Advanced Statistical Modelling: Linear Models

Joel Cantero Priego and Ricard Meyerhofer Parra

12/10/2019

Introduction

In this assignment, we are going to use the IMDB dataset. This IMDB dataset, contains information of 940 films released between 2000 and 2016. The data has been obtained from the IMDB's webpage. The following is a list where we can see all the variables of the dataset:

Variable name	Description	Values
movietitle	Director of the given title	String
gross	Gross in dollars	Integer
budget	Budget in dollars	Integer
duration	Film duration in minutes	Integer
titleyear	The release year of the title	Integer
directorfl	Director Facebook likes	Integer
actor1fl	Actor 1 Facebook likes	Integer
actor2fl	Actor 2 Facebook likes	Integer
actor3fl	Actor 3 Facebook likes	Integer
castfl	Cast Facebook likes	Integer
facenumber_in_poster	Number of faces that appears in the poster	Integer
genre	Genre film	Action/Comedy/Drama/Terror

As we can see we have that all our variables are numerical in exception genre. This dataset is complete which means that it has no missing values. However, this does not imply that there are no outliers. Because this dataset is a simplification from an IMDB dataset we are not going to perform inputations on we might think are outliers. Furthermore, we have to take into consideration that this dataset has a very wide variety of movies so the spectrum of values can be really different.

As required in the assignment, we are going to create a categorical variable: **yearcat** which is the categorical substitution of titleyear with 3 levels: 2000-2005, 2006-2010 and 2011-2016. Therefore, we will have two categorical variables (genre and titleyear).

```
dataset$yearcat<-as.factor(cut(dataset$titleyear,
                      c(2000,2005,2010,2016),
                      include.lowest = TRUE,
                      labels=c("2000-2005", "2006-2010", "2011-2016")))
```

Exploratory Data Analysis

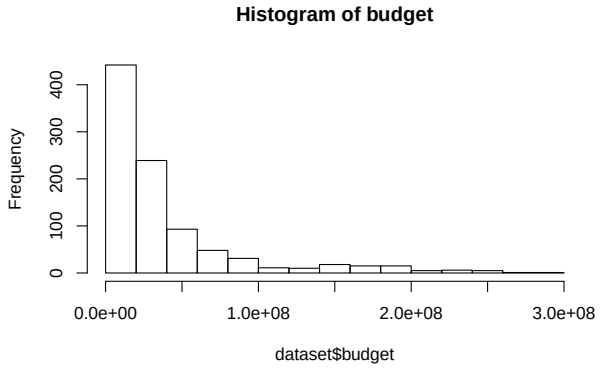
In this section we are going to focus in explaining the most interesting conclusions of our data, perform an univariate and multivariate analysis of the variables in order to find outliers and see how each of these variables is related with the gross. We are also going to modify some variables in order to make the linear model perform better on them.

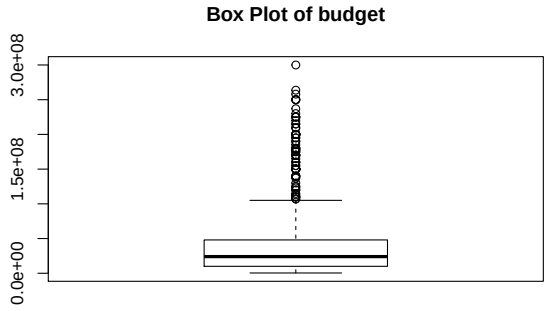
Movie title

We have performed a cloud of the most relevant words that appear in the movies. To do so, we have removed stopwords and punctuation. We could have also done a stemming process but is not so important for us to do so. We can see that the top words are words such as: man, love, movie, house, american, life, big, etc.

Budget

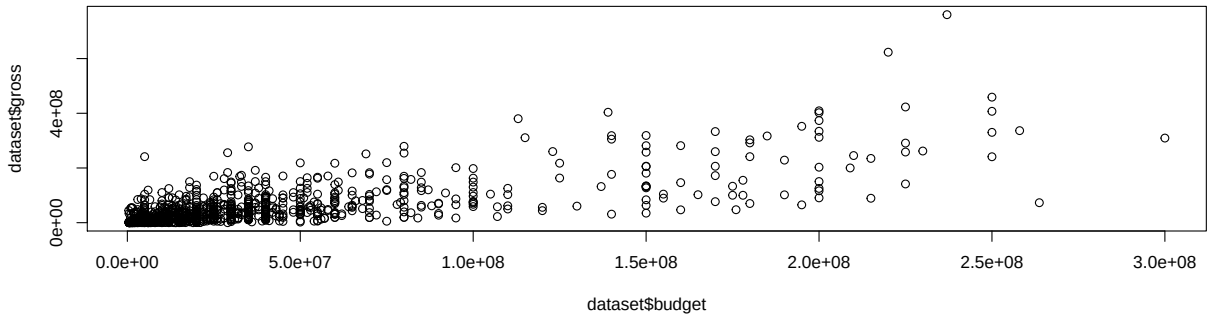
We can see that there is a very disperse amount of values regarding the budget that range from a minimum of 400 thousand dollars (Napoleon Dynamite) to 300 million dollars (Pirates of the Caribbean: At World's End). Despite how crazy this numbers can appear to be, we have revised them by looking at the budget of this two movies on the internet and are correct. Note that this does not imply that all the budgets we have are corrects but it implies that we have to deal with such a range of different values in a same variable.





