StatisticalAnalysis

Kathryn Sam

2025-12-11

Data setup

Load file

```
dfSpacy<-read_excel('MovieData_Spacy.xlsx')
```

Change variable types and recode variables.

```
dfSpacy$Gender = as.factor(dfSpacy$Gender)
dfSpacy$WriterGender= as.factor(dfSpacy$WriterGender)
dfSpacy$MovieYear = as.numeric(dfSpacy$MovieYear)

# Recode for interpretation
dfSpacy$CharIsWoman <- ifelse(dfSpacy$Gender == "M", 0, 1)

# Man = 0, Both = 1, Woman = 2
dfSpacy$WriterGender <- ifelse(dfSpacy$WriterGender == "M", 0,
                               ifelse(dfSpacy$WriterGender == "B", 1, 2))
```

Check and remove movies which have no text in either character gender category

```
# Replace intensifiers with NA for rows with no word count
dfSpacy$intensifiers <- ifelse(dfSpacy$NW == 0, NA, dfSpacy$intensifiers)

# Inspect NAs
sum(is.na(dfSpacy$intensifiers))
```

```
## [1] 132
```

```
# Check if NAs are related to Char Sex code
table(is.na(dfSpacy$intensifiers), dfSpacy$CharIsWoman)
```

```
##
##           0   1
##   FALSE 570 524
##   TRUE   43  89
```

```r
# Count movies with no Text for both Male and Female
movies_no_text <- dfSpacy %>%
  group_by(MovieID) %>%
  summarise(
    no_text_male = all(is.na(Text[CharIsWoman == 0])),
    no_text_female = all(is.na(Text[CharIsWoman == 1]))
  ) %>%
  filter(no_text_male & no_text_female)

# Number of movies with no Text data
print(paste("Number of movies with no Text data:", nrow(movies_no_text)))
```

```
## [1] "Number of movies with no Text data: 19"
```

```r
#Drop if movies have no text for either M or F
dfSpacy2 <- dfSpacy %>%
  filter(!MovieID %in% movies_no_text$MovieID)

# Final data
length(unique(dfSpacy2$MovieID))
```

```
## [1] 594
```

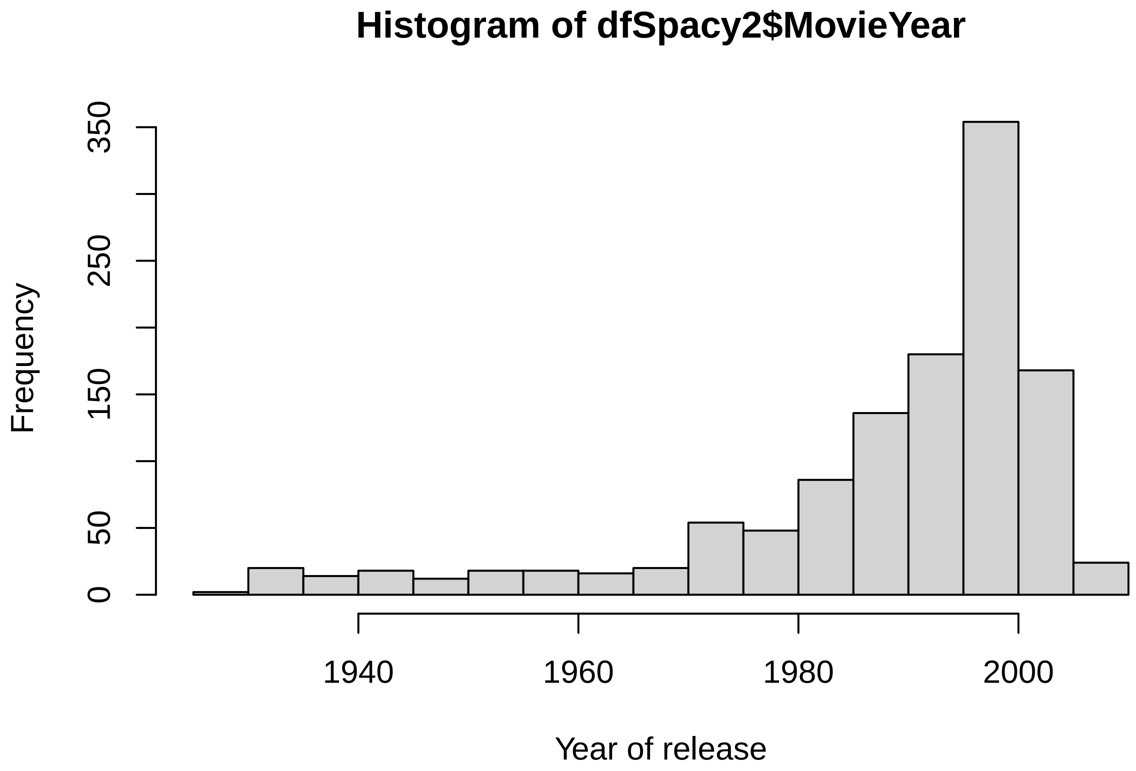
Descriptives

Movie Year

```r
describe(dfSpacy2$MovieYear) # Descriptives
```

```
##    vars    n    mean    sd median trimmed mad  min  max range  skew kurtosis
## X1    1 1188 1988.53 16.36   1994 1991.63 8.9 1927 2009    82 -1.75     2.73
##      se
## X1 0.47
```

```r
hist(dfSpacy2$MovieYear, xlab = "Year of release") # Histogram
```

