# Report Template coursework assignment A - 2019

CS4125 Seminar Research Methodology for Data Science

*Alexander Bieniek, Wesley Quispel, David Nyrnberg*

*04/02/2019*

## Contents

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

## 1.1 Motivation

Students, researchers, and most people beyond these buckets believe that caffeine improves their productivity. Many even insist on having their morning coffee to do their work each day. We design an experiment to empirically evaluate the effects of caffeine as well as other intake trends on one's mental acuity. Perhaps caffeine truly improves one's performance in their studies or work. However, we may also investigate the existence of a placebo effect on one's work ethic. We also investigate more complex effects of caffeine with its relationship on mental acuity, such as the influence of the caffeine-induced "crash", and caffeine tolerance, and possible relationships between caffeine effectiveness and amount of sleep.

## 1.2 Theory

Academics have performed a host of analyses on the effects of caffeine on cognitive performance. For example, a study by Nehlig at UDS found that caffeine changes memory performance in nuanced ways, and it likely does not change the aggregation of long-term memory (Nehlig 2010). The study concluded that "caffeine cannot be considered a 'pure' cognitive enhancer", although it may indirectly influence, and possibly enhance, one's cognitive performance.(Nehlig 2010). Another study by Pasman et al. found that, when taking cognition tests, scores of subjects did not improve, but the tests were completed "approximately 10% faster" (Pasman et al. 2017). The findings of these studies suggest that, while caffeine may not directly improve cognitive performance, indirect factors may still lead users to enjoy increased efficiency when doing their work.

## 1.3 Research Question

The question which we investigate with our true experiment is the following: Does caffeine intake improve a student's testing ability?

## 1.4 Participants

For convenience and consistency, the experiment will use students from TUDelft as participants. A sample from this population would likely generalize to broader student populations for the effects of caffeine intake. We can also find various levels of caffeine intake habits and regular sleep amounts. Lastly, because the students belong to the same university, we can expect that, with a lower variance in mental acuity, a smaller sample size could suffice for testing of statistical significance. One thing to consider is that, because TUDelft is a linguistically diverse university, we can make no assumptions about the language backgrounds of the students. As such, it is important to test subjects with means independent of reading ability, domain experience, etc. Administering of caffeine shall be transparent and consentual. Caffeine will be administered in commercially available forms and otherwise ordinary forms, with the possibility of caffeine-free doses.

## 1.5 Conceptual Model

**Dependent Variable:** IQ Test Score

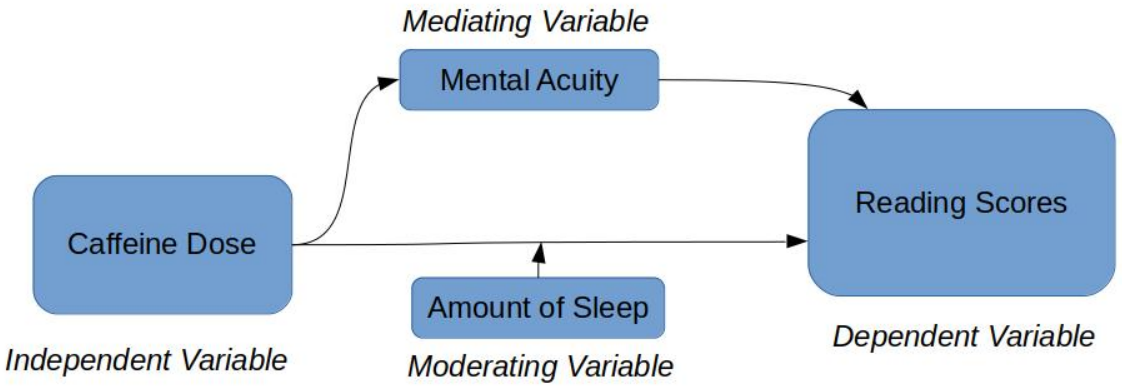
We would like to see if caffeine intake improves one's test taking abilities. In particular, the subject will be given an IQ test, which can be used as an unbiased approximation of the mentual acuity of a subject. An IQ test would be the most fair form of judging a difference in test performance, independent of the background of the subject.

**Independent Variable:** Caffeine Administration (real or placebo)

**Mediating Variables:** Mental Acuity

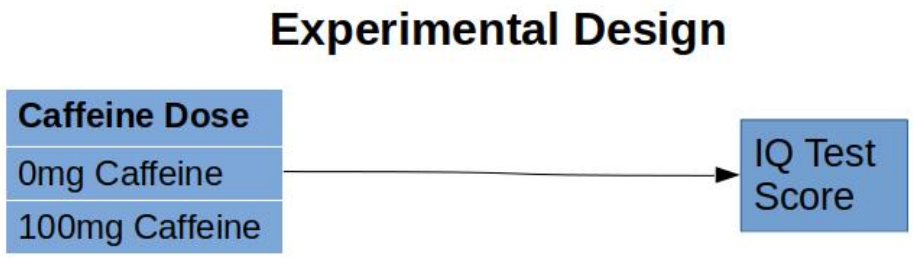
Mental acuity is perhaps the variable that we're more interested in, but there's no real way to measure this.

**Moderating Variable:** Amount of Sleep



### 1.6 Experimental Design

Our experiment is designed to watch test scores as related to caffeine ingestion. There will be two randomly assigned treatments, where subjects receive coffee with either 0mg of caffeine or 100mg of caffeine.




### 1.7 Experimental Procedure

For the colection of data from test subjects:

1. Randomly sample a test subject from a place of study, and preemptively, randomly assign treatment to the subject

- Treatment is either a 100mg dose of caffeine or a 0mg dose of caffeine, both served through a cup of coffee
- We would need to find people who have not had any caffeine yet on the given day

2. Collect some preliminary information about the subject, including estimated sleep amount
3. administer a waiting period for the subject to go about other activities and allow the caffeine to kick in

- Duration of waiting period will be influenced by academic literature about caffeine

4. After the waiting period, administer the IQ test

- The test should be realistically impossible to complete in the given amount of time to avoid statistically meaningless score distributions

### 1.8 Experiments and Suggested Statistical Analyses

To test the effects of caffeine intake on mental acuity, we would use a two-sample t-test.

## 2 Part 2 - Generalized linear models

### 2.1 Question 1: Twitter Sentiment Analysis (Between groups - single factor)

Set up libraries...

```
if (F) {
  install.packages("base64enc", dependencies=T)
  install.packages("twitteR", dependencies=T)
  install.packages("plyr", dependencies=T)
```

```
  install.packages("stringr", dependencies=T)
  install.packages("container", dependencies=T)
}

library(container)
```

```
##
## Attaching package: 'container'
```

```
## The following object is masked from 'package:base':
##
##     remove
```

```
library(base64enc)
library(twitteR)
library(plyr)
```

```
##
## Attaching package: 'plyr'
```

```
## The following object is masked from 'package:twitteR':
##
##     id
```

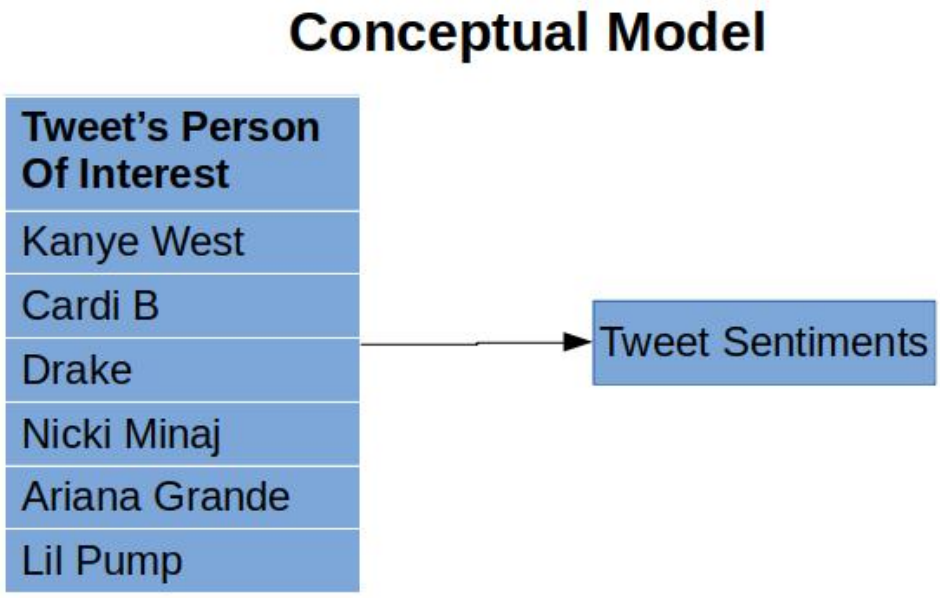
```
## The following objects are masked from 'package:container':
##
##     count, empty
```

```
library(stringr)
library(car)
```

```
## Loading required package: carData
```

### 2.1.1 Conceptual Model

Our tweet analysis seeks to answer the following question: Is there a difference in the sentiment of the tweets related to the different celebrities? This question is explored in the following conceptual model.




### 2.1.2 Collecting Preliminary Data

Collect positive and negative words and related functions, set up twitter oauth...

```
# load local textfiles listing key positive and negative words
#taken from https://github.com/mjhea0/twitter-sentiment-analysis
positive_words = scan('./data/positive-words.txt', what = 'character', comment.char=';') #read the posi
negative_words = scan('./data/negative-words.txt', what = 'character', comment.char=';') #read the nega
source("sentiment3.R")

# set up twitter sesssion
source("twitter_keys.R") # imports consumer_key, consumer_secret, access_token, and access_secret
setup_twitter_oauth(consumer_key=consumer_key,
                    consumer_secret=consumer_secret,
                    access_token=access_token,
                    access_secret=access_secret)
```

```
## [1] "Using direct authentication"
```

Collect tweets about celebrities, calculate sentiment scores...

```
tweetsFilename = 'data/tweets.csv'
n_tweets = 1e3
celebHandles = Dict$new(c("Kanye West" = "@KanyeWest",
                          "Drake" = "@Drake",
                          "Ariana Grande" = "@ArianaGrande",
                          "Cardi B" = "@IAmCardiB",
                          "Lil Pump" = "@LilPump",
                          "Nicki Minaj" = "@NickiMinaj")
                        )
if (file.exists(tweetsFilename)) {
  tweets = read.csv(tweetsFilename, header=T)
} else {
  tweets = NULL
    for (name in celebHandles$keys()) {
    handle = celebHandles[name]
    retTweets = searchTwitter(handle, n=n_tweets, lang="en", resultType="recent")
    out = data.frame("text"=laply(retTweets, function(t)t$getText()))
    out$name = name
    out$handle = handle
    tweets = rbind(tweets, out)
  }
  write.csv(tweets, file='data/tweets.csv', row.names=T)
}

tweets$score = score.sentiment(tweets$text, positive_words, negative_words)[[1]]
```

### 2.1.3 Tweet Sentiment Analysis

Is there a difference in the sentiment of the tweets related to the different celebrities? We inspect this question using context independent sentiment analysis of tweets about them. The data was collected as follows: 1. We used the twitter api to collect the 1000 most recent tweets which contain the twitter tag of the celebrity of interest 2. Punctuation and links were removed from the tweets 3. Upon these "parsed" tweets, we counted the number of "positive" and "negative" words present in the tweet, as provided in lists for the assignment, and the tweets were given a sentiment score as a difference of the positive and negative word counts

Tweets about the following American music artists were collected: - Kanye West - Drake - Ariana Grande - Cardi B - Lil Pump - Nicki Minaj

### 2.1.4 Statistical Considerations:

Making no assumptions, we would expect that all tweets about our celebrities have similar sentiment scores. That is, the sentiment scores found in tweets about celebrities have come from the same, broad distribution of sentiment scores in tweets about all celebrities. The alternative hypothesis would be that tweet sentiment scores come from different distributions when we sample tweets about the various celebrities.

