Part 2 - Generalized linear models

Code ▾

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

Hide

```
head(semFrame)
```

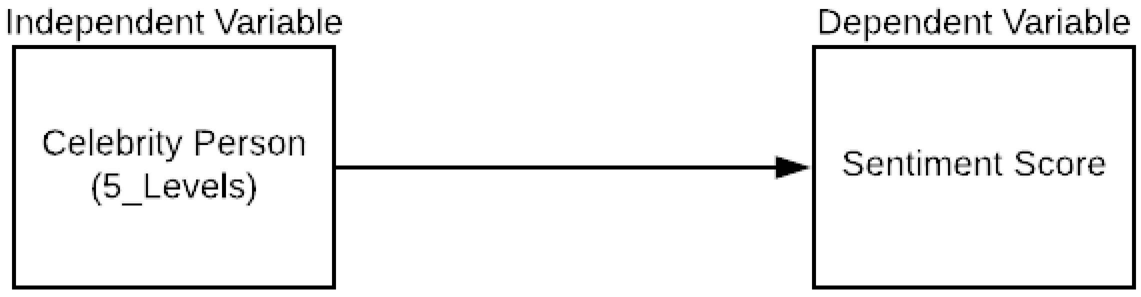
	Candidate <fctr>	score <int>
1	Donald Trump	0
2	Donald Trump	0
3	Donald Trump	1
4	Donald Trump	0
5	Donald Trump	0
6	Donald Trump	3

6 rows

Here we see how the data has been stored and collected using the code provided to analyse tweets. We shall now begin the analysis.

Conceptual model

The conceptual model contains the celebrity type as an independent variable and the sentiment score as the dependent variable. We wish to study if there is a significant effect on the sentiment scores based on the different celebrities.



Homogeneity of variance analysis

The homogeneity of variance is a common assumption underlying many statistical tests such as the T test or the Anova analysis. Here we wish to examine if the variance or the spread around the mean for the different celebrities is the same or not.

Hide

```
pander(leveneTest(semFrame$score, semFrame$Candidate, center=mean))
```

```


---------------------------------------
  &nbsp;      Df    F value    Pr(>F)
----------- ------ --------- -----------
 **group**    4      61.85    4.338e-51

             4995     NA         NA
---------------------------------------


Table: Levene's Test for Homogeneity of Variance (center = mean)
```

The results of our levene test is significant (Degrees of freedom=4,F value=61.85, P value <.001), therefore the null hypothesis or assumption of homogeneity of variance is rejected and so doesn’t hold for this data.

