Analiza danych - filmy Tarantino

Konrad Maciejewski

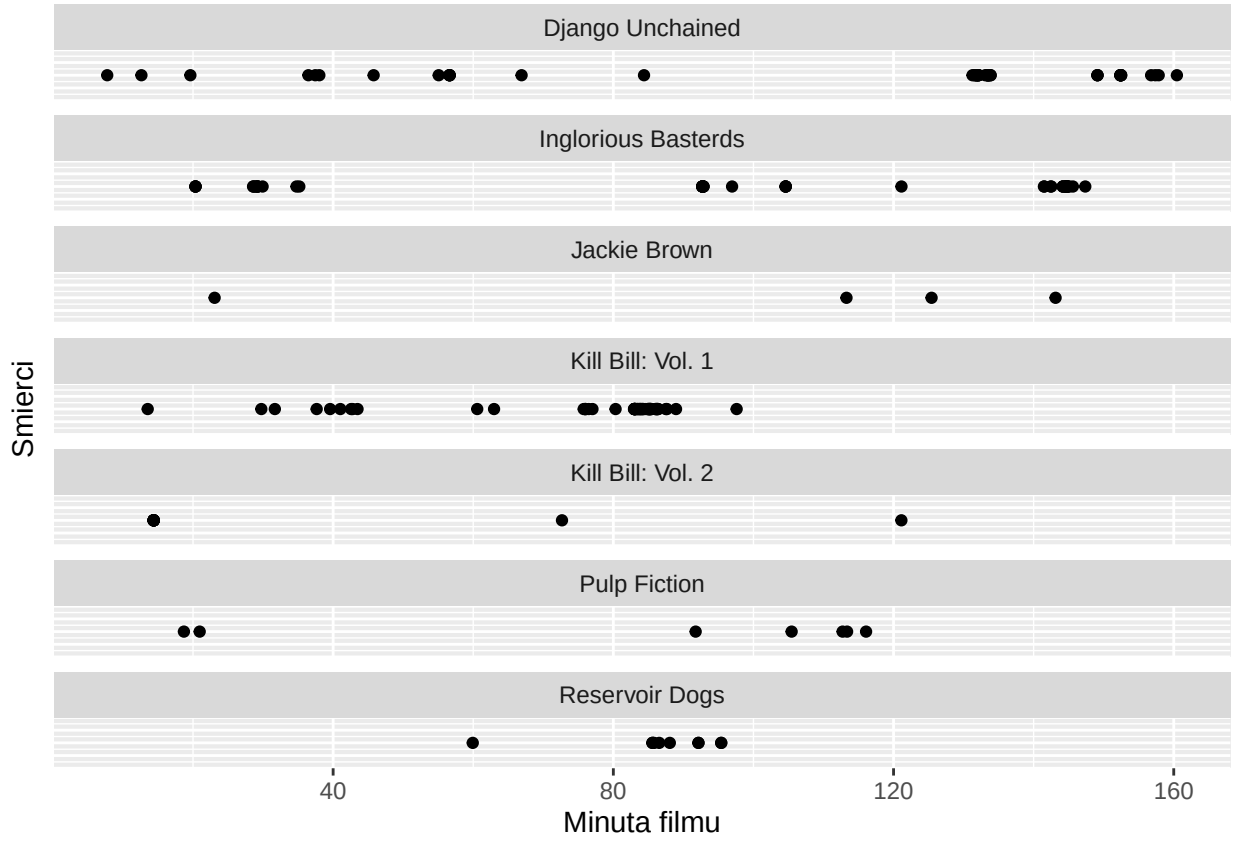
2023-06-03

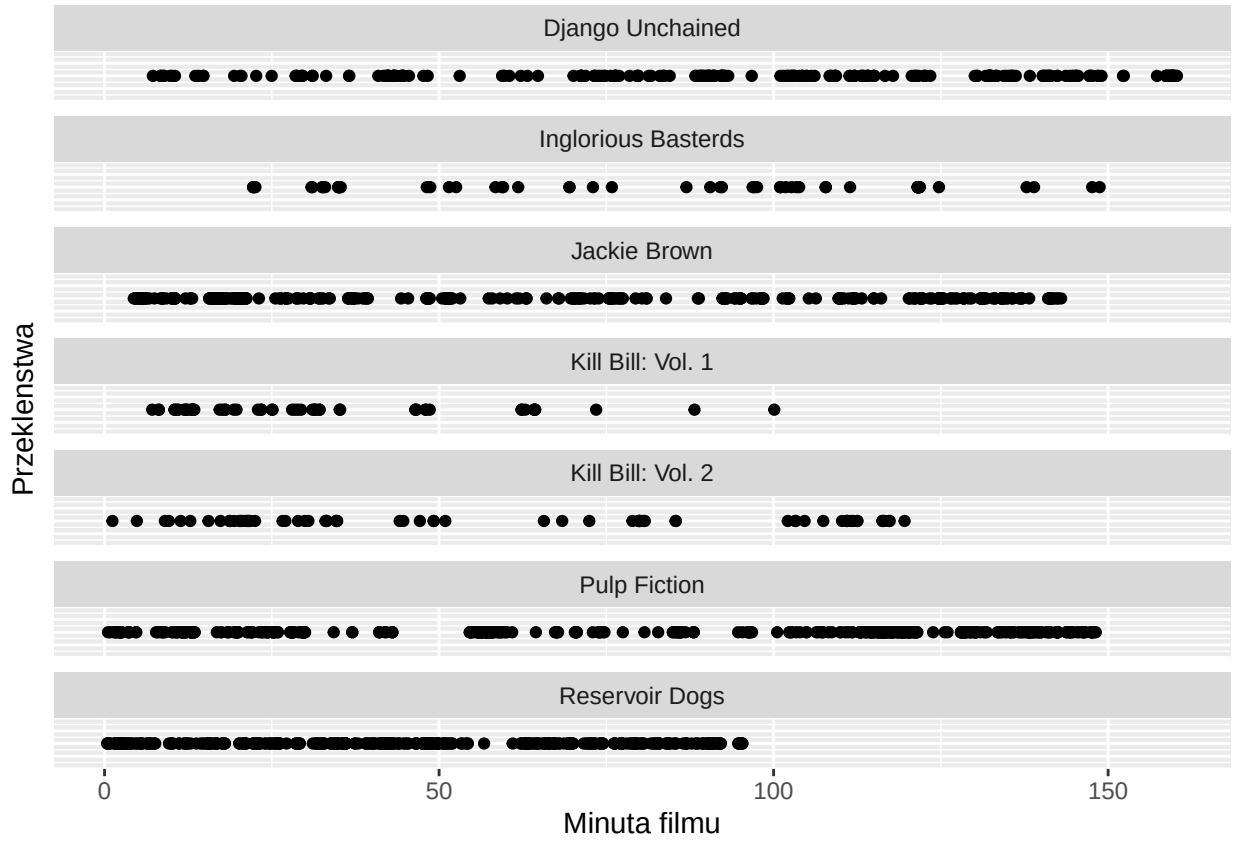
Opis zbioru

Dane są pobrane z repozytorium na GitHubie https://github.com/fivethirtyeight/data/tree/master/tarantino. Są to dane z obserwacji i dotyczą ilości poszczególnych przekleństw oraz śmierci w wybranych filmach reżyserowanych przez Quentina Tarantino.

Analiza eksploracyjna

Dwa przedstawione poniżej wykresy pokazują czas i ilość występowania śmierci bohaterów oraz przekleństw w poszczególnych filmach:





Trzeba zauważyć, że z wykresami są jednak problemy – jeżeli wiele bohaterów umiera w bardzo krótkim czasie, albo jeśli ktoś puści wiązankę, na wykresie widać tylko jeden punkt. Dla przykładu, z wykresu wynika, że w filmie „Kill Bill: Vol. 2" umierają tylko trzy osoby, kiedy w rzeczywistości ginie jedenaście.

