Coursework Assignment A - Group 12
CS4125 - Seminar Research Methodology For Data Science

Nivedita Prasad – 4712099
Aishwarya Shastry – 4743016
Liliana Oliveria – 4767306

March 20, 2018

Contents

1 Part 1 – Design and set-up of true experiment **2**

2 Part 2 - Generalized linear models **3**

2.1 Question 1 Twitter sentiment analysis(Between groups - single factor) . . . 3
2.2 Question 2 – Website visits (between groups – Two factors) . . . 9
2.3 Question 3: Linear regression analysis . . . 11
2.4 Question 4 Logistic regression analysis . . . 16

3 Part 3 - Multilevel Models **18**

Appendices **35**

A Part 2 - Question 1 **35**

A.1 Your Twitter . . . 35
A.2 Sentiment3 . . . 35
A.3 Twitter Analysis . . . 36

B Part 2 - Question 2 Web Page Visits **39**

C Part -2 Question 3: Linear regression analysis **41**

D Part 2 - Question 4 Logistic regression analysis **44**

E Part 3 - Multilevel Models **45**

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

1. **Write a plan for conducting an experiment on group of human test subjects. As a group you are allowed to select your own topic for this experiment.**

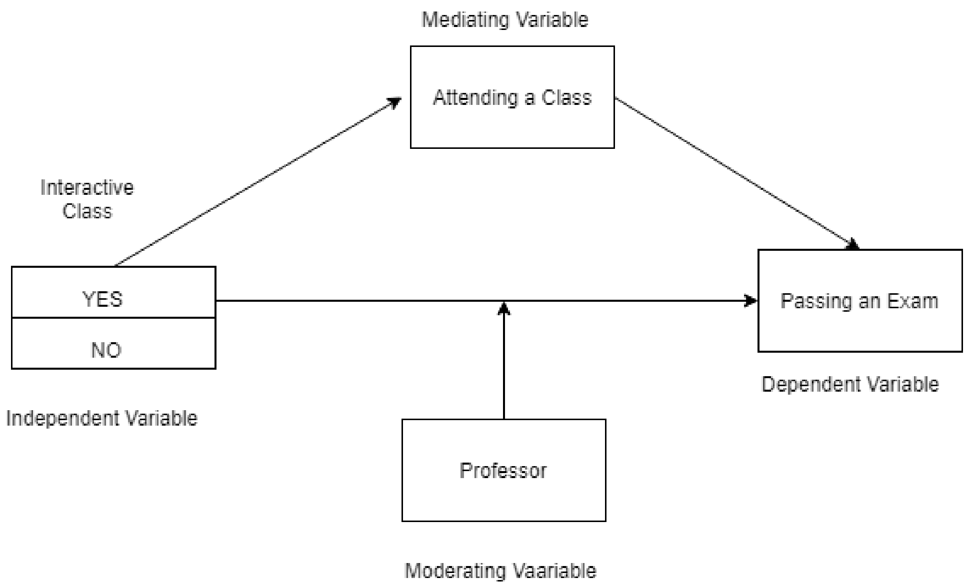

Figure 1: Conceptual Model

The motivation for this research is that the knowledge of the passing rate of students in any class helps us understand the importance of factors such as the class being interactive or not is influencing the student's performance.The interactiveness kindles a motivation to learn the subject or concept better with positive thoughts. The performance on a final exam is directly related to a student passing an exam as it contributes to the overall grade of the course and to get a final degree with better understanding of the content taught in class. The student will apply the knowledge gained in an course not only by passing an exam but also by solving real world problems, for example during a good research or for a client. This all sums up to passing an exam and the factors affecting it is eventually making an impact during the course of the subject.

The theory underlying the research question, which is whether interactiveness plays an important role in passing an exam is explained here. Active learning is where the students engage in interactive learning process .The independent variable we choose is interactive learning, and in [1], the authors have discussed on active learning and said that this actually leads to better student attitude and thinking about an approach and on writing skills. Active Learning is better than traditional classes with lectures as it tries to cultivate the habit of asking questions about a concept for better understanding. In [2], the authors have mentioned that on team projects, discussions that are challenging and peer reviews students tend to learn more effectively. For our moderating variable we chose the *Professor's Intelligence Quotient (IQ)* .In [3] the authors state that the content specific knowledge, vocabulary of the professor, expressive language used in class and skills used to help memory retention ,perpetual skills, coordination of non-verbal behavior such as smiling, movement about the classroom, and relaxed body position of the professor contributes his IQ. This in-turn affects the student's attention and interest the professor cultivates in the students. Hence we have decided on the following variables after research as discussed in the next paragraph.

The relationship is best explained with the "attendance in class". This will be a mediating variable which links the independent variable "interactive class" and the dependent variable "Passing the exam". This is because paying attention boosts understanding of the concept taught in class in a relaxed manner when it is revised. The third factor that might influence the strength of the relationship between the interactiveness of a class and the Passing of an exam is the moderating variable, Professor. The Professor's IQ makes a difference as the students may get involved better in the class rather than self study and on-line lectures.The "Professor's" IQ and quality of teaching acts as an moderating variable. The professor can interact with the students by having a diversity in assessments within the student groups and awarding the students with marks or appreciation. This makes the attention of a student to be fully focused on the content the professor is teaching ,influencing better memory retainment during exams resulting in success in the exam.

The Null Hypothesis H0 is the research question. The research question that will be examined in the experiment is *"What is the effect of class interactiveness on student's success in exam"*. The alternative Hypothesis H1 is Otherwise. (The independent variables do not affect the student's success in the exam). The Conceptual Model is shown in Fig. 1.

Hence, the dependent variable is Passing an Exam, the independent variable is Interactiveness of a class which can be an Yes or No, the mediating variable is Attendance of a Pupil in the class, the moderating variable is the Professor's IQ i.e., the quality of the lectures by the professor. The Experimental Design is displayed in Fig. . This figure shows the True Experimental Design. This is a true experiment as the random allocation of participants that is the students is done and we as the experimenter have complete control over all the variables that influence the conceptual model and we are able to control the independent variable by deciding whether the class is interactive or not. The experimental procedure depends on the research question, number of needed participants, based expected

Table 1: Experimental Design

Between Subject	**Condition 1**	**Condition 2**
Interactive Class	Yes	No
Attending Class	Yes	Yes
Professor IQ	High	High
Participants	50	50

effect size, number of available participants and duration of the experiment. This experiment is a Two groups design - Post-test-only randomized controlled trail (RCT) . The first step is to split the group of students into two groups based on their answer to whether they feel that interactiveness of a class matters to them or no (Yes/No).Then as the second step we check the influence of mediating, moderating variable depending on the participant's choice of the class being interactive or not. We *measure* the dependent variable "Passing an exam" in the form of *test scores* . The participants chosen are 100 in number who are students with ages 21-24, Gender- Male and Female and 2 experimental groups with 50 people each based on whether the answer a YES or a NO to the class being interactive would help them. The Suggested statistical analyses is that we may analyze the correlation between mediating and the moderating variables and We want to perform the one way ANOVA test based on the answer a YES or a NO to the class being interactive.

2 Part 2 - Generalized linear models

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

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

   Fig. 2 shows the conceptual model of the sentiment difference of tweets related to different celebrities

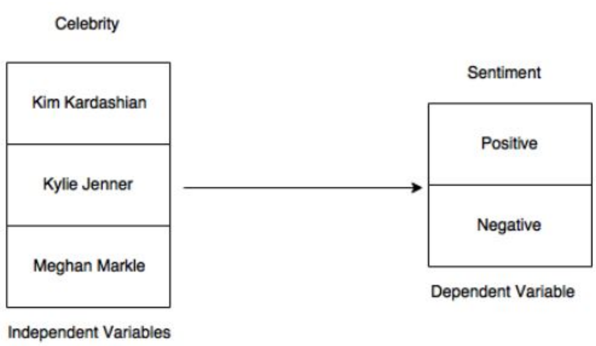

Figure 2: Conceptual Model Twitter

2. **Analyze the homogeneity of variance of sentiments of the tweets of the different celebrities.**

   Homogeneity of variance is used to access the equality of variances of a variable calculated for one or more groups of samples drawn from population. Levene test assess this assumption which is a null hypothesis. The Analysis of Fig 3 is that, the Pr(> F) value is less than 0.05, the standard value of p. Therefore null hypothesis is not true. Hence, there is a difference in variance of the tweets between each of the 3 celebrities i.e. *the variances are not equal.*

```
> leveneTest(semFrame$score, semFrame$Candidate, center = median)
Levene's Test for Homogeneity of Variance (center = median)
        Df F value    Pr(>F)
group    2  185.56 < 2.2e-16 ***
      2997
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 3: Levene Test for Homogeneity of variance

3. **Graphically examine the variance in tweets' sentiments for each celebrity.**

   Variance of tweets : The graphical examination of the celebrity "Kim Kardashian" Fig 4 and 5 show that there is a more number of neutral sentiments and less number of negative sentiments and almost 100 positive tweets. As for the second celebrity, "Kyile Jenner" Fig 6 and 7 show that the there is more number of neutral tweets but more number of positive tweets with sentiment score 1,2 and very less sentiments with score 3, but the negative sentiment score is less compared to the amount of positive tweets which implies that this particular celebrity is more favored by the public. The third celebrity "Meghan Markle" as shown in Fig 8 and 9 show that the neutral sentiment is same as the other two celebrity and the positive sentiment is high with almost 300 tweets with score 1 and nearing 50 for score 3, but this celebrity shows a negligible high sentiment scores of 3,4,and almost 6 also that implies this celebrity is most favored by the people at the moment compared to the other two celebrities under consideration.

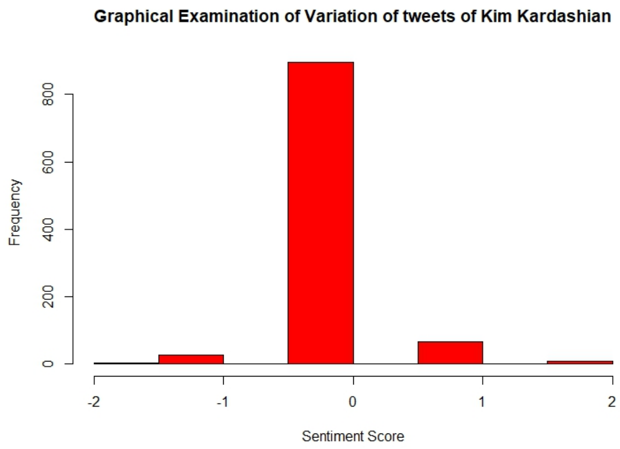

Figure 4: Graphical Examination of Variation of sentiment score - Kim Kardashian

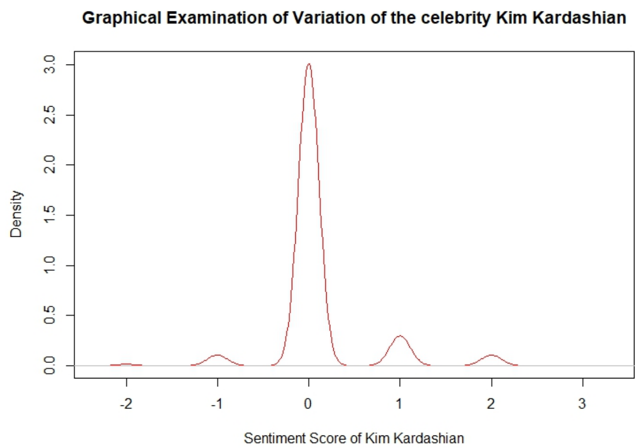

Figure 5: Density Plot - Kim Kardashian

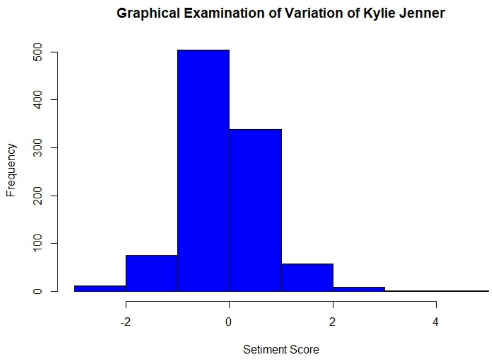

Figure 6: Variation of sentiment score - Kylie Jenner

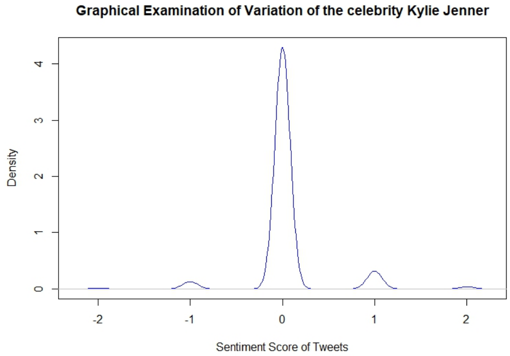

Figure 7: Density plot - Kylie Jenner

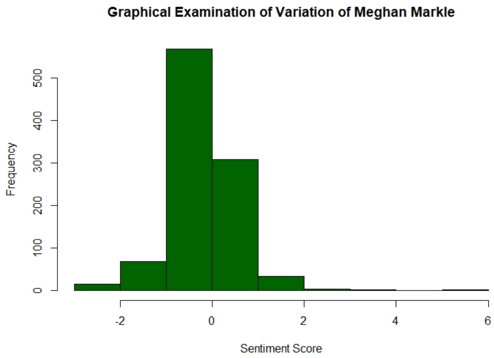

Figure 8: Variation of sentiment score - Meghan Markle

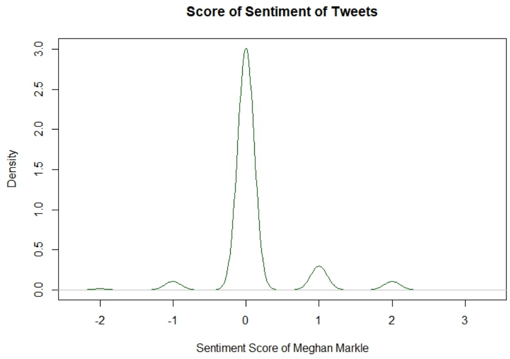

Figure 9: Density plot - Meghan Markle

4. **Graphically examine the mean sentiments of tweets for each celebrity.**

   Mean of Sentiments of Celebrities : As in Fig 10 , 11 , the mean and standard deviation of the all the three celebrities are calculated and plotted. The means of Kim Kardashian is 0.059, Kylie Jenner is 0.414, Meghan Markle is 0.278. This is can be further extended to levene test's result, that since the variance of groups are not same , the means cannot be the same too.

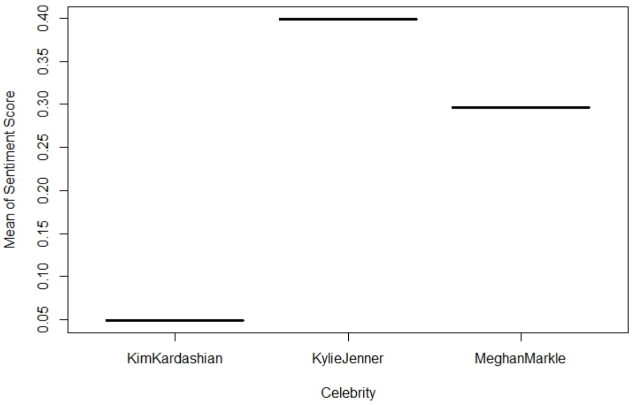

Figure 10: Mean of Sentiment Score

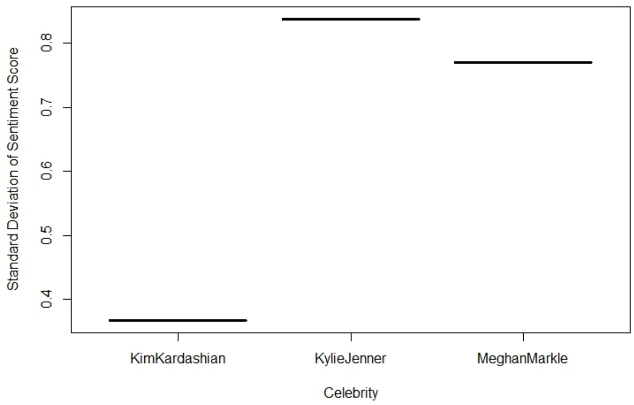

Figure 11: Standard Deviation of Sentiment Score

5. **Use a linear regression model to analyze whether the knowledge to which celebrity a tweet relates has a significant impact on explaining the sentiments of the tweets.**

   Linear Model to Analyze Sentiments : Analysis of Variance (ANOVA) is a statistical method used to test differences between two or more "means"; as inferences about means are made by analyzing variance. Using ANOVA's F test we can check if there is a linear relationship between the independent and dependent variable. We assume the variance is homogeneous for all independent variables (celebrities), then the Interval of data is same, the data is normally distributed and the data is independent from each other. These four assumptions are tested using ANOVA function in R that is applied on models with a predictor and without the predictor. Finally the models are compared to check which model fits the data better. Fig 12, 13, 14, 15, 17, 16, 18. Since the W = 0.74657, p-value $< 2.2e-16$ in Shapiro test, the p value is less that the assumed alpha value which proves that the data is not normally distributed. Sum of the squares of the deviations of observation from their mean i.e the total variance of the observations for models with predictor is 1434.8 and without predictor is 1499. If this total variance of observations is divided by the degrees of freedom then the means squares is calculated for model with is 32.080. To compare if model1 provide better fit than model0, we check the following after performing anova (analysis of variance test) on 2 models model 0 and model 1. Model 0 has no degrees of freedom i.e there is no predictor independent variable (not compared with any celebrity).Model 1 has 2 degrees of freedom i.e we are comparing with all celebrities. In Model 1, 64.161 is the difference between Residual sum of squares (RSS) of 1499.0 and 1434.8 . The lesser the value of RSS the better the model fits. The mean squares is 32.080. Hence the **Model 1** fits better with predictor (celebrities as independent variable). The F value 67.007 is calculated using RSS values, number of independent variables used in model 1 and model 0 to help us understand the significance of "Candidate" predictor variable on the sentiment score. The F value in combination with p value . The assumed alpha value and p value is 0.05. So a coefficient marked *** is one whose p value < 0.001, one whose coefficient is marked ** is $p < 0.01$ and one whose coefficient is marked * is $p < 0.05$. Our P value is marked with 3 asterisk and so it is lesser than 0.001 and hence the null hypothesis of one candidate affecting the other can be rejected using anova.