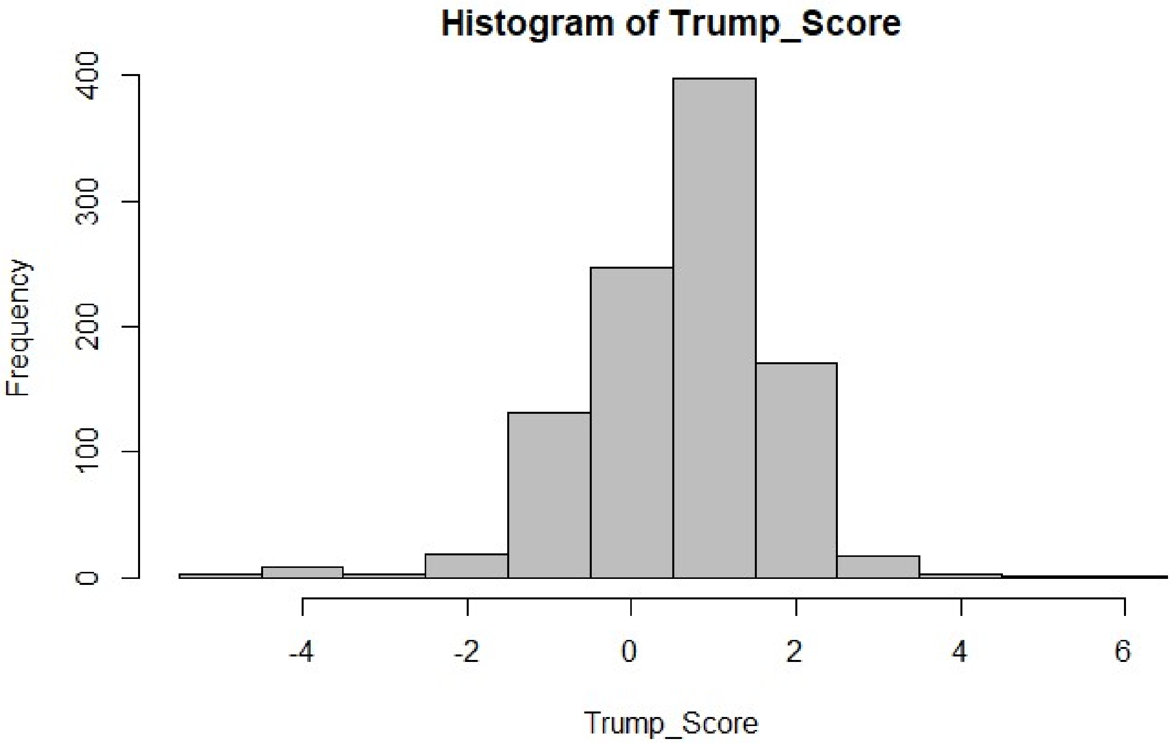
Visual Inspection

Hide

```
Trump_Score <- sem$analysis_Trump.score
Modi_Score <- sem$analysis_Modi.score
Beyonce_Score <- sem$analysis_Beyonce.score
Bill_Gates_Score <- sem$analysis_Gates.score
Rihanna_Score <- sem$analysis_Rihanna.score
```

Hide

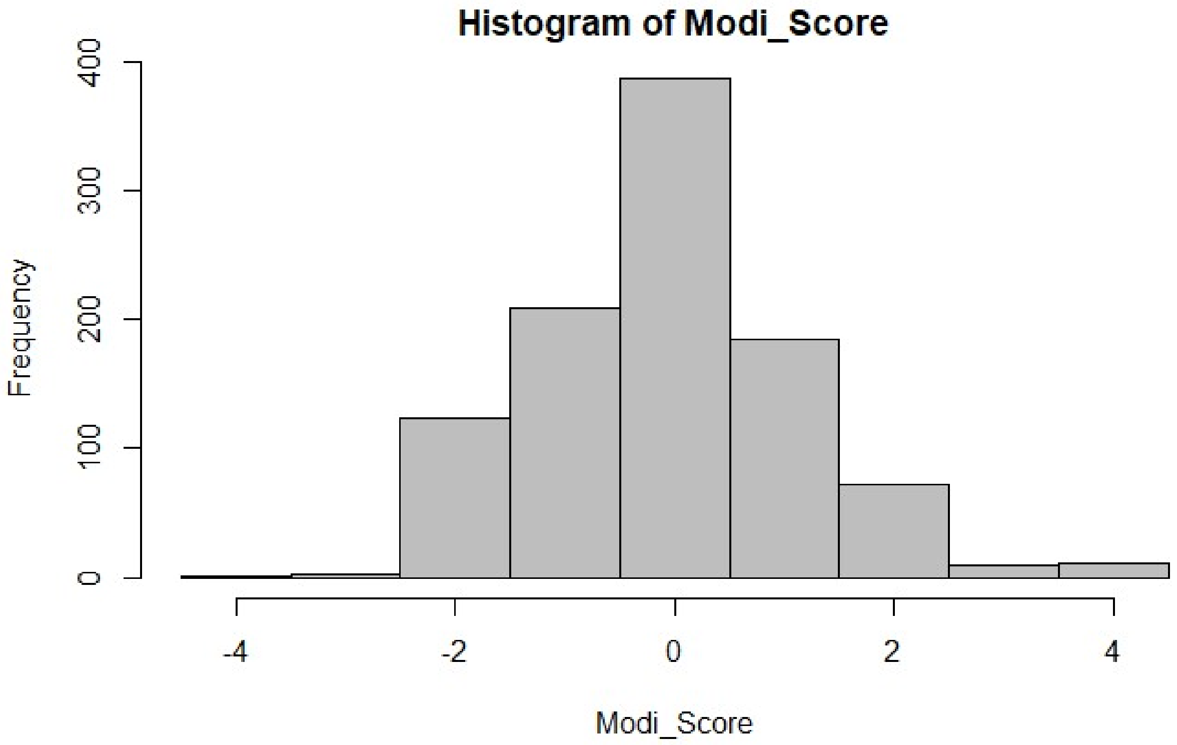
```
h<-hist(Trump_Score, col="grey",breaks=seq(min(Trump_Score)-0.5, max(Trump_Score)+0.5,
by=1))
```