Relationship between budget and gross

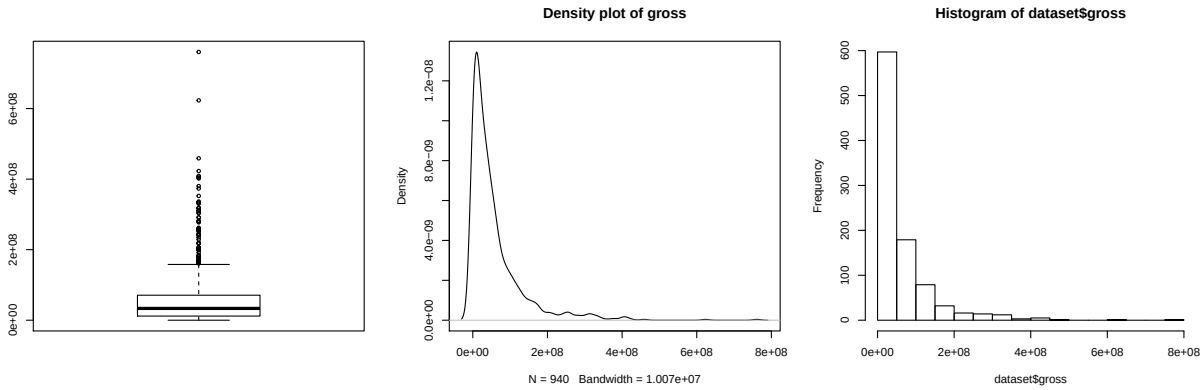
We can see that the revenue is correlated with the budget as we can see in the plot below.



Gross

If we take a look at the gross variable, we can see that in a similar fashion than with budget, we have a range of values that can go from 3330$ with Mi America to Avatar with 760 milions of dollars.

```
par(mfrow=c(1,3))
boxplot(dataset$gross)
plot(density(dataset$gross), main="Density plot of gross")
hist(dataset$gross)
```

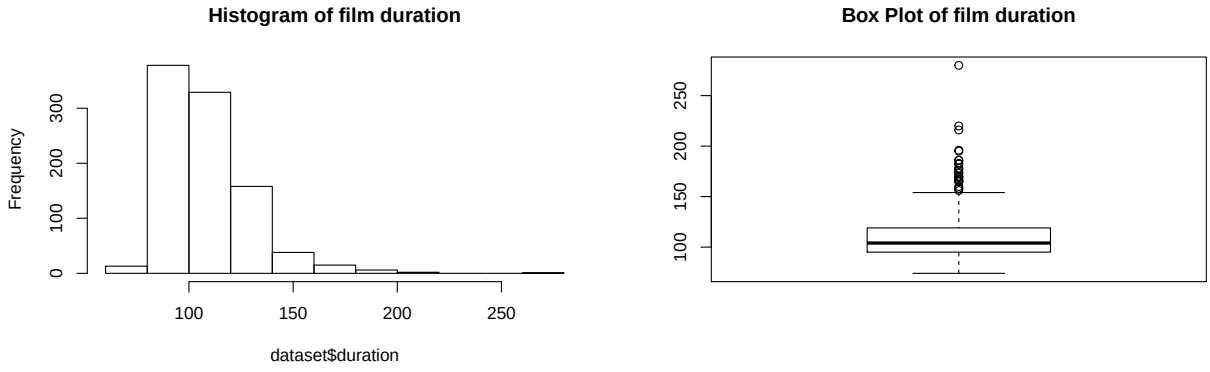


As we have just seen values from budget and gross are in a bigger scale than the rest of our data. This is a problem when performing a linear model since it adds complexity to the model. In order to avoid so, we are going to scale those variables. We decided to apply $\log_{10}$ because it is easier to interprate later when showing (insert justification).

Now we can see that even it does not follow a normal distribution completely it starts to look like one and what is more important is that the range of values is smaller for both, budget and gross.

Duration

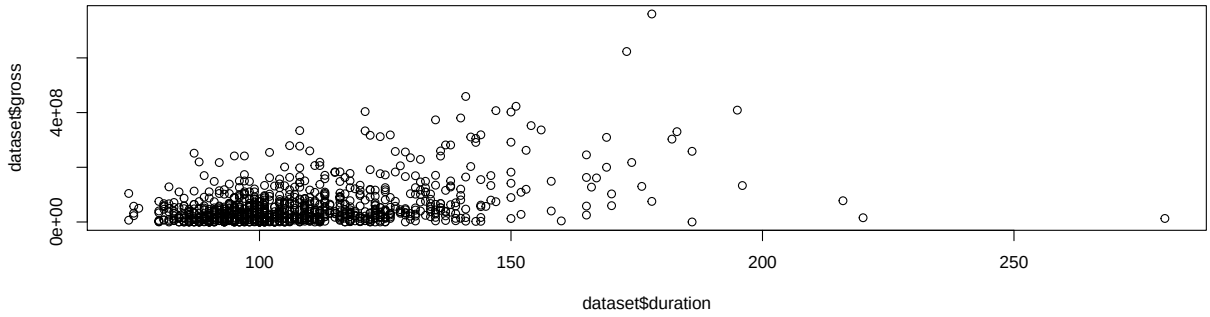
In duration film, we see that there is a certain tendency to normality centered around 100 minutes, we consider it as usual. There is a strange observation of 280 minutes for "Gods and General" film. After we check it, we can say that it is not an error but an extreme value.