```
> model0<- lm(semFrame$score ~ 1, data = semFrame) #model without predictor
> summary(model0)

Call:
lm(formula = semFrame$score ~ 1, data = semFrame)

Residuals:
   Min     1Q Median     3Q    Max 
-3.248 -0.248 -0.248  0.752  5.752 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.24800    0.01288   19.25   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7057 on 2999 degrees of freedom
```

Figure 12: Model 0

```
> model1 <- lm(semFrame$score ~ semFrame$Candidate, data = semFrame)  #model with predictor
> summary(model1)

Call:
lm(formula = semFrame$score ~ semFrame$Candidate, data = semFrame)

Residuals:
   Min     1Q Median     3Q    Max 
-3.399 -0.296 -0.049  0.601  5.704 

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)    
(Intercept)                    0.04900    0.02183   2.244   0.0249 *  
semFrame$CandidateKylieJenner  0.35000    0.03088  11.335  < 2e-16 ***
semFrame$CandidateMeghanMarkle 0.24700    0.03088   7.999 1.77e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6905 on 2997 degrees of freedom
Multiple R-squared:  0.04333,	Adjusted R-squared:  0.04269 
F-statistic: 67.86 on 2 and 2997 DF,  p-value: < 2.2e-16
```

Figure 13: Model 1

```
> anova(model0,model1, test = "F") #compare if model1 provide better fitt than model0
Analysis of Variance Table

Model 1: semFrame$score ~ 1
Model 2: semFrame$score ~ semFrame$Candidate
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)    
1   2999 1493.5                                  
2   2997 1428.8  2    64.706 67.863 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> anova(model1) #print results in anova format
Analysis of Variance Table

Response: semFrame$score
                     Df  Sum Sq Mean Sq F value    Pr(>F)    
semFrame$Candidate    2   64.71  32.353  67.863 < 2.2e-16 ***
Residuals          2997 1428.78   0.477                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 14: Anova - Comparing which model fits better

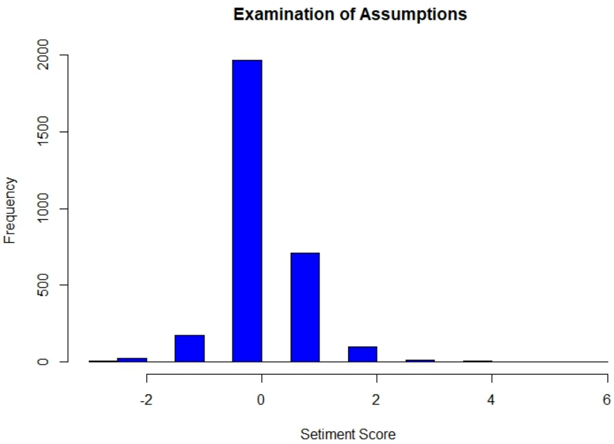

Figure 15: Examination of Assumptions

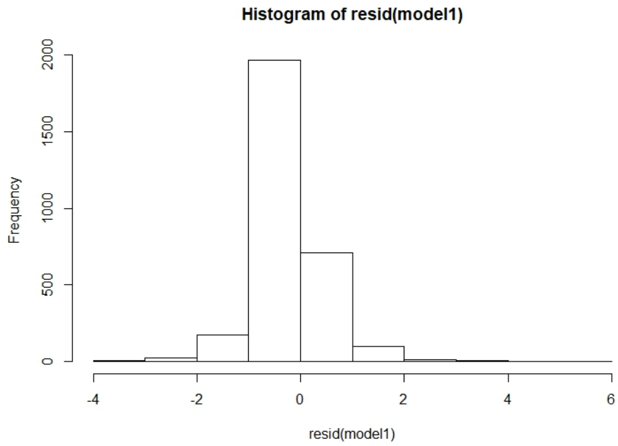

Figure 16: Histogram of Residual Model 1

```
> shapiro.test(semFrame$score)

        Shapiro-Wilk normality test

data:  semFrame$score
W = 0.74897, p-value < 2.2e-16
```

Figure 17: Shapiro Wilk normality test

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

        Shapiro-Wilk normality test

data:  resid(model1)
W = 0.86513, p-value < 2.2e-16
```

Figure 18: Shapiro Wilk normality test resid model

6. **Conduct a post-hoc analysis with Bonferroni correction to examine which of celebrity tweets differs from the other celebrity tweets.**

   Post hoc Analysis : There is a chance of making type 1 error (incorrectly rejecting a null hypothesis). The Bonferroni correction is applied to adjust the p value by testing the individual hypothesis by dividing overall α level with m number of null hypothesis. This calculates pairwise comparisons between group levels with corrections for multiple testing. The results are shown in Fig 19

```
> #post hoc analysis
> 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

             KimKardashian KylieJenner
KylieJenner  < 2e-16       -
MeghanMarkle 5.3e-15       0.0026

