Julia Hruświcka, Kostiantyn Skopych

Sprawozdanie
Analiza danych

24 maja 2022

Spis treści

0.1. Wstęp

Nasze dane pochodzą ze strony: https://figshare.com. Są to dane opisujące dużą ilość otwartych boosterów (pakietów) z kolekcyjnej gry karcianej Magic: The Gathering. Booster to zestaw z 15 kart. Karty są losowe i nie powtarzają się wewnątrz boostera. Każdy booster zawiera 1 rzadką lub mitycznie rzadką kartę (rare oraz mythic rare), 3 niezwykłych karty (uncommon) i resztę zwykłych kart (common).

0.2. Jakie zmienne występują w naszych danych oraz jakie wartości one mogą przyjmować?

— **observer** Imię i nazwisko osoby, która otwierała boostery i wprowadzała dane.
— **card-set** Nazwa setu (zestawu) (możliwe wartości: Ikoria, Core 2021, Zendikar Rising)
— **rep** Numer pudełka (wartości: 1, 2, 3 itd.)
— **pack-rep** Numer boostera w pudełku (wartości wahają się od 1-36 dla każdego pudełka)
— **card-name** Nazwa karty
— **rarity** Rzadkość karty. Możliwe wartości: common, uncommon, rare, mythic
— **card-type** Typ karty: znajduje się na środku karty, może być kilka typów na raz (przykładowe wartości: Creature, Enchantment, Land, Artifact, Sorcery, Instant)
— **creature-type** Typ stworzenia. Dotyczy tylko kart typu Creature (Przykładowe wartości: Cleric, hydra itp.)
— **power** Moc stworzenia. Dotyczy tylko kart typu Creature. Wartość liczbowa znaleziona w prawym dolnym rogu karty stworzenia (0, 1, 2, 3, 4 itd.)

— **toughness** Wytrzymałość stworzenia. Dotyczy tylko kart typu Creature. Druga wartość liczbowa znaleziona w prawym dolnym rogu karty stworzenia (0, 1, 2, 3, 4 itd.)
— **n-colors** Liczba kolorów karty. (0, 1, 2, 3 itd.)
— **color** Kolor karty, w razie potrzeby może mieć kilka (Możliwe wartości: czerwony, zielony, czarny, biały, niebieski i bezbarwny)
— **CMC (Converted mana cost)** Ogólna ilość many (magiczna waluta stosowana w grze) potrzebna dla zastosowania karty w trakcie gry. Przyjmuje wartości liczbowe (3, 5, 8, 2, 0 itd.)
— **finish** Wizualna cecha karty. Wskazuje czy karta ma jakąś wyróżnioną dekorację. Możliwe wartości: Regular (dla karty nie mającej żadnych dodatkowych dekoracji), foil (dla karty mającej pokrycie foliowe),
— **showcase** (dla karty mającej alternatywny (ładniejszy) obrazek, inny niż wykły)
— **n-keywords** ilość słów kluczowych, czyli zdolności karty opisanych 1-2 słowami. Wartości: 0 jeśli nie ma, jeśli jest to 1, 2, 3, 4 itd.
— **keywords** Wyżej wspomniane słowa kluczowe. Zwykle tylko na stworzeniach. Przykładowe wartości: Lifelink, deathtouch, trample itp.)
— **n** zmienna, której sens nie jest do końca zrozumiały i która wśród naszych obserwacji zawsze ma wartość 1
— **price** Cena karty. Autor danych pobierał ceny ze strony FacetoFace. Zmienna przyjmuje wartości liczbowe, oznaczające cenę karty w USD.

0.3. Cele analizy

Skupimy się na wyszukiwaniu zależności pomiędzy zmiennymi. Interesujące nas pytania badawcze to np: Czy zależy cena karty od jej koloru, typu, CMC, n-keywords? Czy zależy cena stworzeń od ich mocy/wytrzymałości? Jaki jest rozkład łącznych cen boosterów?