Kolejnym problemem są różne czasy trwania filmów. Patrząc na wykres, można by (niepoprawnie) wywnioskować, że w filmie „Reservoir Dogs" około setnej minuty bohaterzy przechodzą przemianę i przestają przeklinać. Oczywiście nie dzieje się tak, tylko kończy się film. Poniższa tabela opisuje dokładne wartości danych w celu sprawniejszej interpretacji:

```
##                 Film Przekleństwa Czas trwania Częstotliwość
##       Reservoir Dogs          421        95.40     4.4129979
##         Pulp Fiction          469       148.15     3.1657104
##         Jackie Brown          368       143.13     2.5710892
##     Django Unchained          262       160.45     1.6329074
##    Kill Bill: Vol. 2           69       121.12     0.5696830
##    Kill Bill: Vol. 1           57       100.10     0.5694306
##  Inglorious Basterds           58       148.73     0.3899684
```

```
## Średnia ilość przekleństw: 243.4286
```

```
## Odchylenie standardowe ilości przekleństw: 181.5864
```

```
##                 Film Śmierci Czas trwania Częstotliwość
##    Kill Bill: Vol. 1      63       100.10    0.62937063
##  Inglorious Basterds      48       148.73    0.32273247
##     Django Unchained      47       160.45    0.29292615
##       Reservoir Dogs      10        95.40    0.10482180
##    Kill Bill: Vol. 2      11       121.12    0.09081902
##         Pulp Fiction       7       148.15    0.04724941
```

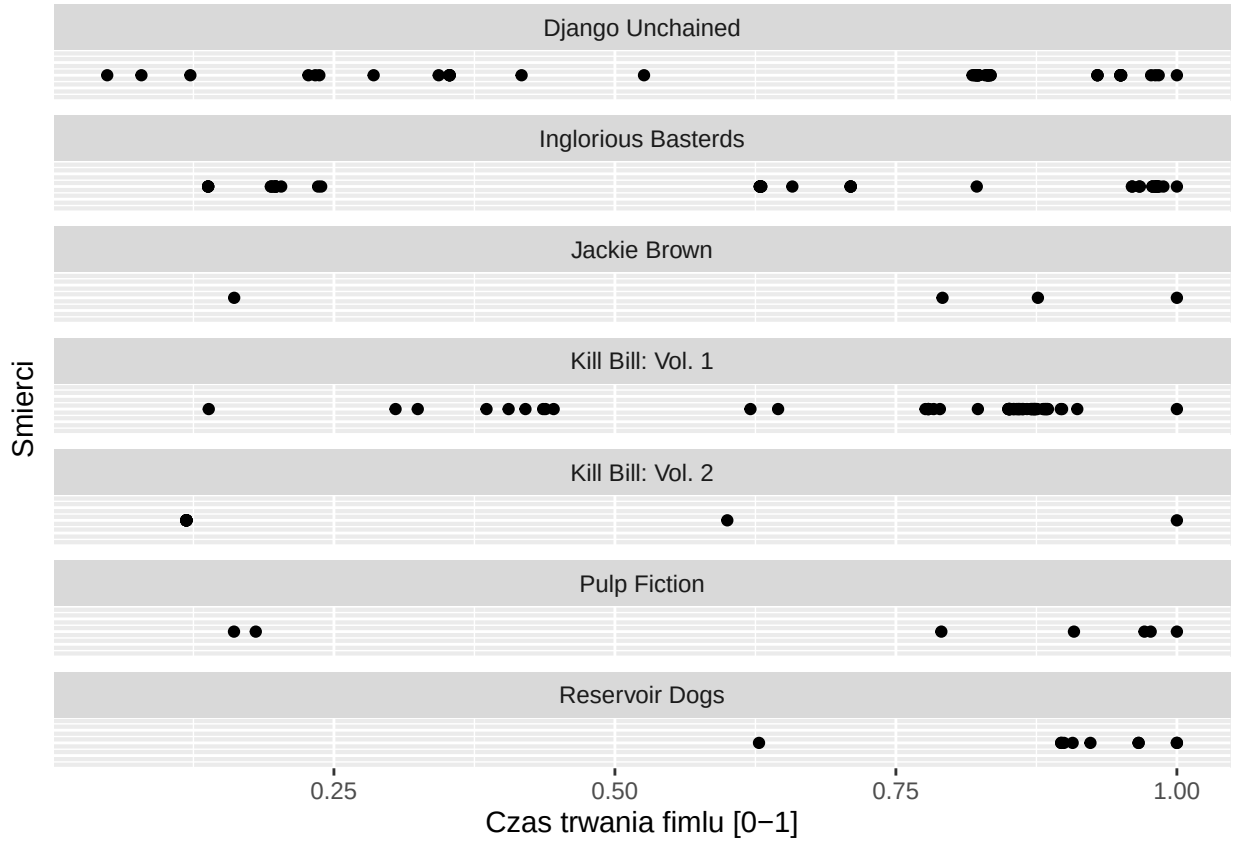
```
##          Jackie Brown       4       143.13    0.02794662
```

