Netflix case study: Genre vs decade

true

Affiliation

Netflix case study: Genre vs decade

Contents

Citations in papaja, detele appropriately later

Add the bibtex entry in the .bib file. You can find the entries in Google scholar, but double check since it is not always correct.

Call the citations in the text:

Citation within parentheses (Aust and Barth 2020)

Multiple citations (Aust and Barth 2020; R Core Team 2021)

In-text citations Aust and Barth (2020)

Year only (2021)

Only if your citation appears in the text it will also show up in the Reference list. Don't manually modify the Reference list.

Executive Summary

(150 words) – 0.3 POINTS Summarize the report. Write this as the very last thing.

What is the main topic you are addressing?

what are your research questions and hypotheses?

what are your results and the main conclusion?

Introduction

Main topic. In this paper, we are going to study the presence of social networks within a movie streaming platform. We're focusing on the structure of links among a group of social players, which consist of users watching and rating movies on Netflix.

Users of Netflix's movie recommendation algorithms are frequently given specific questions about their interests for certain items (which they provide by liking or disliking them, for example). These choices are then immediately integrated into the underlying learning system for future suggestions. If a recommender system starts promoting unwanted products after incorporating new preferences, the user may try to steer the system in the future by correcting it or supplying alternate preference information.

Importance. It is important to study the presence of these socials networks because this could potentially improve the recommender engine that this currently in place. For example, if you know that a user is likely to like a movie that other users with the same "liking profile" also like, you can recommend that movie to the user. When these connections are studied thoroughly, you could have a high probability that the recommendation is successful. This could have a large impact on a movie streaming platform.

Existing studies. In this paper, we will be looking into the Netflix Price Dataset. In 2006 Netflix decided to start a competition with a grand prize of 1 million US dollars. The goal of the competition was to create a collaborative filtering algorithm to predict user ratings for films, based on previous ratings without any other information about the users or films. In order to win you had to at least improve on Netflix's own algorithm by 10%. During the competition a lot of literature has emerged about the dataset and the competition. (Bell and Koren 2007; Takács et al. 2008; Narayanan and Shmatikov 2006)

However, no papers or any other literature can be found on network analysis on this dataset. We like to fill this gap in the literature by analysing the network and network structure that arises from the subset of the data that we will use. (Guillory and Bilmes 2011) developed an active movie recommendation system for Netflix. They found that a recommender system should not constantly ask questions to a user, because those reduces the user's mental image of how the recommendation system learns, prompting some participants to "lose track of what they were teaching." According to Amershi et al. (2014), this was because users are not always eager to act as simple oracles (repeatedly telling the recommendation system whether they like something or not). This is interesting to take into account for our research, because this would mean that a social network within a movie recommendation system can never be fully exposed.

Questions and hypotheses. In this dataset, we can easily see connections between users and movies, but not between just the users or just the movies. At least not, when we do not include one or the other.

Methodology

Dataset. During this study, the data that was shared by Netflix during the Netflix Prize open competition is used. The competition was about developing the best algorithm to predict user ratings for content on Netflix. The contest was started in order to improve their recommender system.

The data consisted of movies and account holders on Netflix who rated the movies on a 5 point scale. Also, the date the rating that was given and the year of the movie release are included in the dataset. Data was collected between October 1998 and December 2005 and the data consists of all ratings that were given during this period. (REF: KAGGLE) The initial dataset contains about 100 million ratings from over 480k users on almost 18k movies. (REF: Matrix Factorization and Neighbor Based Algorithms for the Netflix Prize Problem)

In order to answer research questions about liking behaviour of customers on streaming platforms (such as Netflix) this dataset provides us with a great opportunity. The dataset contains a unusual amount of real, user generated data. Datasets with similar types of data usually contain a lot less data and the large dataset is thus very useful in order to answer research questions about liking behaviour. (REF: Matrix Factorization and Neighbor Based Algorithms for the Netflix Prize Problem)

As stated in the initial distribution of the data, Netflix can and will not guarantee the correctness of the data. As no perfect documentation of the data collection exists, this can cause inaccuracies in the results. Also, Netflix uses algorithms that determine what users see and this effect can influence the results of this research. We could interpret the effect of the algorithm as a effect caused by user behaviour.

Descriptives. Looking at the histogram of the amount of ratings given on a certain date we can see that a peak in ratings exists always in the middle of the week (around wednesday/thursday). Looking at the distribution of the ratings over time we don't see great differences. Over time, people give consistent ratings. arrange(p1, p2, ncol=1)

Data analysis (Research Rationale). (about 500 words) – 1 POINTS * Why are these two methods suitable for your data?

- Why are these two methods suitable for your research questions?
- Are there other methods to address these questions? If yes, why are the methods you chose better for this case?

Results

(about 2000 words)

Wat voor tekst moeten wij hier typen??

QAP Model.

(about 1000 words) – 2.5 POINTS

Constructing data and the model. During preprocessing of the data, an interesting artifact was encountered; When users were removed from the dataset that either liked or disliked a movie, it was clearly shown that it occurred much more often that user gave a 4/5 star rating to a movie than a 1/2 star rating. This is interesting to keep in mind, as this might pose a potential bias in the dataset.

To give a weight to the edges in terms of importance, we calculate the weight according to a formula:

Results.