P value adjustment method: bonferroni
```

Figure 19: Post Hoc Analysis

7. **Write a small section for a scientific publication, in which you report the results of the analyses of point 2-6, and explain the conclusions that can be drawn.**

The analysis drawn by executing points 2 to 6 are as follows. While applying *Levene Test - Homogeneity of Variance :* we found that there is a difference in variance of the tweets between each of the 3 celebrities . We inferred that our third celebrity-Meghan Markle is most favored by the people at the moment compared to the other two celebrities under consideration. The *mean and standard deviation* of the all the three celebrities are further extended to levene test's result and since the variance of groups are not same , the means cannot be the same too. The *Linear Model to Analyze Sentiments* uses The **Model 1** fits better with predictor (celebrities as independent variable). The analysis using *Post hoc Analysis :* is that there is a chance of making type 1 error (incorrectly rejecting a null hypothesis). The Bonferroni correction is applied to adjust the p value by testing the individual hypothesis by dividing overall α level with m number of null hypothesis. This calculate pairwise comparisons between group levels with corrections for multiple testing. The results are shown in Fig 19

8. **Include the annotated R script (excluding your personal Keys and Access Tokens information) in the appendix of the report.** - Code Available in Appendix A

2.2 Question 2 – Website visits (between groups – Two factors)

1. **Make a conceptual model underlying this research question :**

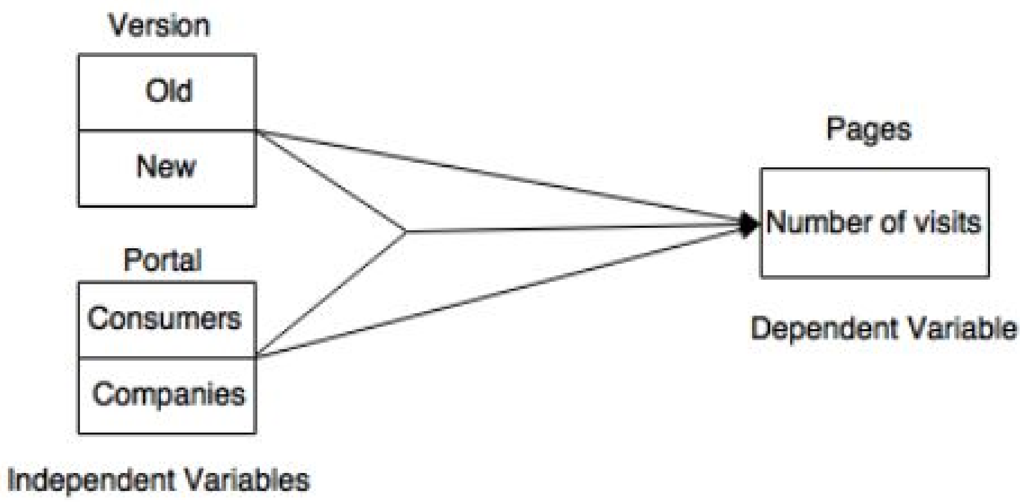

Figure 20: Conceptual Model for Web Page visits

2. **Graphically examine the variation in page visits for different factors levels.**

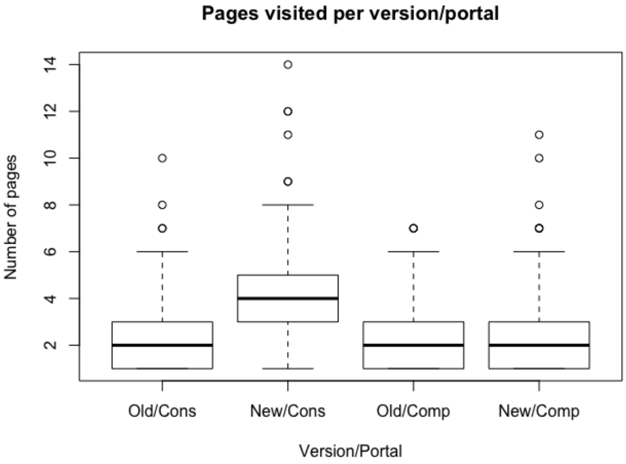

Figure 21: Pages visited per Version/Portal

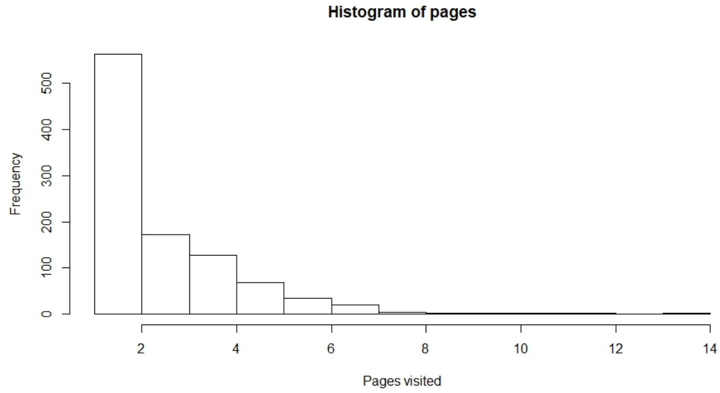

Figure 22: Pages distribution

We analyze that most of the people just visit 1 page and while the number of pages increases, the number of visits decreases. We can also observe that there is a increasing number of pages visited by the consumers when comparing between old and new version. However, when considering companies' portal the number of visits tends to be the same independently of the version. The data was treated and values from version and portal were changed to not lead to misleading conclusions such as mean of those values, which does not make sense.

3. **Statistically test if variable page visits deviates from normal distribution**

   Even though it is clear to see from Figure 22 that the variable Pages do not follow a Normal distribution, a normality test will be performed. The null hypothesis, H0 is that *Variable pages follow a normal distribution* and the alternate hypothesis, H1 is Otherwise. The α value is 0.05 . The results of the Shapiro-Wilk normality test we performed is from the dataset "Question2" of the column pages with the value of W is 0.83076, p-value < 0.05. Our interpretation of whether the variable web page visits deviates from normal distribution is that, the p-value $<= 0.05$, then the NULL hypothesis is rejected which consists in samples coming from a Normal distribution.If p-value > alpha (significance level) it means that there is no evidence to reject null hypothesis. Otherwise we reject that is our data is not normally distributed. We *reject H0* and conclude that there is no strong evidence that the variable pages follows a normal distribution. Having in consideration the distribution of the variable Pages, it looks like a Poisson distribution.

4. **Conduct a model analysis, to examine the added values of adding 2 factors and interaction between the factors in the model to predict page visits.**

   For this we consider the null hypothesis, H0 that *the version of the website and its portal do not influence the number of pages visited* and the alternate hypothesis H1 is that there is a relationship between the version and portal of a website and the number of pages visited. Model 0 is the basic model without any predictor. Model 1 and 2 add the version and the portal to Model 0, respectively. Model 3 adds both portal and version, and model 4 considers all the mentioned combinations and on top of that the combination between portal and version. The analysis found a significant main effect ($z(995) = 11$, $p < 0.05$) for the version and a significant two-way interaction effect ($z(995) = 5$, $p < 0.05$) between version and portal factors. We analyzed that there was no statistically significant difference in mean in pages' visit between portals ($p = .281$), but there were statistically significant differences between versions ($p < .001$).

5. **If the analysis shows a significant two-way interaction effect, conduct a Simple Effect analysis to explore this interaction effect in more detail.**

   A generalized linear model was fitted on the number of pages visited, taking the version and the portal as independent variables, including the two-way interaction between these factors. The analysis found a significant main effect (z(995) = 11, $p < 0.05$) for the version and a significant two-way interaction effect (z(995) = 13, $p < 0.05$) between version and portal factors.

6. **Write a small section for a scientific publication, in which you report the results of the analyses of point 2-6, and explain the conclusions that can be drawn.**

   A two-way ANOVA was conducted that examined the effect of a portal and version of an website in the number of pages visited. There was a statistically significant interaction between the effects of version and portal of the page in the number of pages visited, z(995) = 5, $p < 0.05$. Simple main effects analysis showed that new versions had significantly more webpage's visits than old versions, but there were no differences between portals.

7. **Include annotated R script in the appendix of the report.** - R Code available in Appendix B

2.3 Question 3: Linear regression analysis

1. **Make a conceptual model underlying this research question**

   Research question : How does costs and earnings affect the Profit for RKO movies?

   The *Independent variables* are Re-release that is if a movie has been released again value=1 or 0, Production cost which is the cost of the producing a movie , Total revenue that refers to total earnings made from sale of tickets both in the same nation and abroad , the Distribution cost for making the film available in the market and it includes distribution in movie theaters, CD's and promotion and finally Year that is the Year movie was released-in. The independent variables are of interval(ratio) level. The *Dependent variable* are the Profits obtained by RKO movies in the year 1930-1941. The fig. 23 shows the conceptual model of Profits obtained from costs and earning of 155 RKO movies between 1930-1941. We examine profits obtained from 155 RKO films from 1930-1941 from our dataset.[4]

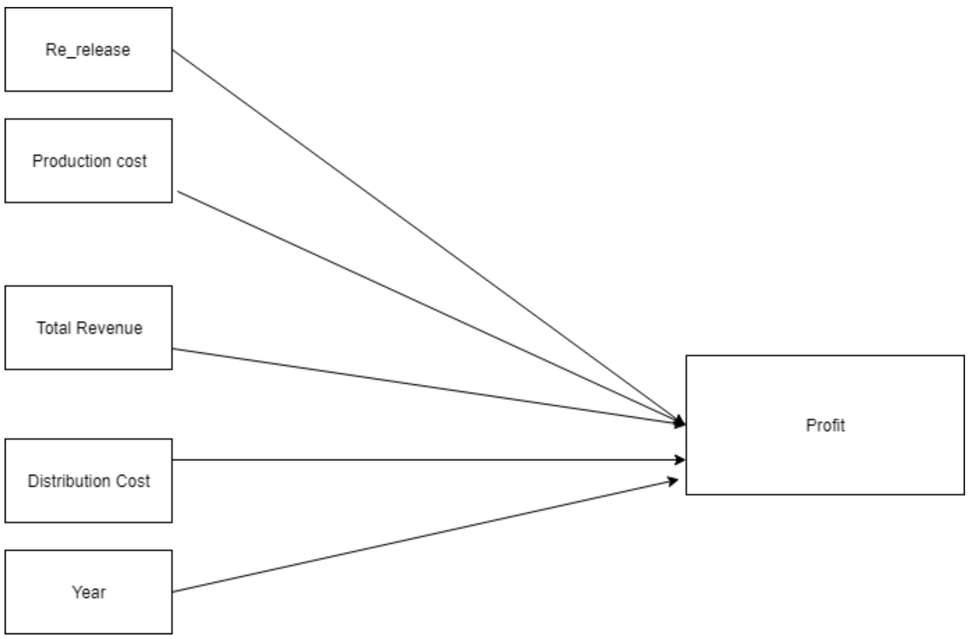

Figure 23: Profit Conceptual Model

2. **Graphical analysis of the distribution of the dependent variable, e.g. histogram, density plot**
   The figures fig 24 ,25 shows the density plot and the histogram of the dependent variable and we can infer that most of the profit values are around 0-200k$.

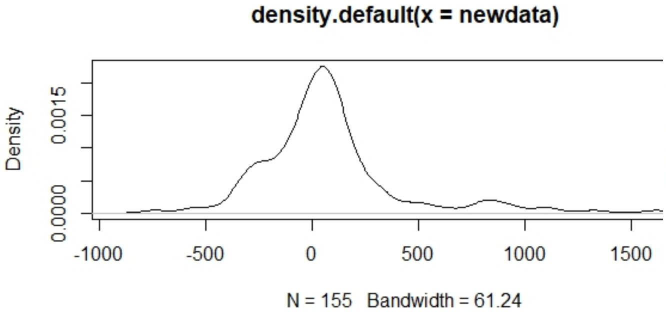

Figure 24: Distribution of the Profits - Density Plot

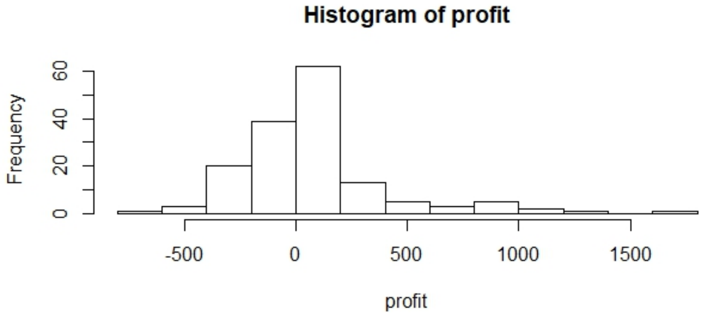

Figure 25: Distribution of the Profits - Histogram

3. **Scatter plots between dependent variable and the predictor variables**
   The *Analysis*, we can infer from Fig 26 that Profit increases with production cost till a value between 500k and 750 k and then decreases which goes well with the assumption that High budget movies earn more profits but as Production cost increase beyond a value, profits decrease. Fig 28 shows that profits normally increase with Total revenue.Same can be inferred for fig 30 except some outliers.

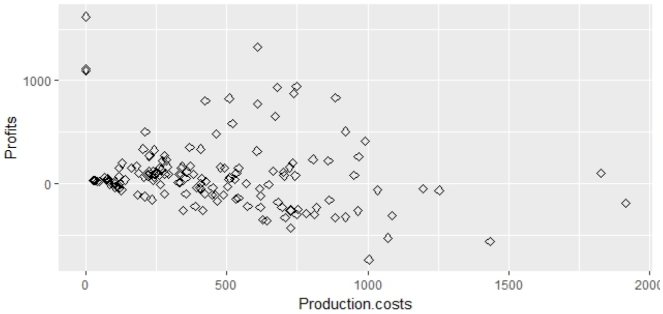

Figure 26: Profit Vs Production Cost

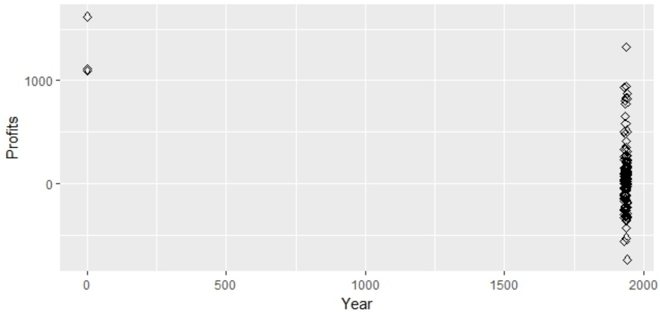

Figure 27: Profit Vs Year

Figure 28: Profit Vs Total Revenue

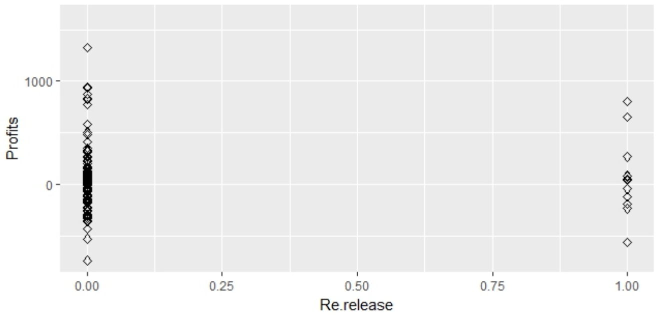

Figure 29: Profit Vs Re-release

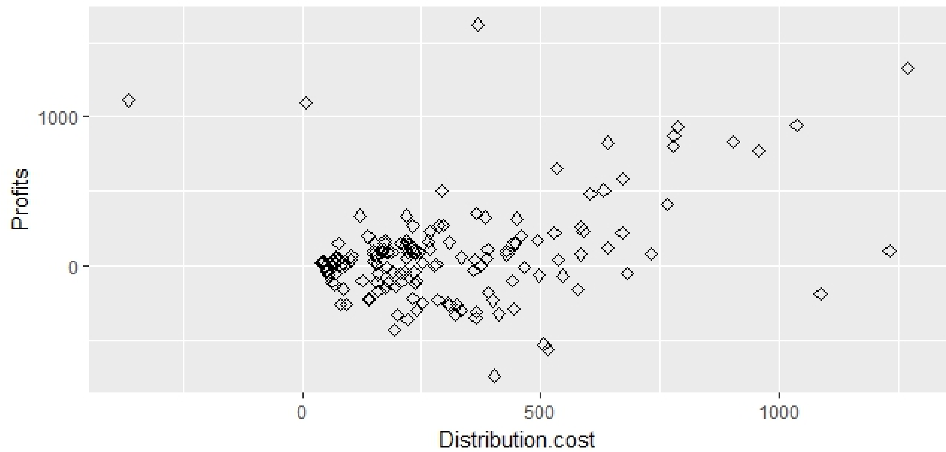

Figure 30: Profit Vs Distribution Cost

4. **Conduct a multiple linear regression (including confidence intervals, and beta-values)**

   We first conduct multiple linear regression on our model to calculate confidence intervals and beta values. We observe that significant beta coefficients are:Production costs and total revenue. For every 1 unit increase Total Revenue, profit increases by 2.178737378 and for every 1 unit decrease in production cost, the profit increases by 1.130865991. The results are displayed in Fig 31, 32.