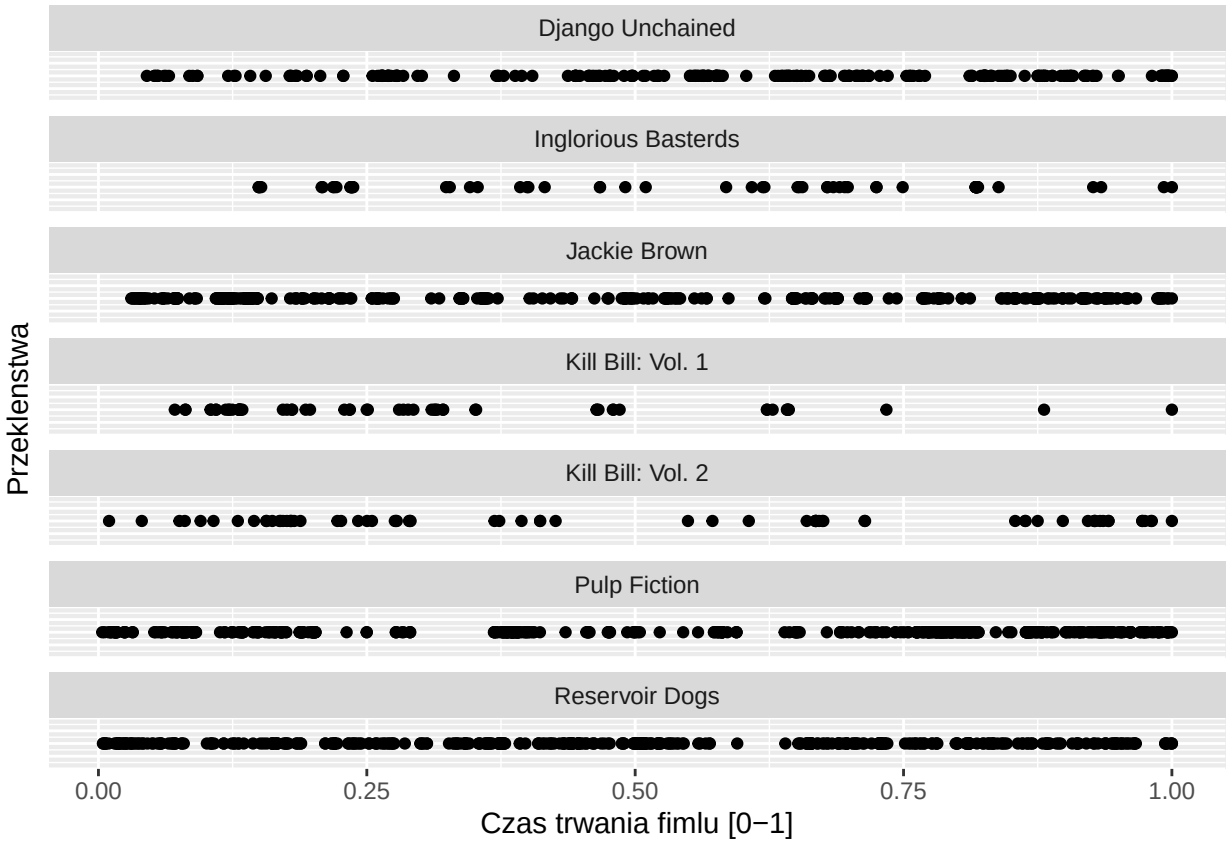
```
## Średnia ilość śmierci: 27.14286
```

```
## Odchylenie standardowe ilości śmierci: 24.53181
```