Trumps sentiment scores are suprisingly skewed towards positive with a peak at +1, suggesting he might be doing well with the Corona situation in the US. The data doesn't seem to look too normally distributed and we can see long tails on both sides.

Hide

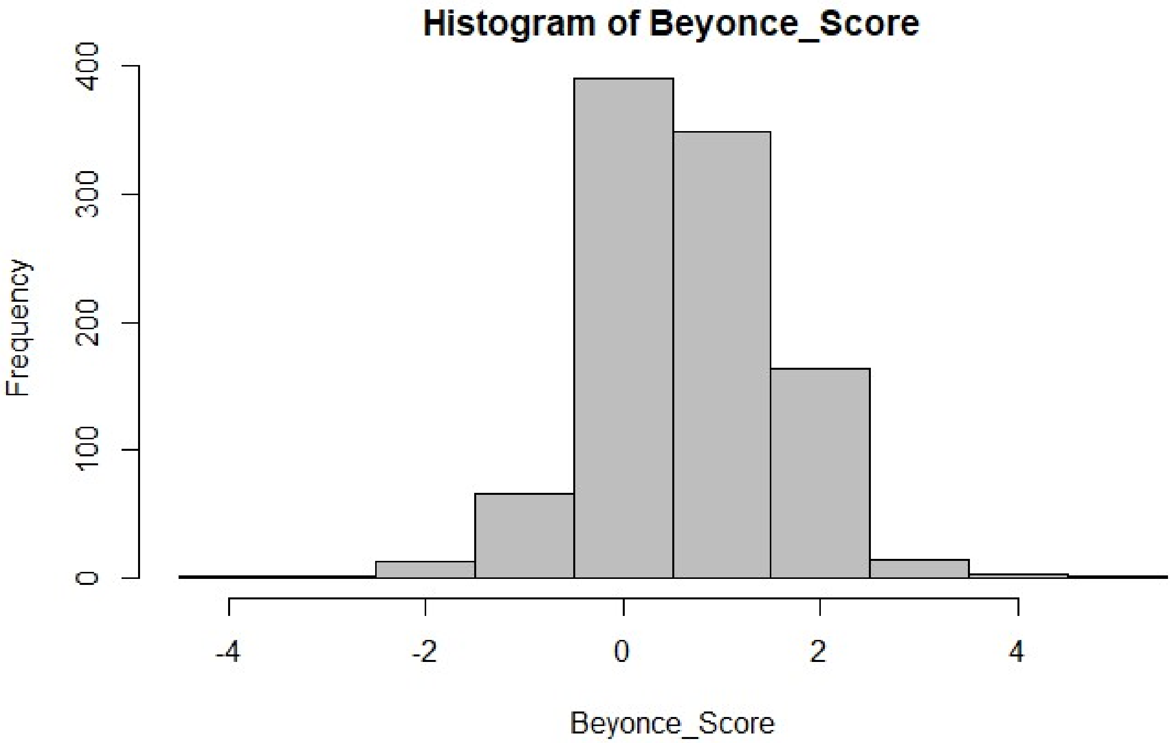
```
h<-hist(Modi_Score, col="grey",breaks=seq(min(Modi_Score)-0.5, max(Modi_Score)+0.5, by
=1) )
```



PM of India, Mr Modi seems to be showing somewhat normally distributed sentiment scores with a sharp peak at 0, which could mean his actions against the corona virus could be showing an overall neutral response by the indian community.