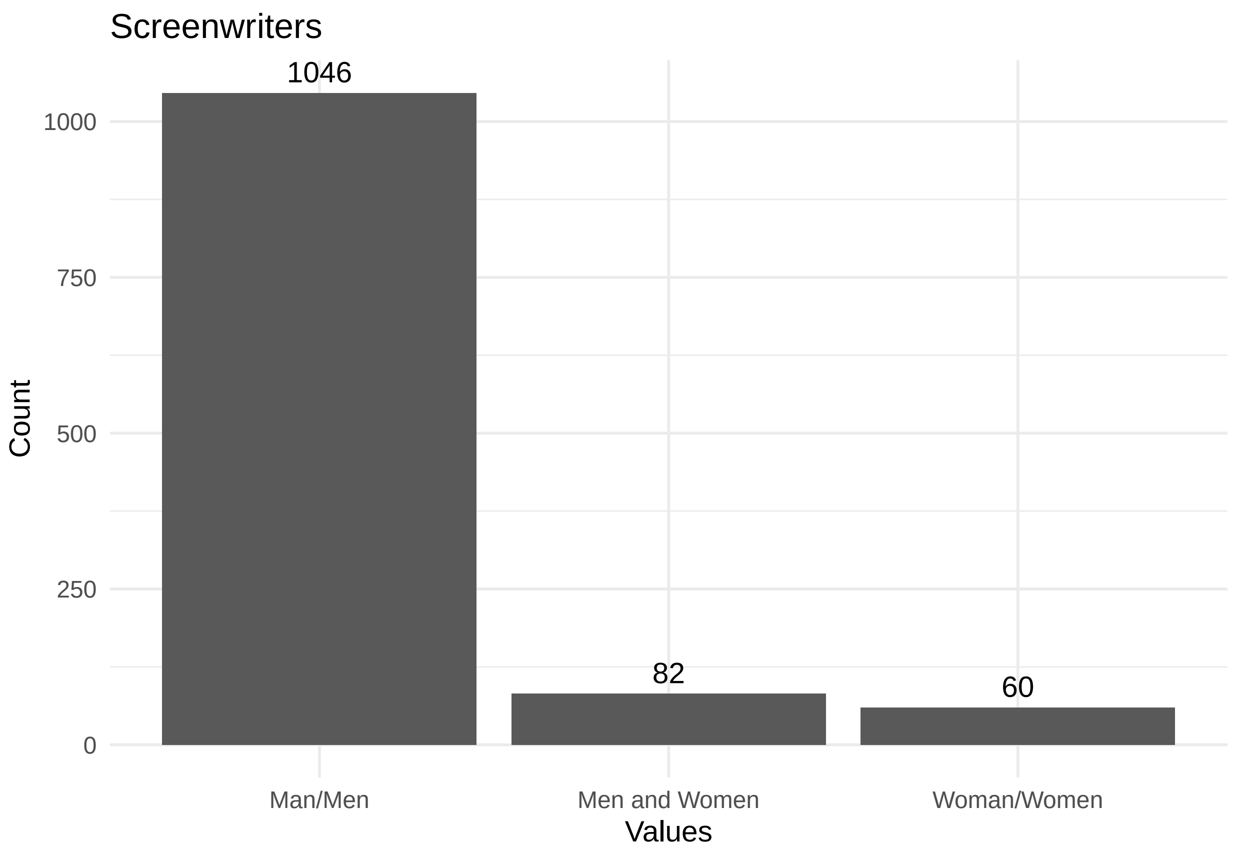


Writer Gender

```
describe(dfSpacy2$WriterGender) # Descriptives
```

```
##    vars    n mean   sd median trimmed mad min max range skew kurtosis   se
## X1    1 1188 0.17 0.49      0    0.03   0   0   2     2 2.89     7.21 0.01
```

```
ggplot(dfSpacy2, aes(x = as.factor(WriterGender))) + # Bar chart
  geom_bar() +
  geom_text(stat = "count",
            aes(label = after_stat(count)),
            vjust = -0.5) +
  labs(x = "Values", y = "Count", title = "Screenwriters") +
  scale_x_discrete(labels = c("Man/Men", "Men and Women", "Woman/Women"))+
  theme_minimal()
```