W tabeli kolumna „Czas trwania" nie jest rzeczywistym czasem trwania filmu, tylko ostatnim wystąpieniem przekleństwa/śmierci. Nie stanowi to jednak istotnej różnicy w przypadku większości filmów. Tarantino lubi kończyć swoje filmy czyjąś śmiercią, a bohaterzy przeklinają nieustannie. Największą różnicą pomiędzy rzeczywistym czasem trwania, a tym z tabelki jest 11 minut dla filmu „Kill Bill: Vol. 1". Nie zmienia to jednak kolejności po posortowaniu. Kolumna „Częstotliwość" jest wynikiem podzielenia drugiej kolumny przez trzecią.

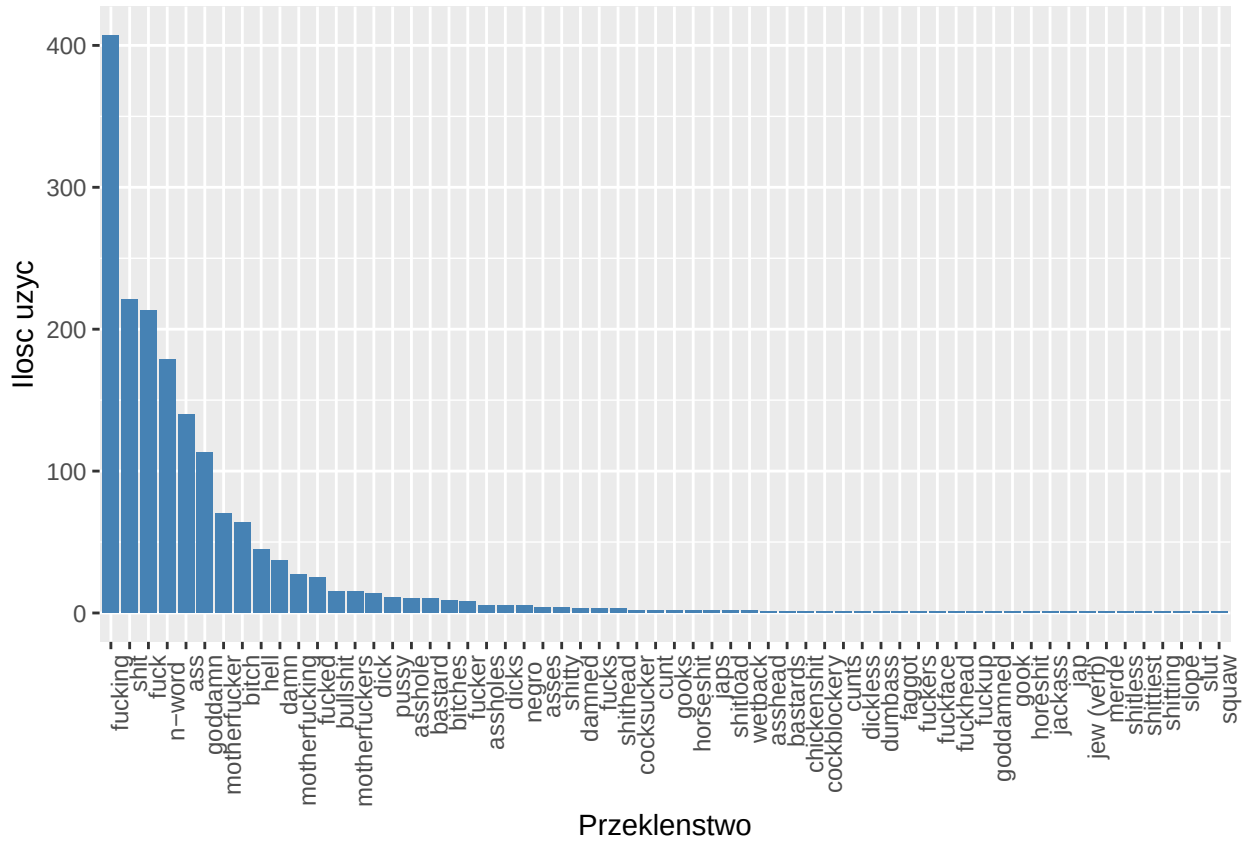
Poniższe wykresy zostały przeskalowane według kolumny „Czas trwania" do skali (0-1: od początku do końca filmu) na osi X:





Warto zauważyć, że analizując te wykresy można wyodrębnić poszczególne akty filmów.

