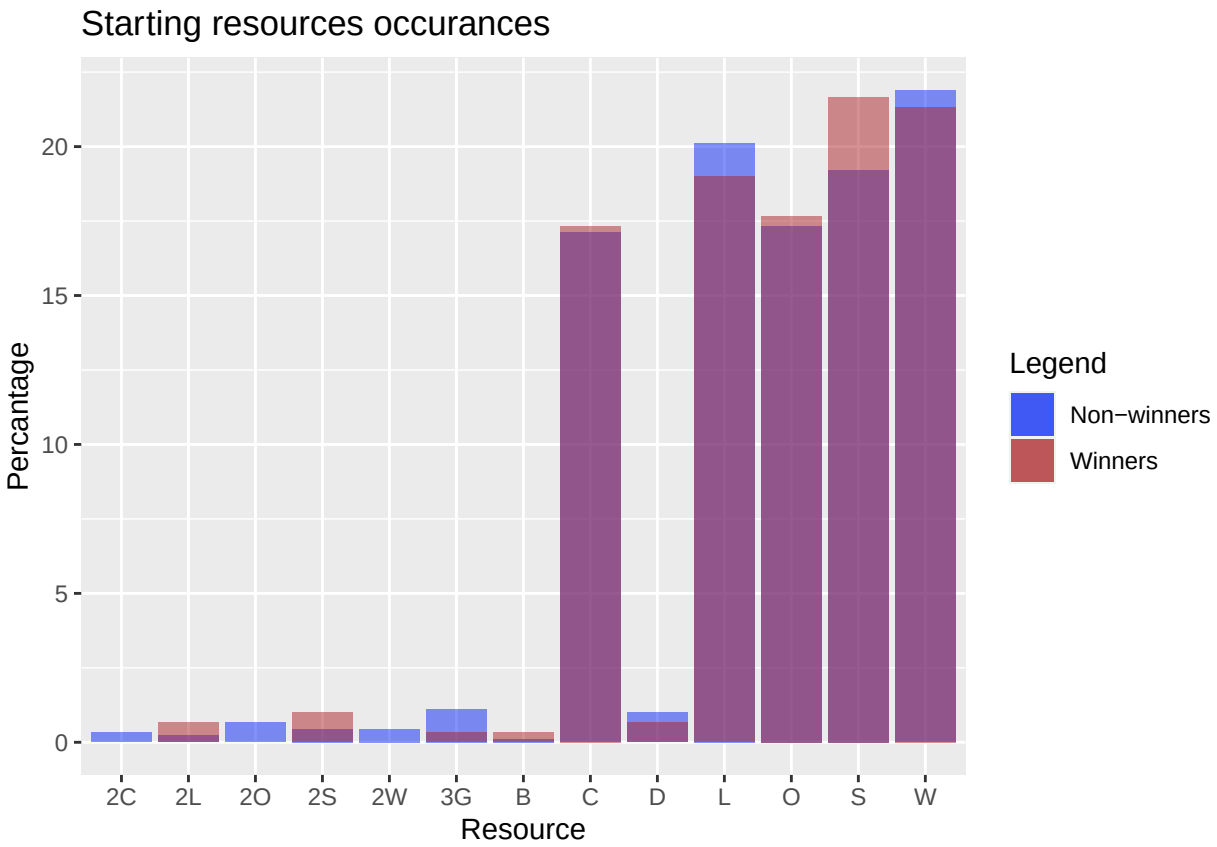
D = desert (hrv. pustinja)