```
Residuals:

    Min      1Q  Median      3Q     Max
-10.144  -4.675  -2.632  -0.326 193.550

Coefficients:
                     Estimate Std. Error  t value Pr(>|t|)
(Intercept)         87.981316  20.915680    4.206 4.51e-05 ***
Re_release.c        -2.433902   6.605553   -0.368    0.713
Production_costs.c  -0.993990   0.007826 -127.016  < 2e-16 ***
Total_Revenue.c      0.986485   0.011973   82.391  < 2e-16 ***
Distribution_cost.c -0.975275   0.031868  -30.604  < 2e-16 ***
Year.c              -0.569278   0.557609   -1.021    0.309
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 31: Residual and Coefficient

```
Residual standard error:

 21.66 on 146 degrees of freedom
  (6 observations deleted due to missingness)
Multiple R-squared:  0.9947,    Adjusted R-squared:  0.994      6
F-statistic:  5514 on 5 and 146 DF,  p-value: < 2.2e-16


confidence interval

  2.5 %      97.5 %
(Intercept)          46.6447032 129.3179297
Re_release.c        -15.4887572  10.6209536
Production_costs.c   -1.0094562  -0.9785236
Total_Revenue.c       0.9628223   1.0101487
Distribution_cost.c  -1.0382567  -0.9122931
Year.c               -1.6713062   0.5327509

Beta coeff

lm.beta(mod1)
       Re_release.c  Production_costs.c     Total_Revenue.c
       -0.002243635        -1.130865991         2.178737378
Distribution_cost.c              Year.c
       -0.809095196        -0.006584569
```

Figure 32: Residual standard error-confidence interval-Beta coeff

5. **Examine assumptions underlying linear regression. E.g collinearity and analyses of the residuals, e.g. normal distributed (QQ plot), linearity assumption, homogeneity of variance assumption. Where possible support examination with visual inspection.**

   For Co linearity, we test the assumptions underlying linear regression. We infer that the high VIF Values of Total Revenue and Distribution Cost exceeds our rule of thumb of 10 indicating that these variables are highly co -related. The low values of Distribution cost and Total revenue go with our assumption that these variables are correlated.

```
Re_release.c  Production_costs.c      Total_Revenue.c
          1.027724            2.197179           19.382482
Distribution_cost.c             Year.c
         19.373570            1.152992
```

Figure 33: Co-linearity

```
Tolerance

       Re_release.c  Production_costs.c     Total_Revenue.c
         0.97302392          0.45512903          0.05159298
Distribution_cost.c              Year.c
         0.05161671           0.8673087
```

Figure 34: Distribution cost and Total revenue

Analysis of residuals:

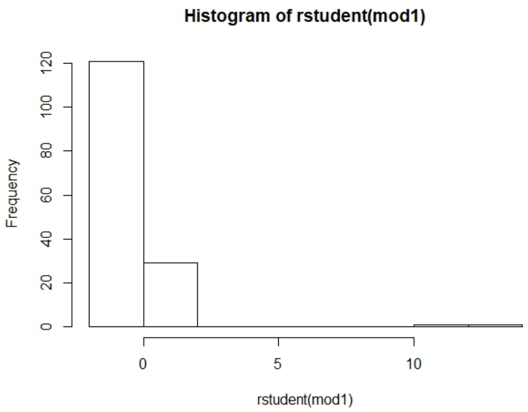

Figure 35: Histogram of rstudent - model 1

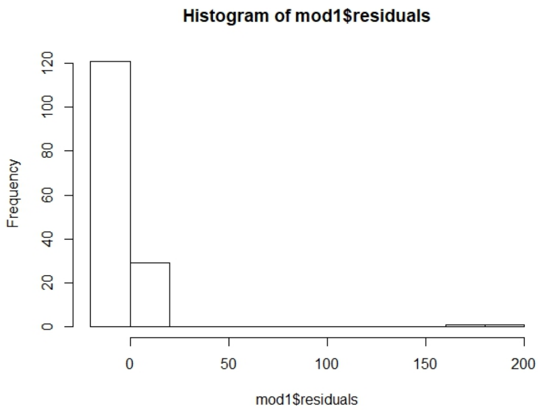

Figure 36: Residual Histogram of rstudent - model 1

Normal QQ plot:

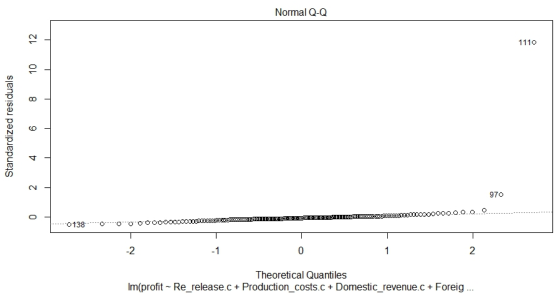

Figure 37: Normal QQ Plot

We can see few outliers with values 97 and 111. We tend to ignore them.

6. **Examine effect of single cases on the predicted values (e.g. DFBeta, Cook's distance)** We finally examine effect of single cases on predicted values through DFbeta and Cook's distance. From the Cook's plot we can tell that 90th and 107th observation strongly influences our fitted model. Then from the DFbeta plot , we can see that 93rd and 107th observation strongly influences our model.

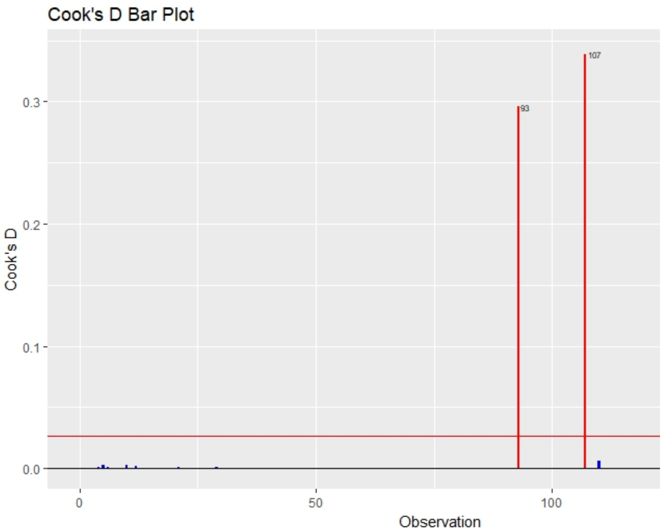

Figure 38: Cook's Plot

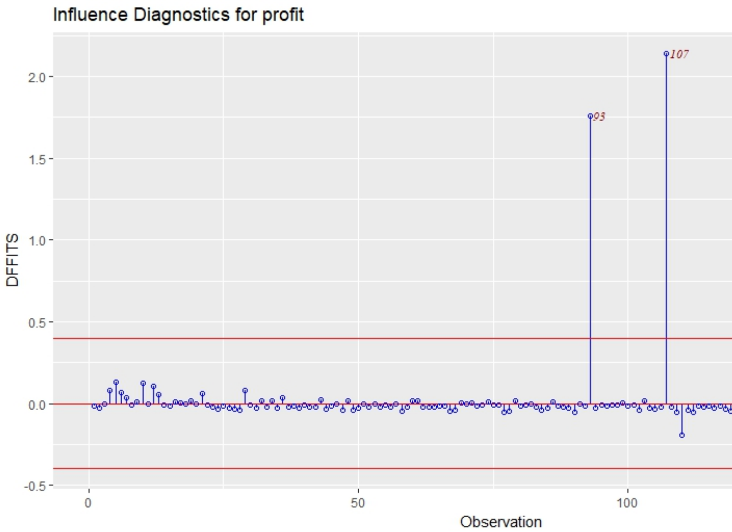

Figure 39: DFbeta for Observations

7. **Write a small section for a scientific publication, in which you explain the data set examined, report the results of the analyses of point 2-6, and explain the conclusions that can be drawn.**

   The data set we took was Costs and Earnings of RKO films from 1930-1941. The dataset had attributes Re-release, Production costs, Domestic Revenue ,Foreign Revenue, Total Revenue, Profits, Distribution Cost($1000s), Dist Cost/Revenue ,Dist Cost/Prod Cost, Year 100*Profit/(Prod+dist cost) . In our model we consider the 5 attributes that affect Profits as seen in the conceptual model. We can observe that most of the profits obtained are in the range 0-200k$. We can also observe that Profits increase with Distribution costs. We can thus conclude that Profit highly depends on Total Revenue and Production costs. Our model almost follows a linear relationship as seen in Normal qq plot.

8. **Include annotated R script in the appendix of the report**

   Refer Appendix C

2.4 Question 4 Logistic regression analysis

1. **Make a conceptual model underlying this research question**

   Our dataset has the following variables The *dependent* variables is Reason of discharge of patient calculated on two levels: 1. By Physician 2. Other Reason. The *independent* variables is Length of Stay that is the number of days that the patient stayed at the center, Month Admitted-The month that the patient was admitted to the center: 1–6. Our independent variables are in Interval level according to our dataset.

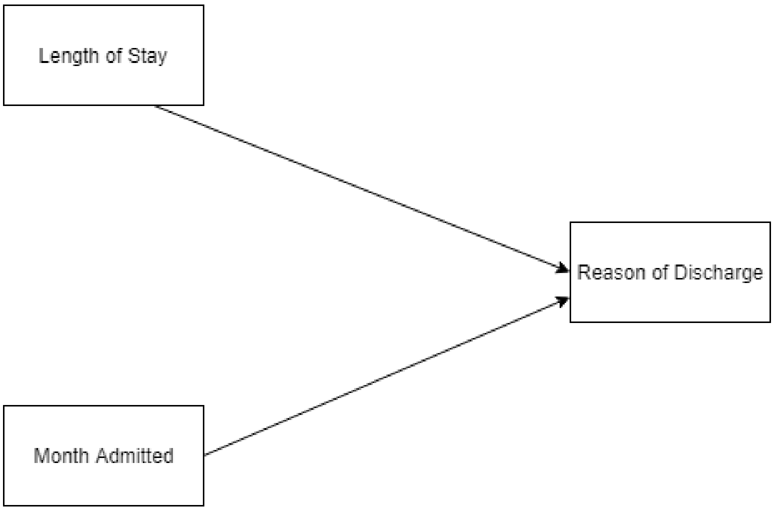

Figure 40: Patient Length of Stay Model

2. **Conduct a logistic regression, examine whether adding individual indicators in the model improves the model compared to Null model. Make a final model with only significant predictor(s). For this model, calculate the pseudo R-square. Calculate the odd ratio for the predictors and their confidence interval**

   We conducted logistic regression and found the below values from ANOVA. Model 3 is the best as it has $p<0.05$ as shown in the figure Fig 41. We take Model 3 for our further analysis.

```
Model 1: dataset$Reason ~ 1
Model 2: dataset$Reason ~ length_mean
Model 3: dataset$Reason ~ month_mean + length_mean
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1     57     68.324
2     56     65.991  1   2.3329  0.12667
3     55     61.199  1   4.7921  0.02859
```

Figure 41: ANOVA results

```
Pseudo R^2 for logistic regression
Hosmer and Lemshow R^2:   0.104
Cox and Snell R^2:        0.116
Nagelkerke R^2:           0.167
```

Figure 42: PesuedoR2

   The Psuedo R square value for the above model are shown in fig refPesuedoR2. Our pseudo r square shows Variance of 10.4 percent(Hosmer and Lemshow), 11.6 percent (Cox and Snell) and 16.7 percent (Nagelkerke) in whether patients were discharged by the physician or left because of other reasons. For one increase in month mean the odds of being released increase by 1.53 and for one increase in length the odds increase by 0.99. For our model 3, we get that the 95 percent confidence interval for Length of stay lies between 0.865 to 1.00 and for Month being admitted lies between 1.04 to 2.298.

```
(Intercept)  month_mean length_mean
  0.3239156   1.5131336   0.9440781
```

Figure 43: Odd Ratio

```
confidence interval
  2.5 %  97.5 %
(Intercept) 0.1555069 0.603067
month_mean  1.0429835 2.298468
length_mean 0.8650280 1.000595
```

Figure 44: Confidence Interval

3. **Make a cross table of the predicted and observed response** We can infer from the cross-table that Total cases predicting Reason of Leave by Physician are 50, and only 38 are correct. Total cases predicting Reason of leave because of other reasons is 8 and only 4 are correct.

```
=============================================================
                  dataset$Reason
dataset$reason_pred    By Physician   Other Reason   Total
-------------------------------------------------------------
By Physician                     38             12      50
                          0.760          0.240   0.862
-------------------------------------------------------------
Other Reason                      4              4       8
                          0.500          0.500   0.138
-------------------------------------------------------------
Total                            42             16      58
```

Figure 45: Cross Table

4. **Write a small section for a scientific publication, in which you explain the data set examined, report the results of the analyses of point 2 and 3, and explain the conclusions that can be drawn.**

   Our Data set Patient Length of Stay data tries to analyze the relationship between time a patient stays in the hospital and reason to leave. Our model predicts for 38 cases Reason of leave by Physician correctly and 4 cases of Other Reason correctly.