Relationship between duration and gross

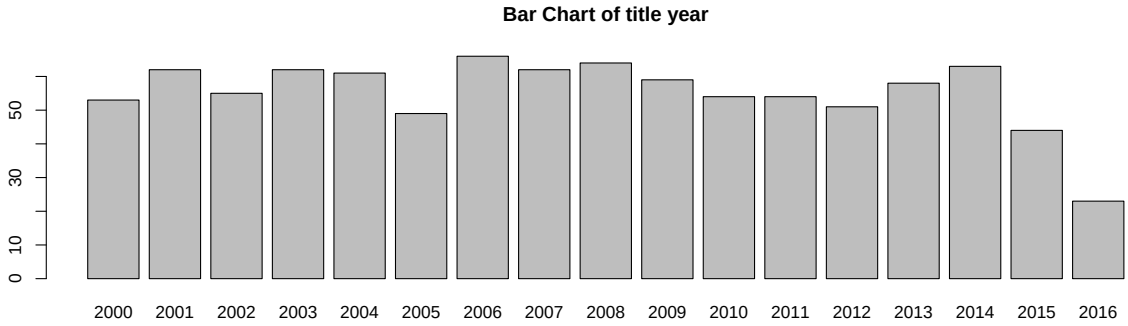
We can see that there is not a clear correlation between the duration of a film and its revenue since we can see that transversally to the duration, we have the similar results in gross. It is true that the

tendence over 150 minutes might be different but is not significant enough to take any conclusion from it.



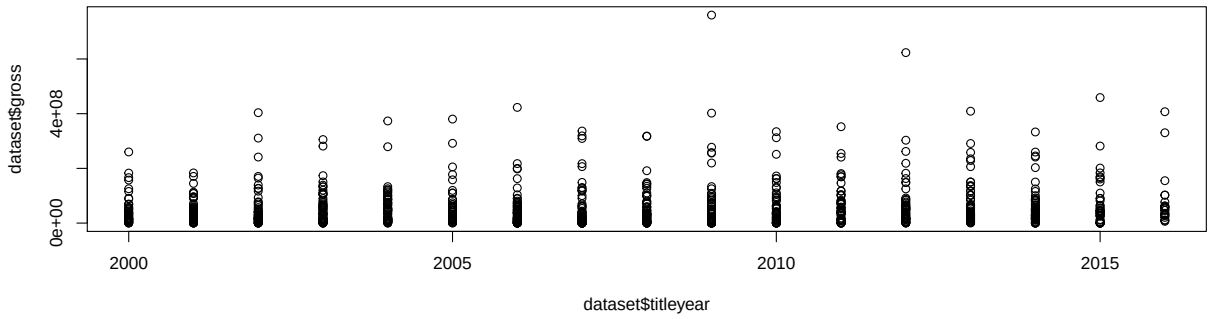
Title Year

No problems for year, there is a certain expected balanced in years proportion even that we can see a decay in the number of films for 2016.



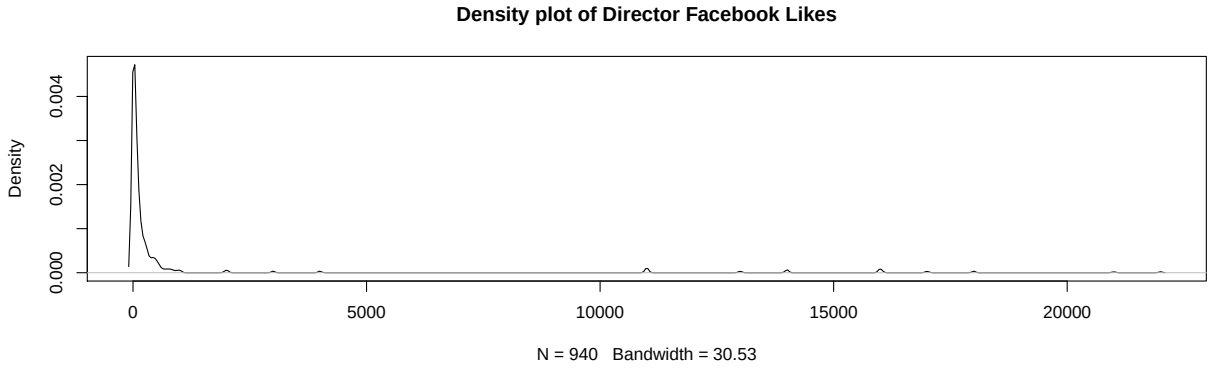
Relationship between years and gross

We can see that there is not a strong correlation between the year of release and the gross obtained from those years. Even that there are some years better than others the trend seems to be the same across the years.