Hide

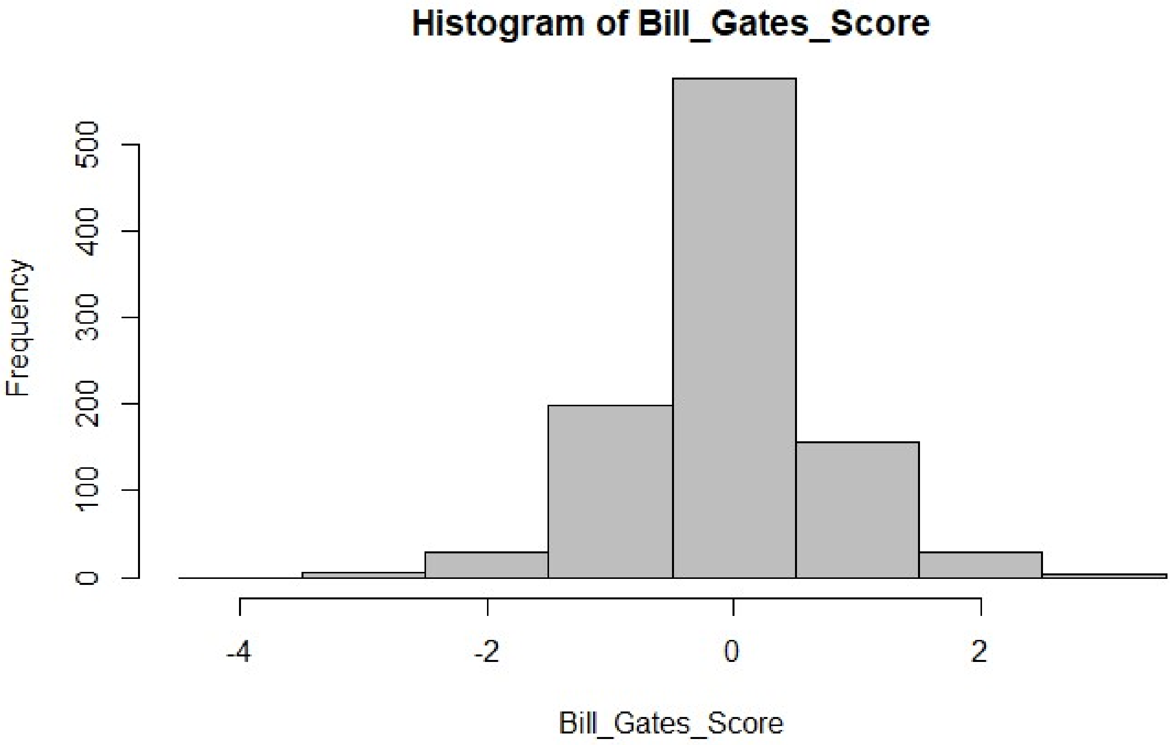
```
h<-hist(Beyonce_Score,col="grey", breaks=seq(min(Beyonce_Score)-0.5, max(Beyonce_Scor
e)+0.5, by=1))
```



Beyonce’s sentiment scores can be seen to be mostly positive as one would expect.The data doesnt look normally distributed at all.

