

# Report coursework assignment A - 2021

## CS4125 Seminar Research Methodology for Data Science

Nikki Bouman (4597648), Anuj Singh (), Gwennan Smitskamp (4349822)

20/04/2021

## Contents

<b>1</b>	<b>Part 1 - Design and set-up of true experiment</b>	<b>1</b>
1.1	The motivation for the planned research . . . . .	1
1.2	The theory underlying the research . . . . .	1
1.3	Research questions . . . . .	1
1.4	The related conceptual model . . . . .	1
1.5	Experimental Design . . . . .	2
1.6	Experimental procedure . . . . .	2
1.7	Measures . . . . .	2
1.8	Participants . . . . .	2
1.9	Suggested statistical analyses . . . . .	2
<b>2</b>	<b>Part 2 - Generalized linear models</b>	<b>2</b>
2.1	Question 1 Twitter sentiment analysis (Between groups - single factor) . . . . .	2
2.1.1	Conceptual model . . . . .	2
2.1.2	Collecting tweets, and data preparation . . . . .	3
2.1.3	Homogeneity of variance analysis . . . . .	3
2.1.4	Visual inspection Mean and distribution sentiments . . . . .	3
2.1.5	Frequentist approach . . . . .	5
2.1.6	Bayesian Approach . . . . .	6

## 1 Part 1 - Design and set-up of true experiment

### 1.1 The motivation for the planned research

(Max 250 words)

### 1.2 The theory underlying the research

(Max 250 words) Preferable based on theories reported in literature

### 1.3 Research questions

The research question that will be examined in the experiment (or alternatively the hypothesis that will be tested in the experiment)

### 1.4 The related conceptual model

This model should include: *Independent variable(s)* *Dependent variable* *Mediating variable (at least 1)* *Moderating variable (at least 1)*

## 1.5 Experimental Design

Note that the study should have a true experimental design

## 1.6 Experimental procedure

Describe how the experiment will be executed step by step

## 1.7 Measures

Describe the measure that will be used

## 1.8 Participants

Describe which participants will recruit in the study and how they will be recruited

## 1.9 Suggested statistical analyses

Describe the statistical test you suggest to carry out on the collected data

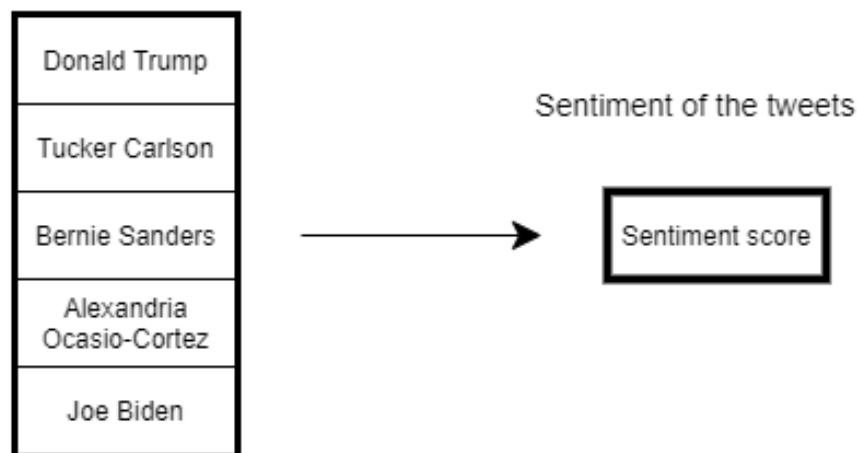
# 2 Part 2 - Generalized linear models

## 2.1 Question 1 Twitter sentiment analysis (Between groups - single factor)

### 2.1.1 Conceptual model

Make a conceptual model for the following research question: Is there a difference in the sentiment of the tweets related to the different celebrities?

Relation to different celebrities



**Is there a difference in the sentiment of the tweet related to the different celebrities?**

Figure 1: The conceptual model for the research question: Is there a difference in the sentiment of the tweets related to the different celebrities?

### 2.1.2 Collecting tweets, and data preparation

We found 5 celebrities in US politics: Donald Trump, Tucker Carlson, Bernie Sanders, Alexandria Ocasio-Cortez, Joe Biden. As dutch students we are not well-versed in the popular English twitter celebrities, so US politics was the best option for us to find celebrities that had enough recent Tweets for the Twitter API.