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

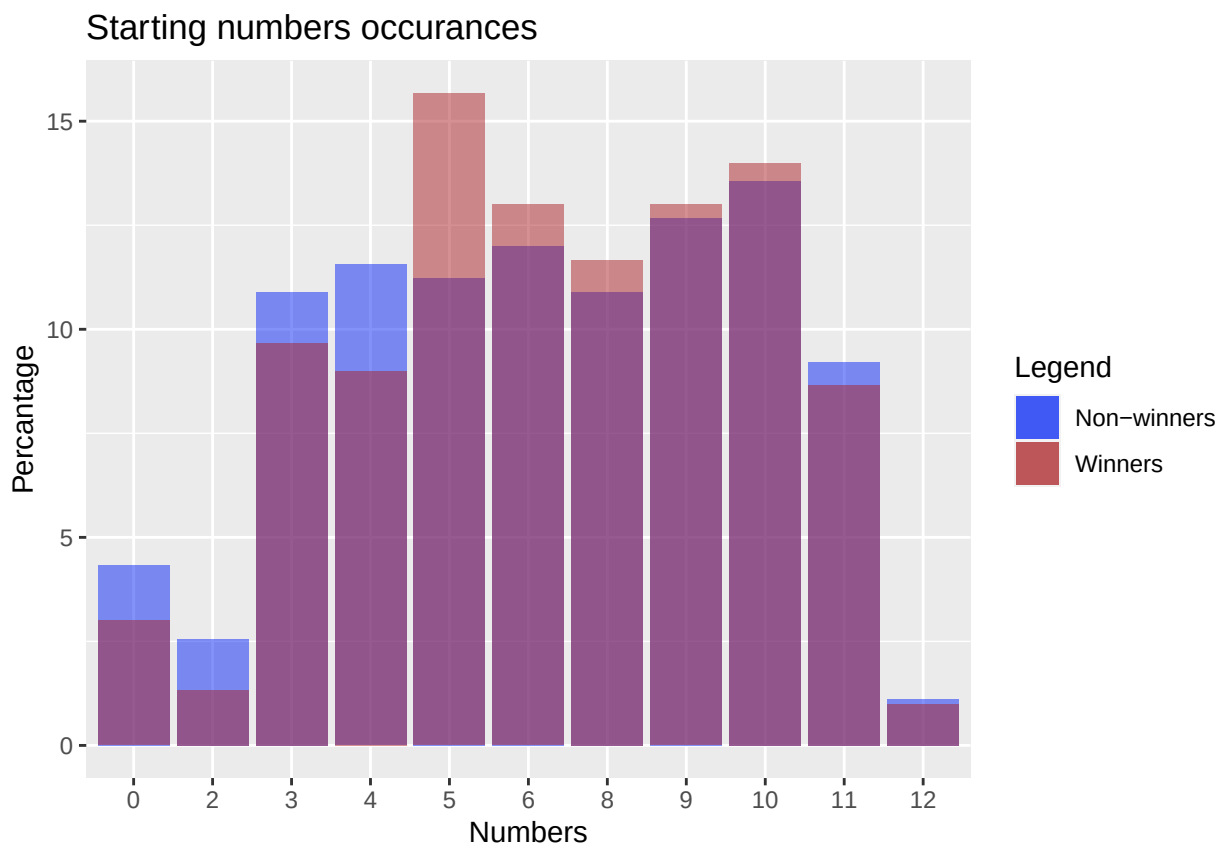
Čini se da pšenica i ovcu imaju najveću vrijednost pri postavljanju početna 2 naselja, nakon čega slijedi drvo te veliku vrijednost, no manju od prethodna 3 resursa, imaju i kamen i cigla. Luke imaju izuzetno malu vrijednost, te pobjednici jedino stavljaju bitno veći prioritet na ovce od ostalih igrača (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