### 2.1.5 Assessing Homogeneity of Variance

```
leveneTest(tweets$score, group=tweets$name)
```

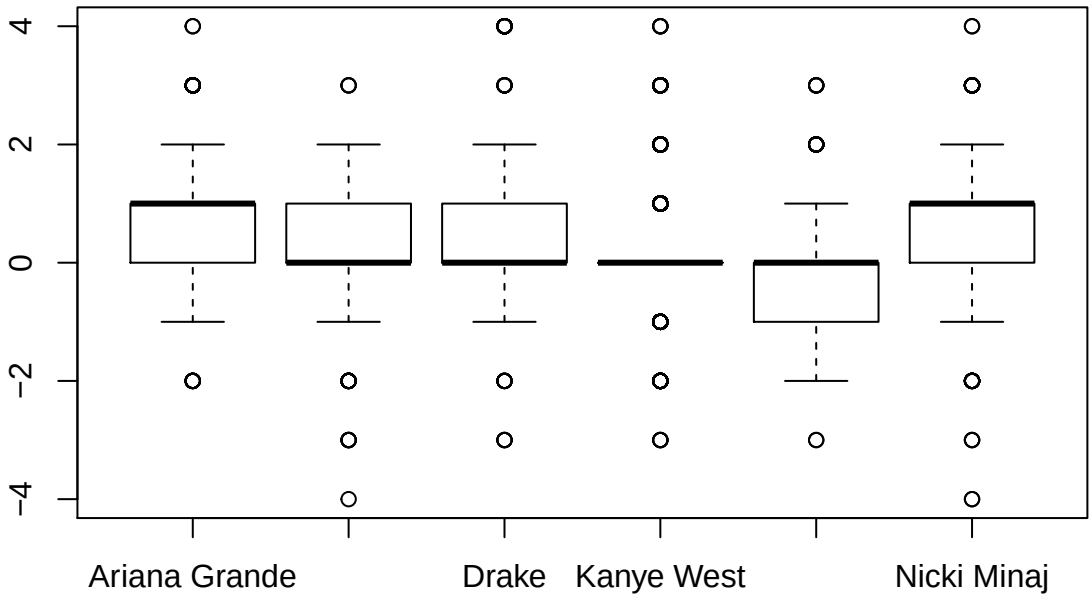
```
## Levene's Test for Homogeneity of Variance (center = median)
##         Df F value   Pr(>F)
```

```
## group    5  3.0336 0.009731 **
##       5994
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

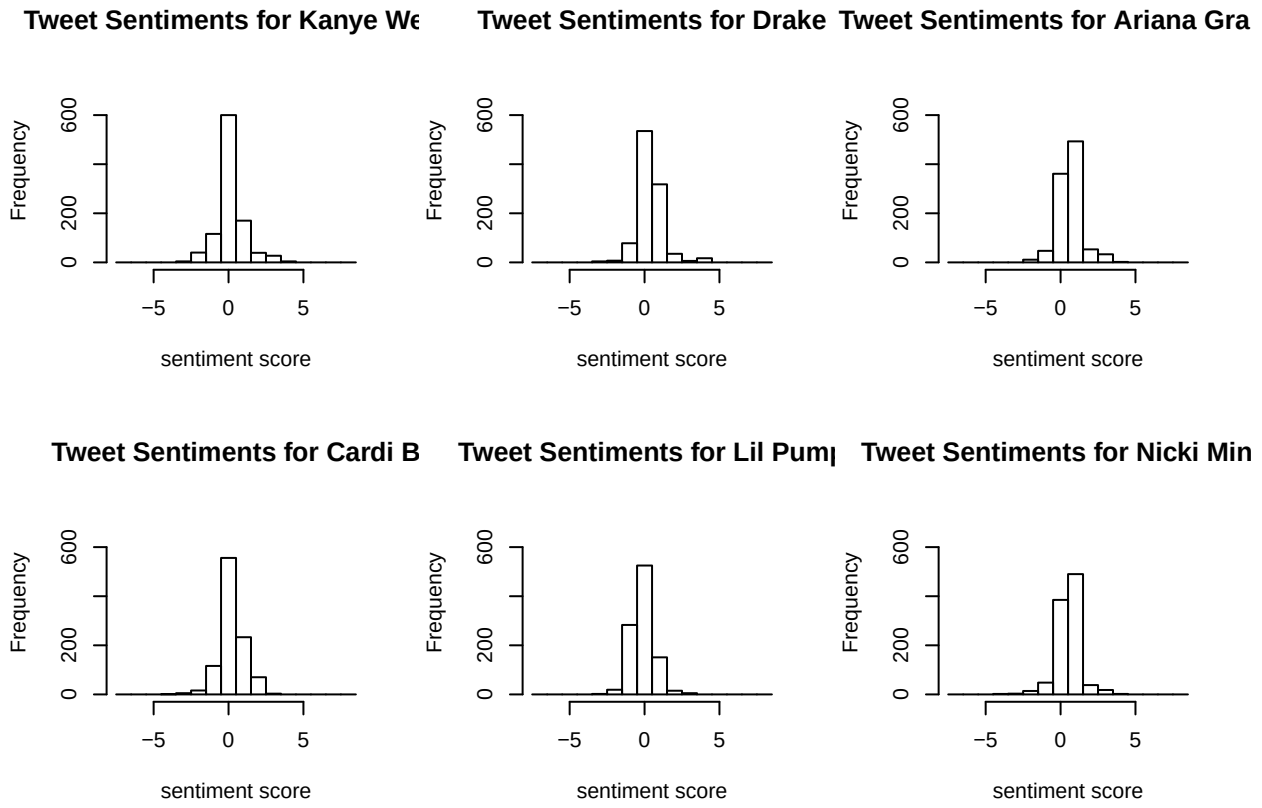
We use the Levene Test to assess homogeneity of variance. At an alpha value of .05, our p-value is less than our alpha value, so we reject the assumption that the tweet sentiment distributions from the various celebrities have homogeneous variances.

### 2.1.6 Graphical Examination of Means and Variations of Tweet Sentiment by Celebrity

```
boxplot(score ~ name, data=tweets)
```




```
par(mfrow=c(2,3))
for (name in unique(tweets$name)) {
  hist(tweets[tweets$name == name,]$score, breaks = -8:8 + .5,
       main=paste("Tweet Sentiments for", name), xlab="sentiment score", ylim=c(0,750))
}
```




Above, we have produced visualizations of the sampled distributions of tweet sentiments about various American celebrity artists. They have different medians, levels of spread, and numbers of outliers. In particular: - All celebrities were found to have a neutral sentiment score for the mode for their sampled tweet scores for Ariana Grande and Lil Pump, who have a mode tweet sentiment of +1 - Most celebrities have an interquartile range of 1, while Drake's tweet scores have an IQR of 2, and Kanye's tweet scores have an IQR of 0 - The boxplots show that a lot of celebrities have many outliers in their tweet score distributions, as defined by the boxplot function - Distributions tend to center around neutral tweets, but some celebrities recieve far less neutral tweets than others: - Kanye West appeared to recieve about ~650 neutral tweets ~ Ariana Grande recieved only about ~250 neutral tweets, and most of her tweets were at a score of +1

### 2.1.7 Tweet Knowledge as Used to Describe Tweet Sentiment

```
tweets_lm0 <- lm(score ~ 1, data = tweets,
na.action = na.exclude)
tweets_lm1 <- lm(score ~ name, data = tweets,
na.action = na.exclude)
anova(tweets_lm0, tweets_lm1)
```

```
## Analysis of Variance Table
##
## Model 1: score ~ 1
## Model 2: score ~ name
##   Res.Df  RSS Df Sum of Sq      F    Pr(>F)
## 1   5999 4891
## 2   5994 4498  5    393.02 104.75 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Using the ANOVA function in R, we find a p-value which is less than 0.05. With that, we find that the distributions of seniments of tweets about various artists come from different distributions.

