CaseStudy1_STA610

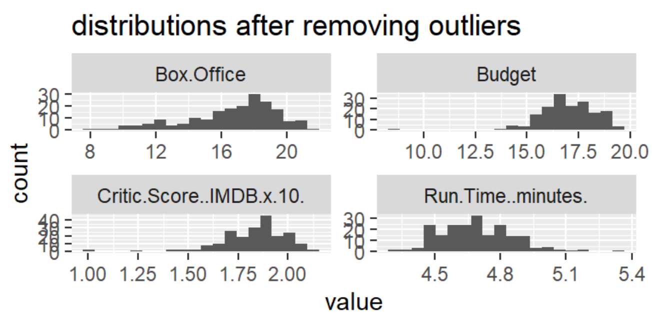
ElenaW.

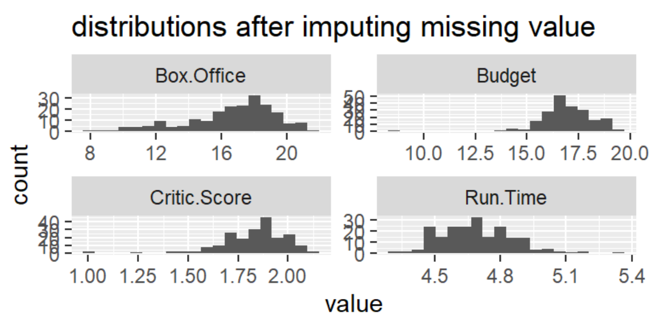
10/12/2022

Cleaning Dataset & EDA

The above summary statistics for numerical variables showing that we only have missing value in `Box Office` and `Budget`. If both of `Box office` and `Budget` are missing, this observations should be dropped.
Since most of `company`, `lead.Cast 1-3`, and `Director` are only for one specific movie, they are individual characterized, and don't significant affect across the group. Therefore, these columns will be dropped for controlling overfitting and model complexity. Moreover, we don't care about the movie title and release date in terms of profit investigation, thus, `Title` and `Release Date` will also be dropped.

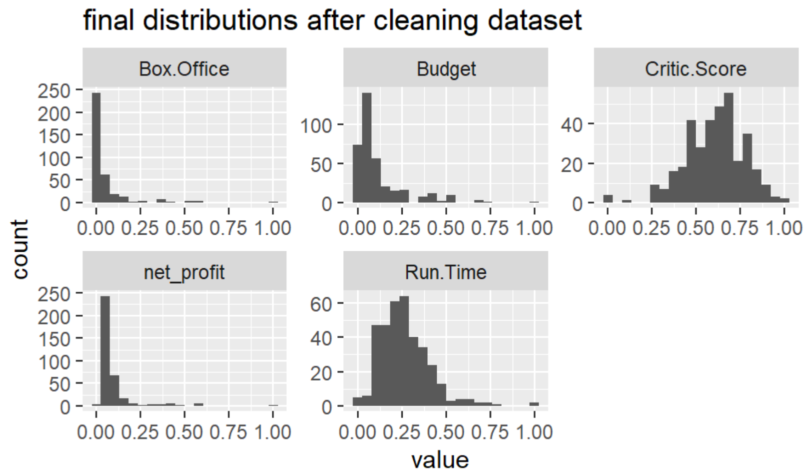
From Business perspective, both of the revenue of a film and budget are determined by running time, IDMB score, and genre. In order to take thoughtful consideration of all variance and to be more accurate for imputation, here I would like to do linear regression imputation for `Box.Office` and `Budget`. This way would be more powerful than simpler imputation methods such as mean imputation or zero substitution and here it will be imputed by Robust Linear Regression through M-estimation, in which the minimization of the square of residuals replaced with an alternative convex function of the residuals that decrease the influence of outliers and imputes the multiple variables simultaneously. Moreover, this method would be employed numerical and/or categorical predictors. However, `Genre` is more individually characteristic and will not be considered as a predictor this time.
Before imputation, we should rescale and normalize predictors. From the original distribution after taking log below, There are obvious outliers affect the overall distributions, and we will drop it at this time, afterwards, the variables are more normalized.
After cleaning and reorganize dataset, we will double check the final distribution for variables. The plots below confirm that the NA imputation works well and does not affect the overall original distribution at all, and there are no NA values any more, we are good to go.





`Genre` is a significant variables for revenue and will be considered as a group to investigate the relationship between net profits and budget as well as net profits and critical score across different `Genre`.

The dataset now is ready to use, and `net_profit`, the response variable should be calculated by inverting log transformation back and then taking log again after calculation. However, we can't guarantee all net profit is positive, in this case, rescaling by log transformation doesn't work. Therefore, I would do normalization for numerical variables to standardize the values between 0 and 1 to reduce bias and variance. The final dataset is below.
The final distributions below demonstrate they seems like normalized and concentrated, although the data is kind of skewed.

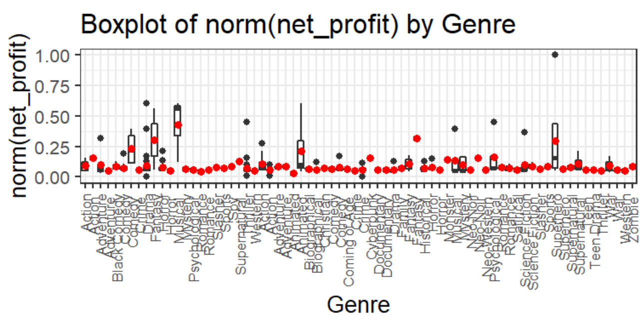


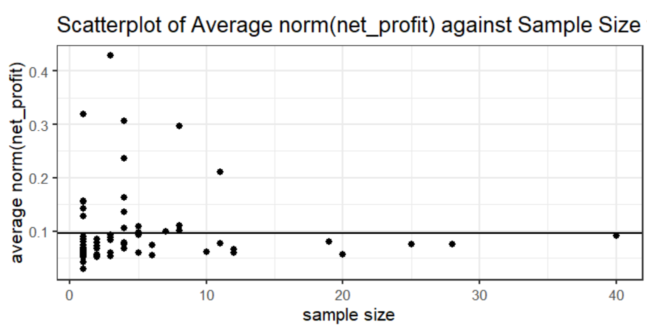
Relationships between response variable and other predictors

net_profit vs Genre

Key observations:

1. Checked all the observations that are in the same `Genre` , observed differences across Genres and motivated potentially modeling random intercepts.
2. The second plot observes the typical hierarchical data pattern and tells that larger sample sizes in Genres will have means closer to the grand mean, which means that we could consider using `Genre` as grouping variable.