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

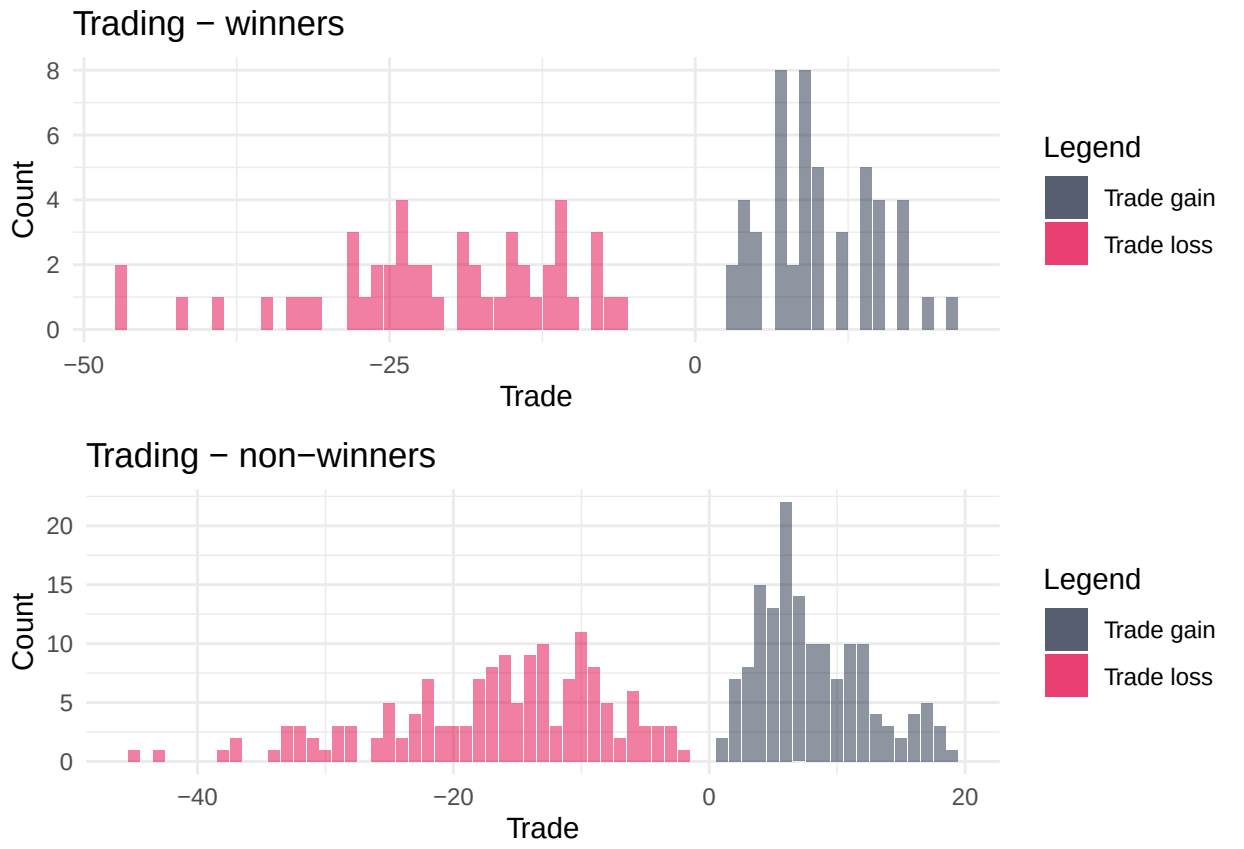
Čini se da je puno bitnije odabrati polja koja imaju "dobre" brojeve (brojeve koji imaju veću vjerojatnost da padnu) na početku igre, nego fokusirati se na resurse, tako da sljedeći put kad igrate dobro razmislite kako postaviti svoja naselja tako da maksimizirate vjerojatnost da dobijete nešto u svakom 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. Utjecaj trgovanja