Hide

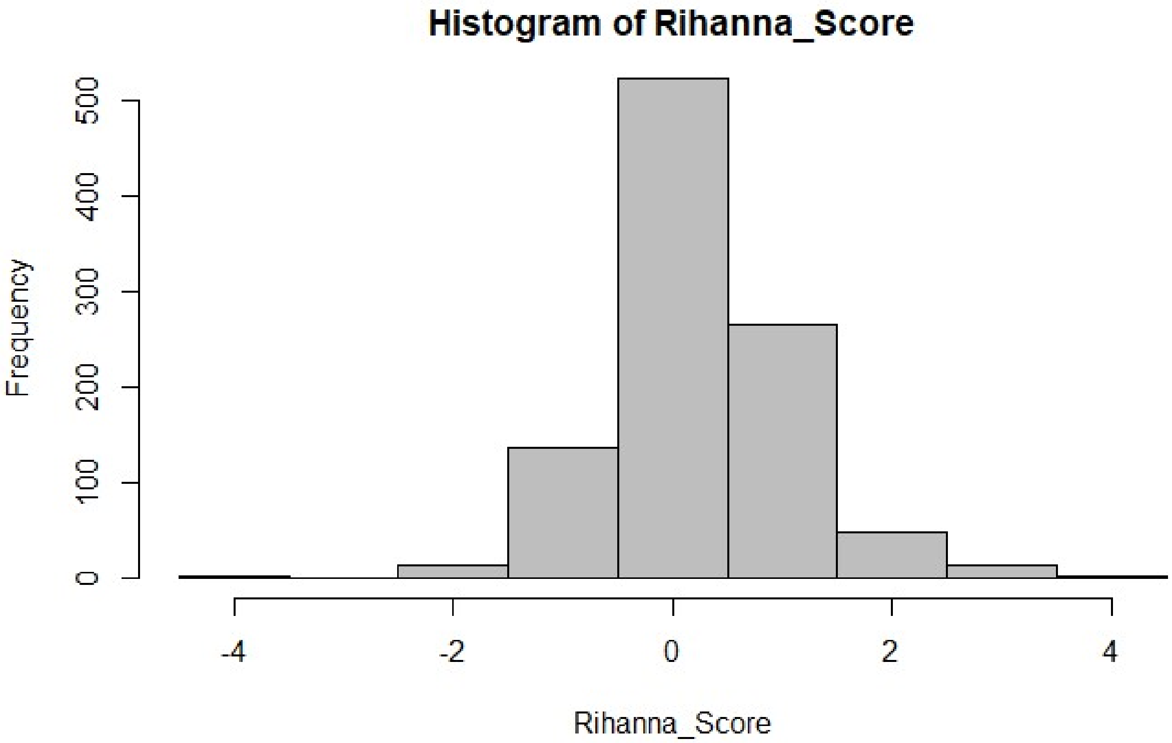
```
h<-hist(Bill_Gates_Score,col="grey", breaks=seq(min(Rihanna_Score)-0.5, max(Bill_Gates
_Score)+0.5, by=1))
```



Bill Gates sentiment scores are mostly neutral as one would expect with a sharp peak at 0. The data looks somewhat normally distributed.

Hide

```
h<-hist(Rihanna_Score,col="grey", breaks=seq(min(Rihanna_Score)-0.5, max(Rihanna_Scor
e)+0.5, by=1))
```



As for Rihanna’s sentiment scores, she has mostly positive sentiment score as well with a sharp peak at 0. The data doesn’t look normally distrubuted at all however.

Normality Analysis

We also wished to check for normality after performing the visual inspection.

Hide

```
shapiro.test(Trump_Score)
```

```


    Shapiro-Wilk normality test

data:  Trump_Score
W = 0.89356, p-value < 2.2e-16
```

Hide

```
shapiro.test(Modi_Score)
```

```

    Shapiro-Wilk normality test

data:  Modi_Score
W = 0.92457, p-value < 2.2e-16
```

Hide

```
shapiro.test(Bill_Gates_Score)
```

```

    Shapiro-Wilk normality test

data:  Bill_Gates_Score
W = 0.84999, p-value < 2.2e-16
```

Hide

```
shapiro.test(Rihanna_Score)
```

```

    Shapiro-Wilk normality test

data:  Rihanna_Score
W = 0.86213, p-value < 2.2e-16
```