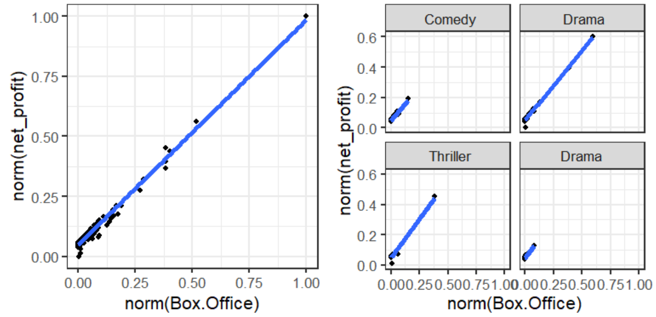




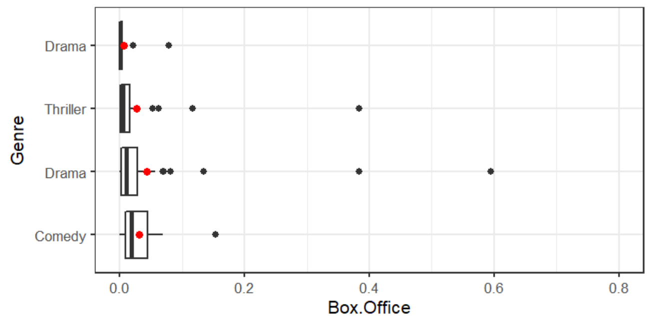
net_profit vs Box.Office & Box.Office vs Genres

Key observations:

1. The graph on the left plots `Box.Office` vs `net_profit` for all the observations. We could see that there is a significantly obvious positive association between `Box.Office` and `net_profit` whether in all or subgroups (here the top 4 gourps in Genre are selected). This means that `Box.Office` should be considered as main effect in this hierarchical model.
2. We could also observe clear differences in average `Box.Office` across different `Genres` types which motivates using `Box.Office` as one of the main effects.



Box.Office vs net_profit

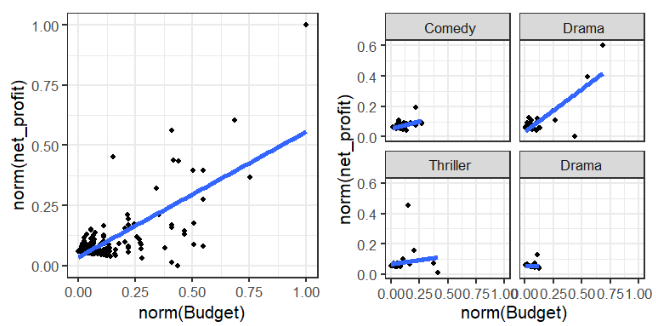


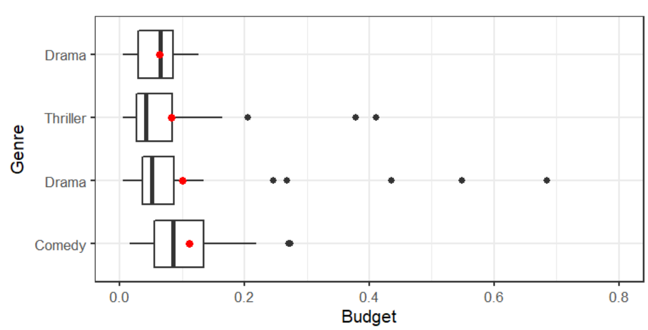
Box.Office vs Genre

net_profit vs Budget & Budget vs Genres

Key observations:

1. Similar as the `Box.Office` observations above, the graph on the left plots `Budget` vs `net_profit` for all the observations. We could see that there is a positive association between `Budget` and `net_profit` in all, however, I suspect that it is affected by outliers and most of points are concentrated around the beginning. Moreover, from the observations of the top 4 Genres, obvious associations are not across all Genres rather some of them. In this case, I would say that there are just potentially different trend across Genres and this main effect should be considered as potential candidate.
2. We could also observe potential differences in average `Budget` across different `Genres` types which means `Budget` could be considered as one of the main effects.



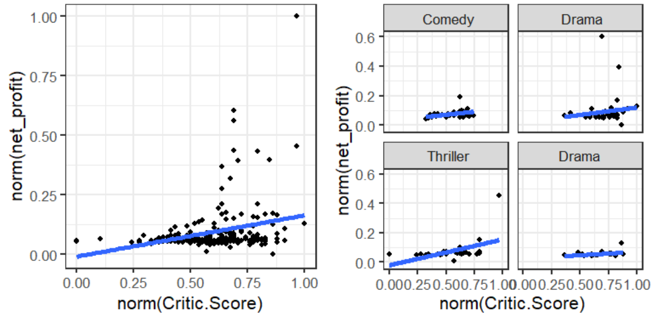


Budget vs Genre

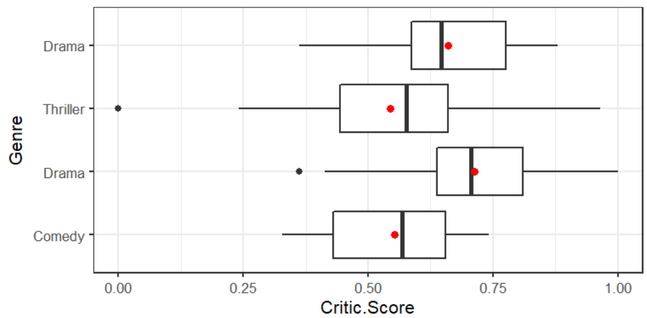
net_profit vs Critic.Score & Critic.Score vs Genres