#### 2.1.8 Assessing Model Quality Using Tweet Knowledge as a Predictor

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

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  tweets$score and tweets$name
##
##             Ariana Grande Cardi B Drake   Kanye West Lil Pump
## Cardi B     < 2e-16       -       -       -          -
## Drake       9.1e-11       0.00096 -       -          -
## Kanye West  < 2e-16       0.66175 2.9e-08 -          -
## Lil Pump    < 2e-16       < 2e-16 < 2e-16 7.6e-11    -
## Nicki Minaj 0.12731       2.7e-15 0.00031 < 2e-16    < 2e-16
##
## P value adjustment method: bonferroni
```

The Bonferroni Correction analysis shows that the tweet distributions for most pairs of celebrities differ from eachother. However, for some pairs, it appears that we fail to reject that the tweets come from different distributions. In this example, assuming an alpha value of 0.05, we see that we fail to reject the given null hypothesis for the following pairs of celebrities: - Drake and Kanye West - Nicki Minaj and Cardi B - Nicki Minaj and Kanye West

#### 2.1.9 Findings

Through statistical significance tests and inspection of visualizations, we believe that the sentiment score distribution from tweets pertaining to various American celebrity artists do not come from the same distribution. The sampled distributions of sentiment scores about artists have different variances and median sentiment scores. However, some pairs of celebrities appear to come from similar distributions.

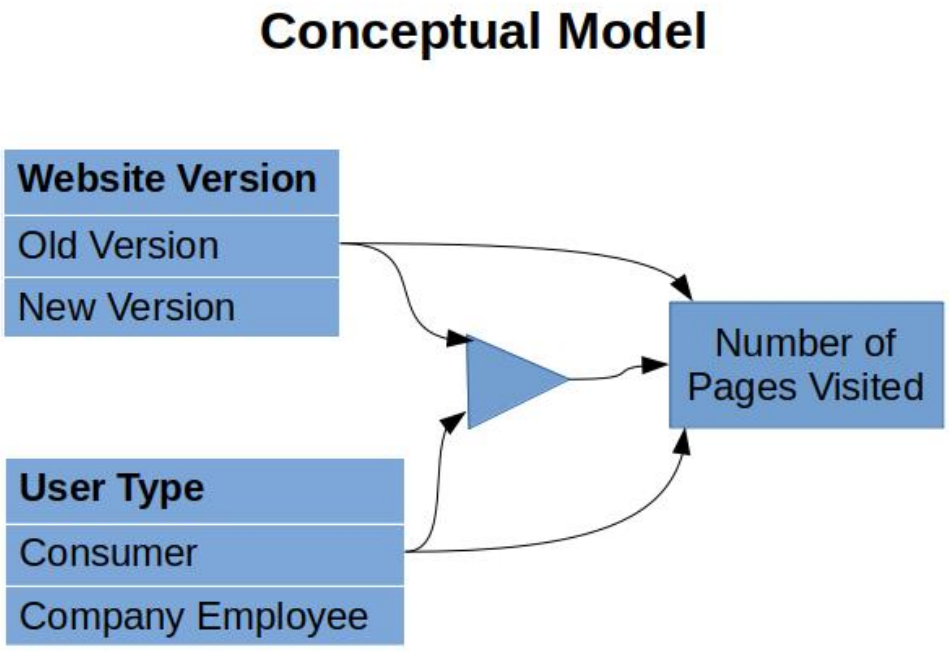
### 2.2 Question 2 - Website visits (between groups - Two factors)

Set up libraries, load data...

```
if (F) {
  install.packages("gmodels", dependencies=T)
}

library(gmodels)

visits = read.csv("data/webvisit1.csv", header = TRUE)
visits$version[visits$version==0] = "Old"
visits$version[visits$version==1] = "New"
visits$portal[visits$portal==0] = "Consumer"
visits$portal[visits$portal==1] = "Company"
```

### 2.2.1 Conceptual model

We are tasked with analyzing the results of an A-B study of a webserver as administered in two different versions to two different groups. Also, notice that we are using the webvisit dataset 1. Through this analysis, we investigate linear modeling between groups of two different factors:




**Independent Variables:**

- Version of webserver (Old or New)
- Type of User (0=consumer, 1=company)
- All combinations of the previously listed independent variables

**Dependent Variables:**

- Number of pages visited

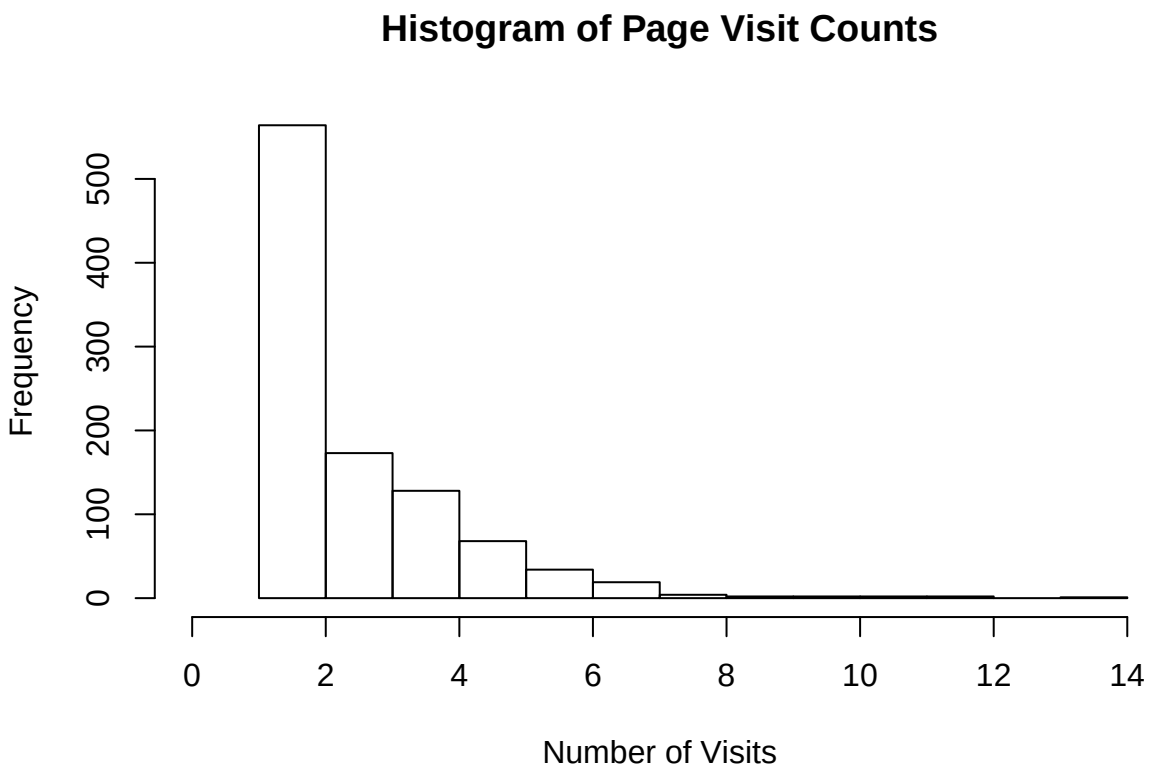
In this analysis, we inspect whether the independent variables, individually and/or in combination, effected the numper of pages visited:

**Null hypothesis:** There is no observed difference in the number of pages visited based on either the versions, the portals, or a combination thereof used. The observed difference in the sample is based on a sampling error and there is no observed difference in the entire population.

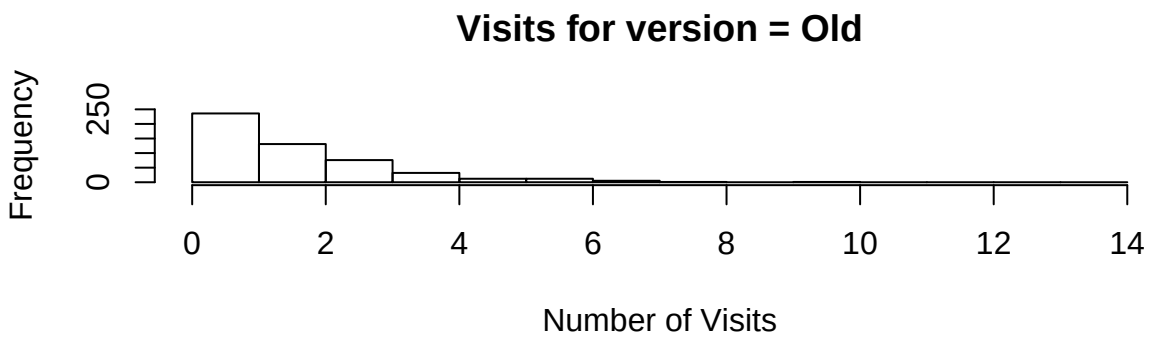
**Alternative hypothesis:** The observed difference in the sample is a real effect plus some change variation.

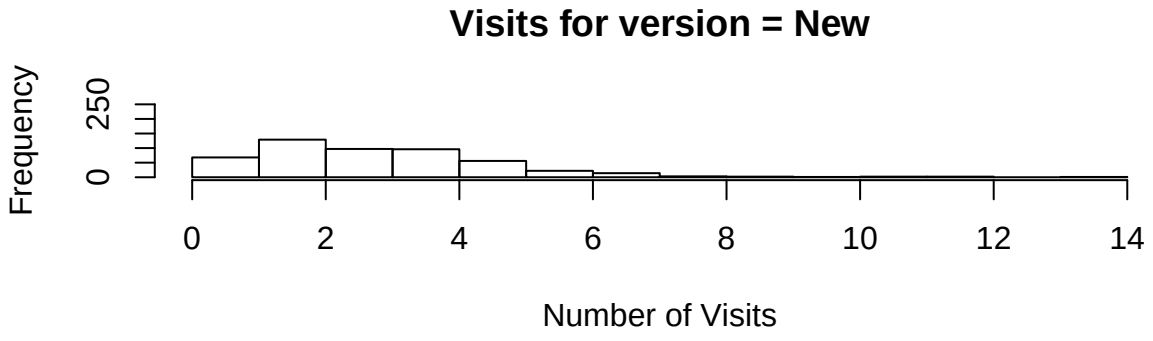
### 2.2.2 Visual inspection

```
xlim = c(0, max(visits$pages))
ylim = c(0, 250)
breaks=max(visits$pages)

# histogram of all page visits
hist(visits$pages, xlab="Number of Visits", main="Histogram of Page Visit Counts", xlim=xlim)
```

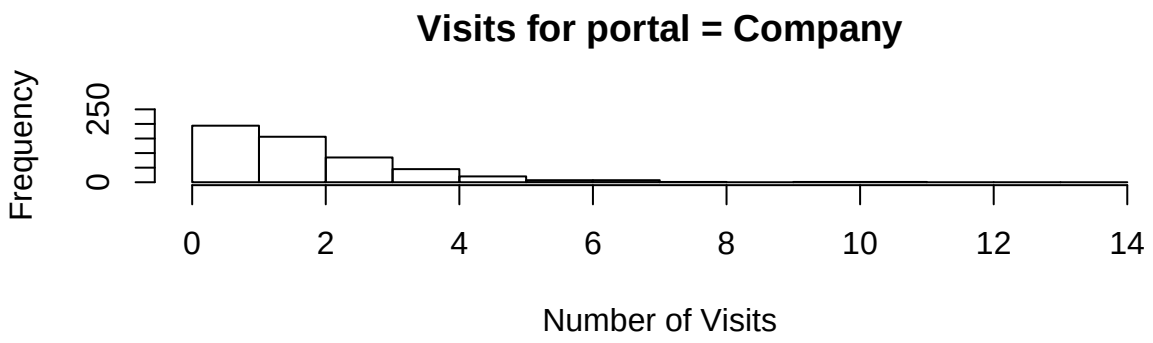


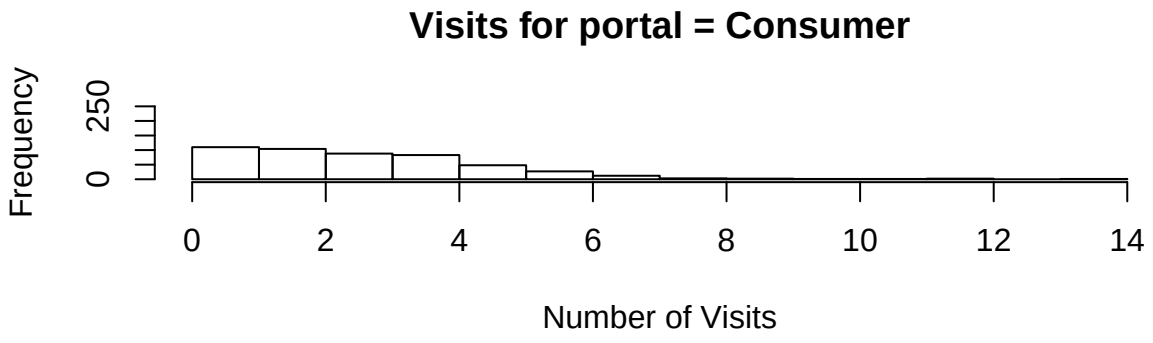