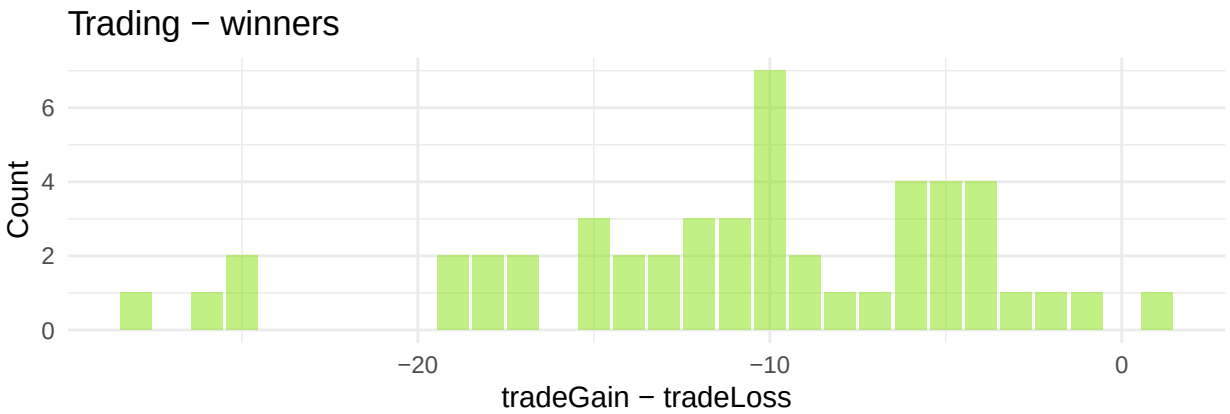
U igri Catan moguće je trgovati s bankom, pri čemu igrač banci daje 4 karte istog resursa za 1 kartu nekog resursa, ili razmjenjivati u boljim uvjetima ako ima odgovarajuću luku. Osim toga, moguće je direktno trgovati s drugim igračima, pri čemu igrači sami dogovaraju koje će karte razmjeniti. U ovom dijelu zanima nas utječe li kako broj dobivenih i izgubljenih karti trgovanjem na ishod igre.

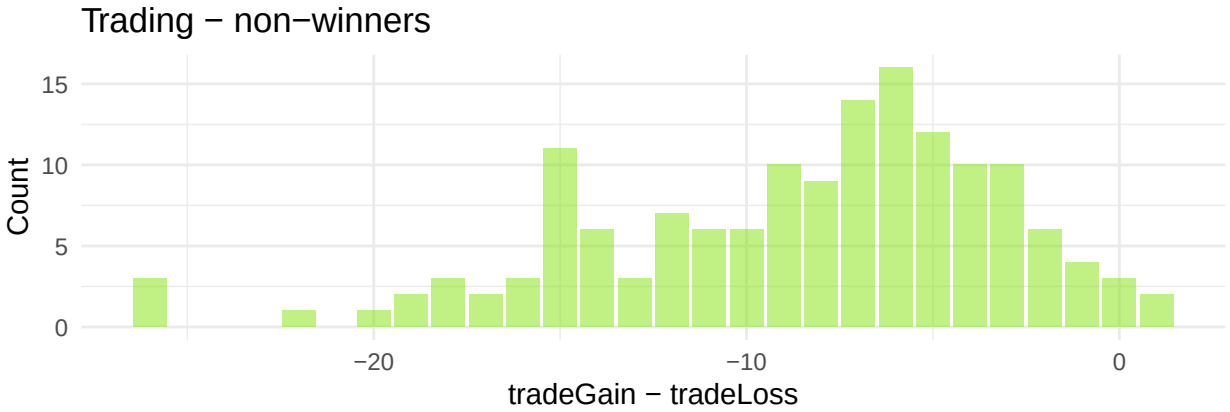
```
g1 <- ggplot(game_winner_data, aes(x = tradeGain)) +
  geom_bar(aes(fill = "Trade gain"),
           stat = "count",
           position = "stack",
           alpha = 0.5) +
  geom_bar(aes(x = -tradeLoss, fill = "Trade loss"),
           stat = "count",
           position = "stack",
           alpha = 0.5) +
  scale_fill_manual(name = "Legend",
                    breaks = c("Trade gain", "Trade loss"),
                    values = c("Trade gain" = "#202A44", "Trade loss" = "#E40046")) +
  labs(title = "Trading - winners", x = "Trade", y = "Count") +
  theme_minimal()

g2 <- ggplot(game_winner_data, aes(x = tradeGain - tradeLoss)) +
  geom_bar(fill = "#84E30B", stat = "count", alpha = 0.5) +
  labs(title = "Trading - winners", y = "Count") +
  theme_minimal()

g3 <- ggplot(non_game_winner_data, aes(x = tradeGain)) +
  geom_bar(aes(fill = "Trade gain"),
           stat = "count",
           position = "stack",
           alpha = 0.5) +
  geom_bar(aes(x = -tradeLoss, fill = "Trade loss"),
           stat = "count",
           position = "stack",
           alpha = 0.5) +
  scale_fill_manual(name = "Legend",
                    breaks = c("Trade gain", "Trade loss"),
                    values = c("Trade gain" = "#202A44", "Trade loss" = "#E40046")) +
  labs(title = "Trading - non-winners", x = "Trade", y = "Count") +
  theme_minimal()

g4 <- ggplot(non_game_winner_data, aes(x = tradeGain - tradeLoss)) +
  geom_bar(fill = "#84E30B", stat = "count", alpha = 0.5) +
  labs(title = "Trading - non-winners", y = "Count") +
  theme_minimal()
```

```
grid.arrange(g1, g3)
```




```
grid.arrange(g2, g4)
```