Key observations:

1. Similar as the observations above, the graph on the left plots `Critic.Score` vs `net_profit` for all the observations. We could see that there is a obvious positive association between `Critic.Score` and `net_profit` in all, even though we have weak correlation in just few Genres according to the right plots. Thus, `Critic.Score` could be considered as main effect.
2. We could also observe clear differences in average `Critic.Score` across different `Genres` types which motivates using `Critic.Score` as one of the main effects.



Critic.Score vs net_profit

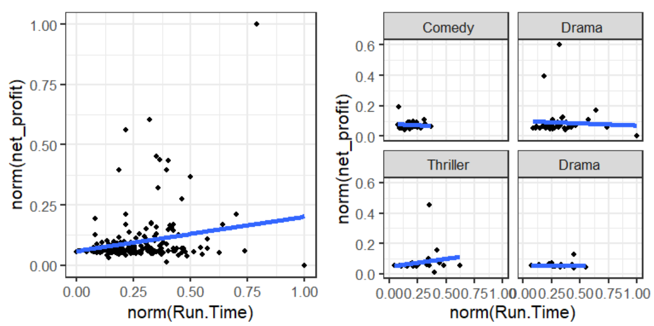


Critic.Score vs Genre

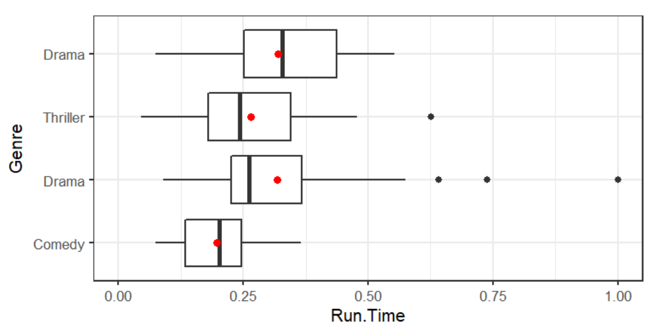
net_profit vs Run.Time & Run.Time vs Genres

Key observations:

1. `Run.Time` observation is really similar as `Critic.Score`. The graph on the left plots `Run.Time` vs `net_profit` for all the observations. We could see that there is a obvious positive association between `Run.Time` and `net_profit` in all, even though we have weak correlation in just few Genres according to the right plots. Thus, `Run.Time` could be considered as main effect.
2. We could also observe clear differences in average `Run.Time` across different `Genres` types which motivates using `Run.Time` as one of the main effects.



Run.Time vs net_profit



Run.Time vs Genre

EDA summerized

1. Imputed missing values by Robust Linear Regression through M-estimation, in which the minimization of the square of residuals replaced with an alternative convex function of the residuals that decrease the influence of outliers and imputes the multiple variables simultaneously. The reasonable imputation also be confirmed by similar distribution before and after imputation
2. Removing outliers and taking normalization transformation to all variables to reduce bias and variance
3. Reasonable to use `Genre` as grouping variable and model random intercepts across the groups.
4. `Box.Office`, `Critic.Score` and `Run.Time` should be determined as main effects in the hierarchical models, and `Budget` will be tested in the following models. We may also test if `Budget` could be modeled as random slope.

Model Selection

The derivation of BIC under the Bayesian probability framework means that if a selection of candidate models includes a true model for the dataset, then the probability that BIC will select the true model increases with the size of the training dataset. This cannot be said for the AIC score. Morevoer, unlike the AIC, the BIC penalizes the model more for its complexity, meaning that more complex models will have a worse (larger) score and will, in turn, be less likely to be selected. (Resource: https://machinelearningmastery.com/probabilistic-model-selection-measures/ (https://machinelearningmastery.com/probabilistic-model-selection-measures/))
Both lower AIC or BIC value indicates a better fit, however, in this case, BIC is better to be considered for model selection to control model complexity and overfitting as well we focusing more on finding TRUE model among the set of candidates.

Test for Genre groups and random effect acorss Genres

null model: model1 = lm(net_profit ~ 1, data=df)
single ANOVA model for Genre: model2 = lm(net_profit ~ Genre, data=df)
random effects Anova for Genre: model3 = lmer(net_profit ~ (1|Genre), data=df)
The negative value of AIC and BIC doesn't affect the selection. Here model3 has lower AIC and BIC, which means that model3, random effect across Genres has a better performance.
(Resource: https://www.statology.org/negative-aic/ (https://www.statology.org/negative-aic/))

```
## Analysis of Variance Table
##
## Model 1: net_profit ~ 1
## Model 2: net_profit ~ Genre
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    358 3.9839
## 2    287 2.6515 71    1.3324 2.0313 2.367e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Data: df
## Models:
## model3: net_profit ~ (1 | Genre)
## model2: net_profit ~ Genre
##        npar     AIC     BIC logLik deviance  Chisq Df Pr(>Chisq)
## model3    3 -618.15 -606.50 312.08  -624.15
## model2   73 -597.24 -313.76 371.62  -743.24 119.09 70  0.0002307 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Test for main effects: `Box.Office`, `Critic.Score` and `Run.Time`

Main Effects model: lmer(net_profit ~ Box.Office + Critic.Score + Run.Time + (1|Genre), data=df)
Below analysis, we could see that the choice of these 3 variables main effects improves the model performance significantly.