5. **Include annotated R script in the appendix of the report.**

   Refer Appendix D

3 Part 3 - Multilevel Models

1. **Use graphics to inspect the distribution of the score, and relationship between session and score**

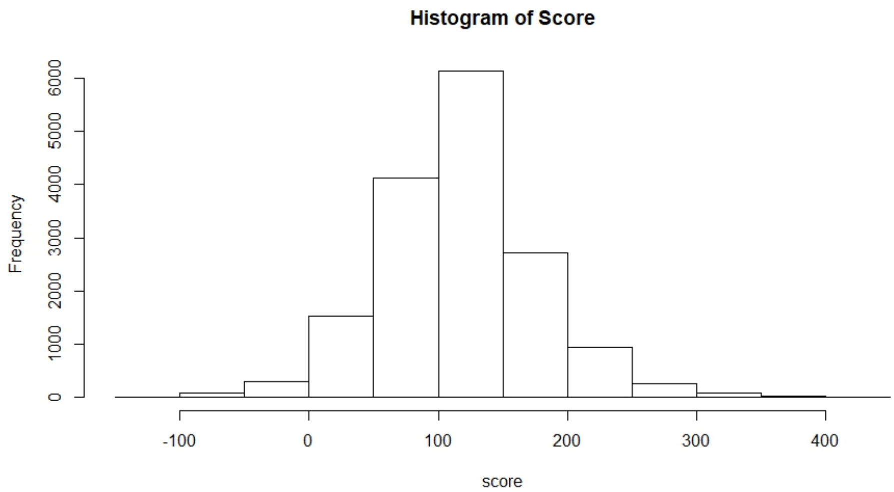

Figure 46: Frequency of Scores

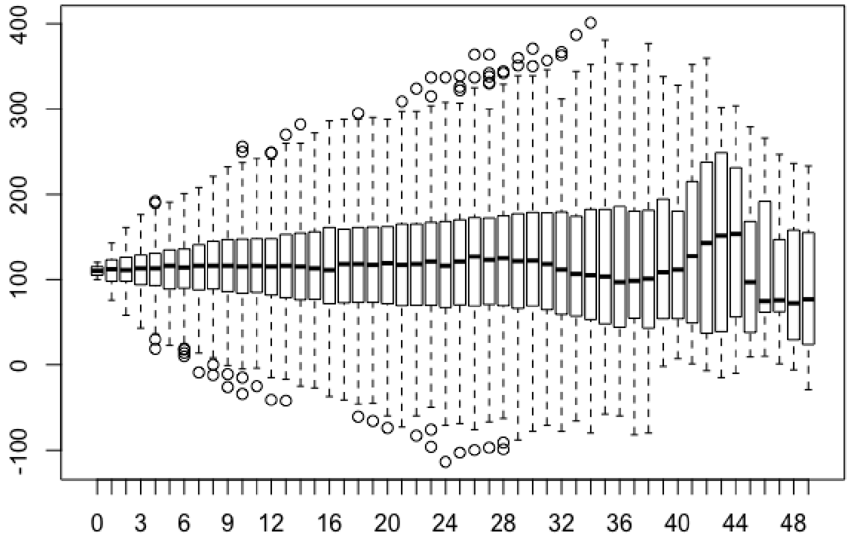

Figure 47: Session versus Score

We can see from Figure 46 and Figure 47 that the average score is between 100-150. In both figures, we can observe that the range of the score is from -100 to 400. On 47 we can observe that the median tens to be around value 100 except for sessions above 40 in which the variances are high. It is also possible to observe in Figure 47 that at the beginning overall subjects have a high level of "agreement" with each other's scores but this similarity tends to disappear as the number of sessions increases.Since we are not considering the subject which might be the reason why some boxes plots are much higher or lower than others. The fact that the box plot is comparatively tall suggests subjects have quite different scores by session. The fact that the 4 sections of the box plot are uneven in size allow us to conclude that many subjects had the same score on the same parts of the scale. The longs upper whisker means that subjects' scores are varied amongst the most positive quartile group, the same happens for the lower when is longer. If it is small means that they had similar scores in a specific session.

2. **Conduct multilevel analysis and calculate 95 percent confidence interval**

   (a) If session has impact on people score

   (b) If there is significant variance between the participants in their score

```
> anova(baselinemodel,sessionModel)
              Model df      AIC      BIC    logLik   Test  L.Ratio p-value
baselinemodel     1  3 162710.9 162734.0 -81352.45
sessionModel      2  4 162545.2 162575.9 -81268.58 1 vs 2 167.7298  <.0001
```

Figure 48: Comparison between Model Baseline and Session model

The following models were considered:

Baseline Model : score $\sim$ (1 | Subject)
Session Model : score $\sim$ (1 | Subject) + Session

The two models were created, one without any fixed effects, and other with session as a fixed effect. While comparing these models, it is possible to conclude that with a significance of p-value < 0.05 the session model has a impact on the score because the *logLik* value increases when adding session as a fixed effect.

```
> intervals(baselinemodel,0.95)
Approximate 95% confidence intervals

 Fixed effects:
               lower     est.    upper
(Intercept) 112.7019 116.8139 120.9259
attr(,"label")
[1] "Fixed effects:"

 Random Effects:
  Level: Subject
                   lower     est.    upper
sd((Intercept)) 43.68637 46.52747 49.55334

 Within-group standard error:
   lower     est.    upper
34.86891 35.25763 35.65067
```

Figure 49: Confidence Interval for Baseline model

In Figure 49 it is possible to conclude that there is a variance of around 4 between the participants' scores by observing the fixed effects table.

3. **Write a small section for a scientific publication, in which you explain the data set examined, report the results of the analyses of point 2 and 3, and explain the conclusions that can be drawn.**
   After analyzing Figure 47 a lot of discrepancy was found among the subjects' score per session. The main hypothesis put forward was that engaging in repeated sessions enhanced scores.
   A null-model (*baselinemodel*) described as *score* $\sim$ *(1 | Subject)* was created. This model was then expanded by adding the fixed effect of session (*sessionmodel*) described as *score* $\sim$ *(1 | Subject) + session*.
   Both models were compared through anova method and for each model was determined the 95 percent confidence interval. This approach allowed to conclude that the session has an impact on the score and that there is a variance of around 4 points in score between the subjects.

4. **Include annotated R script in the appendix of the report** Refer Appendix E

References

[1] Charles C Bonwell and James A Eison. *Active Learning: Creating Excitement in the Classroom. 1991 ASHE-ERIC Higher Education Reports.* ERIC, 1991.

[2] Arthur W Chickering and Zelda F Gamson. Seven principles for good practice in undergraduate education. *AAHE bulletin*, 3:7, 1987.

[3] Arum Oommen. Factors influencing intelligence quotient. *Journal of Neurology and Stroke*, 1:1–5, 2014.

[4] Rko movies.

Part 2 { Generalized linear models - Question 1 Twitter sentiment analysis (Between groups { single

Output:

```
[1] "Using direct authentication"
Levene's Test for Homogeneity of Variance (center = median)
        Df F value    Pr(>F)
group    2  9.6668 6.535e-05 ***
      2997
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Call:
lm(formula = semFrame$score ~ 1, data = semFrame)

Residuals:
   Min     1Q Median     3Q    Max
-4.172 -0.172 -0.172  0.078  3.828

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.17200    0.01446   11.89   <2e-16 ***
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 0.792 on 2999 degrees of freedom

Analysis of Variance Table

Model 1: semFrame$score ~ 1
Model 2: semFrame$score ~ semFrame$Candidate
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1   2999 1881.2
2   2997 1866.2  2    15.014 12.056 6.099e-06 ***
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Call:
lm(formula = semFrame$score ~ semFrame$Candidate, data = semFrame)

Residuals:
    Min      1Q  Median      3Q     Max
0.25 -0.2710 -0.1350  0.0998  3.8650

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                     0.13500    0.02495   5.410  6.8e-08 ***
semFrame$CandidateKylieJenner  -0.02500    0.03529  -0.708 0.478745
semFrame$CandidateMeghanMarkle  0.13600    0.03529   3.854 0.000119 ***
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 0.7891 on 2997 degrees of freedom
Multiple R-squared:  0.007981,Adjusted R-squared:  0.007319
F-statistic: 12.06 on 2 and 2997 DF,  p-value: 6.099e-06
```

```
Analysis of Variance Table

Response: semFrame$score
                    Df  Sum Sq Mean Sq F value    Pr(>F)
semFrame$Candidate    2   15.01  7.5070  12.056 6.099e-06 ***
Residuals          2997 1866.23  0.6227
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Shapiro-Wilk normality test

data:  semFrame$score
W = 0.79581, p-value < 2.2e-16


Shapiro-Wilk normality test

data:  resid(model1)
W = 0.84687, p-value < 2.2e-16

  semFrame$Candidate semFrame$score
1      KimKardashian          0.135
2        KylieJenner          0.110
3       MeghanMarkle          0.271
  semFrame$Candidate semFrame$score
1      KimKardashian      0.7289076
2        KylieJenner      0.7774989
3       MeghanMarkle      0.8557402
Levene's Test for Homogeneity of Variance (center = median)
        Df F value    Pr(>F)
group    2  9.6668 6.535e-05 ***
      2997
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Pairwise comparisons using t tests with pooled SD

data:  semFrame$score and semFrame$Candidate

             KimKardashian KylieJenner
KylieJenner  1.00000       -
MeghanMarkle 0.00036       1.6e-05

P value adjustment method: bonferroni

Part 2 { Generalized linear models -
Question 2 { Website visits (between groups { Two factors)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   1.000   2.000   2.665   4.000  14.000

Shapiro-Wilk normality test

data:  Question2$pages
```

```
W = 0.83076, p-value < 2.2e-16

[1] 1.620463
[1] 7.424128

The downloaded binary packages are in
/var/folders/ck/0trn5d_90j963cb_vh_81ysc0000gn/T//RtmpNOVJzJ/downloaded_packages
     lambda
  2.66466466
 (0.05164622)
Fitting of the distribution ' pois ' by maximum likelihood
Parameters:
       estimate Std. Error
lambda 2.664665 0.05164621

Call:
glm(formula = pages ~ 1, family = poisson(link = "log"), data = Question2,
    na.action = na.exclude)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.1701  -1.1701  -0.4262   0.7610   4.8766

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.98008    0.01938   50.57   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1008.9  on 998  degrees of freedom
Residual deviance: 1008.9  on 998  degrees of freedom
AIC: 3721.3

Number of Fisher Scoring iterations: 5

Analysis of Deviance Table

Model 1: pages ~ 1
Model 2: pages ~ version
  Resid. Df Resid. Dev Df Deviance
1       998    1008.89
2       997     864.07  1   144.82

Call:
glm(formula = pages ~ version, family = poisson(link = "log"),
    data = Question2, na.action = na.exclude)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4836  -0.8172  -0.1635   0.3773   4.3721

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
```

```
(Intercept)  0.71948    0.03102   23.19   <2e-16 ***
versionNew   0.47205    0.03973   11.88   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1008.89  on 998  degrees of freedom
Residual deviance:  864.07  on 997  degrees of freedom
AIC: 3578.5

Number of Fisher Scoring iterations: 5

Analysis of Deviance Table

Model 1: pages ~ 1
Model 2: pages ~ portal
  Resid. Df Resid. Dev Df Deviance
1       998    1008.89
2       997     938.25  1   70.648
Analysis of Deviance Table

Model 1: pages ~ version + portal
Model 2: pages ~ version + portal + version:portal
  Resid. Df Resid. Dev Df Deviance
1       996     799.02
2       995     772.78  1   26.244

Call:
glm(formula = pages ~ portal, family = poisson(link = "log"),
    data = Question2, na.action = na.exclude)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3987  -0.9345  -0.1671   0.4807   4.5073

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      1.13594    0.02581  44.010   <2e-16 ***
portalCompanies -0.32695    0.03908  -8.365   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1008.89  on 998  degrees of freedom
Residual deviance:  938.25  on 997  degrees of freedom
AIC: 3652.6

Number of Fisher Scoring iterations: 5


Call:
glm(formula = pages ~ version + portal, family = poisson(link = "log"),
    data = Question2, na.action = na.exclude)
```

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7152  -0.6177  -0.2638   0.3749   4.0087

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)      0.87405    0.03573  24.463  < 2e-16 ***
versionNew       0.46310    0.03975  11.651  < 2e-16 ***
portalCompanies -0.31390    0.03910  -8.028 9.89e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1008.89  on 998  degrees of freedom
Residual deviance:  799.02  on 996  degrees of freedom
AIC: 3515.4

Number of Fisher Scoring iterations: 5

Analysis of Deviance Table

Model: poisson, link: log

Response: pages

Terms added sequentially (first to last)


        Df Deviance Resid. Df Resid. Dev
NULL                      998    1008.89
version  1  144.819       997     864.07
portal   1   65.055       996     799.02

Call:
glm(formula = pages ~ version + portal + version:portal, family = poisson(link = "log"),
    data = Question2, na.action = na.exclude)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8253  -0.7763  -0.0877   0.4452   3.9283

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)                 0.75452    0.04454  16.939  < 2e-16 ***
versionNew                  0.64898    0.05465  11.874  < 2e-16 ***
portalCompanies            -0.06696    0.06207  -1.079    0.281
versionNew:portalCompanies -0.41061    0.08034  -5.111  3.2e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1008.89  on 998  degrees of freedom
```

```
Residual deviance:  772.78  on 995  degrees of freedom
AIC: 3491.2

Number of Fisher Scoring iterations: 5

Analysis of Deviance Table

Model: poisson, link: log

Response: pages

Terms added sequentially (first to last)


               Df Deviance Resid. Df Resid. Dev
NULL                             998    1008.89
version         1  144.819       997     864.07
portal          1   65.055       996     799.02
version:portal  1   26.244       995     772.78

Call:
aov(formula = pages ~ simple, data = Question2, na.action = na.exclude)