Hide

```
shapiro.test(Beyonce_Score)
```

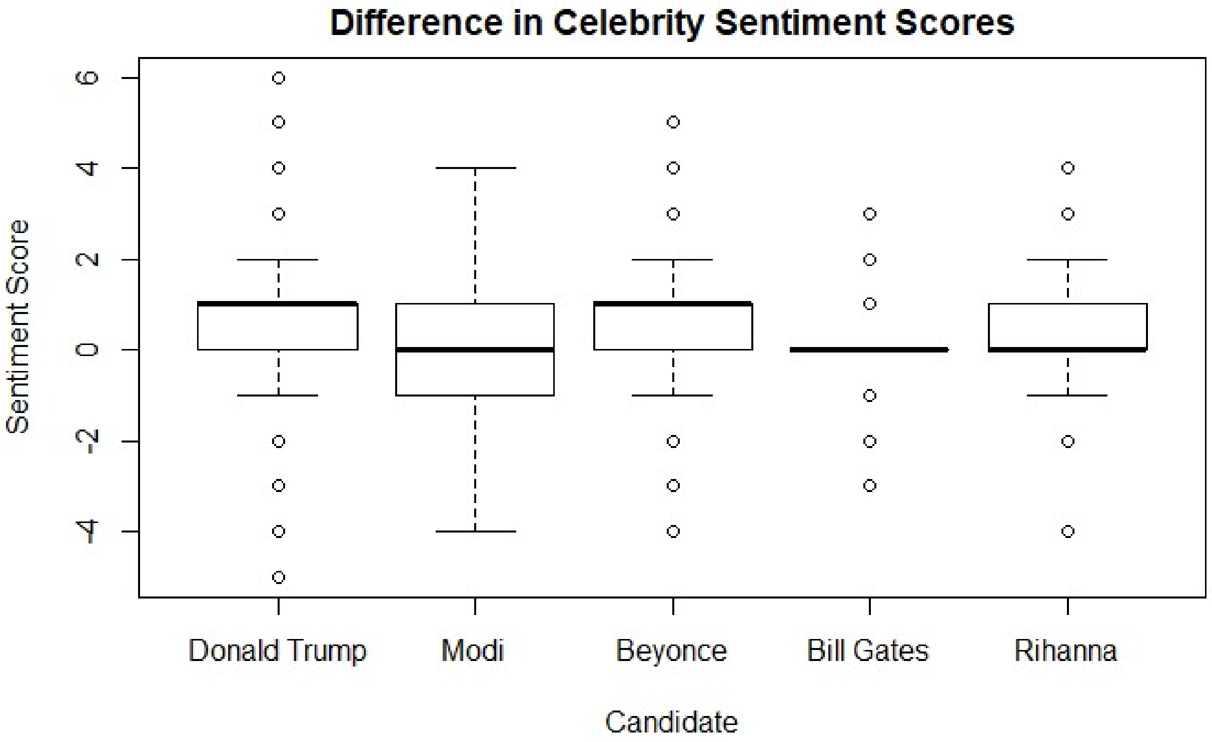
```

    Shapiro-Wilk normality test

data:  Beyonce_Score
W = 0.89648, p-value < 2.2e-16
```

From the shapiro test, we can clearly see that the null hypothesis is rejected(P value<0.001) and can conclude that none of the celebrity sentiment cores are normally distributed.

Mean sentiments

Hide

```
boxplot(score~Candidate,data=semFrame, main="Difference in Celebrity Sentiment Score
s",
        xlab="Candidate", ylab="Sentiment Score")
```



From this boxplot, we can clearly see that there are slight differences in the sentiment scores of the celebrities. We also see that the median for most celebrities is concentrated close to zero except for Donald Trump and Beyonce whose scores are higher. We can also see that Modi's sentiment scores has the widest interquartile range and bill gate's sentiment scores has the smallest. Lastly, we see that the data contains alot of outliers for every celebrity except Modi.

Linear model

Hide

```
model0<- lm(score ~ 1, data = semFrame,na.action = na.exclude) # Model Without Predict
or
model1 <- lm(score ~ Candidate, data = semFrame,na.action = na.exclude) # Model with P
redictor
pander(anova(model0, model1), caption =  "Compare if model1 which includes the indepen
dent variable provides a better fit than without(model0)" )
```

```
--------------------------------------------------
 Res.Df   RSS    Df   Sum of Sq    F     Pr(>F)
-------- ------ ---- ----------- ----- -----------
  4999    5684   NA      NA       NA       NA

  4995    5235   4      448.4     107   1.424e-87
--------------------------------------------------

Table: Compare if model1 which includes the independent variable provides a better fit
than without(model0)
```

Here we use an anova function to compare two linear models, the null model where we have not provided the model with any information about the different celebrities with the extended model which includes the predictor variable or the type of celebrity. We see that the extended model has a much better fit to the data as compared to the null model and can therefore conclude that adding the information about the celebrities is a valid approach.