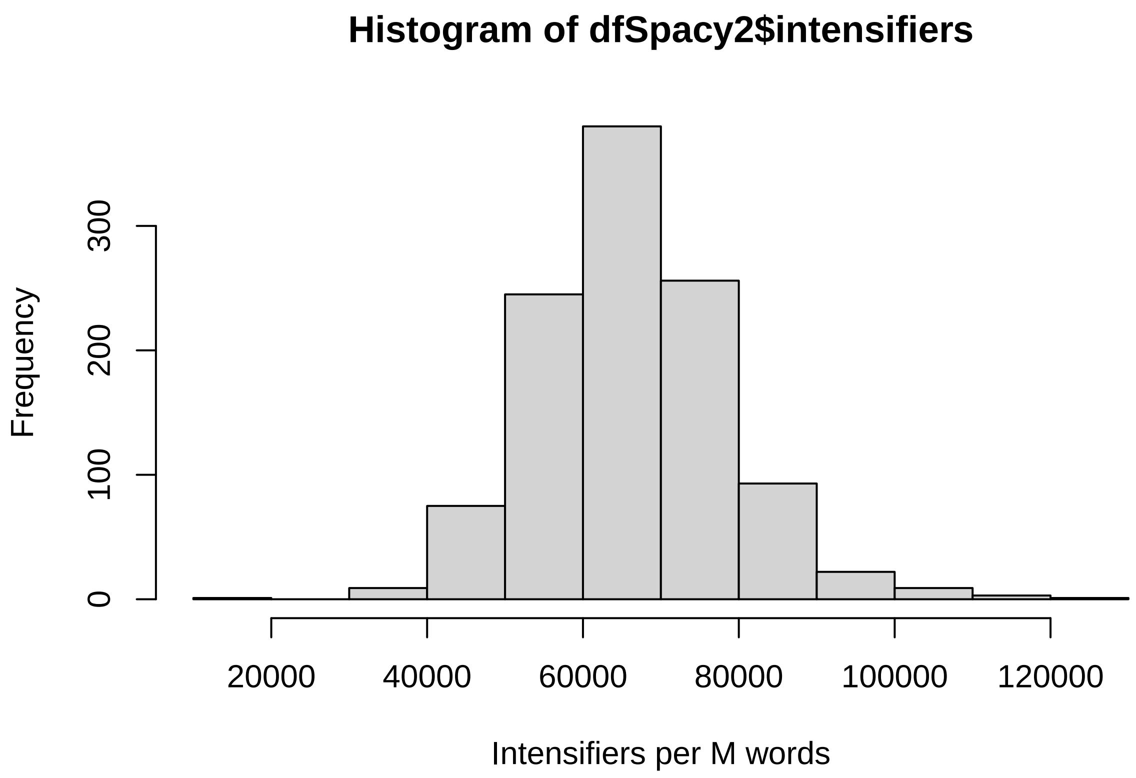


Intensifiers

```
describe(dfSpacy2$intensifiers) # Descriptives
```

```
##    vars    n     mean       sd   median  trimmed      mad      min      max
## X1    1 1094 66358.61 12195.42 65511.27 66015.54 11583.49 13574.66 127423.8
##       range skew kurtosis     se
## X1 113849.2 0.41     1.29 368.71
```

```
hist(dfSpacy2$intensifiers, xlab = "Intensifiers per M words") # Histogram
```

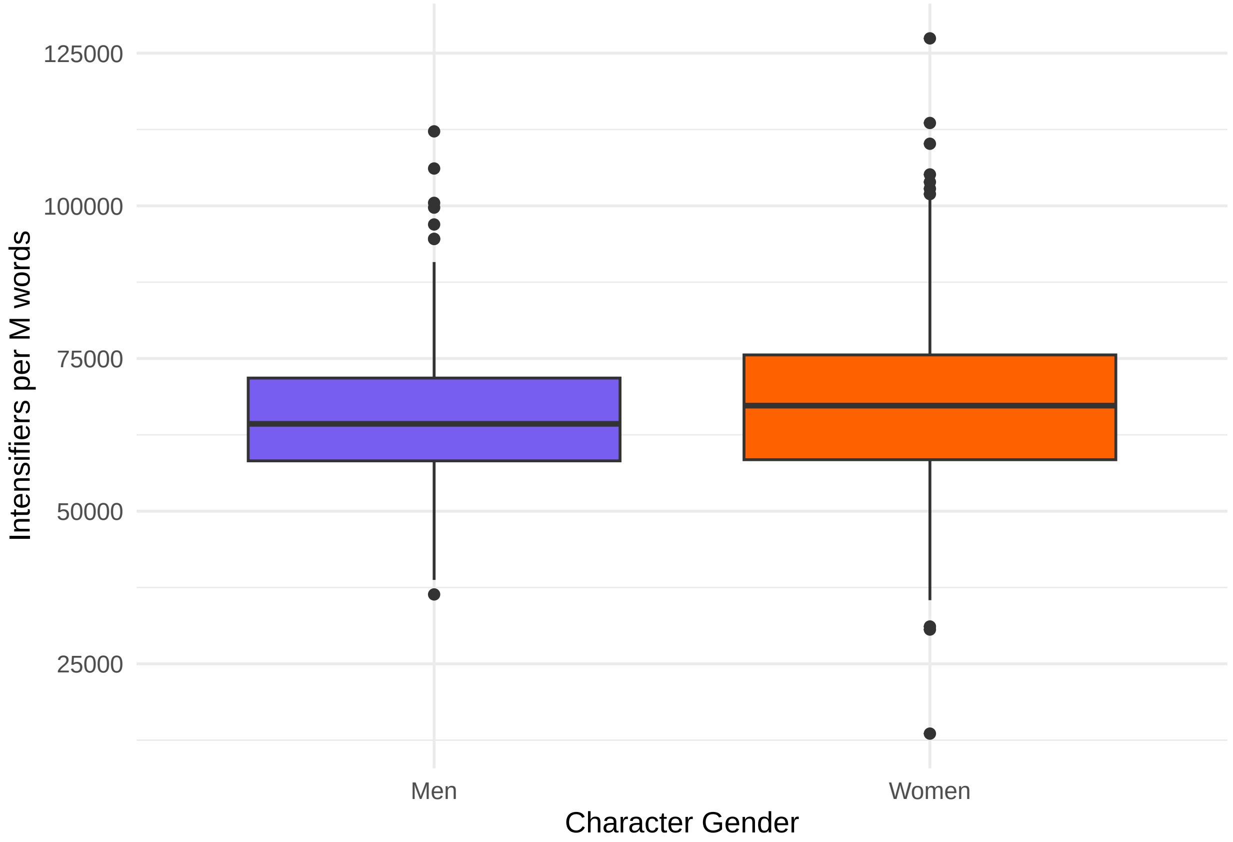


Plots

Intensifiers by Character Gender

```
ggplot(dfSpacy2, aes(x=as.factor(CharIsWoman), y=intensifiers,
                     fill=as.factor(CharIsWoman))) +
  geom_boxplot() + scale_x_discrete(labels = c("Men", "Women"))+
  scale_fill_manual(values = c("#785EF0", "#FE6100"))+
  labs(x = "Character Gender", y = "Intensifiers per M words") +
  theme(axis.text = element_text(color = "black")) +
  theme_minimal()+ theme(legend.position = "none")
```