- Present your results appropriately (plots, tables…) and discuss your findings in plain English

Findings in relation to hypothesis.

- Discuss the meaning of your findings in relation to your hypothesis. (half of the points evaluated in this other part)

age	gender	eyes_col
7	M	BLUE
8	F	BROWN
8	M	GREEN
7	F	PINK

ERGM. (about 1000) – 2.5 POINTS

In order to answer this research question, an ERGM model has been used. The netwerk created consists of movies as nodes. Two movies get an undirected edge if at least one user exists that has watched and liked both movies. An important assumption that is made is that a user liked a movie if the user has given a rating of at least four out of five. Next, in order to assess the influence of the release date, the decade of the release date is used. Furthermore, only the main genre of the movies have been used to assess the influence of two movies being watched by one user.

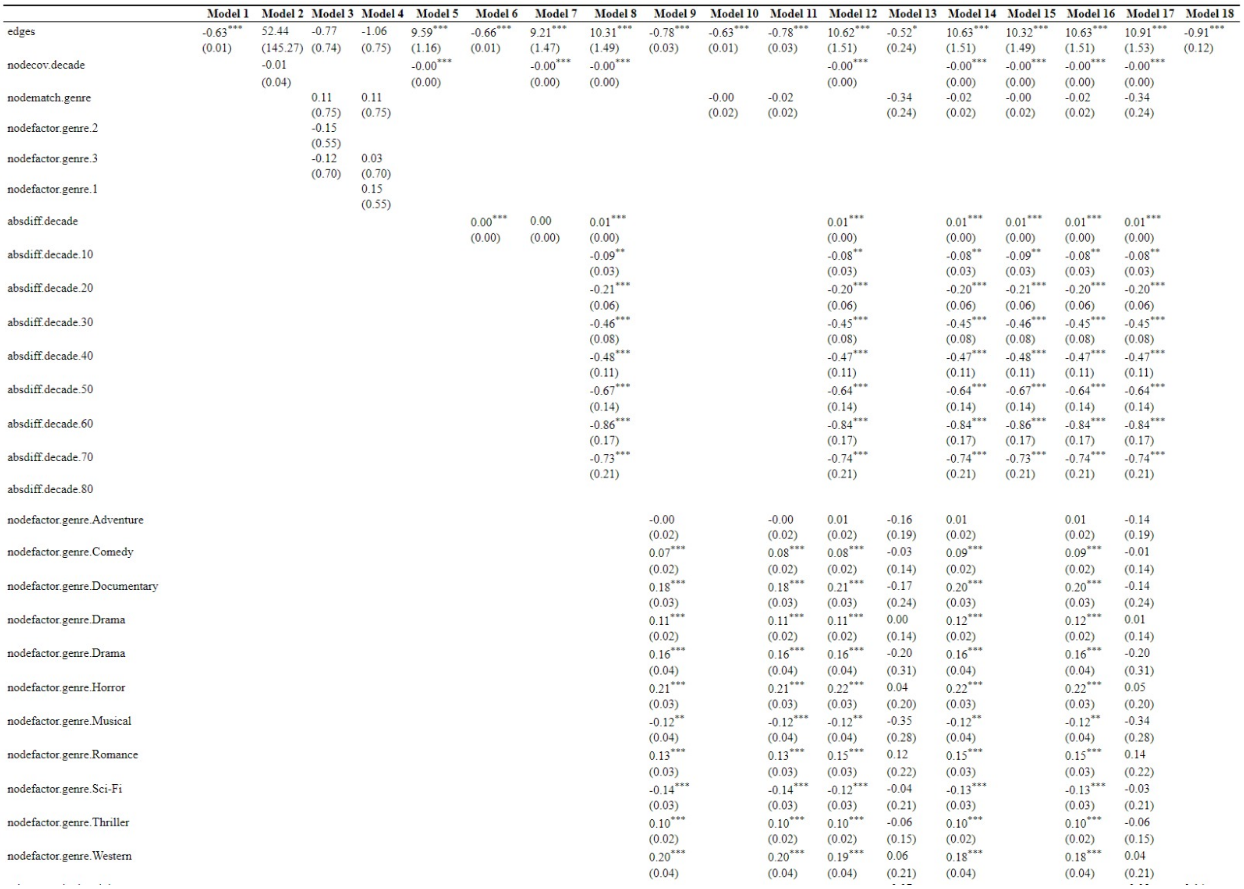
Homophily is hypothesized for genre and decade. Two movies with the same genre are more likely to have an edge and two movies from the same release decade are more likely to have an edge. Furthermore, the likeliness of an edge is even bigger when a movie has both the same genre and is from the same decade.

The Genre. To start off, the effect of genre on the likelihood of an edge between movies will be explored.