```
# histogram of page visits by different explanator variables:
# by design
par(mfrow=c(2,1))
for (v in unique(visits$version)) {
  hist(visits[which(visits$version == v),]$pages,
       xlab="Number of Visits", main=paste("Visits for version =", v),
       xlim=xlim, ylim=ylim, breaks=seq(0,breaks,1))
}
```



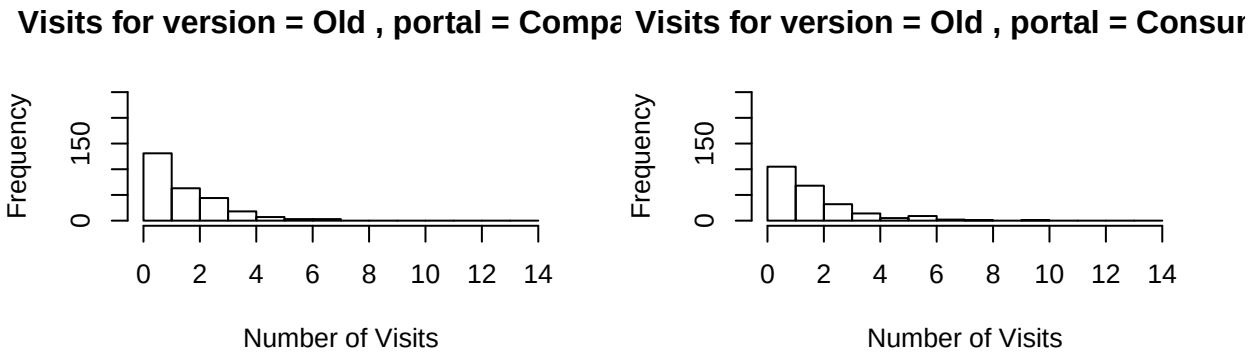



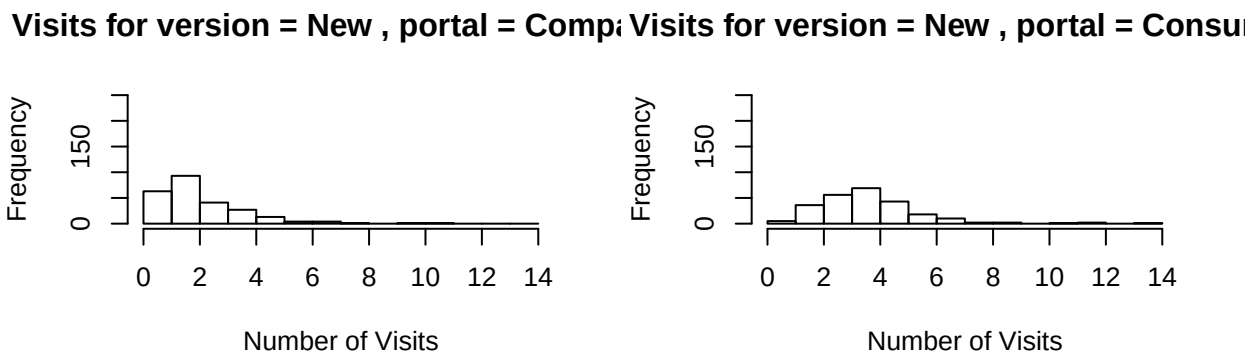

```
# by user type
par(mfrow=c(2,1))
for (p in unique(visits$portal)) {
  hist(visits[which(visits$portal == p),]$pages,
       xlab="Number of Visits", main=paste("Visits for portal =", p),
       xlim=xlim, ylim=ylim, breaks=seq(0,breaks,1))
}
```




```r
# by all combinations:
par(mfrow=c(2,2))
for (v in unique(visits$version)) {
  for (p in unique(visits$portal)) {
    hist(visits[which(visits$version == v & visits$portal == p),]$pages,
         xlab="Number of Visits",
         main=paste("Visits for version =", v, ", portal =", p),
         xlim=xlim, ylim=ylim, breaks=seq(0,breaks,1))
  }
}
```




Upon visual inspection, it appears that the portal type doesn't change the distsribution of page visit counts. However, in both cases, it appears that the version of the website causes the mean page visit count to shift to the right, and the page visit count distributions no longer seem as right skewed.

### 2.2.3 Normality check

Statistically test if variable page visits deviates from normal distribution

```
shapiro.test(visits$pages)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  visits$pages
## W = 0.83076, p-value < 2.2e-16
```

A simple Shapiro-Wilk normality test reveals, with a p-value of 2.2e-16, it is unlikely that the true distribution of page visit counts across all scenarios are sampled from normal distributions.

### 2.2.4 Model analysis

We construct linear models for the number of page visits as described by some factors:

- model0 - page visit count without predicting variables
- model1 - page visit count as linearly described by website version
- model2 - page visit count as linearly described by user type
- model3 - page visit count as linearly described by website version and user type, independently

- model3 - page visit count as linearly described by website version and user type, with interaction effects

```
pages_model0 = lm(pages~1, data=visits,
                  na.action=na.exclude)
pages_model1 = lm(pages~version, data=visits,
                  na.action=na.exclude)
pages_model2 = lm(pages~portal, data=visits,
                  na.action=na.exclude)
pages_model3 = lm(pages~version+portal, data=visits,
                  na.action=na.exclude)
pages_model4 = lm(pages~version+portal+version:portal,
                  data=visits, na.action=na.exclude)
```

```
anova(pages_model0, pages_model1)
```

```
## Analysis of Variance Table
##
## Model 1: pages ~ 1
## Model 2: pages ~ version
##   Res.Df    RSS Df Sum of Sq   F    Pr(>F)
## 1    998 3054.7
## 2    997 2671.5  1    383.16 143 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model 1 attempts to predict the page visit count using page version, and an ANOVA test finds, with p-value = 2.2e-16, that page visit count is not independent of the page version.

```
anova(pages_model0, pages_model2)
```

```
## Analysis of Variance Table
##
## Model 1: pages ~ 1
## Model 2: pages ~ portal
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    998 3054.7
## 2    997 2866.5  1    188.14 65.435 1.734e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model 2 attempts to predict the page visit count using user type, and an ANOVA test finds, with p-value = 1.734-15, that page visit count is not independent of the user type.

```
anova(pages_model3, pages_model4)
```

```
## Analysis of Variance Table
##
## Model 1: pages ~ version + portal
## Model 2: pages ~ version + portal + version:portal
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    996 2498.3
## 2    995 2374.8  1    123.42 51.709 1.264e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Because model1 and model2, which use page version and user type as predictors, can be used to explain trends in page visits, we would expect that a model using both of these predictors would work as well. This is how we constructed model3. Now we use an ANOVA test to see whether including interaction effects would

further help predict the number of page visits, as we construct in model4. Through this ANOVA test with a p-value of 1.264e-12, we see that interaction effects can be further used to explain the number of page visits seen.

### 2.2.5 Simple effect analysis

```
visits$interaction = interaction(visits$portal, visits$version)
allPortalsVersion0 = c(1,-1,0,0)
allPortalsVersion1 = c(0,0,1,-1)
SimpleEff = cbind(allPortalsVersion0, allPortalsVersion1)
contrasts(visits$interaction) = SimpleEff
simpleEffectModel = aov(pages~interaction , data=visits, na.action=na.exclude)
summary.lm(simpleEffectModel)
```

```
##
## Call:
## aov(formula = pages ~ interaction, data = visits, 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 ***
## interactionallPortalsVersion0 -0.77260    0.06958 -11.104   <2e-16 ***
## interactionallPortalsVersion1 -0.06887    0.06882  -1.001    0.317
## interaction                   -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
```

Our analysis shows that, indeed, there is an interaction effect, but only in some cases:

- For version 0, the type of user doesn't change the page visit count. The test specifically finds a p-value of 0.317, so we don't have reason to believe that there is a statistically significant difference in page visit counts for the consumer vs business users when using this version.
- For version 1, the type of user indeed changes the page visit count. The test specifically finds a p-value of 2e-16, so we have reason to believe that there is a statistically significant difference in page visit counts for the consumer vs business users when using this version.

### 2.2.6 Report section for a scientific publication

We analyzed the number of page visits for a given website. In particular, we inspected the effects of the type of user and the version of the website presented to the type of user. In general, we find that the number of page visits is indeed dependent on the type of user visiting the page and the version of the page. We also found that some conditions of the experiment seem to have interaction effects. In particular, we find that for users that are using portal version 0, there is a statistically significant difference in the trends of page visit counts for the business users and the consumer users.