### 2.1.3 Homogeneity of variance analysis

```
pander(leveneTest(semFrame$score, semFrame$Celeb))
```

Table 1: Levene's Test for Homogeneity of Variance (center = median)

	Df	F value	Pr(>F)
<b>group</b>	4	23.64	5.714e-19
	1465	NA	NA

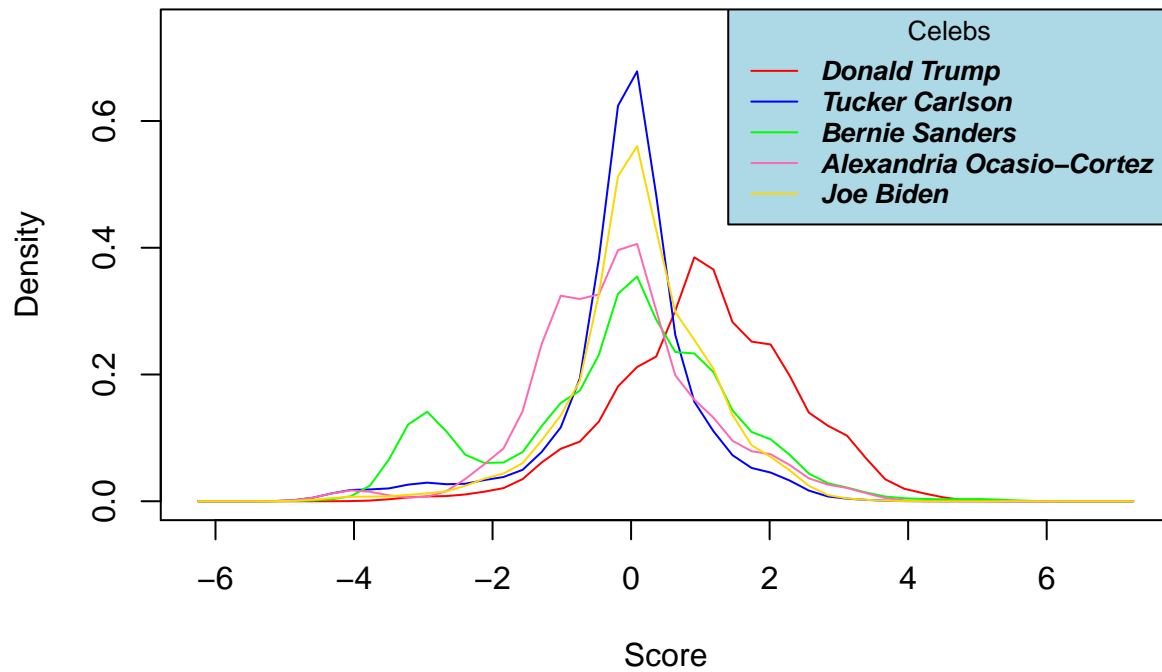
The Levene test reveals a p-value smaller than 0.001, indicating that there is significant difference between the group variances in sentiment score. We conclude that the variance among the five groups is not equal.