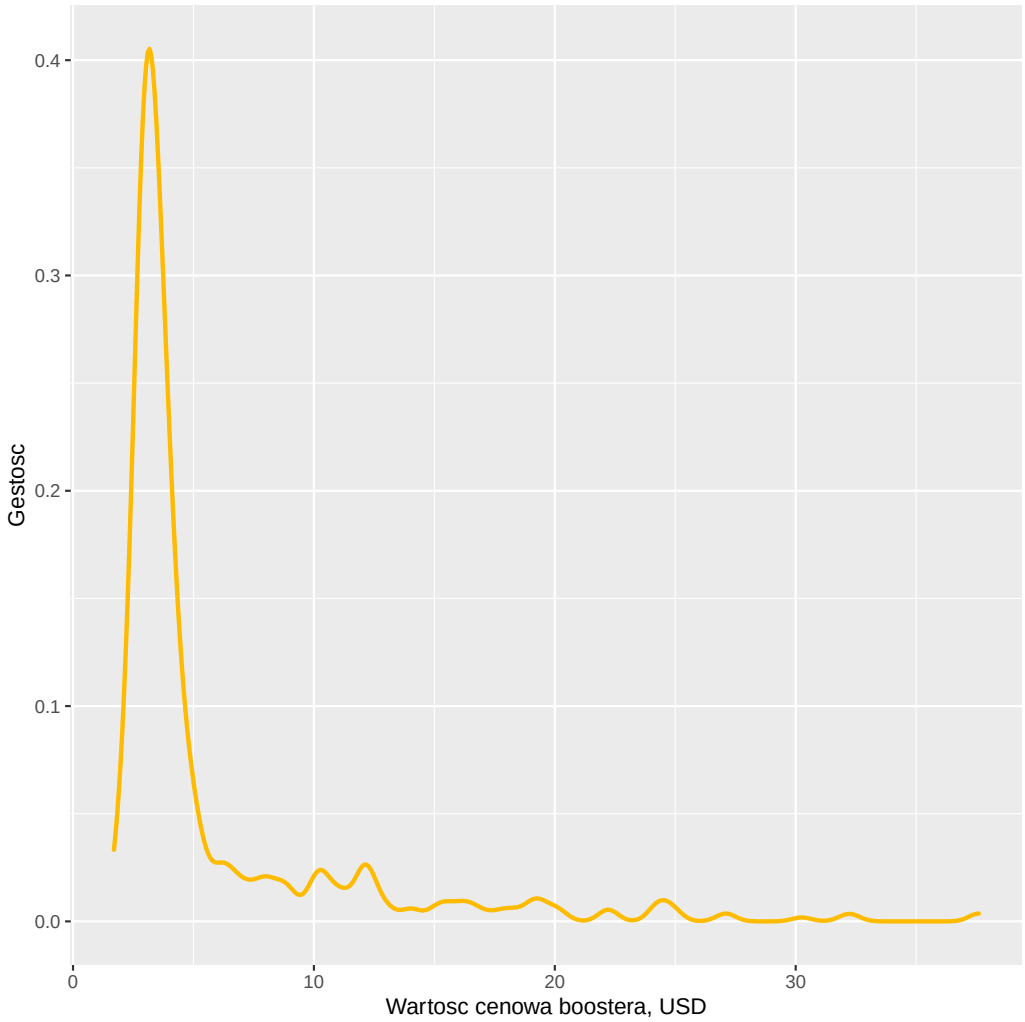
0.4. Analiza

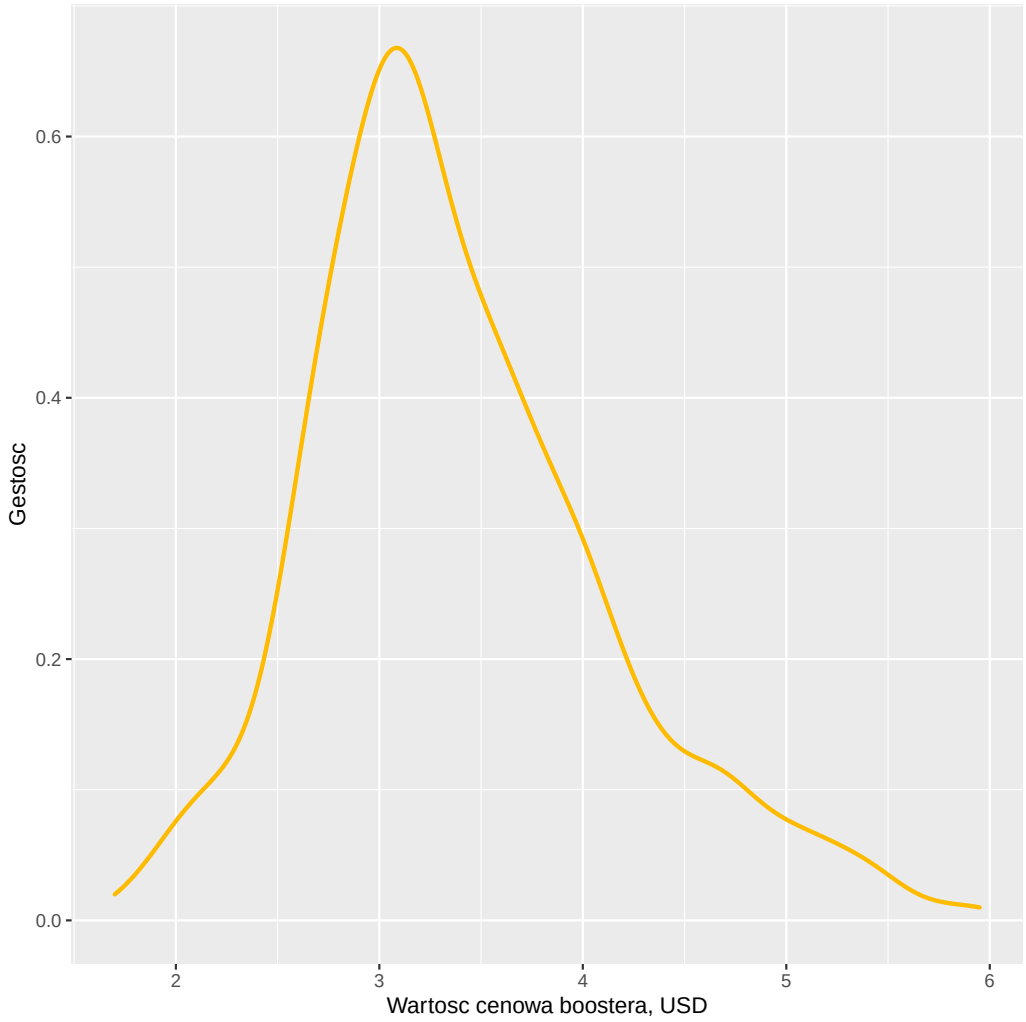
Widzimy, że nasze dane zawierają dużo zmiennych. Nie wszystkie te zmienne będą potrzebne nam do analizy. Dlatego pozbędziemy się niektórych kolumn z tych danych. Np usuniemy kolumnę 'observer', ponieważ ma wszędzie tą samą wartość nie będzie odgrywała w naszej analizie żadnej roli. Usuniemy też kolumnę 'creature type', bo jest ona zbędnym szczegółem w naszej analizie i nie będziemy się skupiali na niej. Ewentualnie można by było zadać kilka ciekawych pytań badawczych odnośnie tej zmiennej, gdyby nasze dane były bardziej różnorodne i mieściły więcej niepowtarzających się kart. Usuniemy kolumnę 'n', ponieważ nie ma ona żadnego praktycznego znaczenia. Usuniemy kolumnę 'keywords', ponieważ będzie nas interesowała tylko liczba słów kluczowych, a nie same słowa. Chcemy też usunąć wszystkie obserwacje, dla których zmienna finish nie ma wartości 'Regular', a później usunąć tą kolumnę.
Zrobimy tak z tego powodu, że karty nie 'Regular' mają takie same charakterystyki jak analogiczne karty 'Regular', jednak drastycznie różnią się w cenie

od nich z powodu swojej dekoracji. To będzie sprawiało nam zbędne obserwacje odstające, których możemy się pozbyć jeszcze do początku analizy.