```
## Data: df
## Models:
## model3: net_profit ~ (1 | Genre)
## model4: net_profit ~ Box.Office + Critic.Score + Run.Time + (1 | Genre)
##        npar      AIC     BIC  logLik deviance  Chisq Df Pr(>Chisq)
## model3    3  -618.15  -606.5  312.08  -624.15
## model4    6 -2177.90 -2154.6 1094.95 -2189.90 1565.8  3  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Test for `Budget`

Model for main effect of Budget:
model5 = lmer(net_profit ~ Box.Office + Critic.Score + Run.Time + Budget + (1|Genre), data=df)
Model for random slopeof Budget:
model6 = lmer(net_profit ~ Box.Office + Critic.Score + Run.Time + Budget + (Budget|Genre), data=df)

Although adding Budget seems like better compared by AIC and BIC in this test, models are failed to converge and have risk of unidentified. In this case, I would like to remove `Budget`, and conclude that `Budget` is not predictive of film' net profits in 2019.

Interactions

Interaction model:
model7 = lmer(net_profit ~ Box.Office * Critic.Score * Run.Time + (1|Genre), data=df)

Now we have determined the fixed effect and random effect across potential models, we could further look at the potential interactions across main effects. Summary shows that they have similar performance whether adding interactions nor not. However, I highly suspect that the model would be overfitted if we add many interactions, so I won't consider interactions at this time.

```
## Data: df
## Models:
## model4: net_profit ~ Box.Office + Critic.Score + Run.Time + (1 | Genre)
## model7: net_profit ~ Box.Office * Critic.Score * Run.Time + (1 | Genre)
##        npar     AIC     BIC logLik deviance  Chisq Df Pr(>Chisq)
## model4    6 -2177.9 -2154.6 1095.0  -2189.9
## model7   10 -2190.6 -2151.8 1105.3  -2210.6 20.705  4  0.0003623 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Final Model

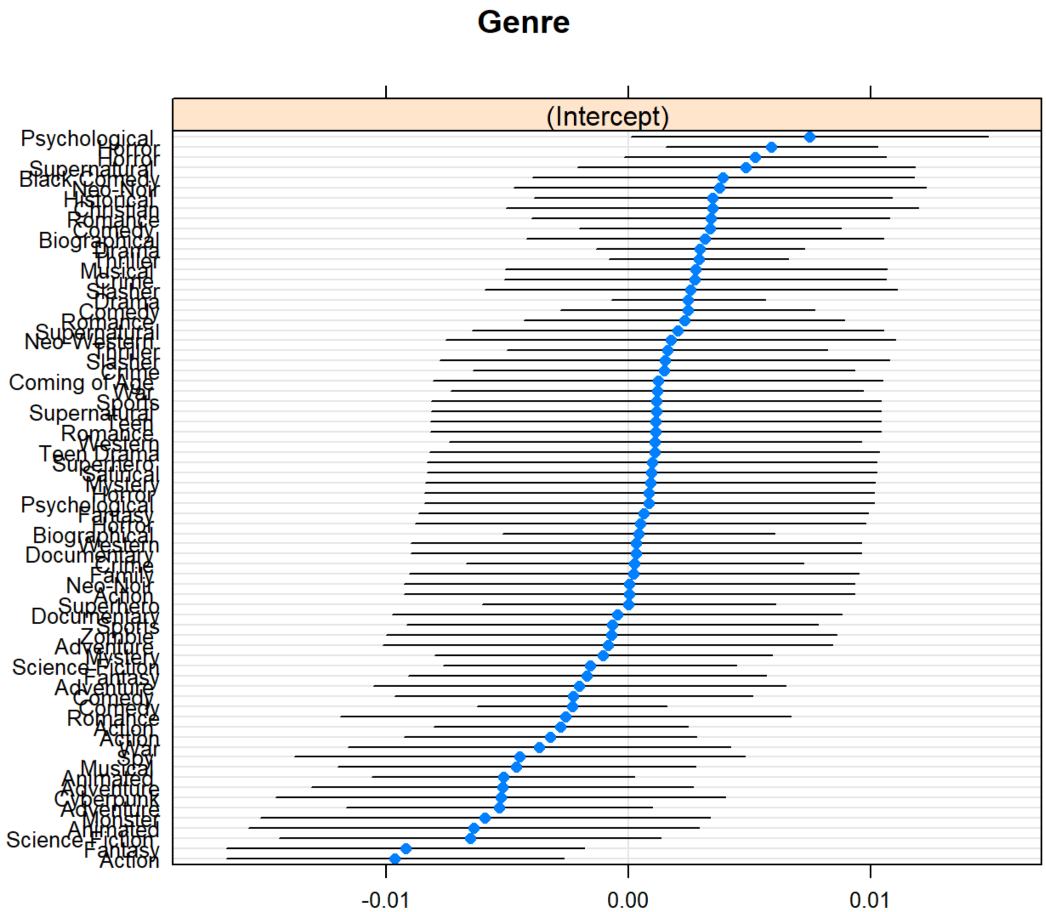
Final model selected:
net_profit ~ Box.Office + Critic.Score + Run.Time + (1|Genre)
Fixed Effects: `Box.Office` , `Critic.Score` and `Run.Time`
Random Effects: Random intercepts across `Genre`

It would be better if we check the difference between MLE and REML. The result is showing that the estimation won't have big difference depending on REML.
REML model:
model8 = lmer(net_profit ~ Box.Office + Critic.Score + Run.Time + (1|Genre), data=df, REML=FALSE)