Hide

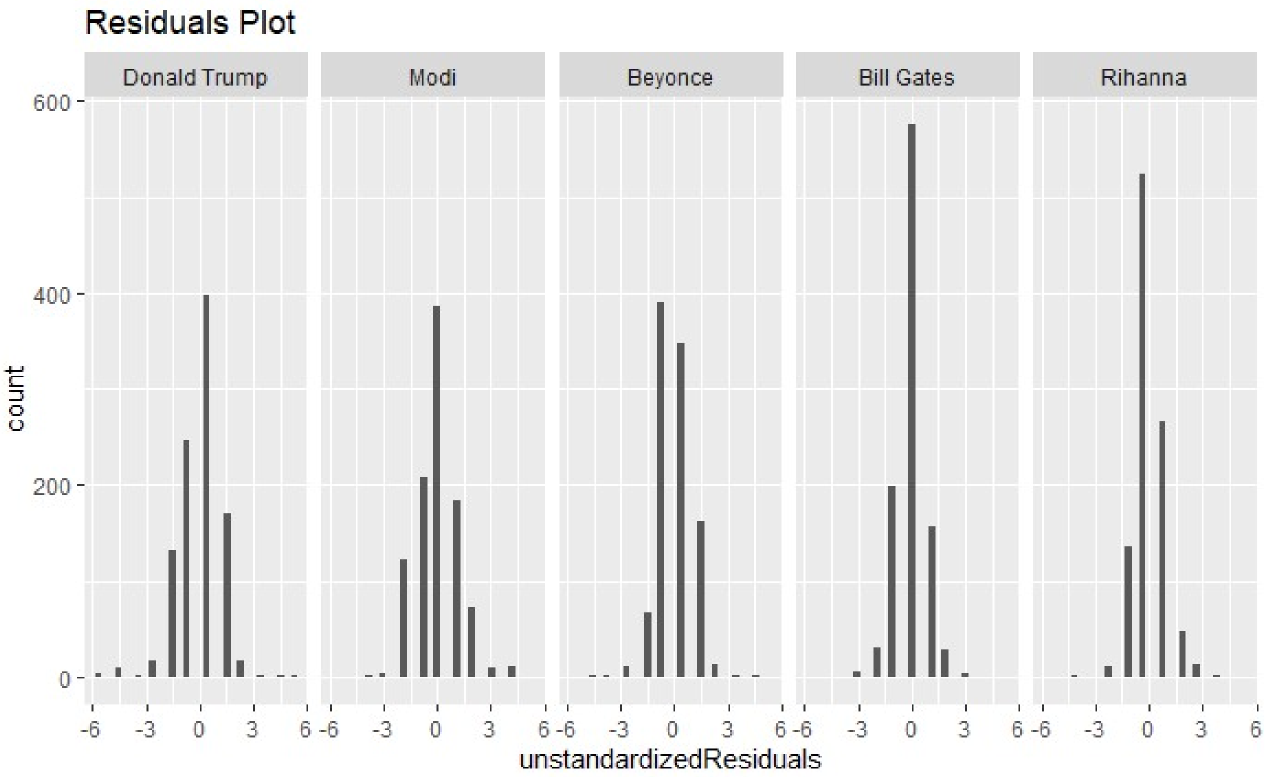
```
pander(oneway.test(score ~ Candidate, data = semFrame, na.action = na.exclude, var.equ
al = FALSE))
```

```
------------------------------------------------------
 Test statistic   num df   denom df       P value
---------------- -------- ---------- -----------------
     109.7          4        2483     3.572e-86 * * *
------------------------------------------------------

Table: One-way analysis of means (not assuming equal variances): `score` and `Candidat
e`
```

Here we run an additional test which allows us to specify that the variance of the different celebrities around their means is not homogoneous(var.equal=FALSE based on the Levene Test performed). We still see that the p value is well below our 5% alpha level and can conclude that there is a significant difference to the fit of our model when celebrity information is included.