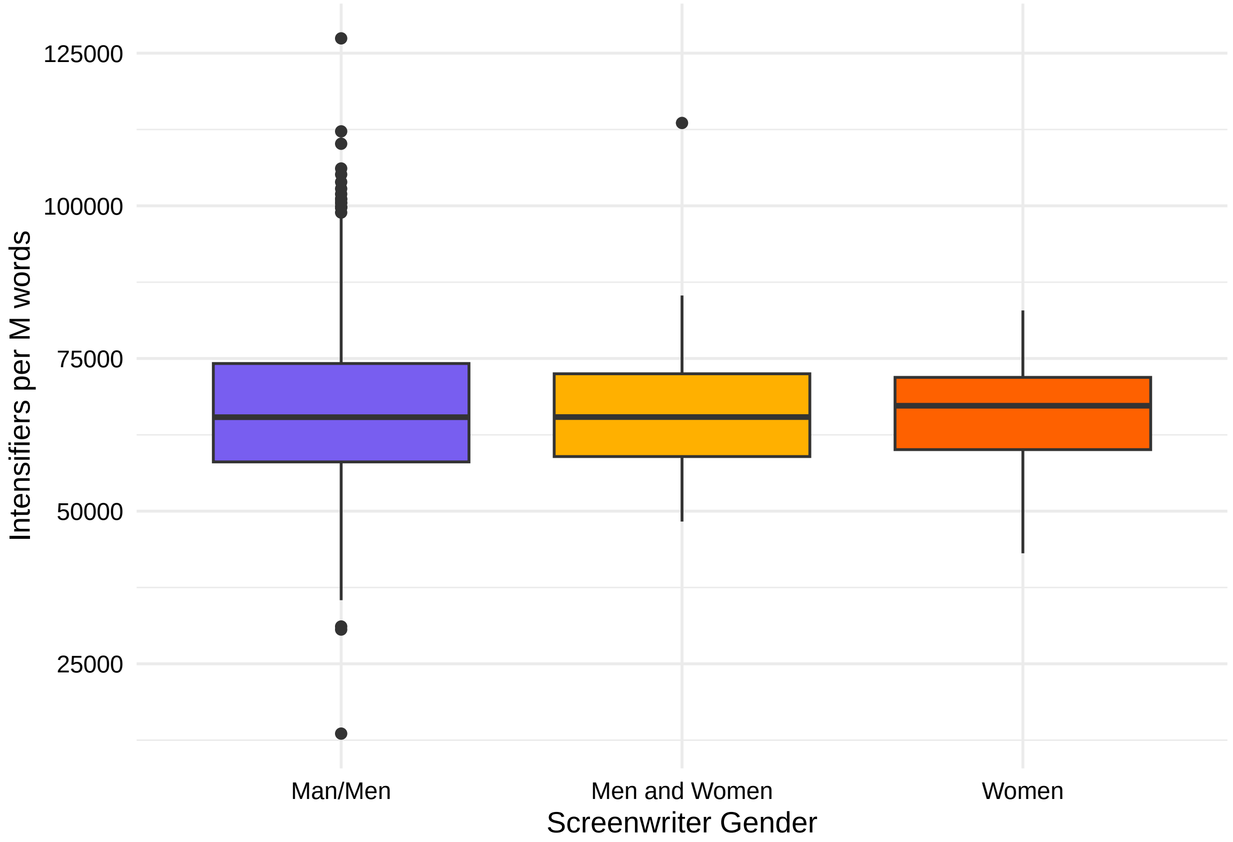
```
## Warning: Removed 94 rows containing non-finite outside the scale range
## (‘stat_boxplot()‘).
```



Intensifiers by Writer Gender

```
ggplot(dfSpacy2, aes(x=as.factor(WriterGender), y=intensifiers,
                     fill=as.factor(WriterGender))) +
  scale_fill_manual(values = c("#785EF0", "#FFB000", "#FE6100"))+
  geom_boxplot() +
  scale_x_discrete(labels = c("Man/Men", "Men and Women", "Women"))+
  labs(x = "Screenwriter Gender", y = "Intensifiers per M words") +
  theme_minimal()+ theme(axis.text = element_text(color = "black"))+
  theme(legend.position = "none")
```

```
## Warning: Removed 94 rows containing non-finite outside the scale range
## (‘stat_boxplot()‘).
```