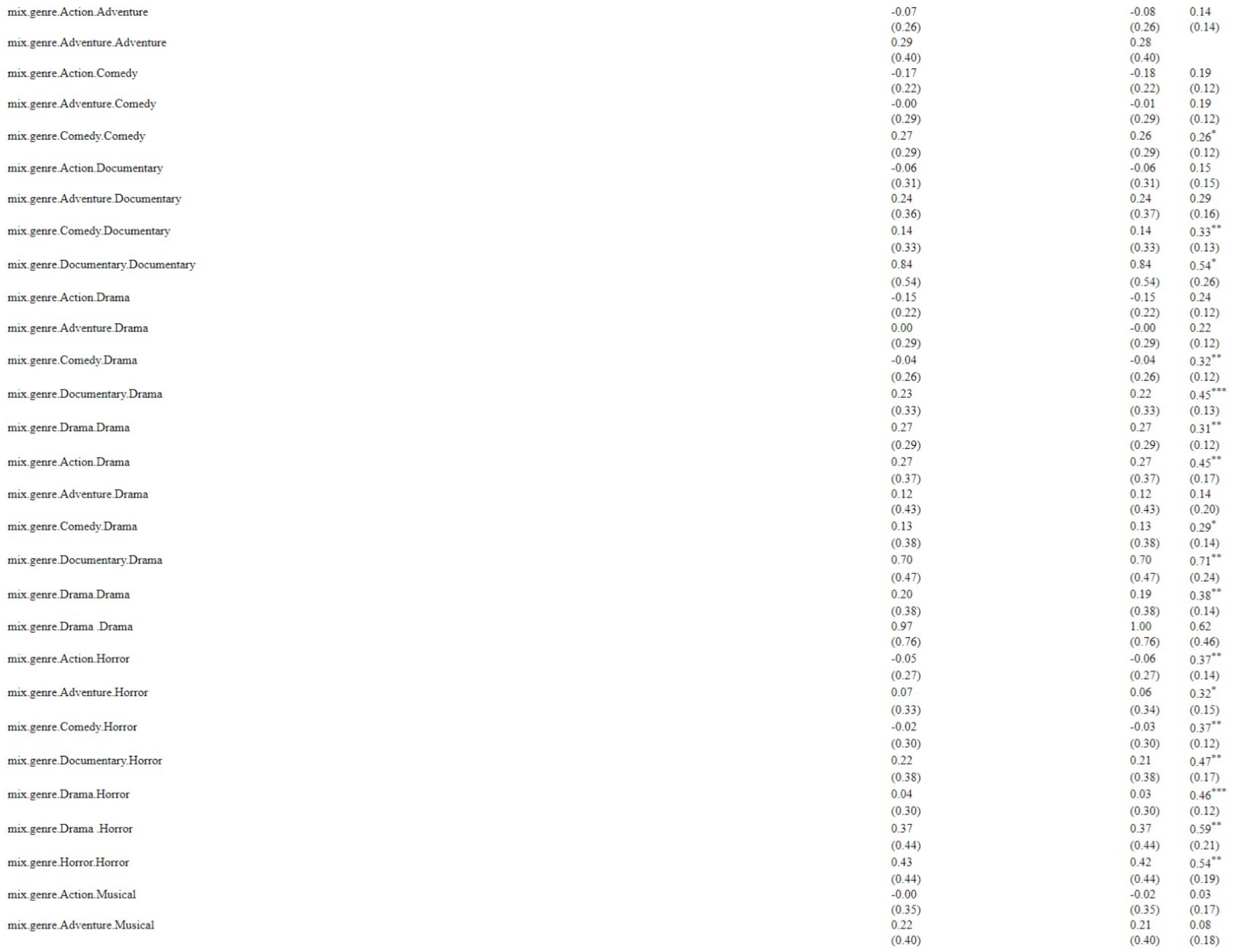
```r
el <- read.csv("data_0512/edgelist.csv", header=T, as.is=T)
attributes <- read.csv("data_0512/nodelist.csv", header=T, as.is=T)
```

```
# Create network
net2 <- network::as.network(el, matrix.type="edgelist", directed=F)

# Add Node attributes
net2 <- network::set.vertex.attribute(net2, 'genre', value = attributes$genre)
net2 <- network::set.vertex.attribute(net2, 'decade', value = attributes$decade)

# Add Edge attribute
net2 <- network::set.edge.attribute(net2, 'number_of_links', value = el$number_of_links)

# ERGM Statistical analyses (dyadic independent terms)
#model.01 <- ergm::ergm(net2 ~ edges)
#summary(model.01)

#model.01.1 <- ergm::ergm(net2 ~ density) # 10-15min run time
#summary(model.01.1)

#model.01.2 <- ergm::ergm(net2 ~ triangles) # 5-10min run time
#summary(model.01.2)
```