Residuals:
    Min      1Q  Median      3Q     Max
-3.0694 -1.0694 -0.1266  0.8734  9.9306

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        2.67725    0.04893  54.714   <2e-16 ***
simplecontrastNew  0.77260    0.06958  11.104   <2e-16 ***
simplecontrastOld  0.06887    0.06882   1.001    0.317
simple             1.23908    0.09786  12.661   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.545 on 995 degrees of freedom
Multiple R-squared:  0.2226,Adjusted R-squared:  0.2202
F-statistic: 94.94 on 3 and 995 DF,  p-value: < 2.2e-16
```

Part2 -Question 3 Linear Regression

	Film	Re.release	Production.costs	Domestic.revenue	Foreign.revenue	Total.Revenue
1	STREET GIRL	0	211	806	198	1004
2	VAGABOND LOVER	0	204	671	85	756
3	SAINT IN NEW YO	0	128	350	310	460
4	BACHELOR MOTHER	0	509	1170	805	1975
5	TOP HAT	0	609	1782	1420	3202
6	LITTLE WOMEN[*]	1	424	1337	663	2000
7	RIO RITA	0	678	1775	625	2400
8	CUCKOOS	0	407	662	201	863
9	FIVE CAME BACK	0	225	441	280	721
10	KITTY FOYLE	0	738	1710	675	2385

11	MAN TO REMEMBER	0	118	293	123	416
12	KING KONG[*]	1	672	745	1111	1856
13	FOLLOW THE FLEE	0	747	1532	1175	2727
14	ANNE OF GREENIG	0	226	573	220	793
15	INFORMER	0	243	455	495	950
16	ROBERTA	0	610	1467	868	2335
17	GAY DIVORCEE	0	520	1077	697	1774
18	EX MRS BRADFORD	0	369	730	354	1084
19	STAR OF MIDNIG[*]	1	280	575	256	831
20	SKY GIANT	0	181	370	148	518
21	SWING TIME	0	886	1624	994	2618
22	FLYING DOWN TO	0	462	923	622	1545
23	MELODY CRUISE	0	163	316	169	485
24	CROSS FIRE	0	26	74	24	98
25	SECOND WIFE	0	68	140	57	197
26	HOOK LINE AND S	0	287	595	185	780
27	COME ON DANGER	0	31	29	27	106
28	PARTNERS	0	33	82	27	109
29	MY FAVORITE WIF	0	921	1452	605	2057
30	BRIDE WALKS OUT	0	289	502	168	670
31	CRACKED NUTS	0	261	505	112	617
32	GUN LAW[*]	1	78	148	47	195
33	PHANTOM OF CRES	0	187	348	88	436
34	FIFTH AVENUE GI	0	607	950	420	1370
35	LUCKY DEVILS	0	117	179	106	285
36	MARSHALL OF ME[*]	1	75	131	49	180
37	IRENE	NA	0	578	845	775
38	29.3	NA	NA	NA	NA	NA
39	GRIDIRON FLASH	0	78	167	32	199
40	SON OF KONG	0	269	331	285	616
41	THAT'S RIGHT YO	0	271	926	92	1018
42	ALICE ADAMS	0	342	574	196	770
43	COMMON LAW	0	339	573	140	713
44	BILL OF DIVORCE	0	250	383	148	531
45	IN PERSON	0	493	496	219	715
46	GHOST VALLEY	0	41	74	27	101
47	MORNING GLORY	0	239	377	205	582
48	SIX GUN GOLD	0	49	98	15	113
49	SEVEN KEYS TO B	0	251	437	80	517
50	SHALL WE DANCE	0	991	1275	893	2168
51	LOVE COMES ALON	0	220	366	112	478
52	SPITFIRE	0	223	492	112	604
53	PACIFIC LINER	0	241	318	190	508
54	MOST DANGEROUS	0	219	263	180	443
55	SEA DEVILS	0	477	580	360	940
56	CAUGHT PLASTERE	0	281	442	107	549
57	YOU'LL FIND OUT	0	371	855	175	1030
58	PEAGH O RENO	0	293	461	109	570
59	MOTHER GAREY'S	0	358	543	160	703
60	LOVING THE LADI	0	207	370	58	428
61	LOST PATROL[*]	1	262	343	240	583
62	LUGKY PARTNERS	0	733	880	510	1390
63	TOM DIGK AND HA	0	806	1223	405	1628
64	CHECK AND DOUBL	0	967	1751	59	1810
65	DIPLOMANIACS	0	242	323	138	461

66	BORN WITH LOVE	0	338	452	117	649
67	YOU CAN'T BUY L	0	86	137	38	175
68	LIFE OF VERGIE	0	331	506	148	654
69	HIT THE DECK	0	542	980	152	1132
70	EVERYTHING'S RO	0	140	205	70	275
71	LOVE AFFAIR	0	860	975	775	1750
72	MAD MISS MANTON	0	383	496	220	716
73	IN NAME ONLY	0	722	926	395	1321
74	ROOKIE COP	0	77	108	54	162
75	THAT GIRL FROM	0	534	683	380	1063
76	PRIMROSE PATH	0	702	898	302	1200
77	DEVIL AND MISS	0	664	921	500	1421
78	ANNIE OAKLEY	0	354	435	185	620
79	RUNAWAY BRIDE	0	103	160	44	204
80	SHOOTING STRAIG	0	238	378	40	418
81	VIVACIOUS LADY[*]	1	703	830	376	1206
82	THREE MUSKETEER	0	512	451	449	900
83	GIRL A GUY AND	0	412	578	270	848

	Dist.Cost.Revenue	Dist.Cost.Prod.Cost	Year	Profit_mean
1	0.292	1.389	1930.000	99.2
2	0.287	1.064	1930.000	79.6
3	0.298	1.070	1938.000	73.6
4	0.324	1.255	1939.000	72.0
5	0.396	2.082	1936.000	70.6
6	0.388	1.830	1934.000	66.7
7	0.328	1.161	1930.000	63.8
8	0.140	0.297	1930.000	63.4
9	0.320	1.027	1939.000	58.1
10	0.326	1.054	1941.000	57.3
11	0.365	1.288	1939.000	54.1
12	0.288	0.795	1933.000	53.9
13	0.380	1.386	1936.000	53.0
14	0.372	1.305	1935.000	52.2
15	0.402	1.572	1935.000	52.0
16	0.409	1.566	1935.000	49.2
17	0.378	1.288	1935.000	49.1
18	0.337	0.989	1936.000	47.7
19	0.344	1.021	1935.000	46.8
20	0.332	0.950	1938.000	46.7
21	0.345	1.018	1936.000	46.4
22	0.390	1.305	1934.000	45.1
23	0.355	1.055	1933.000	44.8
24	0.429	1.615	1933.000	44.1
25	0.360	1.044	1936.000	41.7
26	0.344	0.934	1931.000	40.5
27	0.425	1.452	1933.000	39.5
28	0.422	1.394	1932.000	38.0
29	0.307	0.685	1940.000	32.5
30	0.324	0.751	1936.000	32.4
31	0.334	0.789	1931.000	32.1
32	0.359	0.897	1938.000	31.8
33	0.342	0.797	1933.000	29.8
34	0.328	0.740	1939.000	29.7
35	0.361	0.880	1933.000	29.5
36	0.356	0.853	1940.000	29.5

```
37           675.000              0.417    1.168      1940.0
38                NA                 NA       NA          NA
39             0.392              1.000 1935.000        27.6
40             0.347              0.796 1934.000        27.5
41             0.519              1.948 1940.000        27.4
42             0.343              0.772 1935.000        27.1
43             0.314              0.661 1932.000        26.6
44             0.322              0.684 1933.000        26.1
45             0.105              0.152 1936.000        25.9
46             0.396              0.976 1932.000        24.7
47             0.392              0.954 1934.000        24.6
48             0.372              0.857 1941.000        24.2
49             0.321              0.661 1930.000        24.0
50             0.352              0.771 1937.000        23.5
51             0.351              0.764 1930.000        23.2
52             0.444              1.202 1934.000        23.0
53             0.354              0.747 1939.000        20.7
54             0.336              0.680 1933.000        20.4
55             0.328              0.646 1937.000        19.7
56             0.324              0.633 1932.000        19.6
57             0.478              1.326 1941.000        19.4
58             0.328              0.638 1932.000        18.8
59             0.334              0.656 1938.000        18.5
60             0.364              0.754 1930.000        17.9
61             0.407              0.905 1934.000        16.8
62             0.329              0.623 1940.000        16.8
63             0.361              0.730 1941.000        16.8
64             0.322              0.603 1931.000        16.8
65             0.334              0.636 1933.000        16.4
66             0.341              0.654 1932.000        16.1
67             0.371              0.756 1937.000        15.9
68             0.361              0.713 1934.000        15.3
69             0.393              0.821 1930.000        14.7
70             0.364              0.714 1931.000        14.6
71             0.382              0.778 1939.000        14.5
72             0.342              0.640 1939.000        14.0
73             0.336              0.615 1939.000        13.3
74             0.414              0.870 1939.000        12.5
75             0.403              0.801 1937.000        10.5
76             0.323              0.553 1940.000        10.1
77             0.450              0.964 1941.000         9.0
78             0.352              0.616 1936.000         8.4
79             0.422              0.835 1930.000         7.9
80             0.359              0.630 1930.000         7.7
81             0.355              0.609 1937.000         6.6
82             0.370              0.650 1935.000         6.5
83             0.456              0.939 1941.000         6.1
 [ reached getOption("max.print") -- omitted 75 rows ]

Call:
lm(formula = profit ~ Re_release.c + Production_costs.c + Total_Revenue.c +
    Distribution_cost.c + Year.c)

Residuals:
    Min      1Q  Median      3Q     Max
```

```
-10.144  -4.675  -2.632  -0.326 193.550

Coefficients:
                     Estimate Std. Error  t value Pr(>|t|)    
(Intercept)         87.981316  20.915680    4.206 4.51e-05 ***
Re_release.c        -2.433902   6.605553   -0.368    0.713    
Production_costs.c  -0.993990   0.007826 -127.016  < 2e-16 ***
Total_Revenue.c      0.986485   0.011973   82.391  < 2e-16 ***
Distribution_cost.c -0.975275   0.031868  -30.604  < 2e-16 ***
Year.c              -0.569278   0.557609   -1.021    0.309    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 21.66 on 146 degrees of freedom
  (6 observations deleted due to missingness)
Multiple R-squared:  0.9947,Adjusted R-squared:  0.9946 
F-statistic:  5514 on 5 and 146 DF,  p-value: < 2.2e-16

       Re_release.c  Production_costs.c     Total_Revenue.c Distribution_cost.c              Year.c 
       -0.002243635        -1.130865991         2.178737378        -0.809095196        -0.006584569 
 lag Autocorrelation D-W Statistic p-value
   1     -0.02011917      2.039766   0.914
 Alternative hypothesis: rho != 0
NULL
# A tibble: 5 x 3
            Variables  Tolerance       VIF
                <chr>      <dbl>     <dbl>
1        Re_release.c 0.97302392  1.027724
2  Production_costs.c 0.45512903  2.197179
3     Total_Revenue.c 0.05159298 19.382482
4 Distribution_cost.c 0.05161671 19.373570
5              Year.c 0.86730877  1.152992
       Re_release.c  Production_costs.c     Total_Revenue.c Distribution_cost.c              Year.c 
           1.027724            2.197179           19.382482           19.373570            1.152992 
       Re_release.c  Production_costs.c     Total_Revenue.c Distribution_cost.c              Year.c 
         0.97302392          0.45512903          0.05159298          0.05161671          0.86730877 
       Re_release.c  Production_costs.c     Total_Revenue.c Distribution_cost.c              Year.c 
           1.027724            2.197179           19.382482           19.373570            1.152992 
       Re_release.c  Production_costs.c     Total_Revenue.c Distribution_cost.c              Year.c 
         0.97302392          0.45512903          0.05159298          0.05161671          0.86730877 
[1] 10000
[1] 10000