Directorfl, Actor1fl, Actor2fl, Actor3fl, Castfl

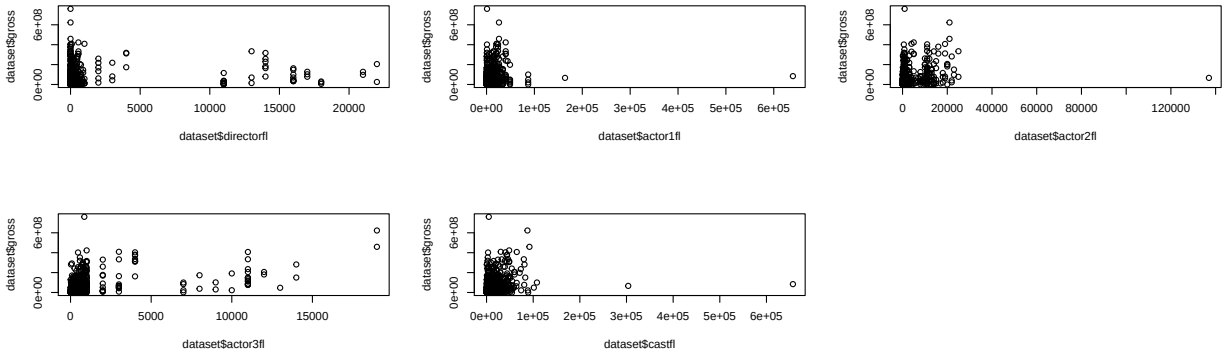
In Director Facebook likes, we see that there is a value which appears in the majority of the cases: In this case, is the 0 value. Apart from this zero value, we see that small number of likes are more common than medium or higher number of likes.



We can see that the other variables Actor1fl, Actor2fl, Actor3fl, Castfl follow a similar fashion.

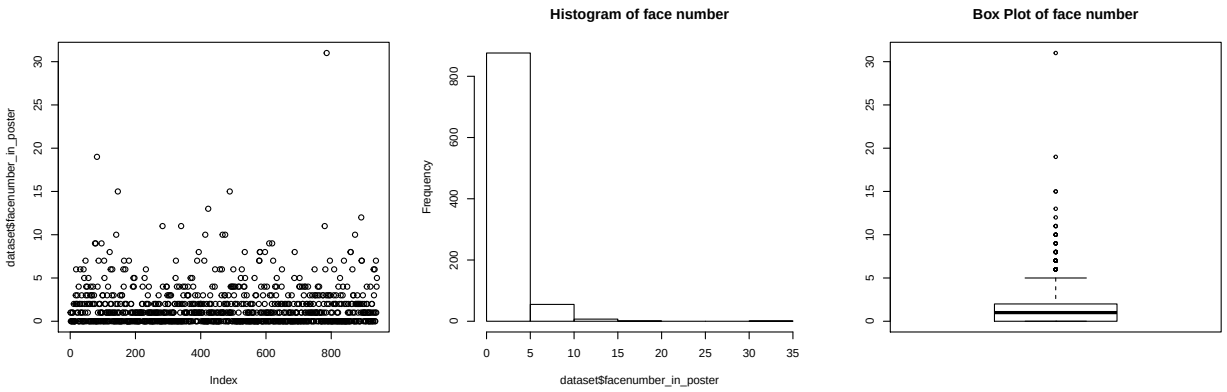
Relationship of Directorfl, Actor1fl, Actor2fl, Actor3fl, Castfl with gross

We can see a trend where the more likes Directors, Acts or Cast have are not correlated with the gross of the movie we even see cases where there is an extreme value of likes which does not project in a high revenue.



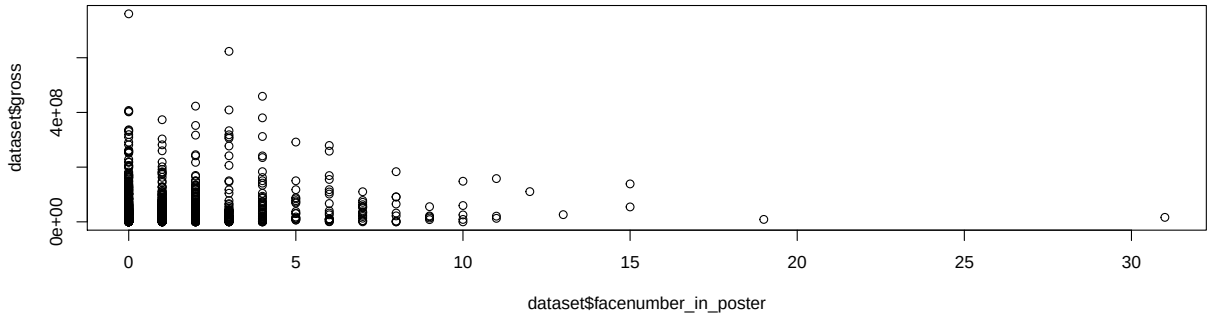
Facenumber in poster

In face number in film poster, the mean is about 1,6 faces and we can observe an extrem value of 31 in “The Master”. We can say once again that it is not an error but an extreme value.



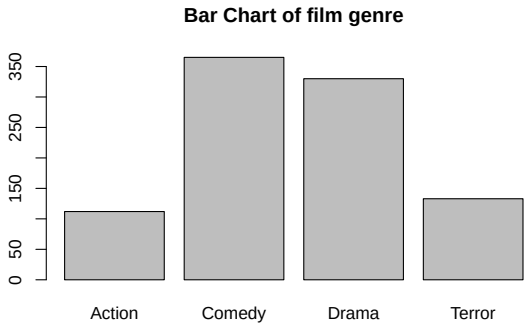
Relationship of Facenumber with Gross

We can see that the number of faces in a poster that appear seems not to be related with the revenue directly. However, we could even say that than 6-7 faces in poster does not seem to be a positive factor to have a good revenue.