```
## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: net_profit ~ Box.Office + Critic.Score + Run.Time + (1 | Genre)
##    Data: df
## 
##      AIC      BIC   logLik deviance df.resid
##  -2177.9  -2154.6   1095.0  -2189.9      353
## 
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -4.9051 -0.2460  0.1208  0.5039  3.4107
## 
## Random effects:
##  Groups   Name        Variance  Std.Dev.
##  Genre    (Intercept) 2.794e-05 0.005286
##  Residual             1.158e-04 0.010759
## Number of obs: 359, groups:  Genre, 72
## 
## Fixed effects:
##               Estimate Std. Error t value
## (Intercept)   0.049494   0.002490  19.879
## Box.Office    0.947477   0.006013 157.560
## Critic.Score  0.012441   0.004362   2.852
## Run.Time     -0.034205   0.004845  -7.060
## 
## Correlation of Fixed Effects:
##             (Intr) Bx.Off Crtc.S
## Box.Office   0.173
## Critic.Scor -0.808 -0.234
## Run.Time    -0.032 -0.122 -0.432
```

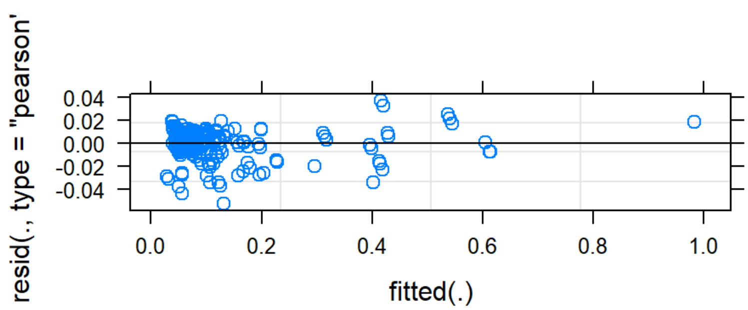
Check Uncertainty - Confidence Interval

From the plots for confidence interval below, we could see that they are relatively overlaped to each other across Genres and most of them have high variability, which means that the model is not really certain. The variance of random effect are not small, which means that the net profit is different across various genders within the fixed effects of revenue, run time, and critic score.



Check and evaluate the adequacy of model fit

From the plot below, we could see that most of residuals are around 0, which is good, however, we have some outliers should be noticed in the future analysis. Moreover, the variance are not stable, which may be against homoscedasticity assumption.



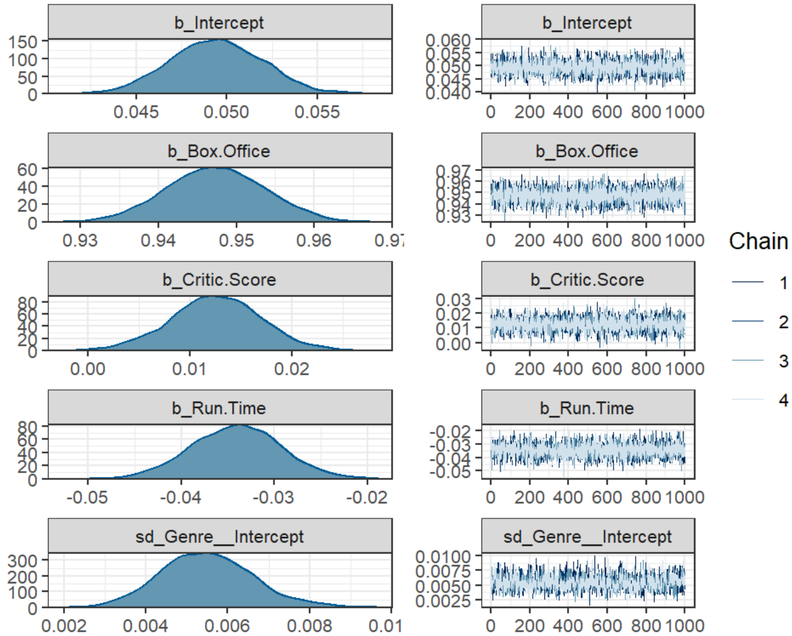
Bayesian model and Diagonostics

Here we would fit the Bayesian model based on the default priors:

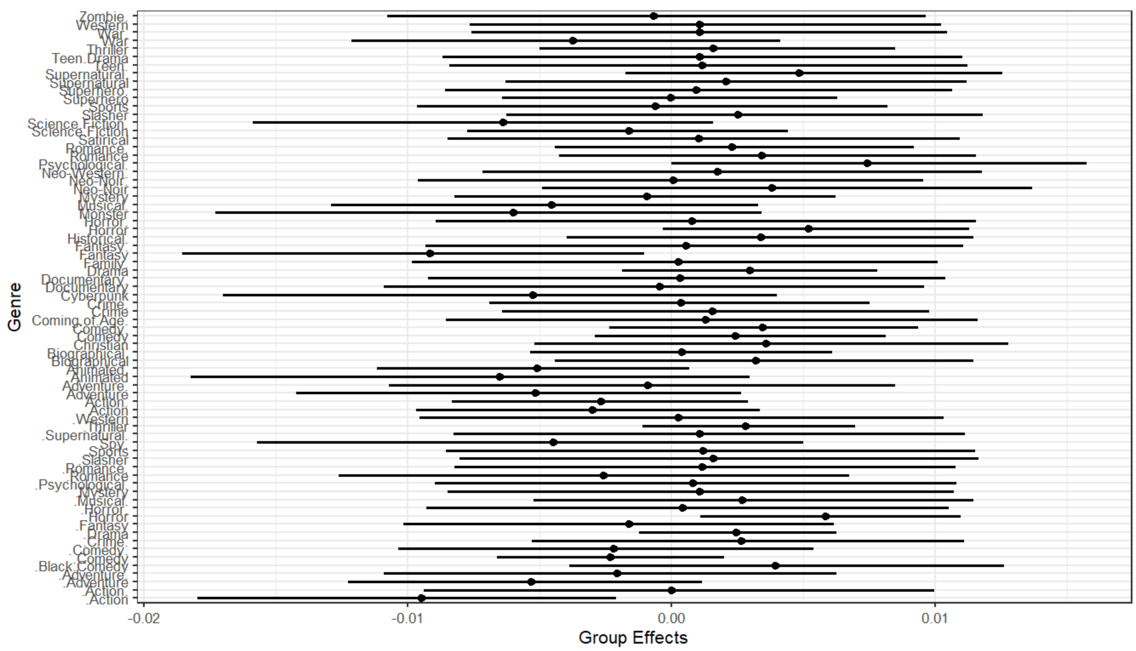
brm(net_profit ~ Box.Office + Critic.Score + Run.Time + (1|Genre), data=df, seed = 1)

The posterior distribution below is showing that the estimations center at the frequentist estimates (the model above), which proves that both of Bayesian and frequency models are good here, and we don't need adjust the informative priors and assumptions. The trace plot does not provide the evidence of non-convergence, which is also good for model results.

The plot of "Genres specific group effects on net profit" displays the corresponding 95% credical intervals for Bayesian model. The plots below also show that these confidence intervals largely overlap each other and have pretty various large range, hence the inference on these overall rankings has high variability.



Bayesian random effects ANOVA model posterior distributions and traceplots



Genres specific group effects on net profit

Summary

This case study thoughtfully analyzed the relationships between net profits and predictors across `Genres` groups. From EDA observation, ANOVA, and Beyesian model analysis, we could conclude that except for the weak relationships between `Budget` and `Net_profit` across `Genres`, other variables has correlations with `Net_profit` in different genres groups such as `Box.Office`, `Critic.Score` and `Run.Time`. After evaluating models' performance, the best model is fixed effect with `Box.Office`, `Critic.Score` and `Run.Time` as well as `Genre` random intercept effects. From both frequentist and Bayesian perspectives, the `net_profit` across all groups has high variability and indicates the model is not really stable. From the adequacy view of frequentist model, some outliers are noticed and should be more careful in the future analysis and unstable variance tend to be against homoscedasticity assumption. Another limitation in this model: since most of `company`, `lead.Cast 1-3`, and `Director` are only for one specific movie, they are individual characterized, and won't significantly affect across the group, we didn't consider these variables at this time, moreover, although `Title` and `Release Date` do not influence the revenue based on the life experience, we could test them if they would facilitate our model.

Appendix