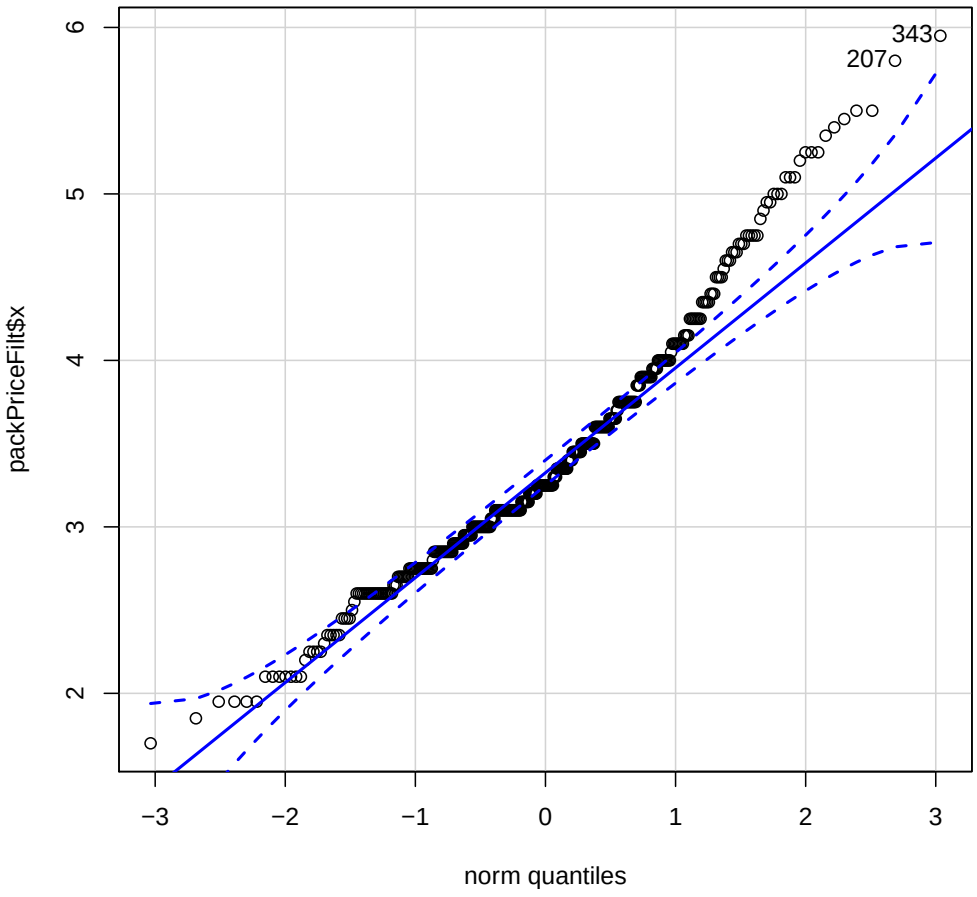
```
##  observer                          set            rep           pack_rep
##  mz:7713   Core 2021                 :2596   Min.   :1.000   Min.   : 1.00
##            Ikoria: Lair of Behemoths:2600   1st Qu.:2.000   1st Qu.:10.00
##            Zendikar Rising           :2517   Median :3.000   Median :19.00
##                                              Mean   :2.995   Mean   :18.51
##                                              3rd Qu.:4.000   3rd Qu.:28.00
##                                              Max.   :5.000   Max.   :36.00
##
##               card_name         rarity                  card_type
##  Capture Sphere    :  36   Common  :5474   Creature           :4020
##  Gloom Sower       :  23   Mythic  :  71   Instant            :1581
##  Infernal Scarring :  23   Rare    : 509   Sorcery            : 897
##  Racking Claws     :  23   Uncommon:1659   Land               : 308
##  Cleansing Wildfire:  22                   Artifact           : 260
##  Deathbloom Thallid:  22                   Enchantment - Aura: 259
##  (Other)           :7564                   (Other)            : 388
##       creature_type      power          toughness         n_colors
##               :3578   Min.   : 0.000   Min.   : 0.000   Min.   :0.0000
##  Human Soldier: 198   1st Qu.: 1.000   1st Qu.: 2.000   1st Qu.:1.0000
##  Beast        : 192   Median : 2.000   Median : 3.000   Median :1.0000
##  Cat          : 159   Mean   : 2.438   Mean   : 2.769   Mean   :0.9653
##  Human Wizard : 154   3rd Qu.: 3.000   3rd Qu.: 3.000   3rd Qu.:1.0000
##  Human Warrior: 145   Max.   :11.000   Max.   :17.000   Max.   :5.0000
##  (Other)      :3287   NA's   :3635     NA's   :3635
##       color           CMC                    finish       n_keywords
##  Red      :1338   Min.   : 0.000   Regular      :7273   Min.   :0.0000
##  White    :1325   1st Qu.: 2.000   Showcase     : 218   1st Qu.:0.0000
##  Green    :1319   Median : 3.000   Foil         : 196   Median :0.0000
##  Black    :1316   Mean   : 2.837   Foil Showcase:  10   Mean   :0.3899
##  Blue     :1222   3rd Qu.: 4.000   Borderless   :   8   3rd Qu.:1.0000
##  Colorless: 696   Max.   :12.000   Alternate art:   2   Max.   :4.0000
##  (Other)  : 497   NA's   :1        (Other)      :   6
##      keywords          n         price
##          :5200   Min.   :1   Min.   : 0.0000
##  Cycling : 365   1st Qu.:1   1st Qu.: 0.1500
##  Kicker  : 299   Median :1   Median : 0.1500
##  Flying  : 293   Mean   :1   Mean   : 0.4832
##  Flash   : 199   3rd Qu.:1   3rd Qu.: 0.2500
##  Landfall: 153   Max.   :1   Max.   :35.0000
##  (Other) :1204
```