## 2.3 Question 3 - Linear regression analysis

Set up libraries, load data...

```
if (F) {
  install.packages("ggpubr", dependencies=T)
  install.packages("ggExtra", dependencies=T)
  install.packages("ppcor", dependencies=T)
  install.packages("mctest", dependencies=T)
}

library(ggpubr)
```

```
## Loading required package: ggplot2
```

```
## Loading required package: magrittr
```

```
##
## Attaching package: 'magrittr'
```

```
## The following object is masked from 'package:container':
##
##     add
```

```
##
## Attaching package: 'ggpubr'
```

```
## The following object is masked from 'package:plyr':
##
##     mutate
```

```
## The following object is masked from 'package:container':
##
##     rotate
```

```
library(ggExtra)
library(car)
library(mctest)
library(ppcor)
```

```
## Loading required package: MASS
```

```
airfare <- read.csv(file="data/airfare.csv", header=T)
```

### 2.3.1 Conceptual model

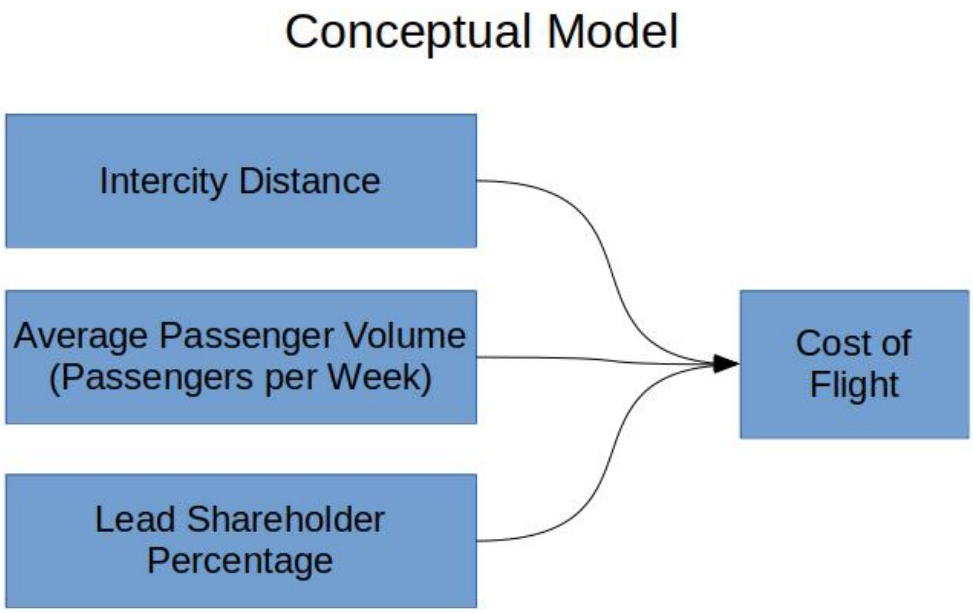
For a self-guided linear regression analysis, we investigate a dataset which records airfares from cities to cities. We would like to see if the chosen independent variables can be used to predict the price of a ticket from one city to nother. For the analysis:

**Dependent Variable: Average Fare**

The average price of the ticket to get from City1 to City2

**Independent Variables:**

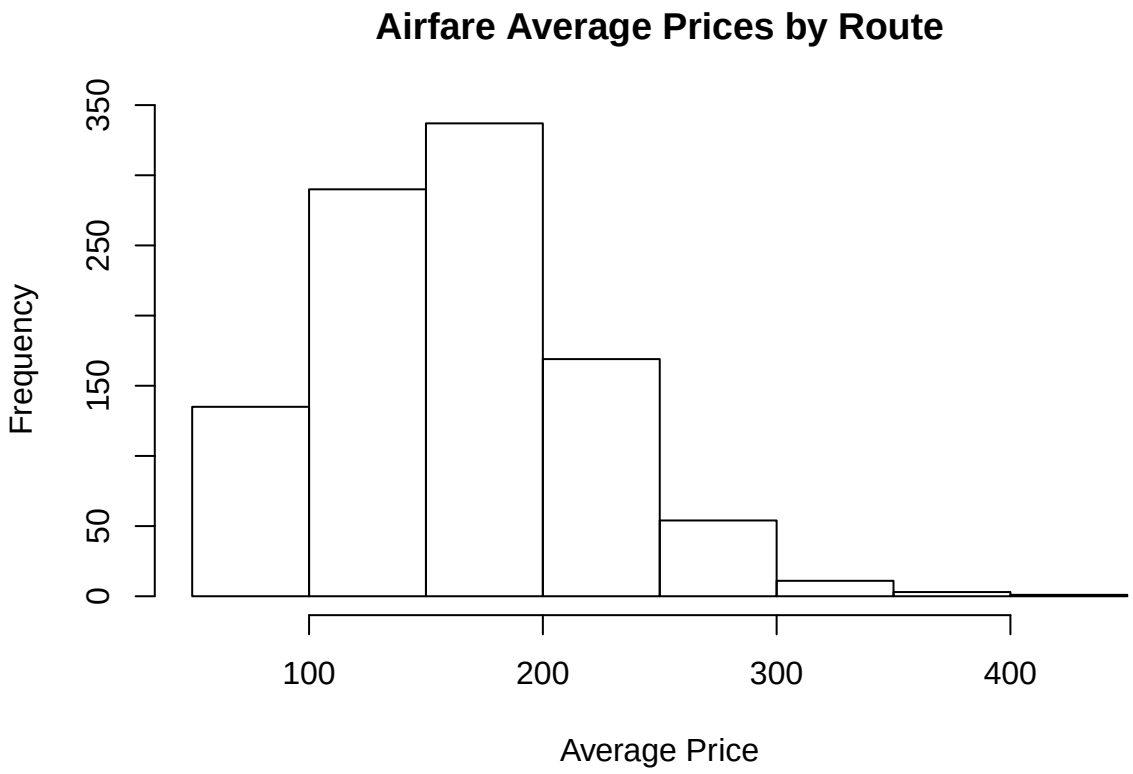
- Distance - the distance between City1 and City2
- Average Weekly Passengers - the average number of passengers that fly from City1 to City2 per week
- Lead Share Percentage - the Percentage of the flights from City1 to City2 which are served through the leading airline of the route




### 2.3.2 Visual inspection

Graphical analysis of the distribution of the dependent variable, e.g. histogram, density plot

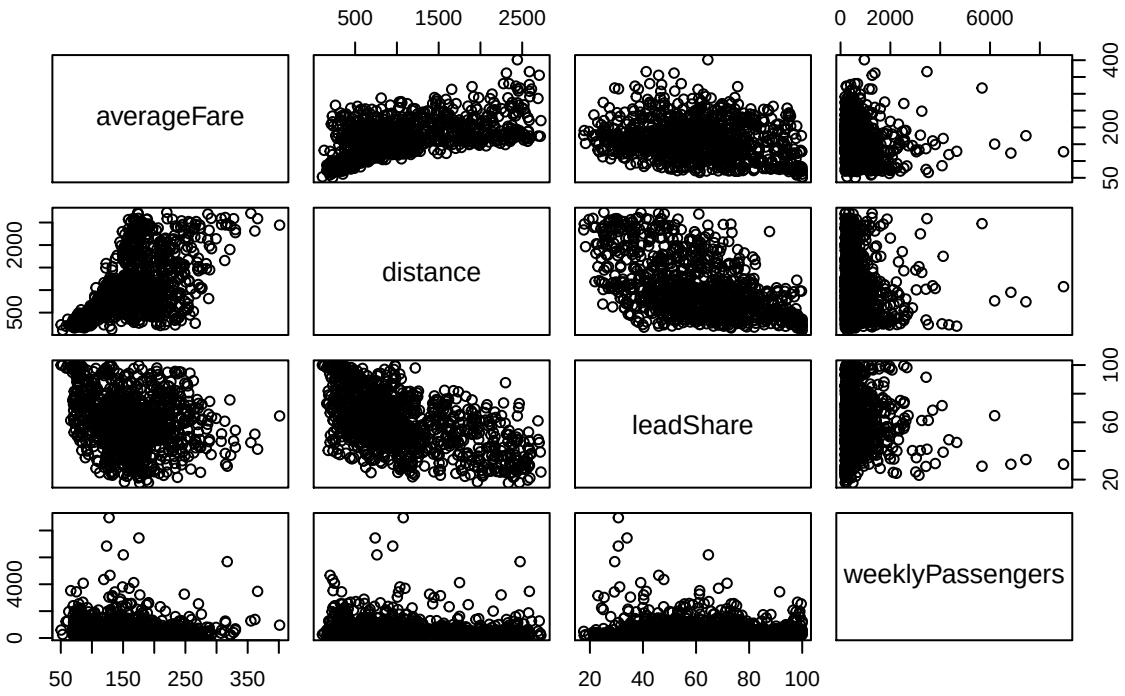
```
hist(airfare$averageFare, main="Airfare Average Prices by Route", xlab="Average Price")
```




The price of airfares appears normally distributed, centered somewhere around 175 units. The prices are right skewed, although this is expected because the price has a lower bound of 0.

### 2.3.3 Scatter plot

```
# Basic Scatterplot Matrix
pairs(~averageFare+distance+leadShare+weeklyPassengers,data=airfare,
   main="Simple Scatterplot Matrix")
```




We produce a scatterplot matrix for our independent and dependent variables. Some of the scatterplots which stand out:

- Average fare seems to increase with distance, which is quite intuitive.
- There doesn't appear to be a strong relationship between the percentage of flights owned by the leading airline and the average fare price.
- It's very hard to see a relationship, visually, between the amount of weekly passengers for a route and the price of the flight. However, the routes which have extremely high weekly passenger counts seem to have lower prices. This trend is visually supported by very few data points, though.

### 2.3.4 Linear regression

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

```
fare_model0 = lm(averageFare ~ 1, data=airfare, na.action=na.exclude)
confint(fare_model0)
```

```
##                2.5 %   97.5 %
## (Intercept) 159.9397 166.8111
```

```
coef(fare_model0)
```

```
## (Intercept)
##    163.3754
```

```
fare_model1 = lm(averageFare ~ distance, data=airfare, na.action=na.exclude)
confint(fare_model1)
```

```
##                     2.5 %       97.5 %
## (Intercept) 104.59931149 115.30817862
## distance      0.04621403   0.05487025
```

```
coef(fare_model1)
```

```
##  (Intercept)     distance
## 109.95374505   0.05054214
```

```
anova(fare_model0, fare_model1)
```

```
## Analysis of Variance Table
##
## Model 1: averageFare ~ 1
## Model 2: averageFare ~ distance
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1    999 3062270
## 2    998 2006500  1   1055770 525.12 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Our ANOVA test finds that the average price of the fare is dependent on the distance of the flight. The test reports a p-value of 2.2e-16, so indeed we reject that the average fare price is independent of the distance between the cities. We also report the confidence interval and the weights of the independent variables above.