```
## ----setup, include=FALSE---------------------------
knitr::opts_chunk$set(echo=F, eval=T, cache=F, warning=F, message=F,
fig.align="center", fig.pos="H")
library(lme4)
library(tidyverse)
library(tidybayes)
library(ggplot2)
library(ggpubr)
library(knitr) # for kable
library(lattice) # dotplot
library(brms) # for Bayesian
options(mc.cores = parallel::detectCores())


## ------------------------------------------------------
df = read.csv("United_States_Film_Releases_2019.csv")


## ------------------------------------------------------
# change $ sign to numeric
df$Box.Office = as.numeric(gsub('[$,]','',df$Box.Office))
df$Budget = as.numeric(gsub('[$,]', '', df$Budget))

# summary statistics
library(skimr)
summary = skim(df)


## ------------------------------------------------------
# drop both are NA
df = df[-which(is.na(df$Box.Office)&is.na(df$Budget)),]

# drop meaningless columns
df = subset(df, select=-c(Title, Release.Date..mmddyyyy., Production.Company,
Lead.Cast.1, Lead.Cast.2, Lead.Cast.3, Director))


## ------------------------------------------------------
# take natural log
df[, c(1,2,3,4)] = log(df[, c(1,2,3,4)])
# original distributions after log
plots = df %>%
  keep(is.numeric) %>%
  gather() %>%
  ggplot(aes(value)) +
    facet_wrap(~ key, scales = "free") +
    geom_histogram() +
    labs(title = "distributions before removing outliers")


## ---- fig.height=2, fig.width=4---------------------
quartiles = quantile(df$Box.Office, probs=c(.25, .75), na.rm = TRUE)
IQR = IQR(df$Box.Office,na.rm = TRUE)
Lower = quartiles[1] - 1.5*IQR
```

```
Upper = quartiles[2] + 1.5*IQR
data_no_outlier <- subset(df, df$Box.Office > Lower & df$Box.Office < Upper)

quartiles = quantile(df$Budget, probs=c(.25, .75), na.rm = TRUE)
IQR = IQR(df$Budget,na.rm = TRUE)
Lower = quartiles[1] - 1.5*IQR
Upper = quartiles[2] + 1.5*IQR
data_no_outlier <- subset(df, df$Budget > Lower & df$Budget < Upper)

quartiles = quantile(df$Run.Time..minutes., probs=c(.25, .75), na.rm = TRUE)
IQR = IQR(df$Run.Time..minutes.,na.rm = TRUE)
Lower = quartiles[1] - 1.5*IQR
Upper = quartiles[2] + 1.5*IQR
data_no_outlier <- subset(df,df$Run.Time..minutes. > Lower &
df$Run.Time..minutes. < Upper)

quartiles = quantile(df$Critic.Score..IMDB.x.10., probs=c(.25, .75), na.rm =
TRUE)
IQR = IQR(df$Critic.Score..IMDB.x.10.,na.rm = TRUE)
Lower = quartiles[1] - 1.5*IQR
Upper = quartiles[2] + 1.5*IQR
data_no_outlier <- subset(df, df$Critic.Score..IMDB.x.10. > Lower &
df$Critic.Score..IMDB.x.10. < Upper)

# distributions after outliers
df %>%
  keep(is.numeric) %>%
  gather() %>%
  ggplot(aes(value)) +
    facet_wrap(~ key, scales = "free") +
    geom_histogram(bins=20)+
    labs(title = "distributions after removing outliers")