Zbadamy czy łączne ceny boosterów (suma cen kart w każdym poszczególnym boosterze) mają rozkład normalny. Za pomocą ggplot narysowaliśmy gęstość tego rozkładu. Widzimy, że napewno nie jest to gęstość rozkładu normalnego. Widzimy też dużo obserwacji odstających dla cen wyższych od 6. Rozpatrzymy dane bez obserwacji odstających. Wykres gęstośći rozkładu przypomina teraz rozkład normalny chociaż też widać, że raczej się różnią. Robimy test Shapiro-Wilka na nprmalność rozkładu. Dostajemy wartość p mniejszą od 0.05 co daje nam podstawę do odrzucenia hipotezy o normalności. Robimy wykres kwantylowy. On też pokazuje nam że dany rozkład różni się od normalnego.



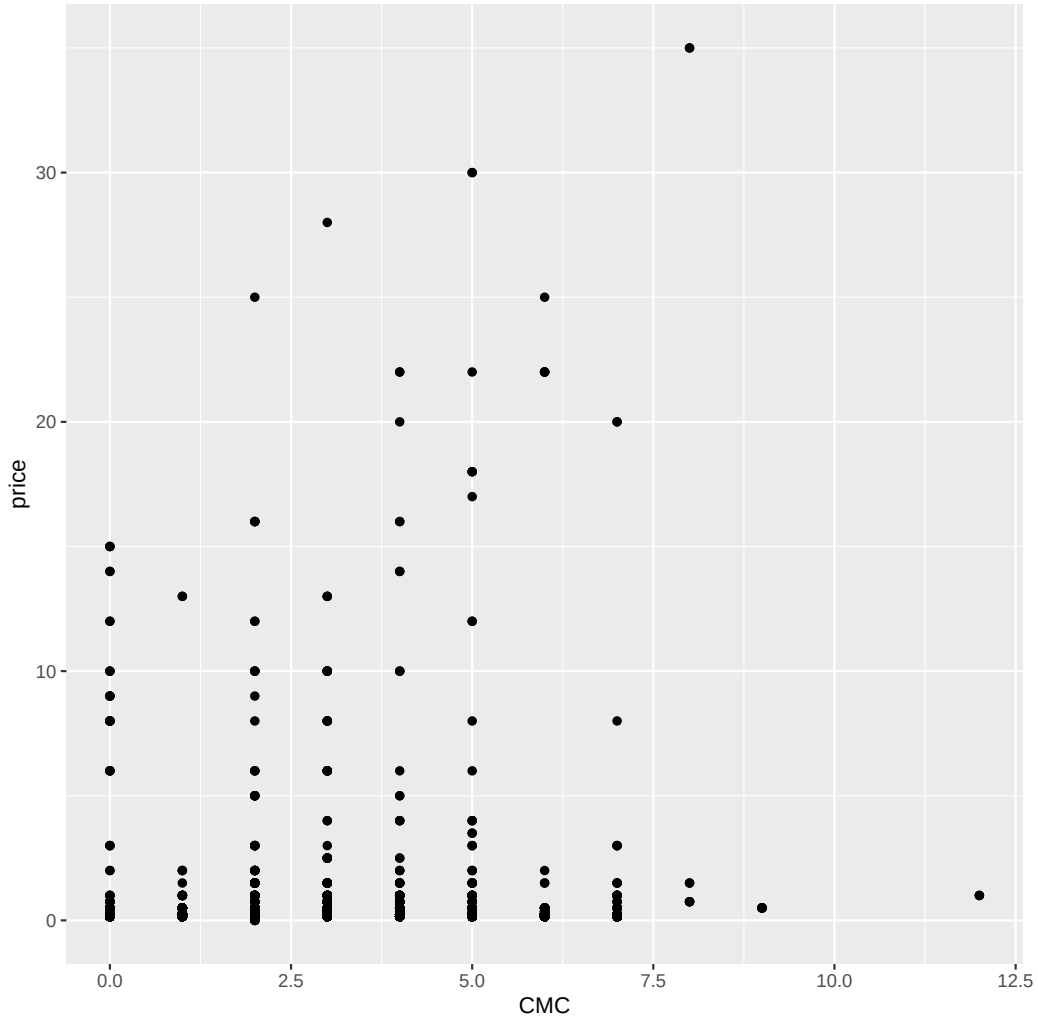



```
##
##  Shapiro-Wilk normality test
##
## data:  packPriceFilt$x
## W = 0.96361, p-value = 1.222e-08
```