Iz grafova čini se kako pobjednici trguju više nego ostali igrači. Pogledati ćemo još srednju vrijednost i medijan tradeGaina i tradeLossa kako bi potvrdili ili odbacili ovu hipotezu.

```
game_winner_data$tradeGain %>%
  mean() %>%
  sprintf("Winners mean trade gain: %.2f", .) %>%
  cat("\n")
```

```
## Winners mean trade gain: 10.10
```

```
game_winner_data$tradeLoss %>%
  mean() %>%
  sprintf("Winners mean trade loss: %.2f", .) %>%
  cat("\n")
```

```
## Winners mean trade loss: 21.06
```

```
non_game_winner_data$tradeGain %>%
  mean() %>%
  sprintf("Non-winners mean trade gain: %.2f", .) %>%
  cat("\n")
```

```
## Non-winners mean trade gain: 8.07
```

```
non_game_winner_data$tradeLoss %>%
  mean() %>%
  sprintf("Non-winners mean trade loss: %.2f", .) %>%
  cat("\n")
```

```
## Non-winners mean trade loss: 16.69
```

```r
game_winner_data$tradeGain %>%
  median() %>%
  sprintf("Winners median trade gain: %.2f", .) %>%
  cat("\n")
```

```
## Winners median trade gain: 9.00
```

```r
game_winner_data$tradeLoss %>%
  median() %>%
  sprintf("Winners median trade loss: %.2f", .) %>%
  cat("\n")
```

```
## Winners median trade loss: 20.00
```

```r
non_game_winner_data$tradeGain %>%
  median() %>%
  sprintf("Non-winners median trade gain: %.2f", .) %>%
  cat("\n")
```

```
## Non-winners median trade gain: 7.00
```

```r
non_game_winner_data$tradeLoss %>%
  median() %>%
  sprintf("Non-winners median trade loss: %.2f", .) %>%
  cat("\n")
```

```
## Non-winners median trade loss: 15.00
```

Iz medijana i srednje vrijednosti vidljivo je da pobjednici češće trguju od ostalih igrača, no vidimo i da je prosječan omjer dobivenih i izgubljenih karti otprilike $\frac{1}{2}$.