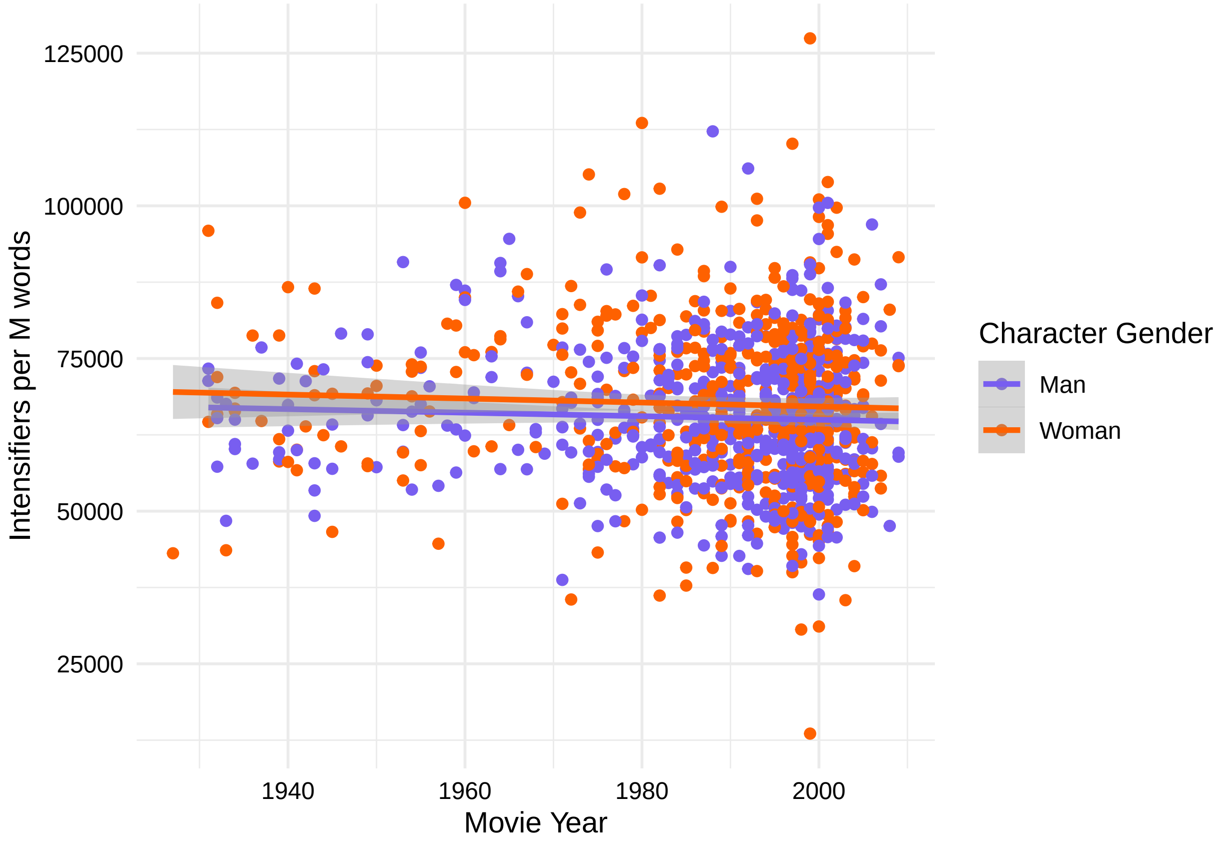
Intensifiers over time by character

```
ggplot(dfSpacy2, aes(x = MovieYear, y = intensifiers,
                     color = as.factor(CharIsWoman))) +
  geom_point() +
  labs(color="Character Gender", x = "Movie Year",
       y = "Intensifiers per M words") +
  geom_smooth(method=lm) +
  scale_color_manual(values = c("#785EF0", "#FE6100"),
                     labels = c("Man", "Woman")) + theme_minimal() +
  theme(axis.text = element_text(color = "black"))
```

```
## ‘geom_smooth()‘ using formula = ’y ~ x’
```

```
## Warning: Removed 94 rows containing non-finite outside the scale range
## (‘stat_smooth()‘).
```

```
## Warning: Removed 94 rows containing missing values or values outside the scale range
## (‘geom_point()‘).
```



Statistical Models

Null model

```
nullmodel <- lmer (intensifiers~(1|MovieID), data = dfSpacy, REML = FALSE)
summary(nullmodel)
```

```
## Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite’s
##   method [lmerModLmerTest]
## Formula: intensifiers ~ (1 | MovieID)
##    Data: dfSpacy
##
##       AIC       BIC    logLik -2*log(L)  df.resid
##   23641.0   23656.0  -11817.5   23635.0      1091
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.7475 -0.5642 -0.0544  0.5350  4.4086
##
## Random effects:
##  Groups   Name        Variance  Std.Dev.
##  MovieID  (Intercept)  47648077  6903
##  Residual             100840389 10042
## Number of obs: 1094, groups:  MovieID, 594
```

```
##
## Fixed effects:
##             Estimate Std. Error      df t value Pr(>|t|)
## (Intercept)  66304.8      417.7   587.6   158.7   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# tab_model(nullmodel,, show.se=TRUE, show.ci=FALSE, show.aicc=TRUE)
```

Model with predictors

A random slopes model is not possible given the data.

```
# model1.1<- lmer (intensifiers~(CharIsWoman|MovieID) + CharIsWoman*MovieYear +WriterGender,
              #  data = dfSpacy2, REML = FALSE, lmerControl(autoscale = TRUE))

#Error: number of observations (=1094) <= number of random effects (=1188) for term (CharIsWoman | Movi
```

Random intercept model

```
model1.2 <- lmer (intensifiers~(1|MovieID) + CharIsWoman*MovieYear +WriterGender,
                data = dfSpacy2, REML = FALSE, lmerControl(autoscale = TRUE))
summary(model1.2)
```