Part2 -Question 4 Logistic Regression
Analysis of Deviance Table

Model 1: dataset$Reason ~ 1
Model 2: dataset$Reason ~ length_mean
Model 3: dataset$Reason ~ month_mean + length_mean
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)  
1        57     68.324                       
2        56     65.991  1   2.3329  0.12667  
3        55     61.199  1   4.7921  0.02859 *
---
```

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Call:
glm(formula = dataset$Reason ~ month_mean + length_mean, family = binomial(),
    data = dataset)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3823  -0.8072  -0.5658   1.0602   1.9501

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.12727    0.33934  -3.322 0.000894 ***
month_mean   0.41418    0.19860   2.086 0.037021 *
length_mean -0.05755    0.03672  -1.567 0.117059
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 68.324  on 57  degrees of freedom
Residual deviance: 61.199  on 55  degrees of freedom
AIC: 67.199

Number of Fisher Scoring iterations: 5

Pseudo R^2 for logistic regression
Hosmer and Lemshow R^2:   0.104
Cox and Snell R^2:        0.116
Nagelkerke R^2:           0.167
(Intercept)  month_mean length_mean
  0.3239156   1.5131336   0.9440781
                2.5 %   97.5 %
(Intercept) 0.1555069 0.603067
month_mean  1.0429835 2.298468
length_mean 0.8650280 1.000595

               By Physician Other Reason
  By Physician           38            4
  Other Reason           12            4
   Cell Contents
|-------------------------|
|                       N |
|           N / Row Total |
|-------------------------|

==========================================================
                       dataset$Reason
dataset$reason_pred    By Physician   Other Reason   Total
----------------------------------------------------------
By Physician                     38             12      50
                              0.760          0.240   0.862
----------------------------------------------------------
Other Reason                      4              4       8
                              0.500          0.500   0.138
```

```
-----------------------------------------------------------
Total                          42             16      58
===========================================================


Part 3 - Multilevel model

Linear mixed-effects model fit by maximum likelihood
  Data: Part3
  Log-likelihood: -81352.45
  Fixed: score ~ 1
(Intercept)
   116.8139

Random effects:
 Formula: ~1 | Subject
        (Intercept) Residual
StdDev:    46.52747 35.25763

Number of Observations: 16128
Number of Groups: 501
Linear mixed-effects model fit by maximum likelihood
 Data: Part3
       AIC    BIC    logLik
  162710.9 162734 -81352.45

Random effects:
 Formula: ~1 | Subject
        (Intercept) Residual
StdDev:    46.52747 35.25763

Fixed effects: score ~ 1
               Value Std.Error    DF  t-value p-value
(Intercept) 116.8139  2.097897 15627 55.68142       0

Standardized Within-Group Residuals:
        Min          Q1         Med          Q3         Max
-4.22644591 -0.61530909  0.01016836  0.62959973  4.10477262

Number of Observations: 16128
Number of Groups: 501
Approximate 95% confidence intervals

 Fixed effects:
               lower     est.    upper
(Intercept) 112.7019 116.8139 120.9259
attr(,"label")
[1] "Fixed effects:"

 Random Effects:
  Level: Subject
                   lower     est.    upper
sd((Intercept)) 43.68637 46.52747 49.55334

 Within-group standard error:
```

```
   lower     est.    upper
34.86891 35.25763 35.65067
Linear mixed-effects model fit by maximum likelihood
  Data: Part3
  Log-likelihood: -81268.58
  Fixed: score ~ session
(Intercept)     session
111.0675622   0.3682005

Random effects:
 Formula: ~1 | Subject
        (Intercept) Residual
StdDev:     46.5146 35.06933

Number of Observations: 16128
Number of Groups: 501
Approximate 95% confidence intervals

 Fixed effects:
                  lower        est.       upper
(Intercept) 106.8665678 111.0675622 115.2685566
session       0.3126229   0.3682005   0.4237781
attr(,"label")
[1] "Fixed effects:"

 Random Effects:
  Level: Subject
                   lower    est.    upper
sd((Intercept)) 43.67456 46.5146 49.53932

 Within-group standard error:
   lower     est.    upper
34.68269 35.06933 35.46028
              Model df      AIC      BIC    logLik   Test  L.Ratio p-value
baselinemodel     1  3 162710.9 162734.0 -81352.45
sessionModel      2  4 162545.2 162575.9 -81268.58 1 vs 2 167.7298  <.0001
Approximate 95% confidence intervals

 Fixed effects:
               lower     est.    upper
(Intercept) 112.7019 116.8139 120.9259
attr(,"label")
[1] "Fixed effects:"

 Random Effects:
  Level: Subject
                   lower     est.    upper
sd((Intercept)) 43.68637 46.52747 49.55334

 Within-group standard error:
   lower     est.    upper
34.86891 35.25763 35.65067
```

Appendices

A Part 2 - Question 1

A.1 Your Twitter

```
consumer_key <-'My_consumer_key'
consumer_scret <- 'My_consumer_secret'
access_token <- 'My_access_token'
access_scret <- 'My_access_secret'
```

A.2 Sentiment3

```
#'
#' score.sentiment() implements a very simple algorithm to estimate
#' sentiment, assigning a integer score by subtracting the number
#' of occurrences of negative words from that of positive words.
#'
#' @param sentences vector of text to score
#' @param pos.words vector of words of postive sentiment
#' @param neg.words vector of words of negative sentiment
#' @param .progress passed to <code>laply()</code> to control of progress bar.
#' @returnType data.frame
#' @return data.frame of text and corresponding sentiment scores
#' @author Jefrey Breen <jbreen@cambridge.aero>
#' https://github.com/mjhea0/twitter-sentiment-analysis
score.sentiment = function(sentences, pos.words, neg.words, .progress='none')
{
  require(plyr)
  require(stringr)

  # we got a vector of sentences. plyr will handle a list or a vector as an "l" for us
  # we want a simple array of scores back, so we use "l" + "a" + "ply" = laply:
  scores = laply(sentences, function(sentence, pos.words, neg.words) {

    # clean up sentences with R's regex-driven global substitute, gsub():
    sentence = gsub('[[:punct:]]', '', sentence)
    sentence = gsub('[[:cntrl:]]', '', sentence)
    sentence = gsub('\\d+', '', sentence)

    # attempt to remove graphic elements, added based on comments on youtube movie
    sentence <- str_replace_all(sentence,"[^[:graph:]]", " ")

    # and convert to lower case:
    sentence = tolower(sentence)

    # split into words. str_split is in the stringr package
    word.list = str_split(sentence, '\\s+')
    # sometimes a list() is one level of hierarchy too much
    words = unlist(word.list)

    # compare our words to the dictionaries of positive & negative terms
```

```
    neg.matches = match(words, neg.words)
    pos.matches = match(words, pos.words)

    # match() returns the position of the matched term or NA
    # we just want a TRUE/FALSE:
    pos.matches = !is.na(pos.matches)
    neg.matches = !is.na(neg.matches)

    # and conveniently enough, TRUE/FALSE will be treated as 1/0 by sum():
    score = sum(pos.matches) - sum(neg.matches)

    return(score)
  }, pos.words, neg.words, .progress=.progress )

  scores.df = data.frame(score=scores, text=sentences)
  return(scores.df)
}
```

A.3 Twitter Analysis

```
# CS4125 Seminar Research Methodology for Data Science
# Coursework assignment A - Part 2, Question 1 - Twitter sentiment analysis
# 2017
#
# This code requires the following file:
# sentiment3.R, negative-words.txt, and positive-words.txt.
#
#
# this is based on youtube https://youtu.be/adIvt_luO1o
# also see https://silviaplanella.wordpress.com/2014/12/31/sentiment-analysis-twitter-an
##########################################################################################

setwd("D:\\Q3_2018\\Seminar research methodolgies in data science\\coursework_A_draft ve
# apple , note use / instead of \, which used by windows


#install.packages("twitteR", dependencies = TRUE)
library(twitteR)
#install.packages("RCurl", dependencies = T)
library(RCurl)
#install.packages("bitops", dependencies = T)
library(bitops)
#install.packages("plyr", dependencies = T)
library(plyr)
#install.packages('stringr', dependencies = T)
library(stringr)
#install.packages("NLP", dependencies = T)
library(NLP)
#install.packages("tm", dependencies = T)
library(tm)
#install.packages("wordcloud", dependencies=T)
#install.packages("RColorBrewer", dependencies=TRUE)
library(RColorBrewer)
library(wordcloud)
#install.packages("reshape", dependencies=T)
```

```
library(reshape)
library(car) #Package includes Levene's test
library(plotly)
################### functions


clearTweets <- function(tweets, excl) {

  tweets.text <- sapply(tweets, function(t)t$getText()) #get text out of tweets


  tweets.text = gsub('[[:cntrl:]]', '', tweets.text)
  tweets.text = gsub('\\d+', '', tweets.text)
  tweets.text <- str_replace_all(tweets.text,"[^[:graph:]]", " ") #remove graphic


  corpus <- Corpus(VectorSource(tweets.text))

  corpus_clean <- tm_map(corpus, removePunctuation)
  corpus_clean <- tm_map(corpus_clean, content_transformer(tolower))
  corpus_clean <- tm_map(corpus_clean, removeWords, stopwords("english"))
  corpus_clean <- tm_map(corpus_clean, removeNumbers)
  corpus_clean <- tm_map(corpus_clean, stripWhitespace)
  corpus_clean <- tm_map(corpus_clean, removeWords, c(excl,"http","https","httpst"))


  return(corpus_clean)
}


## capture all the output to a file.

sink("output.txt")

################# Collect from Twitter

# for creating a twitter app (apps.twitter.com) see youtube https://youtu.be/lT4Kosc_ers
#consumer_key <-'your key'
#consumer_scret <- 'your secret'
#access_token <- 'your access token'
#access_scret <- 'your access scret'

source("your_twitter.R") #this file will set my personal variables for my twitter app, a

setup_twitter_oauth(consumer_key,consumer_scret, access_token,access_scret) #connect to
twitter app


#KimKardashian
tweets_G <- searchTwitter("#KimKardashian", n=1000, lang="en", resultType="recent") #100
#KylieJenner
tweets_B <- searchTwitter("#KylieJenner", n=1000, lang="en", resultType="recent") #1000
#MeghanMarkle
tweets_A <- searchTwitter("#MeghanMarkle", n=1000, lang="en", resultType="recent") #1000
######################### WordCloud
```

```
### This not requires in the assignment, but still fun to do

# based on https://youtu.be/JoArGkOpeU0
corpus_G<-clearTweets(tweets_G, c("kim","amp","Kardashian","kims")) #remove also some ca
wordcloud(corpus_G, max.words=50)

corpus_B<-clearTweets(tweets_B, c("Kylie","amp","Jenner","Kylies"))
wordcloud(corpus_B,  max.words=50)

corpus_A<-clearTweets(tweets_A, c("Meghan", "amp", "Markle","Meghans"))
wordcloud(corpus_A,  max.words=50)

################################


######################### Sentiment analysis

tweets_G.text <- laply(tweets_G, function(t)t$getText()) #get text out of tweets
tweets_B.text <- laply(tweets_B, function(t)t$getText()) #get text out of tweets
tweets_A.text <- laply(tweets_A, function(t)t$getText()) #get text out of tweets



#taken from https://github.com/mjhea0/twitter-sentiment-analysis
pos <- scan('positive-words.txt', what = 'character', comment.char=';') #read the positi
neg <- scan('negative-words.txt', what = 'character', comment.char=';') #read the negati

source("sentiment3.R") #load algorithm
# see sentiment3.R form more information about sentiment analysis. It assigns a integer
# by subtracting the number of occurrence of negative words from that of positive words

analysis_G <- score.sentiment(tweets_G.text, pos, neg)
analysis_B <- score.sentiment(tweets_B.text, pos, neg)
analysis_A <- score.sentiment(tweets_A.text, pos, neg)

sem<-data.frame(analysis_G$score, analysis_B$score, analysis_A$score)


semFrame <-melt(sem, measured=c(analysis_G.score,analysis_B.score, analysis_A.score
))
names(semFrame) <- c("Candidate", "score")
semFrame$Candidate <-factor(semFrame$Candidate, labels=c("KimKardashian", "KylieJenner",

################## Below insert your own code to answer question 1. The data you need ca

################ Comparing two groups ###############

#homogeneity of variance
leveneTest(semFrame$score, semFrame$Candidate, center = median)

# graphical examination of variation of tweet sentiment for each celebrity
semFrameKK <- semFrame[which(semFrame$Candidate == "KimKardashian" ),]
#histogram
hist(semFrameKK$score, xlab="Sentiment Score", col = "red", main = "Graphical Examinatio
#densityplot
```

```
dkk <-density(semFrameKK$score)
xlabel <-"Graphical Examination of Variation of the celebrity Kim Kardashian"
plot(dkk, xlabel, type="l","Score of Sentiment of Tweets",col = "red")

#histogram
semFrameKJ <- semFrame[which(semFrame$Candidate == "KylieJenner" ),]
hist(semFrameKJ$score, xlab="Setiment Score", col = "blue", main = "Graphical Examinatio
#densityplot
dkj <-density(semFrameKJ$score)
xlabel1 <-"Graphical Examination of Variation of the celebrity Kylie Jenner"
plot(dkk, xlabel1, type="l","Sentiment Score of Tweets",col = "blue")

#histogram
semFrameMC <- semFrame[which(semFrame$Candidate == "MeghanMarkle" ),]
hist(semFrameMC$score, xlab="Sentiment Score", col = "darkgreen", main = "Graphical Exar
#densityplot
dmc <-density(semFrameKK$score)
xlabel2 <-"Graphical Examination of Variation of the celebrity Megha Markle"
plot(dkk, xlabel, type="l","Sentiment Score",col = "darkgreen")


#Graphically examine the mean sentiments of tweets for each celebrity
Meanvalues <- aggregate(semFrame$score~semFrame$Candidate, FUN=mean) # mean Postcode for
plot(Meanvalues, type="p",col = "darkblue" , xlab = "Celebrity", ylab = "Mean of Sentime
Standarddeviation <- aggregate(semFrame$score~semFrame$Candidate, FUN=sd)
plot(Standarddeviation, type="b",col = "red", xlab = "Celebrity", ylab = "Standard Devia

#linear model to analyze whether the knowledge to which celebrity a tweet relates has a
#on explaining the sentiments of the tweets
model0<- lm(semFrame$score ~ 1, data = semFrame) #model without predictor
summary(model0)
model1 <- lm(semFrame$score ~ semFrame$Candidate, data = semFrame) #model with predicto
summary(model1)
anova(model0,model1, test = "F") #compare if model1 provide better fitt than model0
anova(model1) #print results in anova format

### examine assumptions
hist(semFrame$score,xlab="Setiment Score", col = "blue", main = " Examination of Assump
shapiro.test(semFrame$score)
hist(resid(model1))
shapiro.test(resid(model1))



#post hoc analysis
pairwise.t.test(semFrame$score, semFrame$Candidate, paired =
                    FALSE, p.adjust.method = "bonferroni")

######## stop redireting output.
sink(NULL)
```