## ------------------------------------------------------
library(simputation)
# impute by linear regression
df = impute_rlm(df, Box.Office + Budget~Run.Time..minutes.
+Critic.Score..IMDB.x.10.)
colnames(df) = c('Box.Office', 'Budget', 'Run.Time', 'Critic.Score', 'Genre')


## ---- fig.height=2, fig.width=4---------------------
# check the sum of missing value and good to go
# sum(is.na(df))

# check the distribution after imputing missing value
df %>%
  keep(is.numeric) %>%
  gather() %>%
  ggplot(aes(value)) +
    facet_wrap(~ key, scales = "free") +
    geom_histogram(bins=20) +
    labs(title = "distributions after imputing missing value")
```

```
## ----------------------------------------------------
# seperate Genre to multiple Geners
df[c('Genre_1', 'Genre_2', 'Genre_3', 'Genre_4')] = str_split_fixed(df$Genre,
'/', 4)
df = df[,-5] %>%
  pivot_longer(cols = starts_with("Genre_"), names_to = "Genre_", values_to =
"Genre") %>%
  select(-Genre_)
df[df==""]=NA
df = na.omit(df)


## ----------------------------------------------------
# log convert back
df[, c(1,2,3,4)] = exp(df[, c(1,2,3,4)])
# calculate net_profit and normalized
df = df %>%
  mutate(net_profit = Box.Office-Budget) %>%
  mutate_if(is.numeric, funs((.-min(.))/max(.-min(.))))


## ---- fig.height=3, fig.width=5---------------------
# distributions for final
df %>%
  keep(is.numeric) %>%
  gather() %>%
  ggplot(aes(value)) +
    facet_wrap(~ key, scales = "free") +
    geom_histogram(bins=20) +
    labs(title = "final distributions after cleaning dataset")


## ---- fig.height=2, fig.width=4---------------------
df %>%
  group_by(Genre) %>%
  ggplot(aes(x=Genre, y=net_profit)) +
  geom_boxplot(outlier.size = 1) +
  stat_summary(fun=mean, geom="point", shape=20, size=2, color="red",
fill="red") +
  theme_bw(base_size = 10) +
  labs(x="Genre", y="norm(net_profit)") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1, size=6))+
  ggtitle("Boxplot of norm(net_profit) by Genre")

df %>%
  group_by(Genre) %>%
  summarise(mean=mean(net_profit), count=n()) %>%
  ggplot(aes(x=count, y=mean)) +
  labs(x="sample size", y="average norm(net_profit)") +
  geom_point(size=1) +
  theme_bw(base_size = 8) +
  geom_hline(yintercept=mean(df$net_profit)) +
  ggtitle("Scatterplot of Average norm(net_profit) against Sample Size for
Genre")
```

```
## ----------------------------------------------------
p1 = df %>%
  ggplot(aes(x=Box.Office, y=net_profit)) +
  geom_point(size=0.8) +
  geom_smooth(se=F, method = "lm") +
  labs(x="norm(Box.Office)", y="norm(net_profit)") +
  xlim(c(0,1)) +
  theme_bw(base_size = 8)

# take top4 Genre to compare
top4 = df %>%
  group_by(Genre) %>%
  summarise(count=n()) %>%
  arrange(desc(count)) %>%
  slice(1:4)

p2 = df %>%
  filter(Genre %in% top4$Genre) %>%
  ggplot(aes(x=Box.Office, y=net_profit)) +
  geom_point(size=0.8) +
  geom_smooth(se=F, method="lm") +
  facet_wrap(~Genre) +
  xlim(c(0,1)) +
  theme_bw(base_size = 8) +
  labs(x="norm(Box.Office)", y="norm(net_profit)")


## ---- fig.height=2, fig.width=4,fig.cap="Box.Office vs net_profit"----
ggarrange(p1, p2, nrow=1)


## ----fig.cap="Box.Office vs Genre", fig.height=2, fig.width=4----
df %>%
  filter(Genre %in% top4$Genre) %>%
  group_by(Genre) %>%
  ggplot(aes(x=Genre, y=Box.Office)) +
  geom_boxplot(outlier.size = 1) +
  ylim(c(0,0.8)) +
  stat_summary(fun.y=mean, geom="point", shape=20, size=2, color="red",
fill="red") +
  coord_flip() +
  theme_bw(base_size = 8)


## ----------------------------------------------------
p1 = df %>%
  ggplot(aes(x=Budget, y=net_profit)) +
  geom_point(size=0.8) +
  geom_smooth(se=F, method = "lm") +
  labs(x="norm(Budget)", y="norm(net_profit)") +
  xlim(c(0,1)) +
  theme_bw(base_size = 8)

p2 = df %>%
  filter(Genre %in% top4$Genre) %>%
```

```
  ggplot(aes(x=Budget, y=net_profit)) +
  geom_point(size=0.8) +
  geom_smooth(se=F, method="lm") +
  facet_wrap(~Genre) +
  xlim(c(0,1)) +
  theme_bw(base_size = 8) +
  labs(x="norm(Budget)", y="norm(net_profit)")


## ---- fig.height=2, fig.width=4---------------------
ggarrange(p1, p2, nrow=1)


## ---- fig.height=2, fig.width=4,fig.cap="Budget vs Genre"----
df %>%
  filter(Genre %in% top4$Genre) %>%
  group_by(Genre) %>%
  ggplot(aes(x=Genre, y=Budget)) +
  geom_boxplot(outlier.size = 1) +
  ylim(c(0,0.8)) +
  stat_summary(fun.y=mean, geom="point", shape=20, size=2, color="red",
fill="red") +
  coord_flip() +
  theme_bw(base_size = 8)