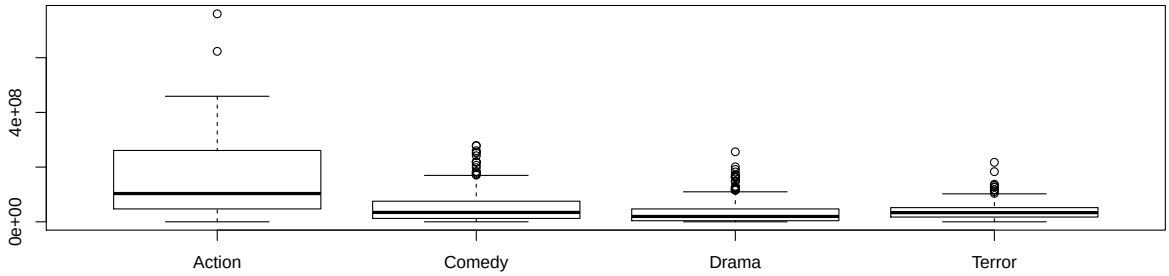
Genre

In genre film we can observe that there are more comedy and drama films than action and terror films.



Relationship of Genre with Gross

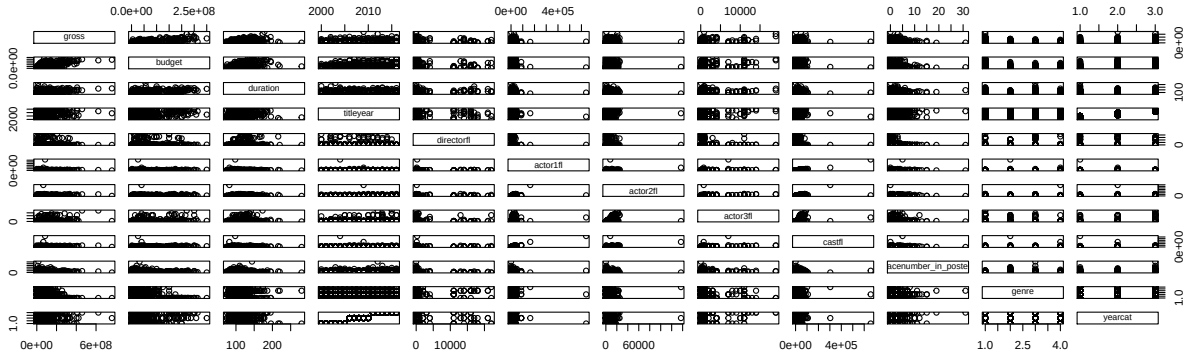
We can see that action films from the dataset tend to have a slightly more income than drama, terror and comedy films. We also can see that there are more outliers in the comedy genre.



Correlation matrix

In this section we can see the correlation between the varibales. It is a bit of what we have seen but in this case is not only focused with the gross but all the variables. Furthermore, by doing the correlation of the variables, we have a numeric value that says us how correlated two variables are which we did not quantify when doing the exploratory analysis.

We can see that there is a positive correlation between gross and budget variable (0,729). On the other hand there is positive correlation between Cast Facebook Likes and Actor Facebook likes.



	gross	budget	duration	directorfl	actor1fl	actor2fl	actor3fl	castfl	facenumber_in_poster
gross	1.0000000	0.7295407	0.4169113	0.1135821	0.1203821	0.2516909	0.3897209	0.2117592	0.0029669
budget	0.7295407	1.0000000	0.4807725	0.1185898	0.1384697	0.2587875	0.3404662	0.2198807	-0.0065147
duration	0.4169113	0.4807725	1.0000000	0.2152645	0.0649505	0.1288276	0.1809332	0.1070346	-0.0123550
directorfl	0.1135821	0.1185898	0.2152645	1.0000000	0.0660325	0.0940824	0.0453623	0.0825869	-0.0843436
actor1fl	0.1203821	0.1384697	0.0649505	0.0660325	1.0000000	0.3491797	0.2377791	0.9618000	0.0551944
actor2fl	0.2516909	0.2587875	0.1288276	0.0940824	0.3491797	1.0000000	0.4591963	0.5725142	0.0258751
actor3fl	0.3897209	0.3404662	0.1809332	0.0453623	0.2377791	0.4591963	1.0000000	0.4160769	0.0847390
castfl	0.2117592	0.2198807	0.1070346	0.0825869	0.9618000	0.5725142	0.4160769	1.0000000	0.0650337
facenumber_in_poster	0.0029669	-0.0065147	-0.0123550	-0.0843436	0.0551944	0.0258751	0.0847390	0.0650337	1.0000000