Dodatkowo, kolejny wykres przedswaia ile razy padło każde przekleństwo:



Hipoteza 1: "Im więcej przekleństw w filmie, tym więcej śmierci."

Poniżej przedstawiona tabela pokazuje ilości przekleństw i śmierci w filmach:

```
##                 Film Przekleństwa Śmierci
##     Django Unchained          262      47
##  Inglorious Basterds           58      48
##          Jackie Brown          368       4
##    Kill Bill: Vol. 1           57      63
##    Kill Bill: Vol. 2           69      11
##          Pulp Fiction          469       7
##       Reservoir Dogs          421      10
```

Na podstawie tabeli można obliczyć współczynnik korelacji:

```
## [1] -0.6674097
```

Trzeba też wykonać test sprawdzający, czy zmienne sa skorelowane:

Hipoteza zerowa: $\rho = 0$
Hipoteza alternatywna: $\rho \neq 0$

Używamy statystyki t z pięcioma stopniami swobody:

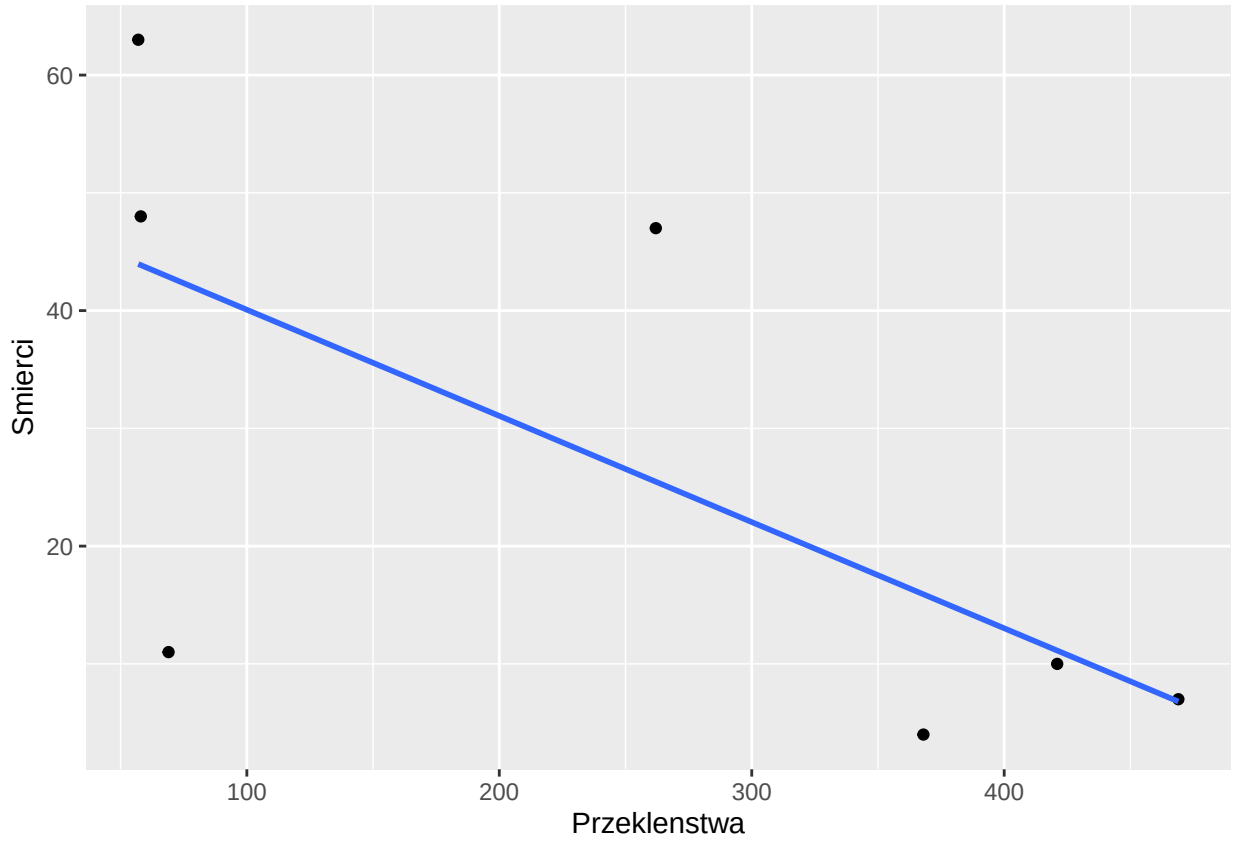
$$T_{n-2} = \frac{r_{xy}}{\sqrt{1 - r_{xy}^2}} \sqrt{n-2}$$