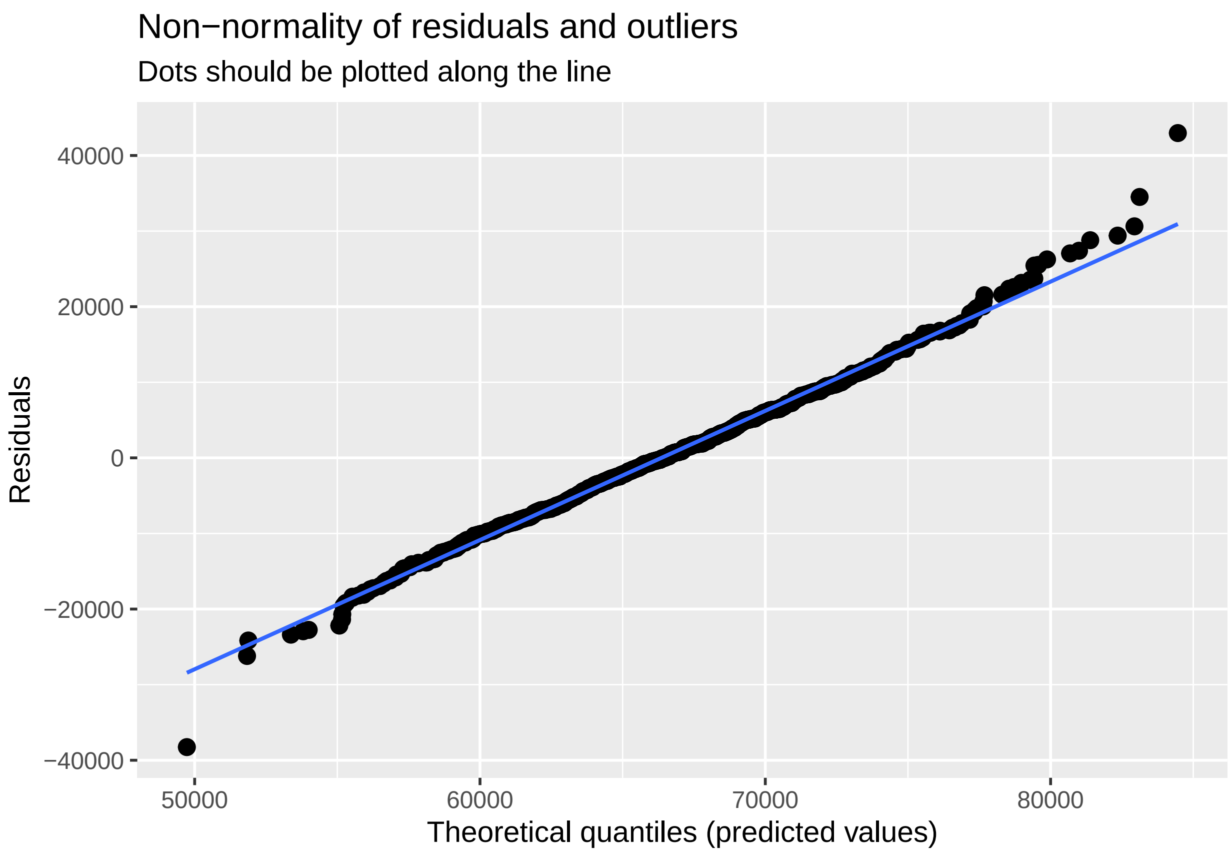
```
## Linear mixed model fit by maximum likelihood . t-tests use Satterthwaite's
##   method [lmerModLmerTest]
## Formula: intensifiers ~ (1 | MovieID) + CharIsWoman * MovieYear + WriterGender
##    Data: dfSpacy2
## Control: lmerControl(autoscale = TRUE)
##
##       AIC       BIC    logLik -2*log(L)  df.resid
##   23634.8   23669.7  -11810.4   23620.8      1087
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -3.8550 -0.5856 -0.0384  0.5284  4.3290
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  MovieID  (Intercept) 48502023 6964
##  Residual             98504136 9925
## Number of obs: 1094, groups:  MovieID, 594
##
## Fixed effects:
##                        Estimate Std. Error         df  t value Pr(>|t|)
## (Intercept)           6.632e+04  6.189e+04  2.855e+11    1.071    0.284
## CharIsWoman           8.800e+03  7.325e+04  8.775e+03    0.120    0.904
## MovieYear            -4.300e+02  3.113e+01  1.369e-02  -13.813    0.928
## WriterGender         -1.279e+02  8.544e+02  1.070e+04   -0.150    0.881
## CharIsWoman:MovieYear -7.707e+03  3.683e+01  5.609e-10 -209.227    1.000
```

```
# tab_model(model1.2, df.method = "satterthwaite", show.se=TRUE, show.ci=FALSE,
# show.icc=FALSE, show.aicc=TRUE)
```

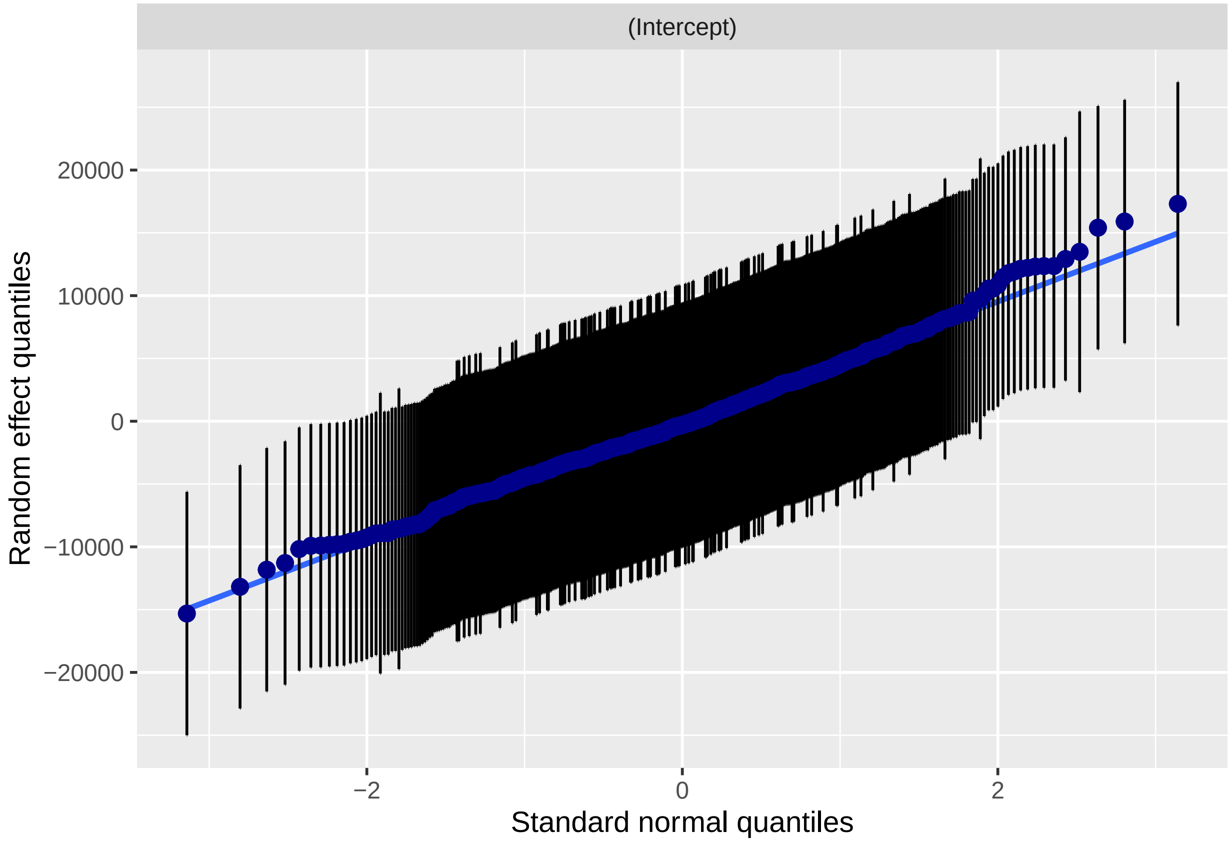
Model assumptions

```
plot_model(model1.2, type='diag')
```