Fitting the complete model

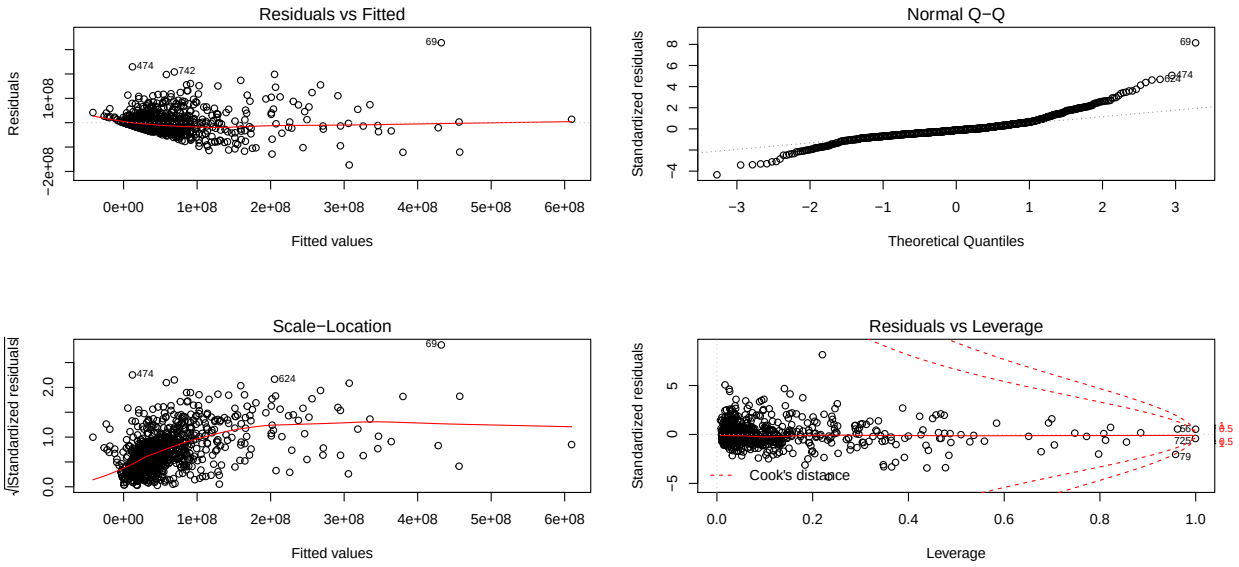
In this section we are going to perform a linear model with all the variables and their combinations so that we can see how it performs. We are going to exclude movietitle because it is a string and titleyear because we are going to user the factor we created previously (yearcat).

```
#gross ~ (. - titleyear - movietitle) * (yearcat + genre),

summary(completeModel<-lm(gross ~ (budget + duration + directorfl + actor1fl + actor2fl
                                   + actor3fl + castfl + facenumber_in_poster
                                   + genre + yearcat)^2, dataset))
```

As we can see in the NormQ-Q our model does not adjust perfectly to the model since we are having an adjusted r-squared of 0.50 and we are performing this model with all the possible combination of variables, which is a huge model. This is not necessary and should be avoided under all circumstances since our aim should be to have the minimal model possible which in the end, will be the one that will generalise the better too.

```
op<-par(mfrow=c(2,2))
plot(completeModel)
```



Use the stepwise procedure, by using the BIC criterion, to select the signicant variables

In order to create a model with the most significant variables, we decided to choose the BIC criterion. We could choose BIC over AIC because BIC is more restrictive therefore we will have less variables which is what we are aiming for. In order to use the standart weight which is 2, we decided to use the log of the size of variables which will be more restrictive and will remove those interactions that are not relevant.

We decided to compare two heuristic strategies when modelling:

- **Forward:** In this strategy we start the case with none available predictor variables and add one at a time
- **Backward:** In this strategy we start with all available predictor variables and delete one at a time

In any case we are later going to validate that the variables obtained from each model are significant.

```
nullModel <- lm(gross ~ 1, dataset)

forwardModel <- step(nullModel,
                     scope = list(upper=completeModel),
```

```
                      direction="both", criterion = "BIC",
                      k=log(nrow(dataset)))

backwardModel <- step(completeModel,
                      scope = list(lower=nullModel),
                      direction="both",
                      criterion = "BIC",
                      k=log(nrow(dataset)))
```

We can see that the forward model is a simpler model that hass less variables and all of them pass the p-value test.

```
kable(summary(forwardModel)$coefficients, format="latex", booktabs=TRUE)  %>%
  kable_styling(position = "center")
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1.926539e+08	2.534451e+07	-7.601407	0.0000000
budget	8.626975e-01	5.979130e-02	14.428473	0.0000000
actor3fl	-1.338824e+04	4.910479e+03	-2.726463	0.0065223
duration	1.755780e+06	2.285396e+05	7.682607	0.0000000
genreComedy	1.605711e+08	3.241567e+07	4.953503	0.0000009
genreDrama	1.974540e+08	2.816839e+07	7.009773	0.0000000
genreTerror	1.871068e+08	3.735289e+07	5.009167	0.0000007
actor3fl:duration	1.533793e+02	3.791719e+01	4.045111	0.0000566
duration:genreComedy	-1.189038e+06	2.997022e+05	-3.967398	0.0000783
duration:genreDrama	-1.715470e+06	2.391787e+05	-7.172338	0.0000000
duration:genreTerror	-1.484417e+06	3.446298e+05	-4.307278	0.0000183

In contraposition with the forward model, here we have a way more complex model and not all their variables are significant.