B Part 2 - Question 2 Web Page Visits

```
######### SAVE THE ANSWERS IN TXT #######
sink("part2question2.txt")
######### PRINT THE ANSWERS ON THE SCREEN #######
```

```
sink(NULL)



############ Libraries ###################
library(sm)

##### P2Q2 Read the dataset file  ########
setwd("~/Desktop/SRMDS/AssignmentA/")
Question2 <- read.csv(file="webvisit2.csv", header=TRUE, sep=",")

###### P2Q2-2. ########
# IV were not in the right format for R
Question2$version = factor(Question2$version, levels=c(0,1),labels=c("Old","New"))
Question2$portal = factor(Question2$portal, levels=c(0,1),labels=c("Consumers","Companie

hist(Question2$pages,xlab="Pages visited",main="Histogram of pages")
d<- density(Question2$pages)
plot(d,xlab="Pages visited",main ="Pages' density")

curve(dnorm(x, mean=m, sd=std),
      col="darkblue", lwd=2, add=TRUE, yaxt="n")

boxplot(Question2$pages ~ Question2$version, main="Pages visited per version", xlab="Ver
boxplot(Question2$pages ~ Question2$portal, main="Pages visited per portal", xlab="Porta
boxplot(Question2$pages ~ Question2$version:Question2$portal, main="Pages visited per ve
summary(Question2$pages)

##### P2Q2-3. ########
shapiro.test(Question2$pages)

qqnorm(Question2$pages)

library(moments)
skewness(Question2$pages)
kurtosis(Question2$pages)
#### test for Poisson
fitdistr(Question2$pages,"Poisson")
fitdist(Question2$pages, 'pois', method = 'mle')

##### P2Q2-4. ########

model0 <-glm(pages ~ 1, data = Question2, family = poisson(link = "log"), na.action = na
model1<-glm(pages ~ version, data = Question2, family = poisson(link = "log"), na.action
model2<-glm(pages ~ portal, data = Question2, family = poisson(link = "log"), na.action
model3<-glm(pages ~ version + portal, data = Question2, family = poisson(link = "log"),
model4<-glm(pages ~ version + portal + version:portal, data = Question2, family = poisso

summary(model0)
anova(model0,model1)
summary(model1)
anova(model0,model2)
anova(model3,model4)
summary(model2)
summary(model3)
```

```r
anova(model3)

summary(model4)
anova(model4)

##### P2Q2-5. ########
Question2$simple <-interaction(Question2$portal,Question2$version )

contrastOld<-c(1,-1,0,0)
contrastNew <-c(0,0,1,-1)

SimpleEff <-cbind(contrastNew,contrastOld)
contrasts(Question2$simple) <-SimpleEff
simpleEffectModel <- aov(pages ~ simple , data = Question2, na.action = na.exclude)

summary.lm(simpleEffectModel)
```

C Part -2 Question 3: Linear regression analysis

```r
library(ggplot2)
library(QuantPsyc)
ourdata<-read.csv(file = 'C:/Users/Aishwarya/Documents/books/books_q3/data_science_semin
ourdata
#histogram and density plot for profit and profit*100...
profit<-data$Profits
#profit contains missing values
newdata<-na.omit(profit)
hist(profit)
plot(density(newdata))

#profit%
profit_mean<-data$Profit_mean
hist(profit_mean)
new_profit_mean<-na.omit(profit_mean)
plot(density(new_profit_mean))
#intiailize everything
Re_release<-data$Re.release
Production_costs<-data$Production.costs
Domestic_revenue<-data$Domestic.revenue
Foreign_revenue<-data$Foreign.revenue
Total_Revenue<-data$Total.Revenue
Distribution_cost<-data$Distribution.cost
Dist_Cost_Revenue<-data$Dist.Cost.Revenue
Dist_Cost_Prod_Cost<-data$Dist.Cost.Prod.Cost
Year<-data$Year
#scatterplot
plot(Re_release,profit)
ggplot(data, aes(x=Re.release, y=Profits)) +
  geom_point(size=2, shape=23)
ggplot(data, aes(x=Production.costs, y=Profits)) +
  geom_point(size=2, shape=23)
ggplot(data, aes(x=Domestic.revenue, y=Profits)) +
  geom_point(size=2, shape=23)
ggplot(data, aes(x=Foreign.revenue, y=Profits)) +
  geom_point(size=2, shape=23)
```

```
ggplot(data, aes(x=Total.Revenue, y=Profits)) +
  geom_point(size=2, shape=23)
ggplot(data, aes(x=Distribution.cost, y=Profits)) +
  geom_point(size=2, shape=23)
ggplot(data, aes(x=Dist.Cost.Revenue, y=Profits)) +
  geom_point(size=2, shape=23)
ggplot(data, aes(x=Dist.Cost.Prod.Cost, y=Profits)) +
  geom_point(size=2, shape=23)
ggplot(data, aes(x=Year, y=Profits)) +
  geom_point(size=2, shape=23)
#multiple linear regression
Re_release.c=scale(data$Re.release, center=TRUE, scale=FALSE)
Production_costs.c=scale(data$Production.costs, center=TRUE, scale=FALSE)
Domestic_revenue.c=scale(data$Domestic.revenue, center=TRUE, scale=FALSE)
Foreign_revenue.c=scale(data$Foreign.revenue, center=TRUE, scale=FALSE)
Total_Revenue.c=scale(data$Total.Revenue, center=TRUE, scale=FALSE)
Distribution_cost.c=scale(data$Distribution.cost, center=TRUE, scale=FALSE)
Dist_Cost_Revenue.c=scale(data$Dist.Cost.Revenue , center=TRUE, scale=FALSE)
Dist_Cost_Prod_Cost.c=scale(data$Dist.Cost.Prod.Cost , center=TRUE, scale=FALSE)
Year.c=scale(data$Year, center=TRUE, scale=FALSE)
mod1=lm(profit~Re_release.c+Production_costs.c+Total_Revenue.c+Distribution_cost.c
        +Year.c)
summary(mod1)
library(car)
scatterplot(profit~Re_release.c+Production_costs.c+Total_Revenue.c+Distribution_cost.c
            +Year.c)
plot(mod1)
#regression coeffeicients for confidence intervals
coef(mod1)
confint(mod1)
#beta
lm.beta(mod1)
#assumptions


shapiro.test(mod1)
qqplot(mod1)
#linear regression


#regression coeffeicients for confidence intervals
#coef(mod2)
#confint(mod2)
#profit_mean can be left-confusion

#examine assumptions
#linear regression
md1<-lm(profit~Re_release, na.action = na.exclude)
md2<-lm(profit~Production_costs, na.action = na.exclude)
md3<-lm(profit~Domestic_revenue, na.action = na.exclude)
md4<-lm(profit~Foreign_revenue, na.action = na.exclude)
md5<-lm(profit~Total_Revenue, na.action = na.exclude)
md6<-lm(profit~Distribution_cost, na.action = na.exclude)
md7<-lm(profit~Dist_Cost_Revenue, na.action = na.exclude)
md8<-lm(profit~Dist_Cost_Prod_Cost, na.action = na.exclude)
```

```
#plot(md1)
#cor(data$Profits,data$Re.release)
#how to find collinearity as it is only performed to find relation b/w different indepen
# so we use multiple linear regression


library(car)
qqPlot(mod1, main="QQ Plot")
# distribution of studentized residuals
library(MASS)
sresid <- studres(mod1)
hist(sresid, freq=FALSE,
     main="Distribution of Studentized Residuals")
xfit<-seq(min(sresid),max(sresid),length=40)
yfit<-dnorm(xfit)
lines(xfit, yfit)
#collinearity
collinear<-durbinWatsonTest(mod1)
collinear
cor<-cor(mod1)

library(Hmisc)
psych.misc()
lowerCor(ourdata)
#vif
library(olsrr)
ols_vif_tol(mod1)
vif_value<-vif(mod1)
vif_value
tolerance_value<-1/vif(mod1)
tolerance_value

library(car)
our_vif<-vif(mod1)
our_vif
our_tolerance<-1/vif(mod1)
our_tolerance

#analysis of residuals
hist(mod1$residuals)
hist(rstudent(mod1))
plot(mod1$residuals, mod1$fitted)
plot(mod1)

#homogenity of variance
#assuming normally distributed and them as outliers
#for memory first
memory.limit()
memory.limit(size=10000)
levene<-leveneTest(profit,interaction(Re_release,Distribution_cost,Year,Production_costs
levene
#(profit,interaction(Re_release,Distribution_cost,Year,Production_costs,Total_Revenue))
library(car)

#errrors
```

```
newdata$predicted.probabilities <-fitted(mod1)
#explore last 20 cases
head(newdata[, c("profit","Re_release.c","Production_costs.c","Total_Revenue.c","Distrib
    "    ,"Year.c", "predicted.probabilities" )], n=20L)

#explore leverage should be around (number of predictors+1)/sample size, so (2+1)/83 = 0
#explore studentized residual only 5% outside 1.96, and 1% outside 2.58
#explore DFBeta should be less than 1
newdata$studentized.residuals<-rstudent(mod1)
newdata$dfbeta<-dfbeta(mod1)
newdata$leverage<-hatvalues(mod1)
newdata[, c("leverage", "studentized.residuals", "dfbeta")]
library(inferr)
library(olsrr)
library(psych)
#leveneTest(mod1)
#cookbook
model1<-lm(profit~Year+Total_Revenue+Distribution_cost+Re_release+Production_costs,data=
cook_dist<-cooks.distance(model1)
plot(cook_dist)
ols_cooksd_barplot(model1)
dfbeta<-dfbeta(model1)
plot(dfbeta)
dfbetaPlots(mod1, data=ourdata)

ols_dffits_plot(model1)
ols_dfbetas_panel(model1)
```

D Part 2 - Question 4 Logistic regression analysis

```
library(foreign)
library(car)
library(ggplot2)
#using as csv as minitab wasn't working on the dataset
dataset<-read.csv("C:/Users/Aishwarya/Documents/books/books_q3/data_science_seminar/proj
#converting it into levels -dichotomous
dataset$Reason<-factor(dataset$Reason, levels = c(1:2), labels = c("By_Physician","Other

#centering the predictors
month<-dataset$Month
length<-dataset$Length

month_mean<-month-mean(month)
length_mean<-length-mean(length)

#creating different models
model0 <- glm(dataset$Reason ~ 1, data = dataset, family = binomial())
model1 <- glm(dataset$Reason ~ length_mean , data = dataset, family = binomial())
model2 <- glm(dataset$Reason ~ month_mean + length_mean, data = dataset, family = binom

anova(model0,model1,model2,test="Chisq")

#model 2 is performing the best with p<0.05
summary(model2)
```

```
residual_deviance<-model2$deviance
null_deviance<-model2$null.deviance
#calculating pseudo r square
logisticPseudoR2s <- function(LogModel)
  #taken from Andy Fields et al. book on R, p.334
{
  dev <- LogModel$deviance
  nullDev <- LogModel$null.deviance
  modelN <- length(LogModel$fitted.values)
  R.l <- 1 - dev / nullDev
  R.cs <- 1 - exp(-(nullDev - dev) / modelN)
  R.n <- R.cs / (1 - (exp(-(nullDev / modelN))))
  cat("Pseudo R^2 for logistic regression\n")
  cat("Hosmer and Lemshow R^2:  ", round(R.l, 3), "\n")
  cat("Cox and Snell R^2:       ", round(R.cs, 3), "\n")
  cat("Nagelkerke R^2:          ", round(R.n, 3), "\n")
}
logisticPseudoR2s(model2)
#R2 value obtained is 0.104281
#Next we have to calculate odd ratio for predictors and their confidence interval
odds<-exp(coef(model2))
odds
#interpretation for one increase in month_mean the odds of being releasd increase by 1.5
the odds increase by 0.99
#calculating confidence interval
confidece_interval<-exp(confint(model2))
confidece_interval


#crosstable of predicted and observed response
dataset$reason_pred[fitted(model2) <= 0.5] <- 0
dataset$reason_pred[fitted(model2) > 0.5] <- 1
dataset$reason_pred<-factor(dataset$reason_pred, levels = c(0:1), labels = c("By Physici
table(dataset$Reason, dataset$reason_pred)

### examine fitt
library(descr)
crosstable<-CrossTable(dataset$reason_pred, dataset$Reason, prop.c=FALSE, prop.t=FALSE,
crosstable
```

E Part 3 - Multilevel Models

```
######### SAVE THE ANSWERS IN TXT #######
sink("part3.txt")
######### PRINT THE ANSWERS ON THE SCREEN #######
sink(NULL)


###
install.packages("ggplot2")
library(foreign)
library(car)
library(ggplot2)
library(nlme)
library(reshape)
```

```
library(graphics)

##### P3 Read the dataset file  ########
setwd("~/Desktop/SRMDS/AssignmentA/")
Part3 <- read.csv(file="set0.csv", header=TRUE, sep=",")

## P3.1

hist(Part3$score,xlab="score",main="Histogram of Score")
boxplot(Part3$score~Part3$session)


## P3.2

baselinemodel <- lme(score ~ 1, data= Part3, random = ~1|Subject, method = "ML", na.acti
baselinemodel
summary(baselinemodel)
intervals(baselinemodel,0.95)

sessionModel <- update(baselinemodel, .~. + session)
sessionModel

intervals(sessionModel,0.95)

anova(baselinemodel,sessionModel)

#Question b
intervals(baselinemodel,0.95)
```