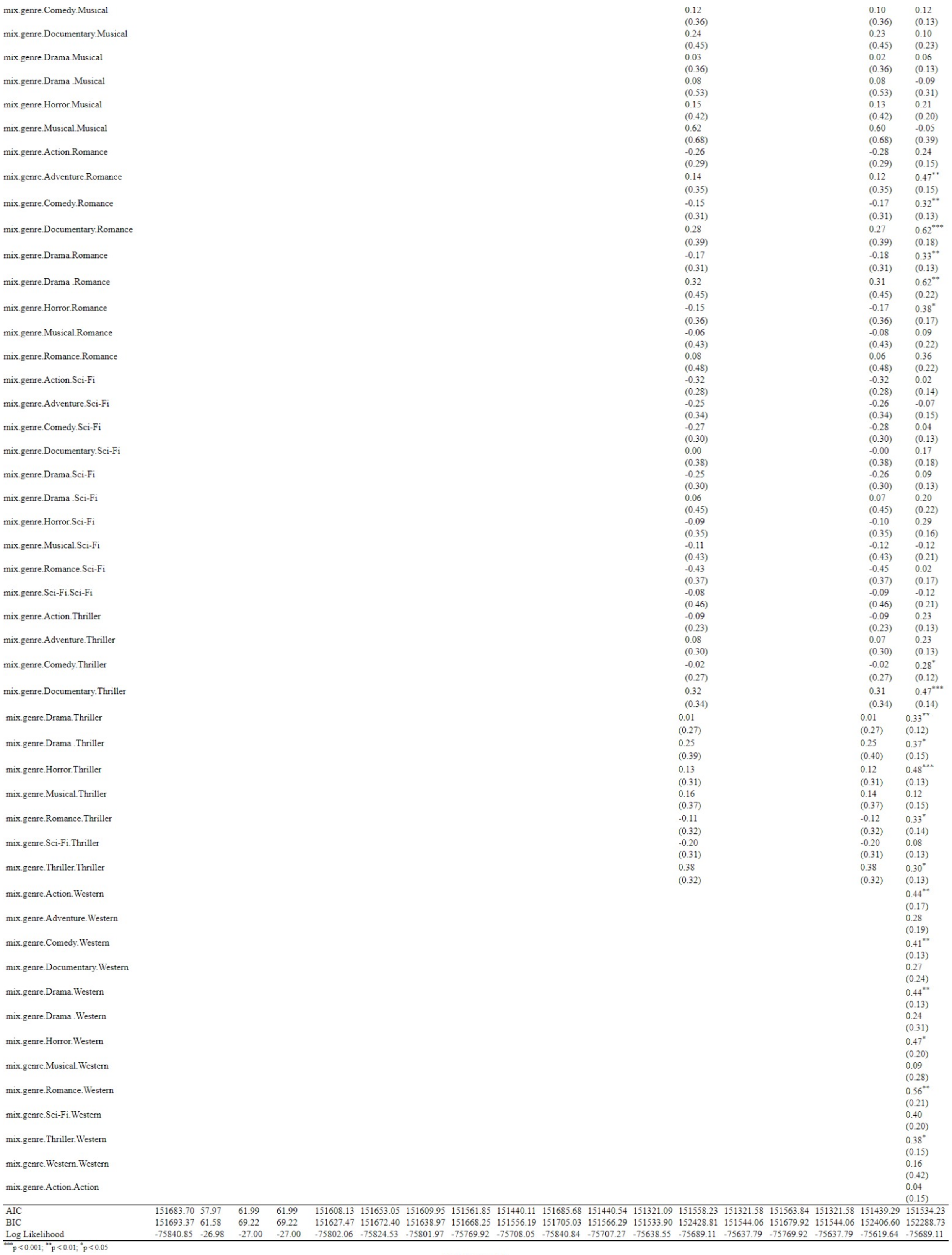
ERGM Model results.