Hide

```
semFrame$unstandardizedResiduals <- resid(model1)
hp <- ggplot(semFrame, aes(x=unstandardizedResiduals))+ geom_histogram() + labs(title
="Residuals Plot")
hp + facet_grid(.~Candidate) # Need to interpret this.
```



We also wanted to verify another assumption of the extended model with respect to the residuals being normally distributed. From the initial visual inspection, it doesn’t seem so.

Hide

```
shapiro.test(resid(model1))
```

```

	Shapiro-Wilk normality test

data:  resid(model1)
W = 0.97175, p-value < 2.2e-16
```

And from the shapiro test, we can confirm that with a p value<.0001, the residuals are not normally distributed.

Lastly, since none of the assumptions were valid, we wanted to see what would happen if we do a non parametric test.

Kruskal Wallis Test One Way Anova by Ranks

Hide

```
kruskal.test(score ~ Candidate, data = semFrame, na.action = na.exclude) # where y1 is
numeric and A is a factor
```

```

    Kruskal-Wallis rank sum test

data:  score by Candidate
Kruskal-Wallis chi-squared = 478.77, df = 4, p-value < 2.2e-16
```

Based on the Kruskal-Wallis rank sum test, we believe that the type of celebrity does have an impact on the sentiment scores(Chi-squared=478.77,degrees of freedom =4, p value <0.001).

Post Hoc analysis

Hide

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

```

    Pairwise comparisons using t tests with pooled SD

data:  semFrame$score and semFrame$Candidate

           Donald Trump Modi    Beyonce Bill Gates
Modi       < 2e-16      -       -       -
Beyonce    1            < 2e-16 -       -
Bill Gates < 2e-16      1       < 2e-16 -
Rihanna    4.9e-13      2.6e-10 < 2e-16 2.3e-09

P value adjustment method: bonferroni
```

Due the fact that we are testing multiple hypothesis by comparing the sentiment scores of 5 different celebrities with one another, we need to correct for our alpha level accordingly. This is because as we perform more hypothesis tests, the probability of getting a significant result simply by chance also increases. Here we have corrected for this by using the famous Bonferroni correction where we simply divide our alpha level with the total number of pairwise hypotheses tests we perform. From this analysis we can see that most of the celebrities still have a significant difference in sentiment scores with a few notable exceptions such as Beyonce vs Donald Trump and Bill Gates vs Modi who do not possess significant differences in mean sentiment scores.

Report section for a scientific publication

As can be seen in our conceptual model, we were interested in looking at the effect of the identity of a celebrity to the sentiment score based on tweets. In order to do this, we compared the base model(without the predictor variable) with a model extended with the predictor variable i.e identity of a celebrity. We saw from our anova analysis of both models that including celebrity identities did have a significant effect (F(4,4995)=107, p.<0.001) on the fit of our model. Furthermore, we also performed a post-hoc analysis correcting for the testing of multiple hypothesis via the bonferonni correction. We found that most celebrities still possessed a significant difference(adjusted p-value < 0.05) in mean sentiment scores with a few exceptions such as (M:0.05,SD:0.8)Bill Gate’s sentiment score as compared to (M:0.06,SD:1.2)Modi(adjusted p-value=1.0) and (M:0.63,SD:0.96)Beyonce’s sentiment score as compared to (M:0.59,SD:1.2)Donald Trump(adusted p-value=1) which was pretty interesting to see. We would also like to mention that we performed a non parametric test due to the fact that the normality and homogeinity of variance assumptions for the sentiment scores of different celebrities couldn’t be validated, we found through the Kruskal Wallis rank sum test that celebrity identities do have a significant effect (χ2(4, N = 5000) = 478.77, p. < 0.001) on the fit of our model to the data.