```
fare_model2 = lm(averageFare ~ distance + weeklyPassengers,
                 data=airfare, na.action=na.exclude)
confint(fare_model2)
```

```
##                          2.5 %        97.5 %
## (Intercept)      108.094615921 120.196476496
## distance           0.045639171   0.054299165
## weeklyPassengers  -0.008967769  -0.001700977
```

```
coef(fare_model2)
```

```
##      (Intercept)         distance weeklyPassengers
##    114.145546209      0.049969168     -0.005334373
```

```
anova(fare_model1, fare_model2)
```

```
## Analysis of Variance Table
##
## Model 1: averageFare ~ distance
## Model 2: averageFare ~ distance + weeklyPassengers
##   Res.Df     RSS Df Sum of Sq      F   Pr(>F)
## 1    998 2006500
## 2    997 1989934  1     16567 8.3003 0.004049 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Our ANOVA test finds that the average price of the fare is dependent on the number of weekly passengers of the route. The test reports a p-value .004049, so indeed we reject that the average fare price is independent of the number of weekly passengers. We also report the confidence interval and the weights of the independent variables above.

```
fare_model3 = lm(averageFare ~ distance + weeklyPassengers + leadShare,
                 data=airfare, na.action=na.exclude)
confint(fare_model3)
```

```
##                          2.5 %        97.5 %
## (Intercept)      77.041006409 106.566226301
## distance          0.049439643   0.059686992
```

```
## weeklyPassengers -0.008156413  -0.000855535
## leadShare         0.111695523   0.451401752
```

```
coef(fare_model3)
```

```
##      (Intercept)         distance weeklyPassengers        leadShare
##     91.803616355      0.054563317     -0.004505974      0.281548638
```

```
anova(fare_model2, fare_model3)
```

```
## Analysis of Variance Table
##
## Model 1: averageFare ~ distance + weeklyPassengers
## Model 2: averageFare ~ distance + weeklyPassengers + leadShare
##   Res.Df     RSS Df Sum of Sq      F   Pr(>F)
## 1    997 1989934
## 2    996 1969017  1     20917 10.581 0.001181 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Our ANOVA test finds that the average price of the fare is dependent on the percentage of flights which are provided by the lead airline. The test reports a p-value .001181, so indeed we reject that the average fare price is independent of the percentage of flights which are controlled by the leading airline of the route. We also report the confidence interval and the weights of the independent variables above.

### 2.3.5 Examine assumption

```
X = airfare[c('distance', 'weeklyPassengers', 'leadShare')]
Y = airfare['averageFare']
imcdiag(x=X, y=Y)
```

```
##
## Call:
## imcdiag(x = X, y = Y)
##
##
## All Individual Multicollinearity Diagnostics Result
##
##                      VIF    TOL       Wi       Fi Leamer   CVIF Klein
## distance          1.4252 0.7016 211.9735 424.3723 0.8376 1.5679     0
## weeklyPassengers 1.0274 0.9733  13.6824  27.3922 0.9866 1.1303     0
## leadShare         1.4201 0.7042 209.4375 419.2952 0.8391 1.5623     0
##
## 1 --> COLLINEARITY is detected by the test
## 0 --> COLLINEARITY is not detected by the test
##
## * all coefficients have significant t-ratios
##
## R-square of y on all x: 0.357
##
## * use method argument to check which regressors may be the reason of collinearity
## ===================================
```

```
pcor(X, method='pearson')
```

```
## $estimate
```

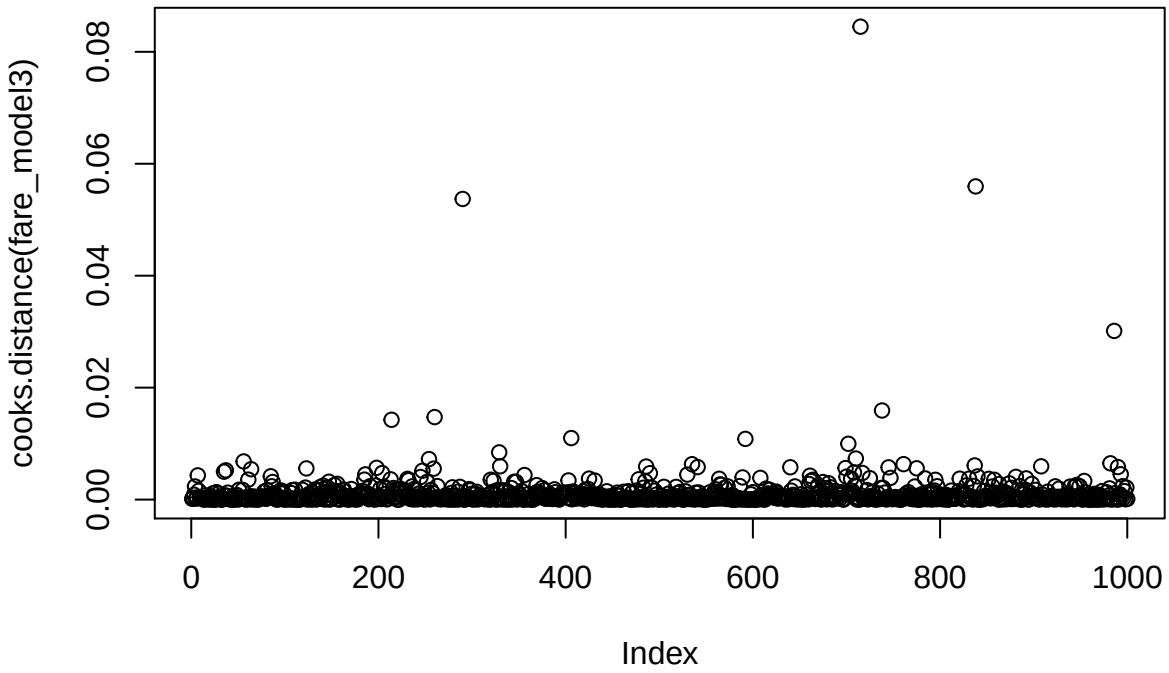
```
##                   distance weeklyPassengers  leadShare
## distance          1.0000000       -0.1491478 -0.5409332
## weeklyPassengers -0.1491478        1.0000000 -0.1369035
## leadShare        -0.5409332       -0.1369035  1.0000000
##
## $p.value
##                      distance weeklyPassengers    leadShare
## distance         0.000000e+00     2.193641e-06 5.292396e-77
## weeklyPassengers 2.193641e-06     0.000000e+00 1.410984e-05
## leadShare        5.292396e-77     1.410984e-05 0.000000e+00
##
## $statistic
##                   distance weeklyPassengers  leadShare
## distance           0.00000        -4.762660 -20.307730
## weeklyPassengers  -4.76266         0.000000  -4.363857
## leadShare        -20.30773        -4.363857   0.000000
##
## $n
## [1] 1000
##
## $gp
## [1] 1
##
## $method
## [1] "pearson"
```

The output of the partial correlation coefficients analysis shows that, for all combinations, the testing of independence of all pairs of independent variables produces p-values close to zero. That is, all pairs of independent variables are found to have some amount of correlation.

### 2.3.6 Impact analysis of individual cases

Examine effect of single cases on the predicted values (e.g. DFBeta, Cook's distance)

```
plot(cooks.distance(fare_model3))
```




The plotting of cook's distance shows that all DFBeta are far less than one. Of the 1000 airfare data points, only a handful of points have values which seem to deviate from the rest, but we do not believe that these values would not reveal any sort of influence of single cases on predicted values.

#### 2.3.7 Report section for a scientific publication

We performed a multiple regression on the average airfare of a route from one city to another. We found that the average airfare is dependent on the distance between cities, the amount of flights on the route owned by a dominant airline, and the average number of weekly passengers of the given route. We also found that each of these independent variables exhibit high degrees of correlation.

### 2.4 Question 4 - Logistic regression analysis

Set up libraries, collect data...

```
if (F) {
  install.packages("caret")
}

library(caret)
```

```
## Loading required package: lattice
```

```
library(gmodels)
library(ggpubr)
library(ggExtra)