	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10	Model 11	Model 12	Model 13	Model 14	Model 15	Model 16	Model 17	Model 18
edges	-0.63***	52.44	-0.77	-1.06	9.59***	-0.66***	9.21***	10.31***	-0.78***	-0.63***	-0.78***	10.62***	-0.52*	10.63***	10.32***	10.63***	10.91***	-0.91***
	(0.01)	(145.27)	(0.74)	(0.75)	(1.16)	(0.01)	(1.47)	(1.49)	(0.03)	(0.01)	(0.03)	(1.51)	(0.24)	(1.51)	(1.49)	(1.51)	(1.53)	(0.12)
nodecov.decade		-0.01			-0.00***		-0.00***	-0.00***				-0.00***		-0.00***	-0.00***	-0.00***	-0.00***	
		(0.04)			(0.00)		(0.00)	(0.00)				(0.00)		(0.00)	(0.00)	(0.00)	(0.00)	
nodematch.genre			0.11	0.11						-0.00	-0.02		-0.34	-0.02	-0.00	-0.02	-0.34	
			(0.75)	(0.75)						(0.02)	(0.02)		(0.24)	(0.02)	(0.02)	(0.02)	(0.24)	
nodefactor.genre.2			-0.15															
			(0.55)															
nodefactor.genre.3			-0.12	0.03														
			(0.70)	(0.70)														
nodefactor.genre.1				0.15														
				(0.55)														
absdiff.decade						0.00***	0.00	0.01***				0.01***		0.01***	0.01***	0.01***	0.01***	
						(0.00)	(0.00)	(0.00)				(0.00)		(0.00)	(0.00)	(0.00)	(0.00)	
absdiff.decade.10								-0.09**				-0.08**		-0.08**	-0.09**	-0.08**	-0.08**	
								(0.03)				(0.03)		(0.03)	(0.03)	(0.03)	(0.03)	
absdiff.decade.20								-0.21***				-0.20***		-0.20***	-0.21***	-0.20***	-0.20***	
								(0.06)				(0.06)		(0.06)	(0.06)	(0.06)	(0.06)	
absdiff.decade.30								-0.46***				-0.45***		-0.45***	-0.46***	-0.45***	-0.45***	
								(0.08)				(0.08)		(0.08)	(0.08)	(0.08)	(0.08)	
absdiff.decade.40								-0.48***				-0.47***		-0.47***	-0.48***	-0.47***	-0.47***	
								(0.11)				(0.11)		(0.11)	(0.11)	(0.11)	(0.11)	
absdiff.decade.50								-0.67***				-0.64***		-0.64***	-0.67***	-0.64***	-0.64***	
								(0.14)				(0.14)		(0.14)	(0.14)	(0.14)	(0.14)	
absdiff.decade.60								-0.86***				-0.84***		-0.84***	-0.86***	-0.84***	-0.84***	
								(0.17)				(0.17)		(0.17)	(0.17)	(0.17)	(0.17)	
absdiff.decade.70								-0.73***				-0.74***		-0.74***	-0.73***	-0.74***	-0.74***	
								(0.21)				(0.21)		(0.21)	(0.21)	(0.21)	(0.21)	
absdiff.decade.80																		
nodefactor.genre.Adventure									-0.00		-0.00	0.01	-0.16	0.01		0.01	-0.14	
									(0.02)		(0.02)	(0.02)	(0.19)	(0.02)		(0.02)	(0.19)	
nodefactor.genre.Comedy									0.07***		0.08***	0.08***	-0.03	0.09***		0.09***	-0.01	
									(0.02)		(0.02)	(0.02)	(0.14)	(0.02)		(0.02)	(0.14)	
nodefactor.genre.Documentary									0.18***		0.18***	0.21***	-0.17	0.20***		0.20***	-0.14	
									(0.03)		(0.03)	(0.03)	(0.24)	(0.03)		(0.03)	(0.24)	
nodefactor.genre.Drama									0.11***		0.11***	0.11***	0.00	0.12***		0.12***	0.01	
									(0.02)		(0.02)	(0.02)	(0.14)	(0.02)		(0.02)	(0.14)	
nodefactor.genre.Drama									0.16***		0.16***	0.16***	-0.20	0.16***		0.16***	-0.20	
									(0.04)		(0.04)	(0.04)	(0.31)	(0.04)		(0.04)	(0.31)	
nodefactor.genre.Horror									0.21***		0.21***	0.22***	0.04	0.22***		0.22***	0.05	
									(0.03)		(0.03)	(0.03)	(0.20)	(0.03)		(0.03)	(0.20)	
nodefactor.genre.Musical									-0.12**		-0.12***	-0.12**	-0.35	-0.12**		-0.12**	-0.34	
									(0.04)		(0.04)	(0.04)	(0.28)	(0.04)		(0.04)	(0.28)	
nodefactor.genre.Romance									0.13***		0.13***	0.15***	0.12	0.15***		0.15***	0.14	
									(0.03)		(0.03)	(0.03)	(0.22)	(0.03)		(0.03)	(0.22)	
nodefactor.genre.Sci-Fi									-0.14***		-0.14***	-0.12***	-0.04	-0.13***		-0.13***	-0.03	
									(0.03)		(0.03)	(0.03)	(0.21)	(0.03)		(0.03)	(0.21)	
nodefactor.genre.Thriller									0.10***		0.10***	0.10***	-0.06	0.10***		0.10***	-0.06	
									(0.02)		(0.02)	(0.02)	(0.15)	(0.02)		(0.02)	(0.15)	
nodefactor.genre.Western									0.20***		0.20***	0.19***	0.06	0.18***		0.18***	0.04	
									(0.04)		(0.04)	(0.04)	(0.21)	(0.04)		(0.04)	(0.21)	



mix.genre.Action.Adventure	-0.07	-0.08	0.14
	(0.26)	(0.26)	(0.14)
mix.genre.Adventure.Adventure	0.29	0.28	
	(0.40)	(0.40)	
mix.genre.Action.Comedy	-0.17	-0.18	0.19
	(0.22)	(0.22)	(0.12)
mix.genre.Adventure.Comedy	-0.00	-0.01	0.19
	(0.29)	(0.29)	(0.12)
mix.genre.Comedy.Comedy	0.27	0.26	0.26^{*}
	(0.29)	(0.29)	(0.12)
mix.genre.Action.Documentary	-0.06	-0.06	0.15
	(0.31)	(0.31)	(0.15)
mix.genre.Adventure.Documentary	0.24	0.24	0.29
	(0.36)	(0.37)	(0.16)
mix.genre.Comedy.Documentary	0.14	0.14	0.33^{**}
	(0.33)	(0.33)	(0.13)
mix.genre.Documentary.Documentary	0.84	0.84	0.54^{*}
	(0.54)	(0.54)	(0.26)
mix.genre.Action.Drama	-0.15	-0.15	0.24
	(0.22)	(0.22)	(0.12)
mix.genre.Adventure.Drama	0.00	-0.00	0.22
	(0.29)	(0.29)	(0.12)
mix.genre.Comedy.Drama	-0.04	-0.04	0.32^{**}
	(0.26)	(0.26)	(0.12)
mix.genre.Documentary.Drama	0.23	0.22	0.45^{***}
	(0.33)	(0.33)	(0.13)
mix.genre.Drama.Drama	0.27	0.27	0.31^{**}
	(0.29)	(0.29)	(0.12)
mix.genre.Action.Drama	0.27	0.27	0.45^{**}
	(0.37)	(0.37)	(0.17)
mix.genre.Adventure.Drama	0.12	0.12	0.14
	(0.43)	(0.43)	(0.20)
mix.genre.Comedy.Drama	0.13	0.13	0.29^{*}
	(0.38)	(0.38)	(0.14)
mix.genre.Documentary.Drama	0.70	0.70	0.71^{**}
	(0.47)	(0.47)	(0.24)
mix.genre.Drama.Drama	0.20	0.19	0.38^{**}
	(0.38)	(0.38)	(0.14)
mix.genre.Drama .Drama	0.97	1.00	0.62
	(0.76)	(0.76)	(0.46)
mix.genre.Action.Horror	-0.05	-0.06	0.37^{**}
	(0.27)	(0.27)	(0.14)
mix.genre.Adventure.Horror	0.07	0.06	0.32^{*}
	(0.33)	(0.34)	(0.15)
mix.genre.Comedy.Horror	-0.02	-0.03	0.37^{**}
	(0.30)	(0.30)	(0.12)
mix.genre.Documentary.Horror	0.22	0.21	0.47^{**}
	(0.38)	(0.38)	(0.17)
mix.genre.Drama.Horror	0.04	0.03	0.46^{***}
	(0.30)	(0.30)	(0.12)
mix.genre.Drama .Horror	0.37	0.37	0.59^{**}
	(0.44)	(0.44)	(0.21)
mix.genre.Horror.Horror	0.43	0.42	0.54^{**}
	(0.44)	(0.44)	(0.19)
mix.genre.Action.Musical	-0.00	-0.02	0.03
	(0.35)	(0.35)	(0.17)
mix.genre.Adventure.Musical	0.22	0.21	0.08
	(0.40)	(0.40)	(0.18)