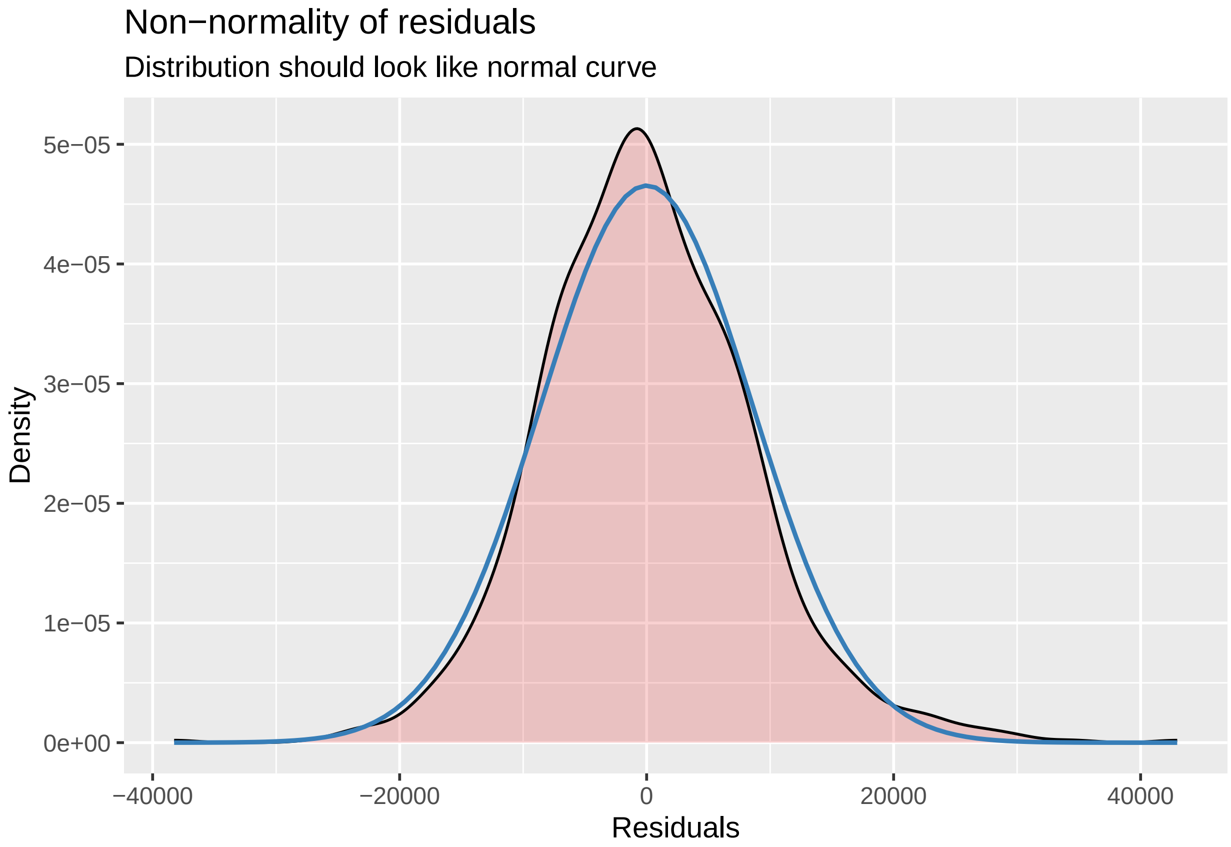
```
## Warning: Using ‘size‘ aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use ‘linewidth‘ instead.
## i The deprecated feature was likely used in the sjPlot package.
##   Please report the issue at <https://github.com/strengejacke/sjPlot/issues>.
## This warning is displayed once every 8 hours.
## Call ‘lifecycle::last_lifecycle_warnings()‘ to see where this warning was
## generated.

## [[1]]

## ‘geom_smooth()‘ using formula = ’y ~ x’
```