shf <- read.csv("data/logisticDataStatureHandFoot.csv")
```

### 2.4.1 Conceptual model

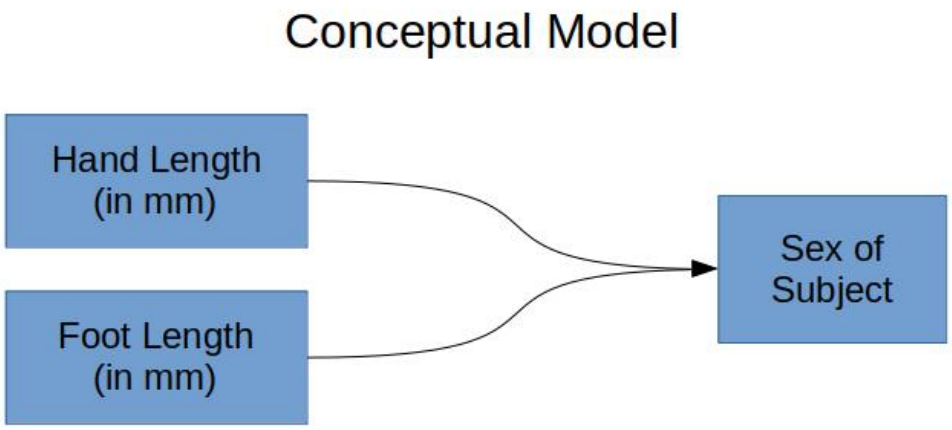
In this logistic regression analysis, we consider some size measurements of subjects and look for a relationship between these measurements and the sex of the subject. Of course, we assume that the lengths of the hands and feet of our subjects as well as their sexes are independent of those observations in other subjects.

**Dichotomous Dependent Variable: sex**

Note: the experiment collected and recorded this variable as a "gender". We shall call this variable "sex" because we believe that this is what the experimenters were actually observing.

**Independent Variables:**

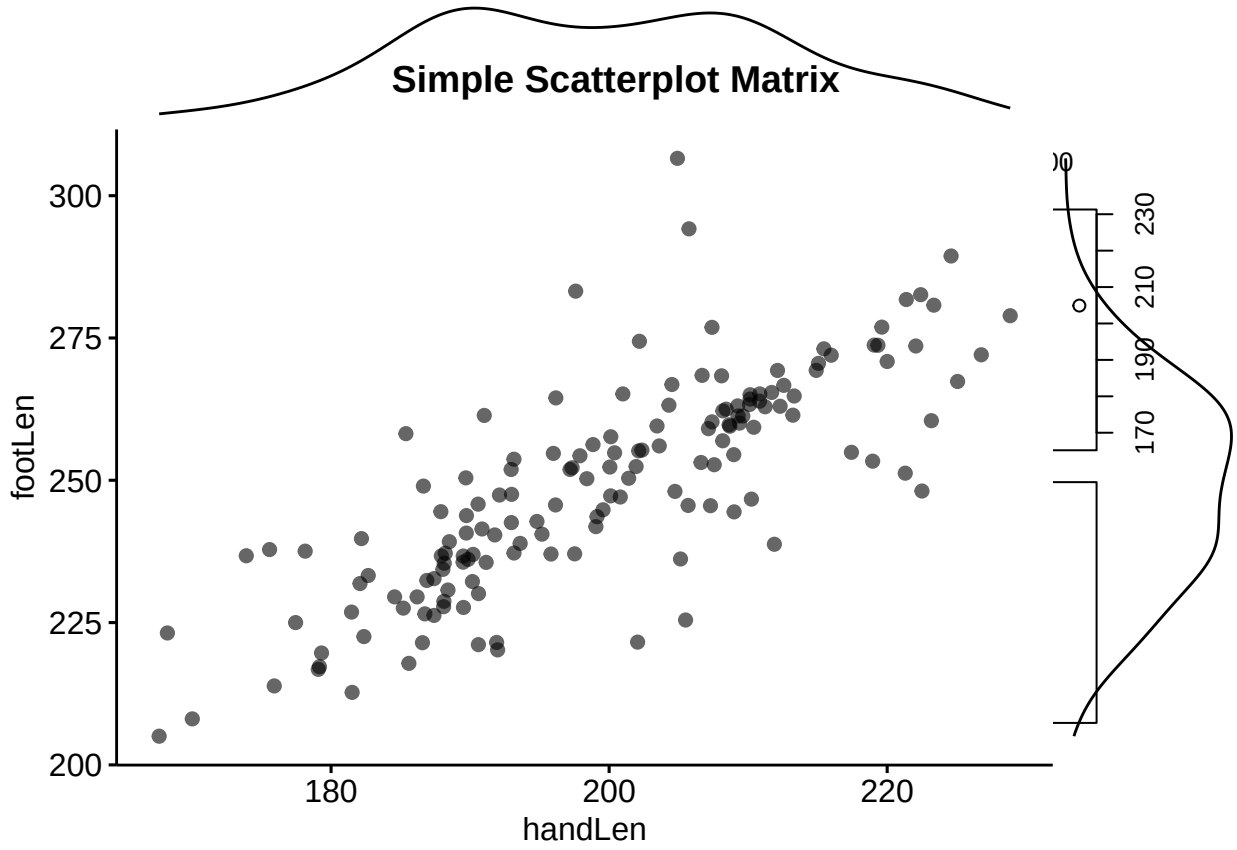
- Hand Length (in mm)
- Foot Length (in mm)



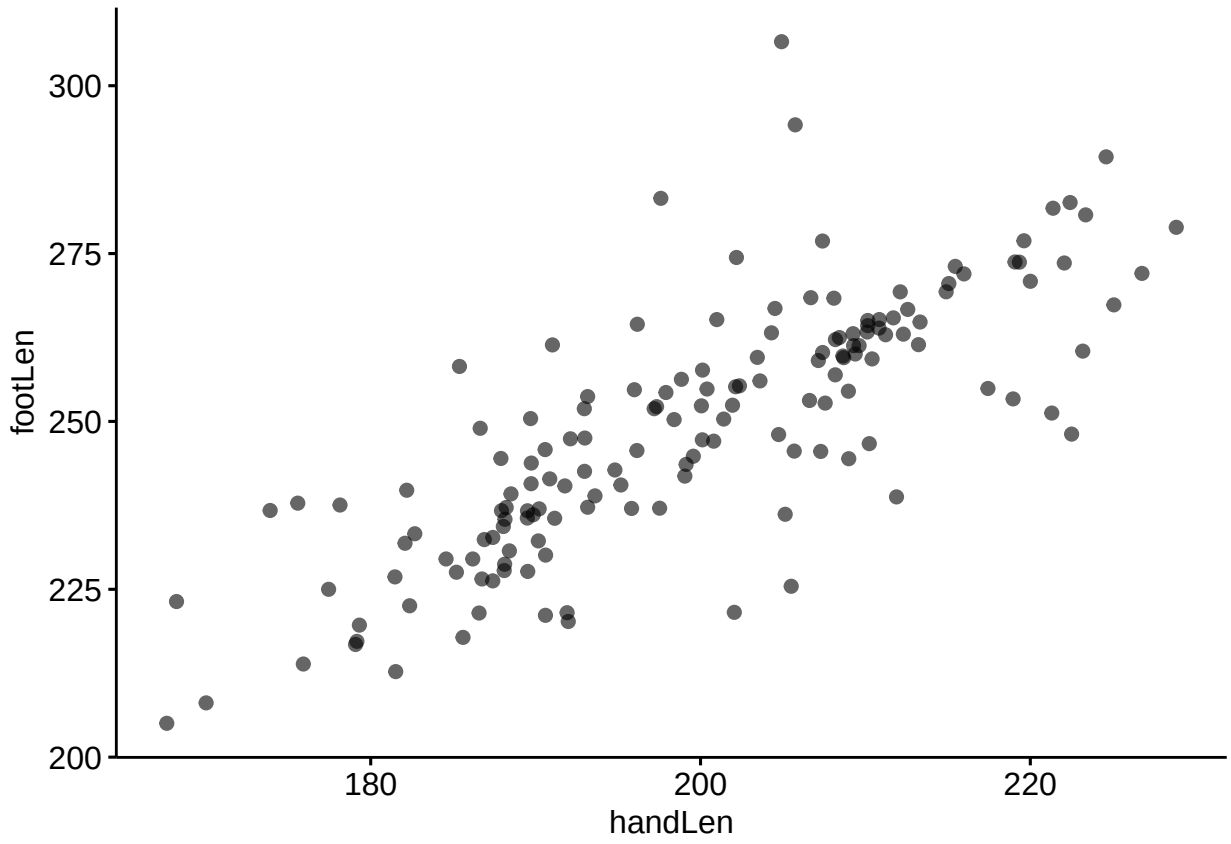

The Null Hypothesis would suggest that the independent variables, which are the lengths of hands and feet, do not have statistically significant relationships with the sex of the subject. The Alternative Hypothesis, then, would be that there is a statistically significant relationship between the lengths of hands and feet and the sex of the subject.

### 2.4.2 Visualization of Data

```
pairs(~handLen+footLen,data=shf,
   main="Simple Scatterplot Matrix")

p1Q4 <- ggscatter(shf, x = "handLen", y = "footLen",
               palette = "jco",
               size = 2, alpha = 0.6)
ggMarginal(p1Q4, type = "density")
```


Simple Scatterplot Matrix
300
275
250
225
200
footLen
180
200
220
handLen
230
210
190
170

```
plot(p1Q4)
```




### 2.4.3 Logistic Regression

```
shf$sex[shf$gender==1] <- 'male'
shf$sex[shf$gender==2] <- 'female'
shf$sex = factor(shf$sex)

model0 = glm(sex ~ 1, data = shf, family = binomial())
model1 = glm(sex ~ handLen, data = shf, family = binomial())
model2 = glm(sex ~ handLen + footLen, data = shf, family = binomial())

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

```
## Analysis of Deviance Table
##
## Model 1: sex ~ 1
## Model 2: sex ~ handLen
## Model 3: sex ~ handLen + footLen
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1       154    214.714
## 2       153    105.843  1  108.872 < 2.2e-16 ***
## 3       152     75.671  1   30.172 3.954e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#pander(anova(model0,model1,model2,test = "Chisq" ),
#       caption = "Model comparison of binominal variable of sex")
```

We use of ANOVA for the comparison of our models:

- From random classification of subjects, we find that adding in Hand Length as a predictor improves our linear model. Statistically, the chance that Hand Length as an indicator improves our model without it being a truly good indicator is 2.2e-16: we find that hand length has a relationship with the sex of a subject.
- Similarly, we find that additionally adding Foot Length as a predictor improves our model. Statistically, the chance that Foot Length as an indicator improves our previous model, with just Hand Length as an indicator, without it being a truly good indicator is 3.954e-8:we find that foot length has a relationship with the sex of a subject.

### 2.4.4 Visualization of Results

```
sexProbs = predict.glm(model2, shf, type="response")
sexPreds = sapply(sexProbs, function(x) if (x<.5) 'male' else 'female')
sexPreds = factor(sexPreds)

dnn = c('predicted', 'observed')
sexTable = table(sexPreds, shf$sex, dnn=dnn)
sexConfusionMatrix = confusionMatrix(sexTable)
sexConfusionMatrix
```

```
## Confusion Matrix and Statistics
##
##          observed
## predicted female male
##    female      9   73
##    male       66    7
##
##                Accuracy : 0.1032
##                  95% CI : (0.0602, 0.1622)
##     No Information Rate : 0.5161
##     P-Value [Acc > NIR] : 1.0000
##
##                   Kappa : -0.7902
##
##  Mcnemar's Test P-Value : 0.6108
##
##             Sensitivity : 0.12000
##             Specificity : 0.08750
##          Pos Pred Value : 0.10976
##          Neg Pred Value : 0.09589
##              Prevalence : 0.48387
##          Detection Rate : 0.05806
##    Detection Prevalence : 0.52903
##       Balanced Accuracy : 0.10375
##
##        'Positive' Class : female
##
```

### 2.4.5 Report section for a scientific publication

We find that, when predicting the sex of an individual, Hand Length and Foot Length can each be used to attempt to predict the sex of the individual. Using these measurements as indicators are both statistically significantly better than attempting to predict the sex of a person by random chance. Our model considering both of these factors achieves a training error of ~89.7%.