### 2.1.4 Visual inspection Mean and distribution sentiments

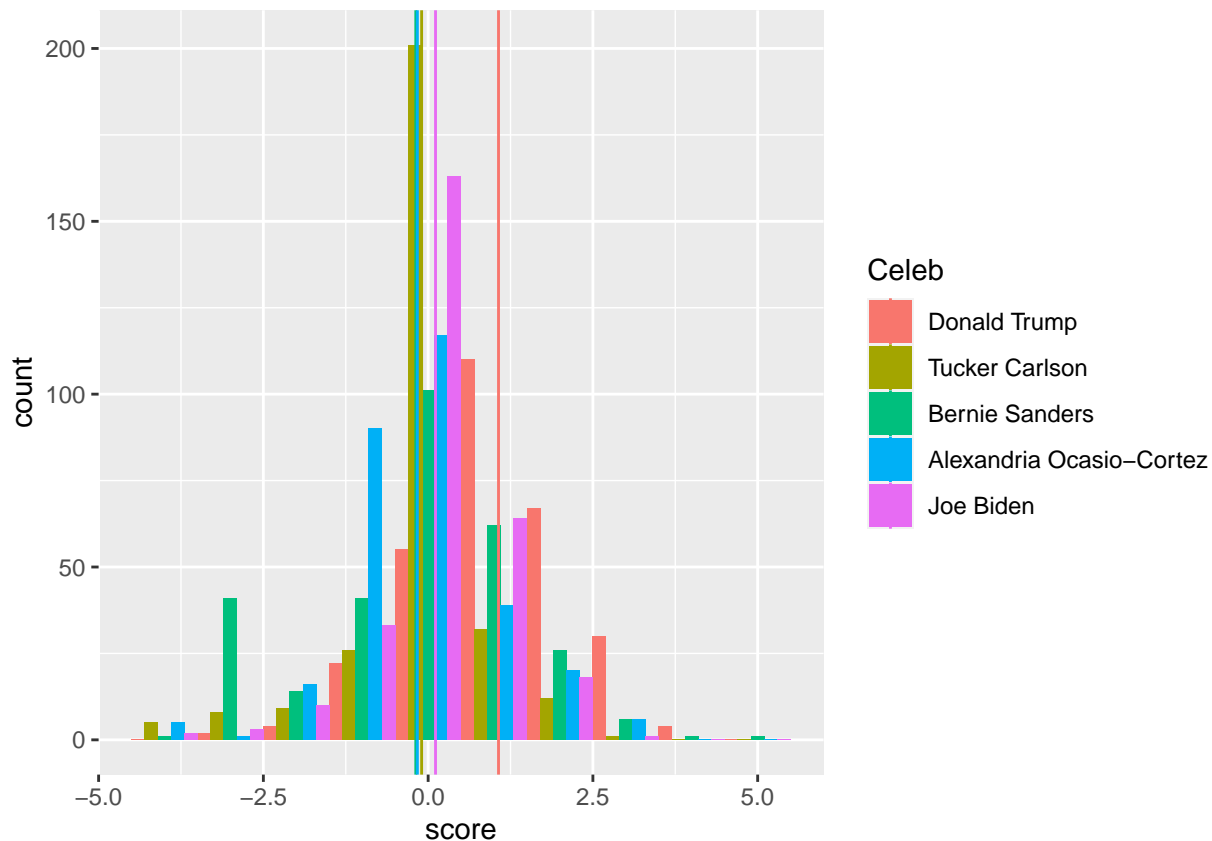
We plot both a line density and a distribution histogram which includes a mean line.

```
#boxplot(score ~ Celeb, data = semFrame, text.font = 4)
sm.density.compare(semFrame$score, semFrame$Celeb, xlab = "Score",
                    col=c('red', 'blue', 'green', 'hotpink', 'gold'), lty=c(1,1,1,1,1))
title(main="Visual inspection Mean and distribution sentiments")
legend('topright', legend = levels(semFrame$Celeb),
       col=c('red', 'blue', 'green', 'hotpink', 'gold'),
       title="Celebs", lty=1, cex=0.8, text.font = 4, bg='lightblue')
```

## Visual inspection Mean and distribution sentiments



```
cdata <- ddpoly(semFrame, "Celeb", summarise, score.mean=mean(score))
ggplot(semFrame, aes(x=score, fill=Celeb)) +
  geom_histogram(binwidth=1, position="dodge") +
  geom_vline(data=cdata, aes(xintercept=score.mean, colour=Celeb),
            linetype="solid", size=0.5)
```



We see all US politic Celebs have a mean around 0. #trump has the highest sentiment mean and the largest difference with the rest.

### 2.1.5 Frequentist approach

**2.1.5.1 Linear model** A one-way between subjects ANOVA was conducted to compare the effect of relation to celebrities on sentiment score in five conditions.

```
model0 <- lm(formula = score ~ 1 , data = semFrame)
model1 <- lm(formula = score ~ Celeb , data = semFrame)
pander(anova(model0, model1, test = "F"))
```

Table 2: Analysis of Variance Table

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1469	2476	NA	NA	NA	NA
1465	2147	4	328.6	56.06	4.859e-44

There was a significant effect of relation to celebrities on sentiment score at the  $p < .001$  level for the five conditions  $[F(4, 1465) = 56.06, p < 0.001]$ .

```
#AIC
models <- list(model0, model1)
model.names <- c("model0", "model1")
pander(aictab(cand.set = models, modnames=model.names),
       caption="Model selection based on AICc.")
```

Table 3: Model selection based on AICc.