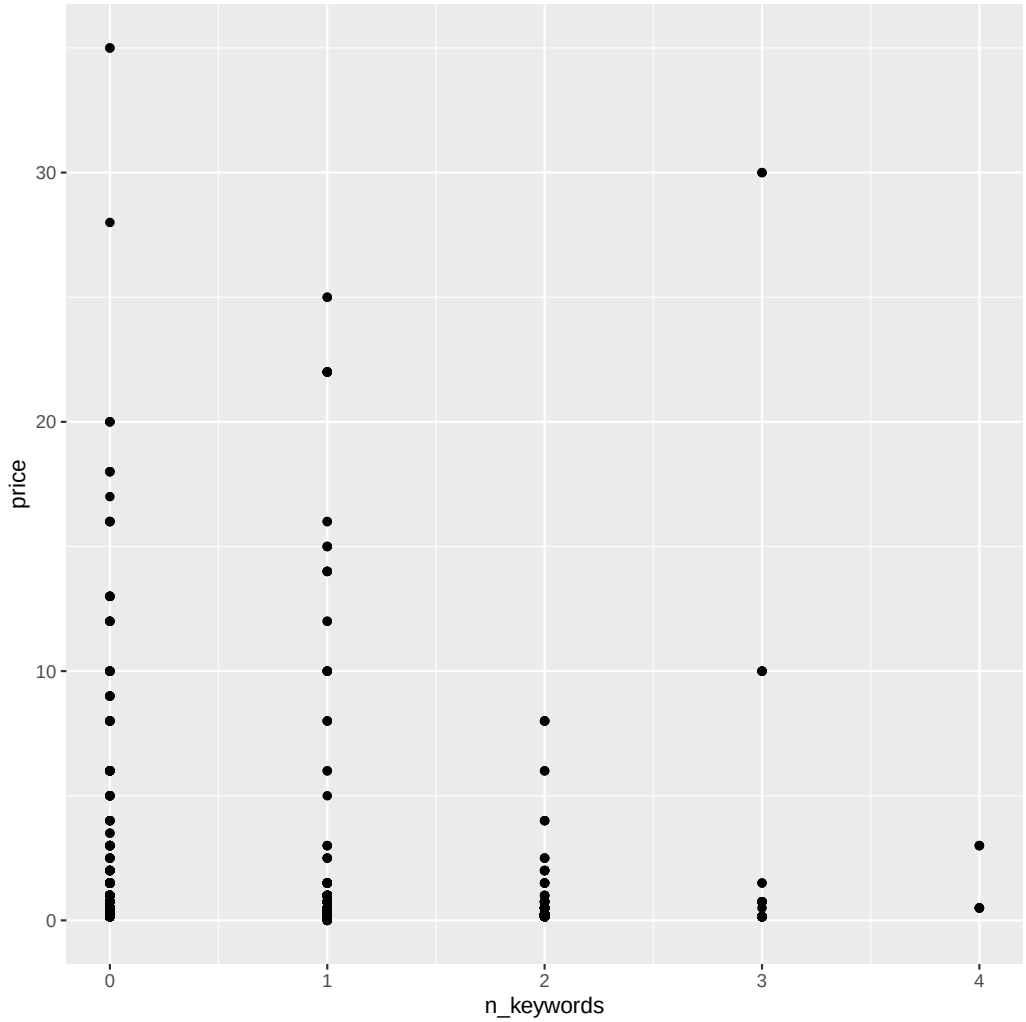



```
## [1] 343 207
```

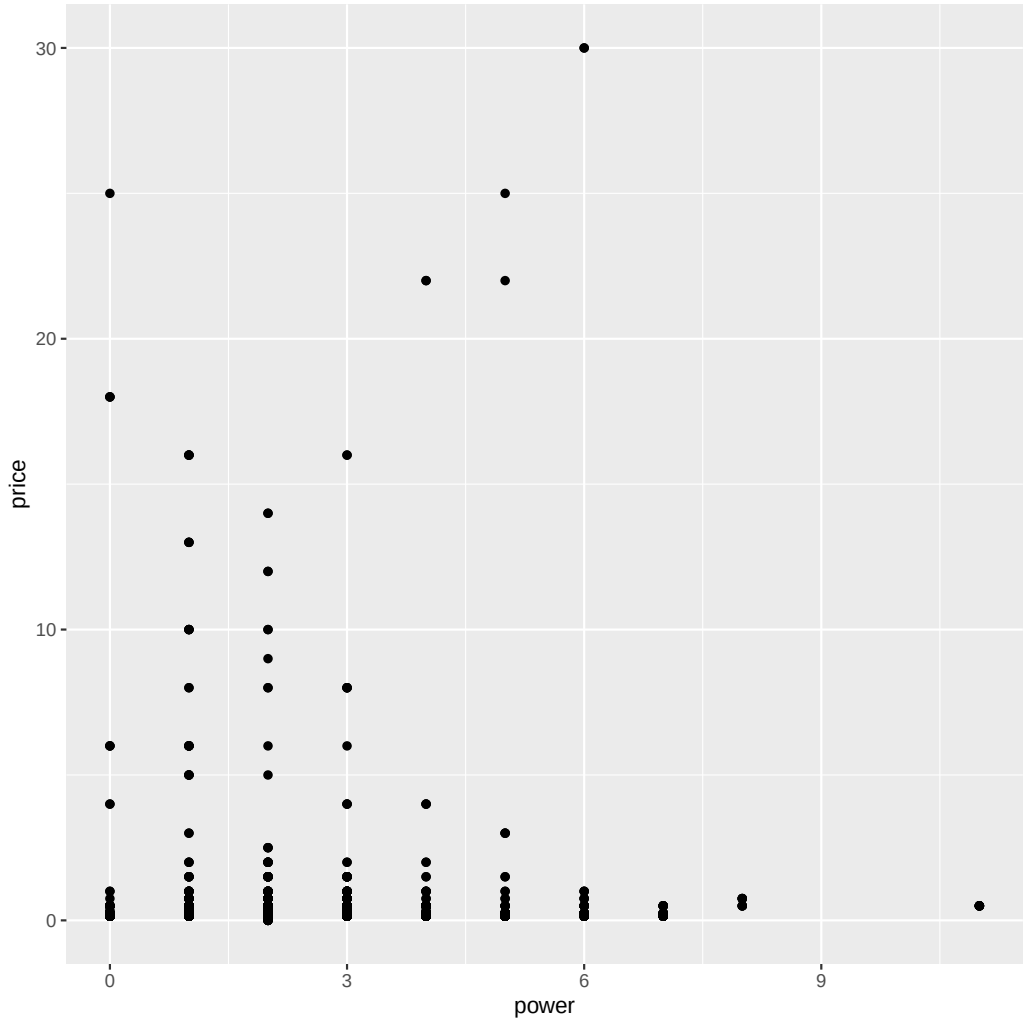
Teraz chcemy zbadać czy cena karty zależy od CMC i nkeywords. Z wykresów żadnego związku nie widać. Sprawdzamy współczynnik korelacji Pearsona. W obu przypadku jest bliski 0, co świadczy o bardzo słabej, prawie nieistotnej korelacji. Chcemy też sprawdzić czy istnieje korelacja między ceną karty, a jej mocą lub wytrzymałością. Te charakterystyki dotyczą tylko kart typu Creature, dlatego odbierzemy te dane w osobny dataframe creatures. Robimy wykresy dla mocy i wytrzymałości, obliczamy współczynniki Pearsona. Zarówno pierwsze jak i drugie wskazują na prawie zerową korelację, czyli nieistiotną. Podsumowując, możemy powiedzieć, że te zmienne nie są skorelowane.