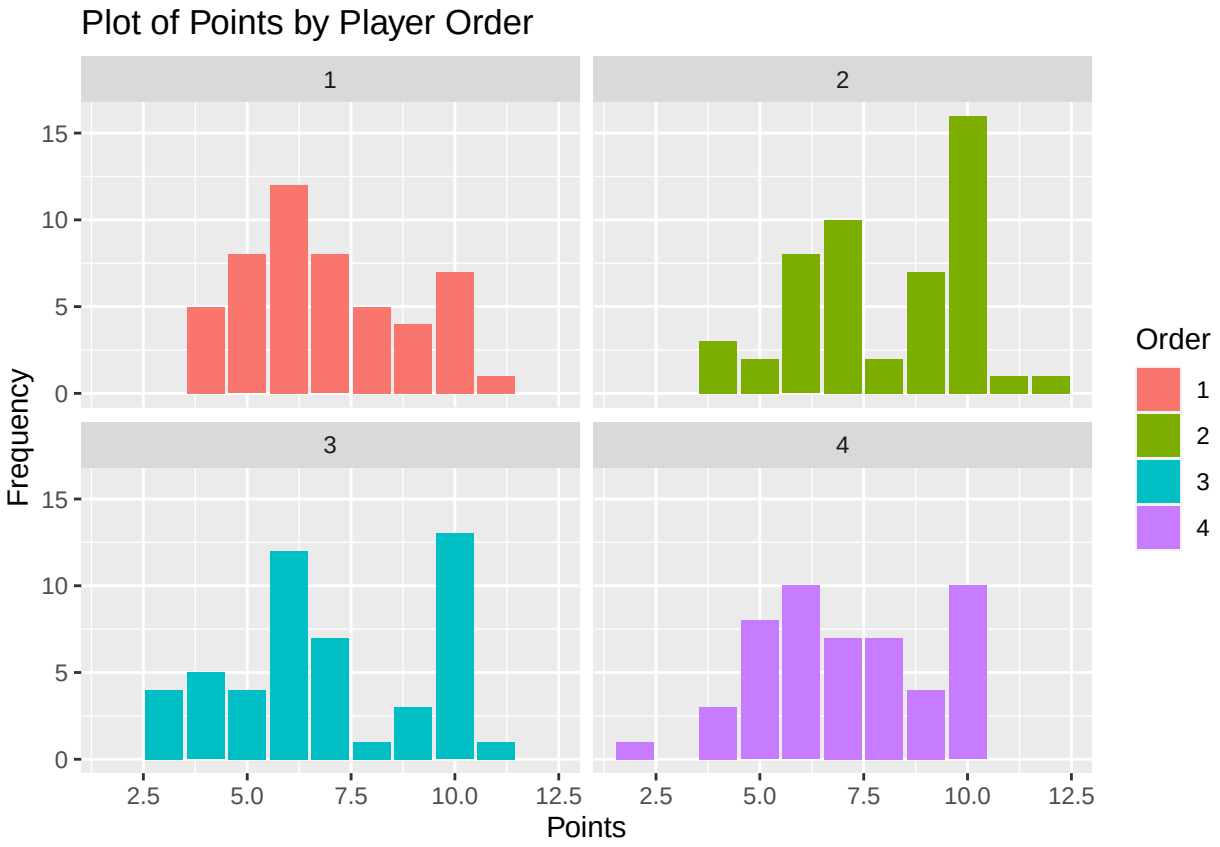
ZAKLJUČAK

Pobjednici u prosjeku dobiju 2 više karte iz trgovanja i izgube 5 više karata, iz čega zaključujemo da pobjednici više trguju od ostalih igrača.