```
kable(summary(backwardModel)$coefficients, format="latex", booktabs=TRUE) %>%
  kable_styling(position = "center")
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-8.490934e+06	1.917817e+07	-0.4427396	0.6580588
budget	2.380804e-01	2.461948e-01	0.9670405	0.3337793
duration	1.458370e+04	1.429500e+05	0.1020196	0.9187634
actor1fl	8.503675e+03	4.270383e+03	1.9913143	0.0467433
actor2fl	9.404936e+03	4.883411e+03	1.9258946	0.0544271
actor3fl	4.454723e+03	3.257593e+03	1.3674890	0.1718079
castfl	-8.254113e+03	4.139204e+03	-1.9941307	0.0464342
genreComedy	1.279004e+07	1.172711e+07	1.0906390	0.2757189
genreDrama	1.304753e+07	1.137359e+07	1.1471781	0.2516079
genreTerror	2.473511e+07	1.216256e+07	2.0337088	0.0422690
yearcat2006-2010	-2.465552e+06	5.052513e+06	-0.4879853	0.6256772
yearcat2011-2016	1.442671e+07	5.163716e+06	2.7938618	0.0053171
budget:duration	7.345900e-03	1.571500e-03	4.6745225	0.0000034
budget:actor1fl	-2.970000e-05	6.900000e-06	-4.2879432	0.0000199
budget:castfl	1.840000e-05	4.900000e-06	3.8001755	0.0001542
budget:genreComedy	3.220768e-01	1.392955e-01	2.3121839	0.0209888
budget:genreDrama	-3.817265e-01	1.425043e-01	-2.6787015	0.0075234
budget:genreTerror	-2.389062e-01	1.938630e-01	-1.2323456	0.2181366
budget:yearcat2006-2010	3.609220e-02	9.704530e-02	0.3719106	0.7100455
budget:yearcat2011-2016	-4.138527e-01	9.541450e-02	-4.3374216	0.0000160
duration:actor1fl	-1.010769e+02	3.310265e+01	-3.0534366	0.0023278
duration:actor2fl	-1.152404e+02	4.008615e+01	-2.8748185	0.0041363
duration:castfl	1.009861e+02	3.115397e+01	3.2415174	0.0012319
actor1fl:actor3fl	2.664938e-01	9.862560e-02	2.7020745	0.0070184
actor3fl:castfl	-2.532358e-01	8.968910e-02	-2.8234851	0.0048536

Both have an R-squared of around 0.5 so their performance is similar. But as we said we are mostly interested in simple models so the forward model looks better to us.

Check the presence of multicollinearity. If there is some non-interaction multicollinearity in the model, make the corresponding corrections.

We are now going to evaluate the multicollinearity of the models, which states if one prediction variable can be linearly predicted with the others in a substantial degree of accuracy. To evaluate the multicollinearity we are going to compute the VIF.

```
kable(car::vif(forwardModel), format="latex", booktabs=TRUE)%>%
  kable_styling(position = "center")
```

	GVIF	Df	GVIF^(1/(2*Df))
budget	3.541305	1	1.881835
actor3fl	37.395201	1	6.115162
duration	9.061749	1	3.010274
genre	70073.898079	3	6.420853
actor3fl:duration	38.483049	1	6.203471
duration:genre	75721.204208	3	6.504335

```
kable(car::vif(backwardModel), format="latex", booktabs=TRUE)%>%
  kable_styling(position = "center")
```

	GVIF	Df	GVIF^(1/(2*Df))
budget	64.039318	1	8.002457
duration	3.781456	1	1.944597
actor1fl	4502.437213	1	67.100203
actor2fl	381.690247	1	19.536894
actor3fl	17.553498	1	4.189689
castfl	5707.834415	1	75.550211
genre	29.194714	3	1.754759
yearcat	3.298294	2	1.347635
budget:duration	53.333935	1	7.303009
budget:actor1fl	39.929605	1	6.318988
budget:castfl	52.402965	1	7.238989
budget:genre	17.672165	3	1.613919
budget:yearcat	15.120965	2	1.971945
duration:actor1fl	2772.166633	1	52.651369
duration:actor2fl	282.260885	1	16.800622
duration:castfl	3553.832230	1	59.614027
actor1fl:actor3fl	100.648550	1	10.032375
actor3fl:castfl	135.877816	1	11.656664

By checking the values of VIF, we see that from the forward model that is the simplest, the only variables that are significant, are budget and duration.

```
finalModel <- lm(gross ~ budget + duration, dataset)
kable(summary(finalModel)$coefficients, format="latex", booktabs=TRUE)%>%
  kable_styling(position = "center")
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-2.018513e+07	9.717161e+06	-2.077266	0.0380489
budget	1.068118e+00	3.930890e-02	27.172454	0.0000000
duration	3.191747e+05	9.392676e+04	3.398123	0.0007071