czego wynikiem jest:

```
##
##  Pearson's product-moment correlation
##
## data:  death_counts$type and word_counts$type
## t = -2.004, df = 5, p-value = 0.1014
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.9453410  0.1721916
## sample estimates:
##       cor
## -0.6674097
```

Według testu nie ma statystycznie istotnej podstawy do odrzucenia hipotezy zerowej. Istnieje jednak jakaś nieistotna korelacja wskazująca, że im więcej przekleństw, tym mniej śmierci. Można ją zobaczyc na wykresie:

```
## `geom_smooth()` using formula = 'y ~ x'
```



Hipoteza 2: "W drugim akcie występuje najmniej przekleństw."

Poniższa tabela pokazuje ilość przekleństw w każdym akcie filmów:

```
## # A tibble: 7 x 4
##   movie                `1`   `2`   `3`
##   <chr>              <int> <int> <int>
## 1 Django Unchained      56    85   121
## 2 Inglorious Basterds   18    13    27
## 3 Jackie Brown         153    84   131
## 4 Kill Bill: Vol. 1     39    15     3
```

```
## 5 Kill Bill: Vol. 2      31     9    29
## 6 Pulp Fiction          135    99   235
## 7 Reservoir Dogs        144   125   152
```

W celu sprawdzenia hipotezy trzeba wykonać dwa testy.

Test pierwszy:

Hipoteza zerowa: „Ilość przekleństw w pierwszym akcie jest równa lub mniejsza, niż w akcie drugim."
Hipoteza alternatywna: „Ilość przekleństw w pierwszym akcie jest większa, niż w akcie drugim."
Poziom istotności: $\alpha = 0.05$
Test t dla dwóch próbek:

```
##
##  Welch Two Sample t-test
##
## data:  pivoted_table$"1" and pivoted_table$"2"
## t = 0.72581, df = 11.511, p-value = 0.2412
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -30.5415      Inf
## sample estimates:
## mean of x mean of y
##  82.28571  61.42857
```

Podsumowując, na podstawie wyników testu t, nie mamy wystarczających dowodów, aby stwierdzić, że istnieje statystycznie istotna różnica między średnimi ilości przekleństw w pierwszym akcie a drugim akcie.

Test drugi:

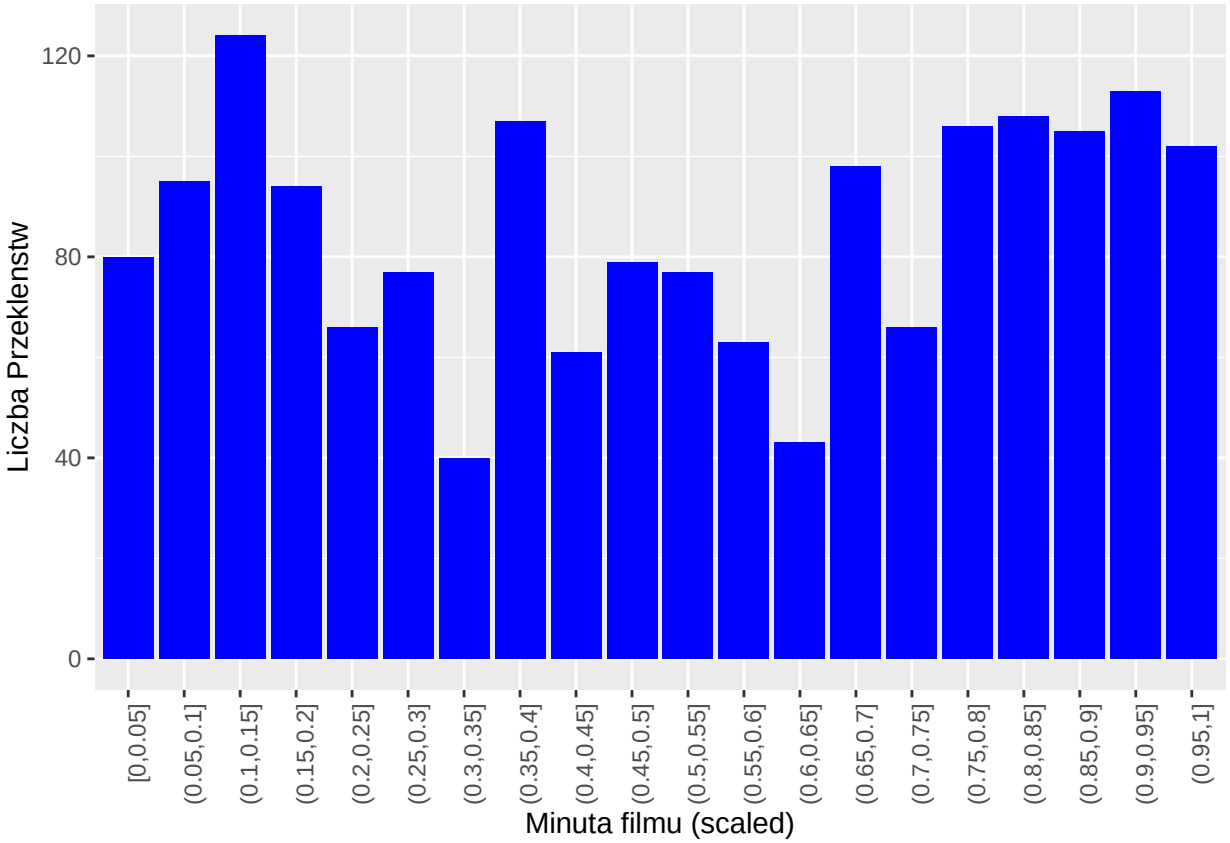
Hipoteza zerowa: „Ilość przekleństw w trzecim akcie jest równa lub mniejsza, niż w akcie drugim."
Hipoteza alternatywna: „Ilość przekleństw w trzecim akcie jest większa, niż w akcie drugim."
Poziom istotności: $\alpha = 0.05$
Test t dla dwóch próbek:

```
##
##  Welch Two Sample t-test
##
## data:  pivoted_table$"3" and pivoted_table$"2"
## t = 1.0496, df = 9.5445, p-value = 0.1599
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -28.14496       Inf
## sample estimates:
## mean of x mean of y
##  99.71429  61.42857
```

Podsumowując, na podstawie wyników testu t, nie mamy wystarczających dowodów, aby stwierdzić, że istnieje statystycznie istotna różnica między średnimi ilości przekleństw w akcie 3 a akcie 2.

Na podstawie obu testów nie można potwierdzić postawionej hipotezy.

Na koniec można spojrzeć na wykres, wyraźnie widać na nim podział na akty:


Liczba Przeklenstw
120
80
40
0
[0,0.05]
(0.05,0.1]
(0.1,0.15]
(0.15,0.2]
(0.2,0.25]
(0.25,0.3]
(0.3,0.35]
(0.35,0.4]
(0.4,0.45]
(0.45,0.5]
(0.5,0.55]
(0.55,0.6]
(0.6,0.65]
(0.65,0.7]
(0.7,0.75]
(0.75,0.8]
(0.8,0.85]
(0.85,0.9]
(0.9,0.95]
(0.95,1]
Minuta filmu (scaled)