mix.genre.Comedy.Musical													0.12				0.10	0.12
													(0.36)				(0.36)	(0.13)
mix.genre.Documentary.Musical													0.24				0.23	0.10
													(0.45)				(0.45)	(0.23)
mix.genre.Drama.Musical													0.03				0.02	0.06
													(0.36)				(0.36)	(0.13)
mix.genre.Drama .Musical													0.08				0.08	-0.09
													(0.53)				(0.53)	(0.31)
mix.genre.Horror.Musical													0.15				0.13	0.21
													(0.42)				(0.42)	(0.20)
mix.genre.Musical.Musical													0.62				0.60	-0.05
													(0.68)				(0.68)	(0.39)
mix.genre.Action.Romance													-0.26				-0.28	0.24
													(0.29)				(0.29)	(0.15)
mix.genre.Adventure.Romance													0.14				0.12	0.47**
													(0.35)				(0.35)	(0.15)
mix.genre.Comedy.Romance													-0.15				-0.17	0.32**
													(0.31)				(0.31)	(0.13)
mix.genre.Documentary.Romance													0.28				0.27	0.62***
													(0.39)				(0.39)	(0.18)
mix.genre.Drama.Romance													-0.17				-0.18	0.33**
													(0.31)				(0.31)	(0.13)
mix.genre.Drama .Romance													0.32				0.31	0.62**
													(0.45)				(0.45)	(0.22)
mix.genre.Horror.Romance													-0.15				-0.17	0.38*
													(0.36)				(0.36)	(0.17)
mix.genre.Musical.Romance													-0.06				-0.08	0.09
													(0.43)				(0.43)	(0.22)
mix.genre.Romance.Romance													0.08				0.06	0.36
													(0.48)				(0.48)	(0.22)
mix.genre.Action.Sci-Fi													-0.32				-0.32	0.02
													(0.28)				(0.28)	(0.14)
mix.genre.Adventure.Sci-Fi													-0.25				-0.26	-0.07
													(0.34)				(0.34)	(0.15)
mix.genre.Comedy.Sci-Fi													-0.27				-0.28	0.04
													(0.30)				(0.30)	(0.13)
mix.genre.Documentary.Sci-Fi													0.00				-0.00	0.17
													(0.38)				(0.38)	(0.18)
mix.genre.Drama.Sci-Fi													-0.25				-0.26	0.09
													(0.30)				(0.30)	(0.13)
mix.genre.Drama .Sci-Fi													0.06				0.07	0.20
													(0.45)				(0.45)	(0.22)
mix.genre.Horror.Sci-Fi													-0.09				-0.10	0.29
													(0.35)				(0.35)	(0.16)
mix.genre.Musical.Sci-Fi													-0.11				-0.12	-0.12
													(0.43)				(0.43)	(0.21)
mix.genre.Romance.Sci-Fi													-0.43				-0.45	0.02
													(0.37)				(0.37)	(0.17)
mix.genre.Sci-Fi.Sci-Fi													-0.08				-0.09	-0.12
													(0.46)				(0.46)	(0.21)
mix.genre.Action.Thriller													-0.09				-0.09	0.23
													(0.23)				(0.23)	(0.13)
mix.genre.Adventure.Thriller													0.08				0.07	0.23
													(0.30)				(0.30)	(0.13)
mix.genre.Comedy.Thriller													-0.02				-0.02	0.28*
													(0.27)				(0.27)	(0.12)
mix.genre.Documentary.Thriller													0.32				0.31	0.47***
													(0.34)				(0.34)	(0.14)
mix.genre.Drama.Thriller													0.01				0.01	0.33**
													(0.27)				(0.27)	(0.12)
mix.genre.Drama .Thriller													0.25				0.25	0.37*
													(0.39)				(0.40)	(0.15)
mix.genre.Horror.Thriller													0.13				0.12	0.48***
													(0.31)				(0.31)	(0.13)
mix.genre.Musical.Thriller													0.16				0.14	0.12
													(0.37)				(0.37)	(0.15)
mix.genre.Romance.Thriller													-0.11				-0.12	0.33*
													(0.32)				(0.32)	(0.14)
mix.genre.Sci-Fi.Thriller													-0.20				-0.20	0.08
													(0.31)				(0.31)	(0.13)
mix.genre.Thriller.Thriller													0.38				0.38	0.30*
													(0.32)				(0.32)	(0.13)
mix.genre.Action.Western																		0.44**
																		(0.17)
mix.genre.Adventure.Western																		0.28
																		(0.19)
mix.genre.Comedy.Western																		0.41**
																		(0.13)
mix.genre.Documentary.Western																		0.27
																		(0.24)
mix.genre.Drama.Western																		0.44**
																		(0.13)
mix.genre.Drama .Western																		0.24
																		(0.31)
mix.genre.Horror.Western																		0.47*
																		(0.20)
mix.genre.Musical.Western																		0.09
																		(0.28)
mix.genre.Romance.Western																		0.56**
																		(0.21)
mix.genre.Sci-Fi.Western																		0.40
																		(0.20)
mix.genre.Thriller.Western																		0.38*
																		(0.15)
mix.genre.Western.Western																		0.16
																		(0.42)
mix.genre.Action.Action																		0.04
																		(0.15)
AIC	151683.70	57.97	61.99	61.99	151608.13	151653.05	151609.95	151561.85	151440.11	151685.68	151440.54	151321.09	151558.23	151321.58	151563.84	151321.58	151439.29	151534.23
BIC	151693.37	61.58	69.22	69.22	151627.47	151672.40	151638.97	151668.25	151556.19	151705.03	151566.29	151533.90	152428.81	151544.06	151679.92	151544.06	152406.60	152288.73
Log Likelihood	-75840.85	-26.98	-27.00	-27.00	-75802.06	-75824.53	-75801.97	-75769.92	-75708.05	-75840.84	-75707.27	-75638.55	-75689.11	-75637.79	-75769.92	-75637.79	-75619.64	-75689.11