## ------------------------------------------------------
p1 = df %>%
  ggplot(aes(x=Critic.Score, y=net_profit)) +
  geom_point(size=0.8) +
  geom_smooth(se=F, method = "lm") +
  labs(x="norm(Critic.Score)", y="norm(net_profit)") +
  xlim(c(0,1)) +
  theme_bw(base_size = 8)

p2 = df %>%
  filter(Genre %in% top4$Genre) %>%
  ggplot(aes(x=Critic.Score, y=net_profit)) +
  geom_point(size=0.8) +
  geom_smooth(se=F, method="lm") +
  facet_wrap(~Genre) +
  xlim(c(0,1)) +
  theme_bw(base_size = 8) +
  labs(x="norm(Critic.Score)", y="norm(net_profit)")


## ---- fig.height=2, fig.width=4, fig.cap="Critic.Score vs net_profit"----
ggarrange(p1, p2, nrow=1)


## ----fig.cap="Critic.Score vs Genre", fig.height=2, fig.width=4----
df %>%
  filter(Genre %in% top4$Genre) %>%
  group_by(Genre) %>%
  ggplot(aes(x=Genre, y=Critic.Score)) +
  geom_boxplot(outlier.size = 1) +
```

```
  ylim(c(0,1)) +
  stat_summary(fun.y=mean, geom="point", shape=20, size=2, color="red",
fill="red") +
  coord_flip() +
  theme_bw(base_size = 8)


## ----------------------------------------------------
p1 = df %>%
  ggplot(aes(x=Run.Time, y=net_profit)) +
  geom_point(size=0.8) +
  geom_smooth(se=F, method = "lm") +
  labs(x="norm(Run.Time)", y="norm(net_profit)") +
  xlim(c(0,1)) +
  theme_bw(base_size = 8)

p2 = df %>%
  filter(Genre %in% top4$Genre) %>%
  ggplot(aes(x=Run.Time, y=net_profit)) +
  geom_point(size=0.8) +
  geom_smooth(se=F, method="lm") +
  facet_wrap(~Genre) +
  xlim(c(0,1)) +
  theme_bw(base_size = 8) +
  labs(x="norm(Run.Time)", y="norm(net_profit)")


## ---- fig.height=2, fig.width=4, fig.cap="Run.Time vs net_profit"----
ggarrange(p1, p2, nrow=1)


## ----fig.cap="Run.Time vs Genre",fig.height=2, fig.width=4----
df %>%
  filter(Genre %in% top4$Genre) %>%
  group_by(Genre) %>%
  ggplot(aes(x=Genre, y=Run.Time)) +
  geom_boxplot(outlier.size = 1) +
  ylim(c(0,1)) +
  stat_summary(fun.y=mean, geom="point", shape=20, size=2, color="red",
fill="red") +
  coord_flip() +
  theme_bw(base_size = 8)


## ---- fig.height=0.1, fig.width=0.1-----------------
# null model
model1 = lm(net_profit ~ 1, data=df)
# single ANOVA model for Genre
model2 = lm(net_profit ~ Genre, data=df)
# random effects Anova for Genre
model3 = lmer(net_profit ~ (1|Genre), data=df)

anova(model1, model2)
anova(model3, model2)
```

```
## ---- fig.height=0.1, fig.width=0.1-----------------
model4 = lmer(net_profit ~ Box.Office + Critic.Score + Run.Time + (1|Genre),
data=df)

anova(model3, model4)


## ---- warning = FALSE, include=FALSE----------------
# main effect for Budget
model5 = lmer(net_profit ~ Box.Office + Critic.Score + Run.Time + Budget + (1|
Genre), data=df)
# random slope for Budget
model6 = lmer(net_profit ~ Box.Office + Critic.Score + Run.Time + Budget +
(Budget|Genre), data=df)

anova(model4, model5)
anova(model4, model6)
anova(model5, model6)

summary(model5)
summary(model6)


## ---- warning = FALSE, fig.height=0.1, fig.width=0.1----
model7 = lmer(net_profit ~ Box.Office * Critic.Score * Run.Time + (1|Genre),
data=df)
anova(model4, model7)


## ---- include=FALSE---------------------------------
summary(model4)


## ---- fig.height=0.001, fig.width=0.001-------------
model8 = lmer(net_profit ~ Box.Office + Critic.Score + Run.Time + (1|Genre),
data=df, REML=FALSE)
summary(model8)


## ----fig.height=6,fig.width=7, warning=F------------
dotplot(ranef(model8, condVar=TRUE))$Genre


## ----fig.height=2,fig.width=5-----------------------
plot(model8)


## ---- message=F, warning=F--------------------------
# default priors
model_bay = brm(net_profit ~ Box.Office + Critic.Score + Run.Time + (1|Genre),
data=df, seed = 1)


## --------------------------------------------------
# get variables: https://cran.r-project.org/web/packages/tidybayes/vignettes/
tidy-brms.html
```

```
#library(tidybayes)
#get_variables(model_bay)


## ----mcmc-plot, fig.height=4, fig.width=5, fig.cap="Bayesian random effects
ANOVA model posterior distributions and
traceplots",include=TRUE,echo=FALSE----
plot(model_bay, variable=c("b_Intercept","b_Box.Office", "b_Critic.Score",
"b_Run.Time", "sd_Genre__Intercept"),
     theme=theme_bw(base_size = 10))


## ----mod-res, fig.height=4, fig.width=7, fig.cap="Genres specific group
effects on net profit",include=TRUE,echo=FALSE----
model_bay %>%
  spread_draws(r_Genre[Genre,]) %>%
  median_qi(`Group Effects` = r_Genre) %>%
  ggplot(aes(y=Genre, x=`Group Effects`, xmin=.lower, xmax=.upper)) +
  geom_pointinterval(orientation="horizontal", size=0.8) +
  theme_bw(base_size = 8)
```