	Modnames	K	AICc	Delta_AICc	ModelLik	AICcWt	LL	Cum.Wt
<b>2</b>	model1	6	4741	0	1	1	-2364	1
<b>1</b>	model0	2	4942	201.3	1.939e-44	1.939e-44	-2469	1

A lower AIC indicates a better fit, which is the model with the predictor.

```
(pairwise.t.test(semFrame$score, semFrame$Celeb, paired = FALSE, p.adjust.method = "bonferroni"))
```

### 2.1.5.2 Post Hoc analysis

```
##
## Pairwise comparisons using t tests with pooled SD
##
## data: semFrame$score and semFrame$Celeb
##
##          Donald Trump Tucker Carlson Bernie Sanders
## Tucker Carlson <2e-16      -              -
## Bernie Sanders <2e-16      1.000          -
## Alexandria Ocasio-Cortez <2e-16      1.000      1.000
## Joe Biden <2e-16      0.348      0.028
##          Alexandria Ocasio-Cortez
## Tucker Carlson -
## Bernie Sanders -
## Alexandria Ocasio-Cortez -
## Joe Biden 0.059
##
## P value adjustment method: bonferroni
```

Post hoc comparisons using the Bonferroni correction indicated that the corrected p-value for the trump condition was significantly different than the other conditions ( $p < 0.001$ ). However, between the others condition it does not show a significant difference.

**2.1.5.3 Report section for a scientific publication** A one-way between subjects ANOVA was conducted to compare the effect of relation to celebrities on sentiment score in five conditions. There was a significant effect of relation to celebrities on sentiment score at the  $p < .001$  level for the five conditions [ $F(4, 1465) = 56.06, p < 0.001$ ]. However, post hoc comparisons using the Bonferroni correction indicated that only the corrected p-value for the trump condition was significantly different than the other conditions ( $p < 0.001$ ), between the others condition it does not show a significant difference. Taken together, these results suggest that some celebrities really do have an effect on the sentiment in Tweets.

### 2.1.6 Bayesian Approach

**2.1.6.1 Model description** The sentiment scores seem to center around 0, and all seem to be single digits.

$$score \sim Norm(\mu, \sigma)$$

$$\mu = \alpha + b * Celeb$$

$$\alpha \sim Norm(0, 10)$$

$$\sigma \sim Uniform(0.001, 10)$$

```
m0 <-map2stan(alist(
  score ~ dnorm(mu, sigma),
  mu <-a,
  a ~ dnorm(0, 10),
  sigma ~ dunif(0.001, 10)),
  data = semFrame , iter= 10000, chains = 4, cores = 4 )
```

### 2.1.6.2 Model comparison

```
## Computing WAIC
```

```
m1 <-map2stan(alist(
  score ~ dnorm(mu, sigma),
  mu <-a[Celeb] ,
  a[Celeb] ~ dnorm(0, 10),
  sigma ~ dunif(0.001, 10)),
  data = semFrame ,iter= 10000, chains = 4, cores = 4 )
```

```
## Computing WAIC
```

```
pander(compare(m0, m1, func=WAIC))
```

	WAIC	SE	dWAIC	dSE	pWAIC	weight
<b>m1</b>	4741	70.18	0	NA	6.724	1
<b>m0</b>	4942	67.72	201.1	28.92	2.556	2.166e-44

Lower WAIC indicates a better performing model, so with predictors is the winning model.

```
pander(precis(m1, depth=2, prob = .95))
```

### 2.1.6.3 Comparison celebrity pair

	mean	sd	2.5%	97.5%	n_eff	Rhat4
<b>a[1]</b>	1.068	0.07052	0.9297	1.204	29693	0.9999
<b>a[2]</b>	-0.09814	0.07089	-0.2377	0.04192	30105	0.9998
<b>a[3]</b>	-0.1871	0.07043	-0.3238	-0.04972	30543	0.9999
<b>a[4]</b>	-0.1628	0.07184	-0.3034	-0.02108	28697	0.9998
<b>a[5]</b>	0.1121	0.07084	-0.02701	0.25	29942	0.9999
<b>sigma</b>	1.212	0.02273	1.168	1.257	29965	1

Looking at the credibility intervals of the celebrities effects, We see the conditions where the mean of a condition does not fall within a credibility interval of an other condition. This holds for the Trump condition and a couple other combinations. We can again conclude that some celebrities really do have an effect on the sentiment in Tweets.