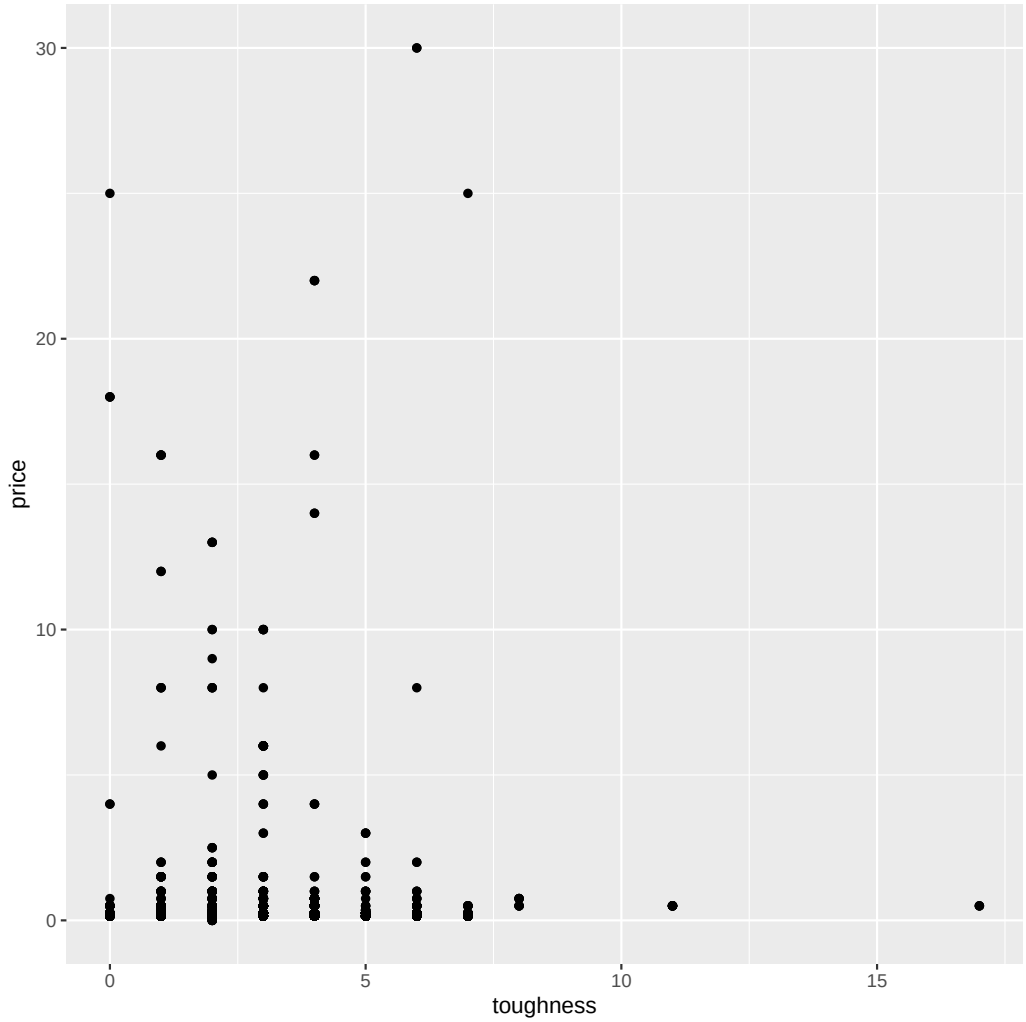

```
##
##  Pearson's product-moment correlation
##
## data:  cards$CMC and cards$price
## t = 4.6999, df = 7270, p-value = 2.651e-06
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.03209413 0.07792387
## sample estimates:
##       cor
## 0.05503798
```




```
##
##  Pearson's product-moment correlation
##
## data:  cards$n_keywords and cards$price
## t = 5.9701, df = 7271, p-value = 2.482e-09
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.04693597 0.09267762
## sample estimates:
##       cor
## 0.06984351
```




```
##
##  Pearson's product-moment correlation
##
## data:  creatures$power and creatures$price
## t = -0.06027, df = 3665, p-value = 0.9519
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.03336268  0.03137367
## sample estimates:
##          cor
## -0.0009955508
```




```
##
##  Pearson's product-moment correlation
##
## data:  creatures$toughness and creatures$price
## t = 0.66416, df = 3665, p-value = 0.5066
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.02140565  0.04332298
## sample estimates:
##       cor
## 0.01097015
```

0.5. Podsumowanie

Na samym początku sprawozdania zadaliśmy sobie pytania: Czy zależy cena karty od jej koloru, typu, CMC, n-keywords? Czy zależy cena stworzeń od ich mocy/wytrzymałości? Jaki jest rozkład łącznych cen boosterów?

Podsumowując, to co udało się nam zauważyć z wykresów jest:

— W przypadku zależności ceny kart od CMC i nkeywords nie zauważalny jest związek. Sprawdzają współczynnik korelacji wynosił on prawie 0. To samo tyczy się przypadku gdy badamy zależność ceny od mocy i wytrzymałości.
— Rozkład łącznych boosterów na pewno nie jest rozkładem normalnym.