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Statistical models

- Present your results appropriately (plots, tables…) and discuss your findings in plain English

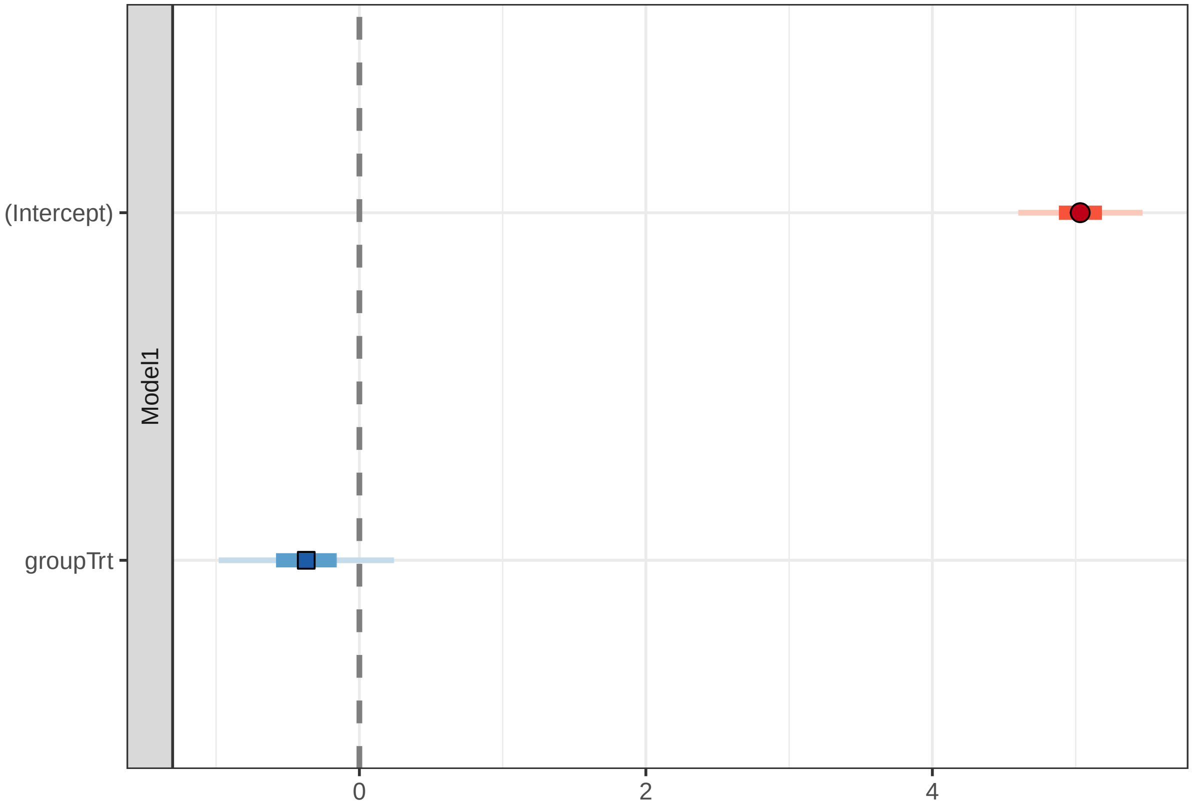
- Discuss the meaning of your findings in relation to your hypothesis. (half of the points evaluated in this other part)