```
finalModel2 <- lm(gross ~ duration + actor3fl + genre
                  + yearcat + budget:genre + budget:yearcat, dataset)
kable(summary(finalModel2)$coefficients, format="latex", booktabs=TRUE)%>%
  kable_styling(position = "center")
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-9.570171e+07	1.477313e+07	-6.4780939	0.0000000
duration	6.421807e+05	1.090401e+05	5.8894013	0.0000000
actor3fl	6.749698e+03	8.887240e+02	7.5948191	0.0000000
genreComedy	4.282683e+07	1.142527e+07	3.7484319	0.0001890
genreDrama	3.288797e+07	1.151621e+07	2.8557989	0.0043888
genreTerror	5.223530e+07	1.205628e+07	4.3326199	0.0000163
yearcat2006-2010	-7.975983e+05	5.280552e+06	-0.1510445	0.8799735
yearcat2011-2016	1.234144e+07	5.410786e+06	2.2808954	0.0227810
genreAction:budget	1.355980e+00	1.182467e-01	11.4673866	0.0000000
genreComedy:budget	1.199462e+00	1.194042e-01	10.0453849	0.0000000
genreDrama:budget	6.852159e-01	1.412826e-01	4.8499671	0.0000014
genreTerror:budget	7.400187e-01	2.001429e-01	3.6974513	0.0002306
yearcat2006-2010:budget	-5.600320e-02	1.001442e-01	-0.5592256	0.5761430
yearcat2011-2016:budget	-4.071085e-01	9.981230e-02	-4.0787422	0.0000492

In this model we have that the variable yearcat2006-2010 does have multicolinearity together with yearcat2006-2010:budget. However, if we revise both models we see that this second model is way more complex and even that it performs slightly better, it is more complex. Therefore, we are going to choose **finalModel** which is the one of the **forwardModel**

Validate the model by checking the assumptions

Once we have our model selected, we are going to validate it with anova comparing the null model with our final model.

```
anova(nullModel, finalModel)
```

```
## Analysis of Variance Table
##
## Model 1: gross ~ 1
## Model 2: gross ~ budget + duration
##   Res.Df        RSS Df  Sum of Sq     F    Pr(>F)
## 1    939 5.5833e+18
## 2    937 2.5799e+18  2 3.0034e+18 545.4 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In this case we can see that our model, **passes** the test since our p-value is lower than 0.05.

Interpret the final model

Even our final model has an adjusted r-squared of 0.4928 which is a low value, we can assume that does not exist multicollinearity in its variables, one of our purposes in this project.

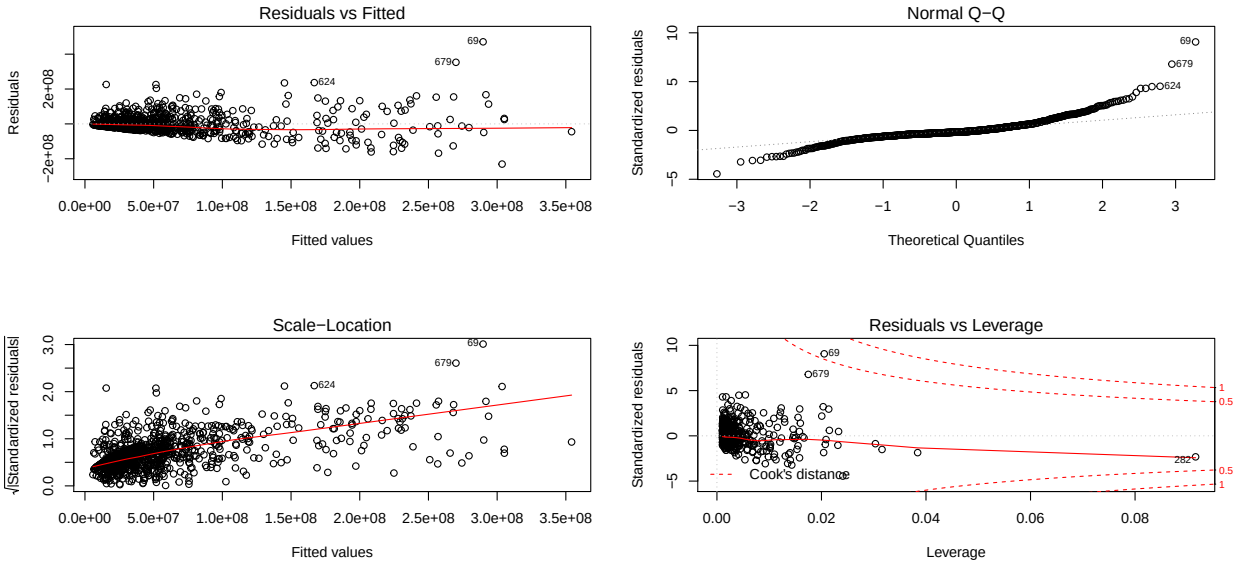
As we have previously said, we would rather a simple model with two predictor variables than a complex model with many of them. It seems difficult to predict gross variable in this dataset due its size (just 940 observations) and the very different data that contains. Probably it would be much wiser to select movies from a certain revenue and classify them and create a model for them. Regardless, we do not have this many variables that can be relevant as we have seen in the exploratory analysis.

We can say that other kind of models should lead to better results than a simple linear regression model. Also performing a PCA would give a first insight on how variables are related to each other.

Finally, we can see that our model now has a worst performance than the complete model but in this case we are not using all variables but rather a very simple model.

We can see from the residuals vs fitted that we still have non-linear patterns but not very differently than in the aforementioned. In the same way, now we have more skewness in the tail of the Q-Q plot. We can see from the Scale-Location that from 1.75e+08 the residuals are wider.

```
op=par(mfrow=c(2,2))
plot(finalModel)
```




```
library(car)
scatterplot(predict(finalModel),dataset$gross,smooth=FALSE)
```