4. Utjecaj Redoslijed postavljanja naselja

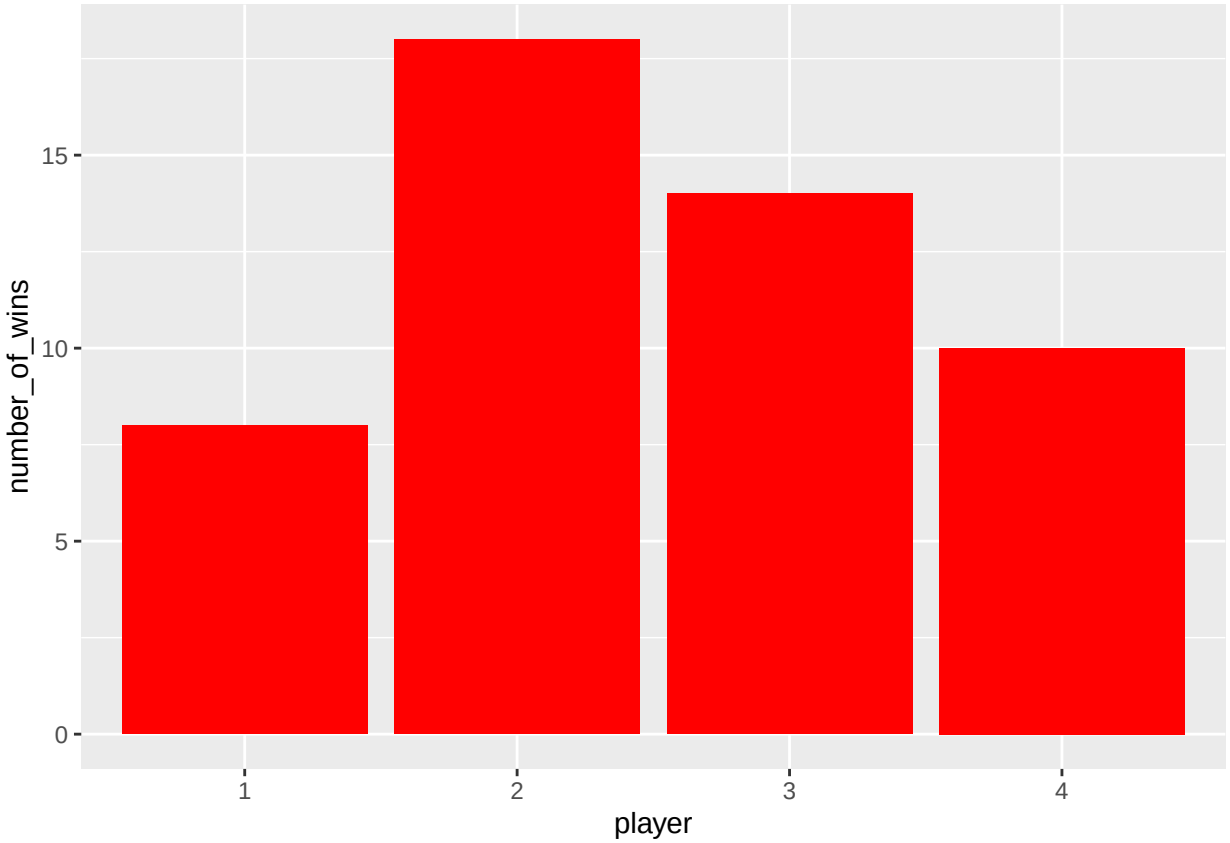
Za većinu društvenih igara, osobe koje ih igraju malo ozbiljnije smatraju da je bolje biti osoba koja započinje igru, odnosno biti prvi na potezu. Budući da je cilj rada pokušati naći informaciju koja može dovesti do povećanja vjerojatnosti pobjede, ispitati ćemo istinitost ove tvrdnje. Konkretnije, tražimo postoji li korelacija između redoslijeda igranja i osvojenog broja bodova te broja pobjeda.

```
catan_data$player <- factor(catan_data$player)

order_points_plot <- ggplot(catan_data, aes(x = points, fill = player)) +
  geom_bar() +
  facet_wrap(~ player) +
  labs(title = "Plot of Points by Player Order", x = "Points", y = "Frequency", fill = "Order")

order_points_plot %>% print()
```



Na temelju analize grafova, možemo zaključiti da redoslijed igranja ne igra veliku ulogu u osvajanju broja bodova, iako je primjetno da nešto više puta pobjeđuje igrač koji je 2. na potezu te da manje bodova osvaja upravo igrač koji započinje igru.. Sada idemo analizirati ovisnost broja pobjeda i redoslijeda igranja.

```
catan_data$winner <- ifelse(catan_data$points >= 10, TRUE, FALSE)
wins_by_order <- catan_data %>% group_by(player) %>% summarize(number_of_wins = sum(winner))

wins_by_order_plot <- ggplot(wins_by_order, aes(x = player, y = number_of_wins)) +
  geom_bar(fill = "red", stat = "identity")
wins_by_order_plot %>% print()
```