Option to showcase a model:

	Model 1	Model 2
(Intercept)	5.03 ***	
	(0.22)	
groupTrt	-0.37	4.66 ***
	(0.31)	(0.22)
groupCtl		5.03 ***
		(0.22)
R^2	0.07	0.98
Adj. R^2	0.02	0.98
Num. obs.	20	20

Option 3

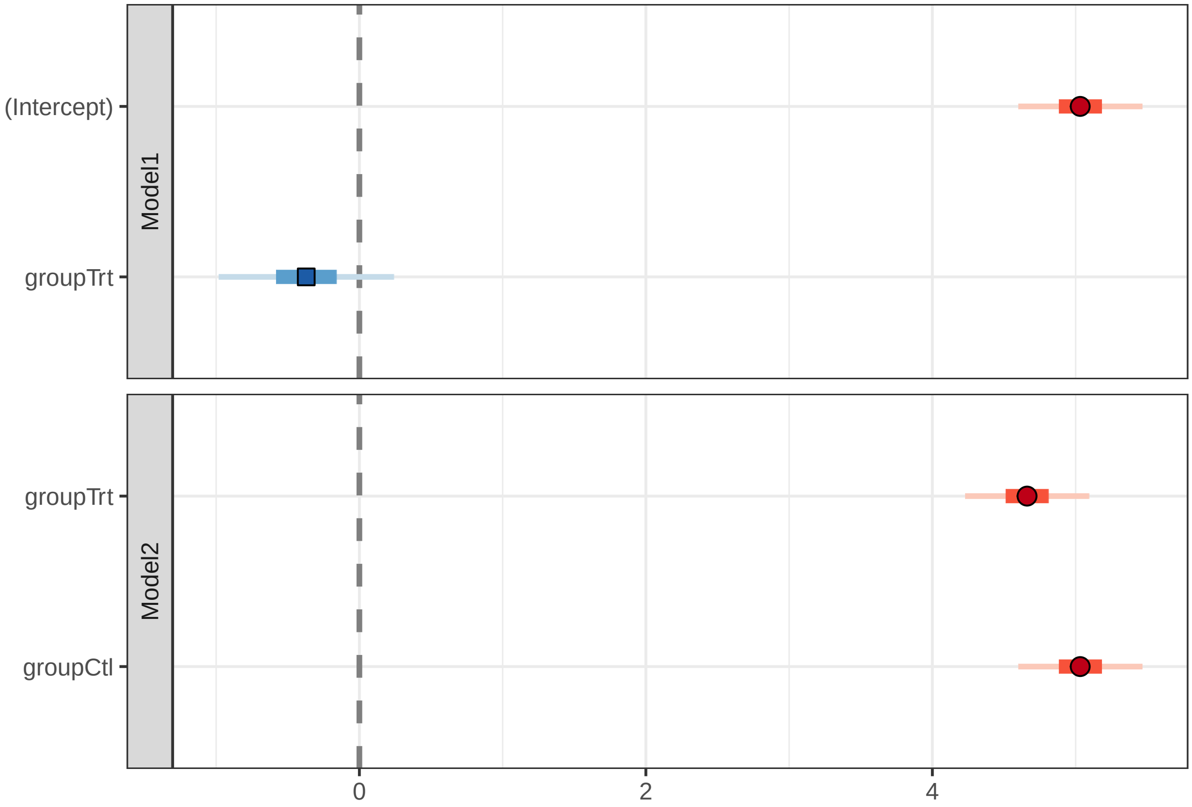
```
## Model: bars denote 0.5 (inner) resp. 0.95 (outer) confidence intervals (computed from
```



Bars denote CIs. Circle points denote significance.

Option 4

```
## Models: bars denote 0.5 (inner) resp. 0.95 (outer) confidence intervals (computed fro
```



Bars denote CIs. Circle points denote significance.

Conclusion

(about 350 words) – 0.7 POINTS What were your topic and research questions again? (1 sentence)

What did you learn from the two analysis you run? *** most important point to address 0.5 POINTS here

Who benefits from your findings?

What does remain an open problem?

Can you give suggestions for future work in this area?

References

Amershi, Saleema, Maya Cakmak, William Bradley Knox, and Todd Kulesza. 2014. “Power to the People: The Role of Humans in Interactive Machine Learning.” *Ai Magazine* 35 (4): 105–20.

Aust, Frederik, and Marius Barth. 2020. *papaja: Create APA Manuscripts with R Markdown*. https://github.com/crsh/papaja.

Bell, Robert M, and Yehuda Koren. 2007. “Lessons from the Netflix Prize Challenge.” *Acm Sigkdd Explorations Newsletter* 9 (2): 75–79.

Guillory, Andrew, and Jeff A Bilmes. 2011. “Simultaneous Learning and Covering with Adversarial Noise.” In *ICML*.

Narayanan, Arvind, and Vitaly Shmatikov. 2006. “How to Break Anonymity of the Netflix Prize Dataset.” *arXiv Preprint Cs/0610105*.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Takács, Gábor, István Pilászy, Bottyán Németh, and Domonkos Tikk. 2008. “Matrix Factorization and Neighbor Based Algorithms for the Netflix Prize Problem.” In *Proceedings of the 2008 ACM Conference on Recommender Systems*, 267–74.