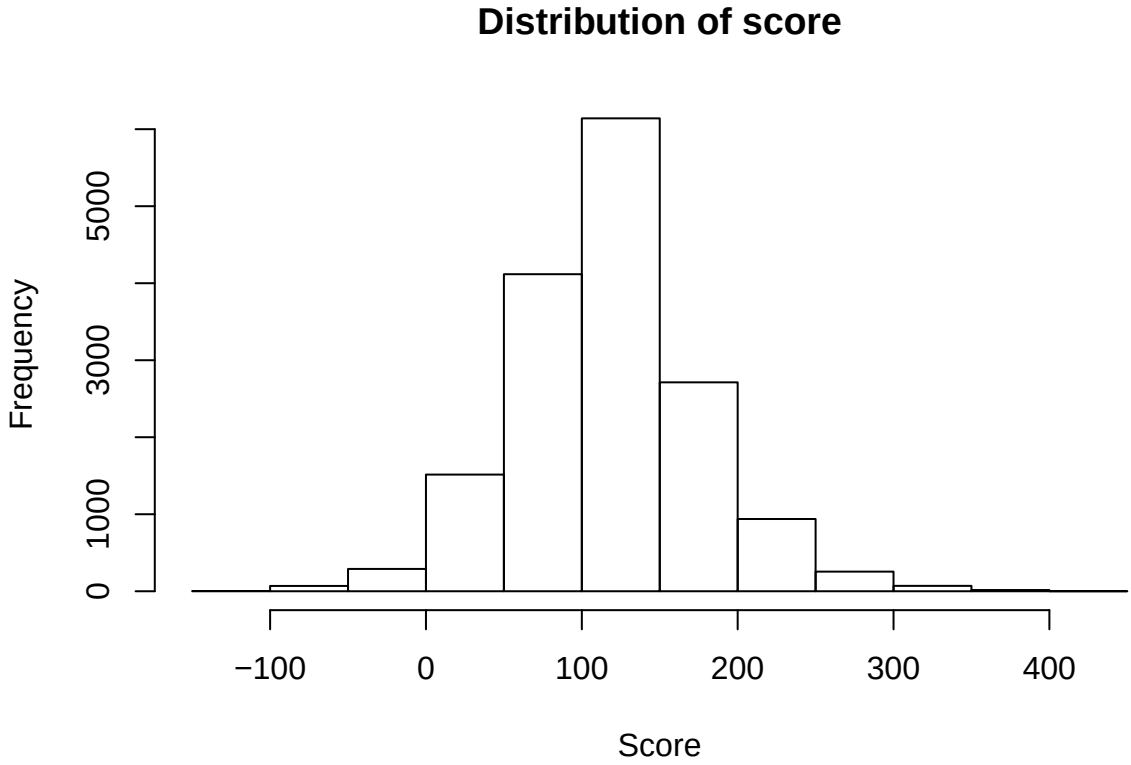
# 3 Part 3 - Multilevel model

Collecting data and setting up libraries...

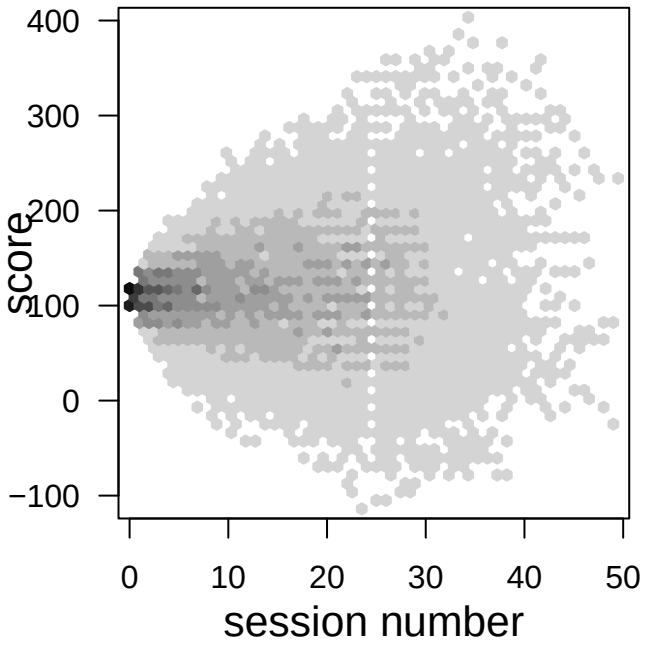
```
library(ggplot2)
library(hexbin)
library(lattice)
library(nlme)

learningData<-read.csv("./data/set1.csv", header = TRUE)
```

## 3.1 Visual inspection

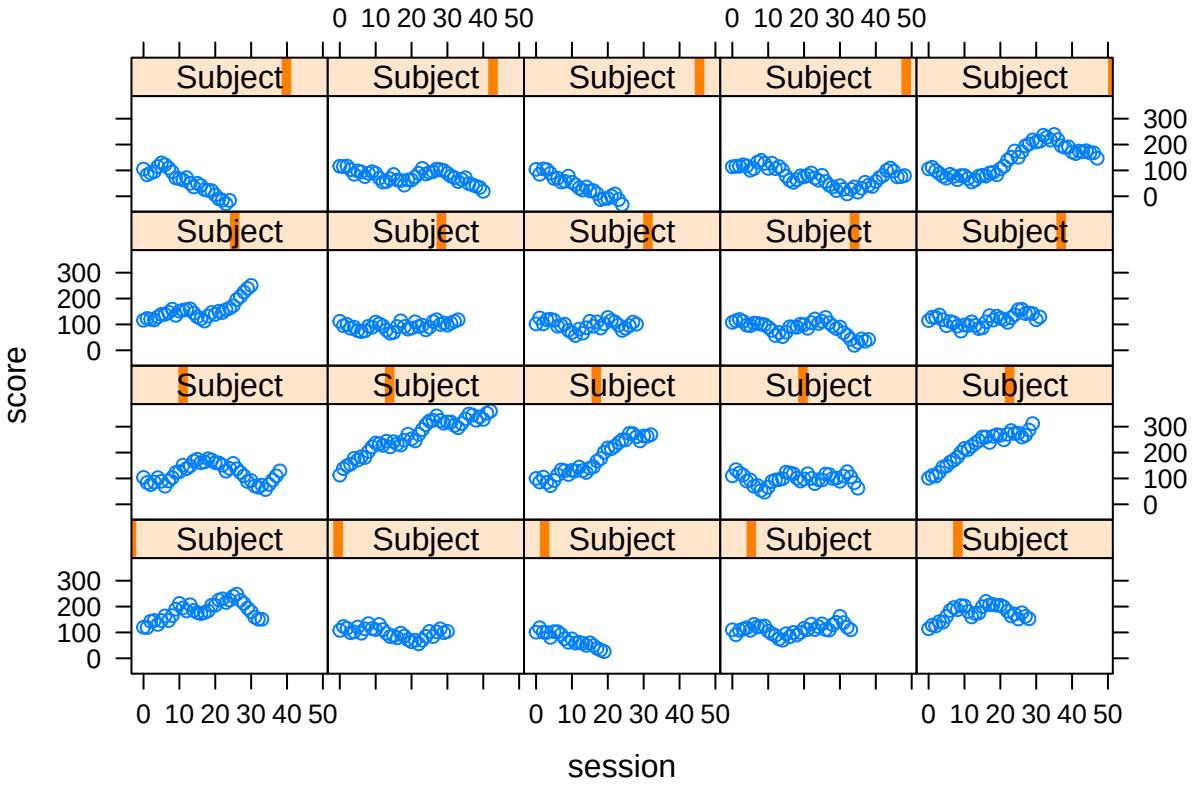
```
hist(learningData$score, xlab="Score", main="Distribution of score")
```




```
plot(hexbin(learningData$score ~ learningData$session, xbins=50, xlab="session number", ylab="score"))
```




```
xyplot(score~session | Subject, data=learningData[learningData$Subject %in% seq(1,191,10),])
```




Through a visual inspection:

- scores alone seem normally distributed

- When plotting a scatterplot of scores against session number, it appears that the score of a testee still centers around ~100 as session number increases, although one might see that the center increases slightly with session number. Also, the spread of scores increases as a function of session number.
- We sample some ~20 subjects from our data and plot their scores as a function of the session after which

they were tested. We see quite inconsistent trends, where some subjects score similarly before and after sessions, some subjects have increasing scores over time, and some subjects even have decreasing scores over time.

## 3.2 Multilevel analysis with scentific findings

Is there significant variance between the participants in their score?

```
randomInterceptOnly <- lme(score ~ 1, data = learningData,
                           random = ~1|Subject, method = "ML")
summary(randomInterceptOnly)
```

```
## Linear mixed-effects model fit by maximum likelihood
##  Data: learningData
##        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
## -4.22644590 -0.61530909  0.01016836  0.62959973  4.10477262
##
## Number of Observations: 16128
## Number of Groups: 501
```

```
intervals(randomInterceptOnly, 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
```

We find that there is very high variance between the scores of each subject on a given session. With a p-value of 0, we find approximately no chance that, if the distributions of scores as a function of lesson number came from the same distribution, we would find a collection of this sort of data. That is, We reject the null

hypothesis that the scores of a subject as a function of the number of lessons he had recieved are not from the same distribution. Subjects' scores respond differently to recieving lessons.

Does session have an impact on people score?

```
randomInterceptSession <- lme(score ~ session,
                              data = learningData, random = ~1|Subject, method = "ML")
summary(randomInterceptSession)
```

```
## Linear mixed-effects model fit by maximum likelihood
##  Data: learningData
##        AIC      BIC    logLik
##   162545.2 162575.9 -81268.58
##
## Random effects:
##  Formula: ~1 | Subject
##         (Intercept) Residual
## StdDev:     46.5146 35.06933
##
## Fixed effects: score ~ session
##                 Value Std.Error    DF  t-value p-value
## (Intercept) 111.0676  2.143371 15626 51.81911       0
## session       0.3682  0.028356 15626 12.98493       0
##  Correlation:
##         (Intr)
## session -0.206
##
## Standardized Within-Group Residuals:
##          Min           Q1          Med           Q3          Max
## -4.120041920 -0.613554431  0.009847298  0.627208531  3.952634923
##
## Number of Observations: 16128
## Number of Groups: 501
```

```
anova(randomInterceptOnly,randomInterceptSession)
```

```
##                         Model df      AIC      BIC    logLik   Test
## randomInterceptOnly         1  3 162710.9 162734.0 -81352.45
## randomInterceptSession      2  4 162545.2 162575.9 -81268.58 1 vs 2
##                          L.Ratio p-value
## randomInterceptOnly
## randomInterceptSession 167.7298  <.0001
```

We perform an analysis to observe the linear regressions of scores on the number of sessions from a subject across all subjects. We create linear models with and without use of session as an explanatory variable. We then compare these linear models, and with a p-value of <.001, we find that it cannot be the case that scores of a subject are independent of the number of sessions they attend. That is, constructing a linear model for the score of a subject is changed when we consider the number of sessions they attend.

Nehlig, A. 2010. "Is Caffeine a Cognitive Enhancer?" Journal Article. https://www.ncbi.nlm.nih.gov/pubmed/20182035.

Pasman, Wilrike J, Ruud Boessen, Yoni Donner, Nard Clabbers, and Andre Boorsma. 2017. "Effect of Caffeine on Attention and Alertness Measured in a Home-Setting, Using Web-Based Cognition Tests." Journal Article. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5608989/.