Iako rezultate treba interpretirati s oprezom, budući da mi pretpostavljamo da su sva 4 igrača jednako vješta u igri, primjetna je ogromna razlika u broju pobjeda, pogotovo u odnosu na ispitanu tvrdnju. Vidljivo je da najviše pobjeđuju igrači koji su 2. ili 3. na potezu, dok 1. i 4. najrjeđe pobjeđuju. Dapače, vjerojatnost pobjede za 2. po redu igrača raste sa očekivanih 25% na 36%.

ZAKLJUČAK

Suprotno pretpostavci da osoba koja je prva na potezu pobjeđuje češće od ostalih, čini se da ustvari najčešće pobjeđuje osoba koja je po redu 2. na potezu. Naravno, ovaj zaključak dolazi uz pretpostavku da su sva 4 igrača jednako dobri.

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

U danom podatkovnom skupu nalaze se podaci o 50 partija igre Catan, tako da ne možemo tvrditi da je skup sasvim reprezentativan te da su rezultati mjerodavni, no rezultati su prilično konzistentni s rezultatima koji se mogu naći na internetu. Na početku ove analize postavili smo 4 pitanja na koja smo tražili odgovor te smo analizom dobili i odgovor na njih.

Je li kocka zaista poštena? Odgovor na zaprepaštenje svakog igrača bilo kakve društvene igre je da (što ću sad kriviti kada ne dobijem ako ne mogu kocku...).

Postoji li neka kombinacija resurs(a) na početnim susjednim poljima koja povećava vjerojatnost pobjede? Ne možemo tvrditi da početni resursi imaju presudnu ulogu u igri, jer vidimo da i pobjednici i ostali igrači imaju slična viđenja o bitnosti pojedinih resursa, tako da je odgovor iz ove analize ne.

Postoji li neka kombinacija broj(eva) na početnim susjednim poljima koja povećava vjerojatnost pobjede? Za razliku od prošlog pitanja, u analizi smo otkrili da je najbolja strategija za povećavanje vjerojatnosti pobjede u igri postavljati naselja tako da se graniče s poljima koja imaju "najbolje" brojeve (brojeve s najvećom vjerojatnosti pojavljivanja). Stoga je odgovor na ovo pitanje u ovoj analizi da.

Utječe li, i ako da kako, tradeanje (trgovanje) na ishod igre? U analizi smo otkrili da pobjednici trguju više od preostalih igrača, tako da je zaključak ove analize da trgovanje utječe na vjerojatnost pobjede i da je bolje češće trgovati.

Utječe li redoslijed postavljanja naselja na pobjednika? Ovo je pitanje koje je možda dalo najodlučniji odgovor. Iz analize vidimo da najveću vjerojatnost pobjede ima 2. igrač, nakon njega 3. igrač, pa 4. igrač te se na posljednjem mjestu nalazi 1. igrač.

Pokušali smo i napraviti linearan model za predviđanje broja bodova na temelju podataka iz igre te nismo dobili najbolje rezultate te je možda prikladnije koristiti neki drugi model (npr. logističku regresiju ili kNN).

Iz ovog skupa podataka otkrili smo da postoje određeni faktori koji utječu na pobjedu, no za dublju analizu, vjerojatno je potreban detaljniji skup podataka koji bilježi podatke o svakom potezu. Slijedeći koraci u analizi ove igre mogli bi biti: pronalazak još detaljnijeg skupa podataka, korištenje prediktivnih modela te pokušati kombinirati vrijednost broja na polju i vrijednost resursa na polju.

To je sve od nas, hvala Vam na čitanju i nadamo se da ste se bar malo zabavili.