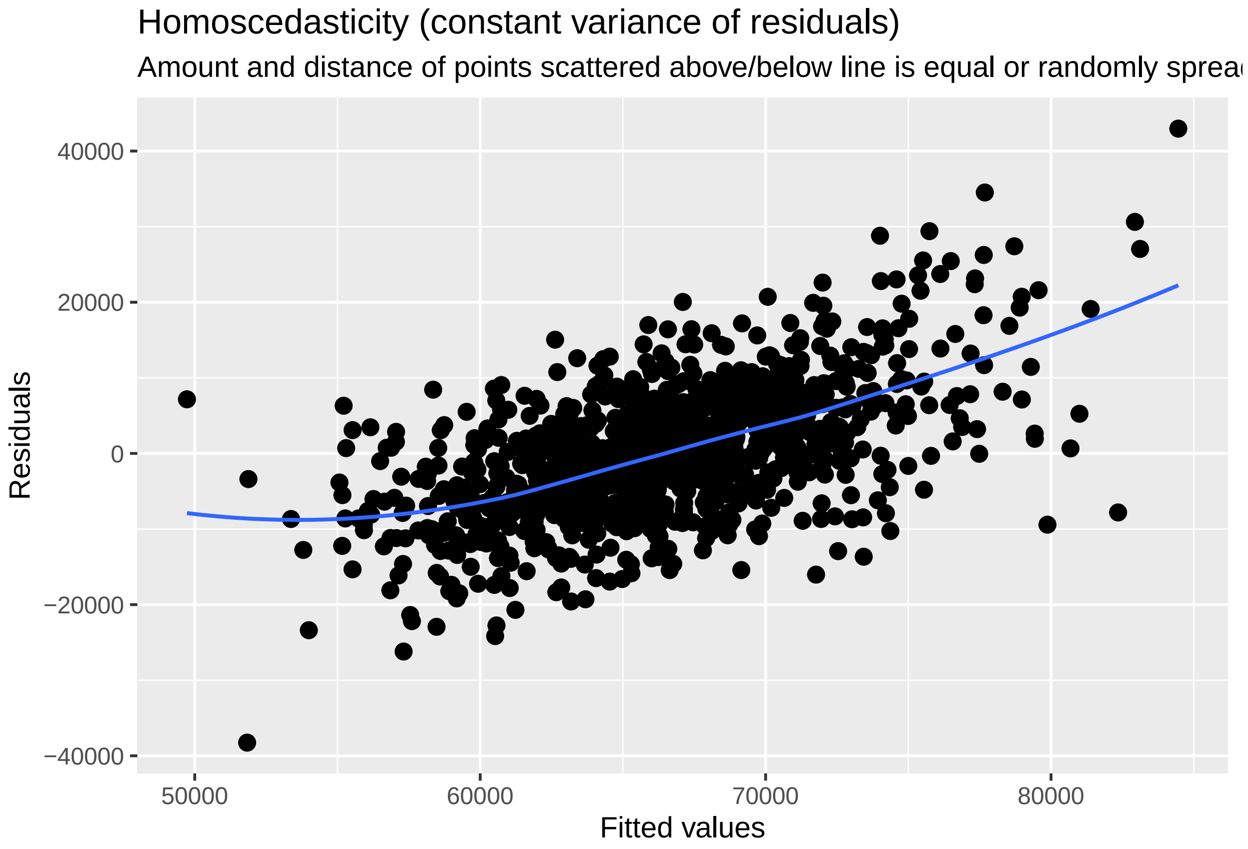
```
##
## [[2]]
## [[2]]$MovieID

## ‘geom_smooth()‘ using formula = ’y ~ x’
```


(Intercept)
Random effect quantiles
20000
10000
0
−10000
−20000
−2
0
2
Standard normal quantiles

```
##
##
## [[3]]
```



```
##
## [[4]]

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



As residuals are not homoscedastic, try running a model using robust lmer. Overall conclusions are the same, so retain model 1.2.

```
model2<- rlmer(intensifiers~(1|MovieID)+CharIsWoman*MovieYear+WriterGender,
               data = dfSpacy2,
               REML = FALSE, lmerControl(autoscale = TRUE))
summary(model2)
```

```
## Robust linear mixed model fit by DAStau
## Formula: intensifiers ~ (1 | MovieID) + CharIsWoman * MovieYear + WriterGender
##    Data: dfSpacy2
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.8161 -0.6192 -0.0325  0.6130  5.4984
##
## Random effects:
##  Groups   Name        Variance  Std.Dev.
##  MovieID  (Intercept)  21439844  4630
##  Residual             106057901 10298
## Number of obs: 1094, groups: MovieID, 594
##
## Fixed effects:
##                       Estimate Std. Error t value
## (Intercept)           66063.79     375.73  175.83
## CharIsWoman            8619.45   38742.21    0.22
## MovieYear              -469.11     491.39   -0.95
## WriterGender             14.69     375.62    0.04
## CharIsWoman:MovieYear -7531.16   38744.31   -0.19
##
## Correlation of Fixed Effects:
##             (Intr) ChrIsW MoviYr WrtrGn
## CharIsWoman  0.000
## MovieYear    0.001  0.642
## WriterGendr -0.001 -0.022 -0.047
## ChrIsWmn:MY  0.000 -1.000 -0.642  0.022
##
## Robustness weights for the residuals:
##  909 weights are ~= 1. The remaining 185 ones are summarized as
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.245   0.638   0.831   0.769   0.931   0.999
##
## Robustness weights for the random effects:
##  All 594 weights are ~= 1.
##
## Rho functions used for fitting:
##   Residuals:
##     eff: smoothed Huber (k = 1.345, s = 10)
##     sig: smoothed Huber, Proposal 2 (k = 1.345, s = 10)
##   Random Effects, variance component 1 (MovieID):
##     eff: smoothed Huber (k = 1.345, s = 10)
##     vcp: smoothed Huber, Proposal 2 (k = 1.345, s = 10)
```

```
# tab_model(model2, show.se=TRUE, df.method = "satterthwaite", show.ci=FALSE,
# show.icc=FALSE) view model
```