**Session 1: 1\_1 Getting your data.R**  
(1) Import the excel sheet Error\_data from Homophone\_ExpData

Store it in an object called Homophone\_Errors

Homophone\_Errors <- read\_excel("Homophone\_ExpData.xls", sheet = "Error\_data")

(2) What is the name of the 5th column of Homophone\_Errors?  
names(Homophone\_Errors)[5]  
  
Answer: "exception\_\_2"  
  
  
(3) What is the value of the 15th row in the 5th column?  
  
Homophone\_Errors[15,5]

Answer: 1

(4) Rename the first column of Homophone Error “Subject”

Students should be able to re-run and modify the code on line 160

names(Homophone\_Errors)[1]<-"Participant"

(5) Rename the first column of Homophone Error “Statistics”

Students should be able to re-run the code on line 162

names(Homophone\_Errors)[2]<-"Statistic"

(6) Check what is in columns 164 and 165

head(Homophone\_RT[,164])

tail(Homophone\_RT[,164])

head(Homophone\_RT[,165])

tail(Homophone\_RT[,165])

Answer: NAs

(7) Check for any junk columns in Homophone\_Errors

Checking through names, there are no extra/junk columns as for RT

(8) Get the Mean and Standard Deviation for reaction time data from the 5th column in Homophone\_RT

names(Homophone\_RT)[5] = “e\_10.control”

mean(Homophone\_RT$ e\_10.control, na.rm=TRUE) = 809.359

sd(Homophone\_RT$ e\_10.control,na.rm=TRUE) = 235.2315

(9) Complete an internet search to find out how to create a histogram in R.

When you have found the code, run it to create a histogram of the data in   
Homophone\_RT$e\_10.homophone

hist(Homophone\_RT$e\_10.homophone)



(10) In the Homophone\_RT data set, each participant is on one row. How would you calculate the average RT for one participant? That is, how would get the average/mean for a row?  
Write out that code and try it for participant 6. You can try an internet search to help!

Note that you have to specify columns 3:162, as columns 1 and 2 have text in them / non-numeric data.

rowMeans(Homophone\_RT[6, 3:162], na.rm = TRUE)   
  
= 805.1284

(11) Generate a histogram, density plot and qq/quantile plots for the Twitches column in Mydata, Make sure they all display on the same window

par(mfrow=c(2,2))

hist(Mydata$Twitches)

plot(density(Mydata$Twitches,na.rm=TRUE))

qqnorm(Mydata$Twitches)

qqline(Mydata$Twitches)

**Session 2: 1\_2 Manipulating Data.R**

1. Rename columns 5 and 6 in the Homophone\_RT\_Long data so that column 5 is labelled “Item” and column 6 is labelled “WordCondition”  
     
   names(Homophone\_RT\_Long)[5]<- "Item”

names(Homophone\_RT\_Long)[6]<-"WordCondition"

1. Can you do the above in ONE line of code?

names(Homophone\_RT\_Long)[5:6]<-c("Item","WordCondition")

1. Use dcast to generate ‘wide format’ data for each participant, with their average reaction time to each Item (collapsing across condition)  
     
   dcast(Homophone\_RT\_Long, Participant ~ Item, value.var = "RT",

fun.aggregate = mean, na.rm = TRUE)

1. Use melt to create a long form data set for the Homophone\_Errors data.   
   You do not need to make the Item/Word Condition labels.  
     
   Homophone\_Errors\_Long <- melt(Homophone\_Errors, id.vars = c("Subject", "Statistic"), variable.name = "Condition", value.name = "Accuracy")
2. From Mydata, select the columns SubNo, Trial, ACC, RT, Congruence and Twitches. Store this in a new object called ‘Mydata\_1’.

Mydata\_1 <- select(Mydata, SubNo, Trial, ACC, RT, Congruence, Twitches)

1. Using Mydata\_1, use melt and dcast to create a file with average RT for each Participant for the Congruent and Incongruent conditions in the ‘Congruence’ column.

dcast(Mydata\_1, SubNo ~ Congruence, value.var = "RT", fun.aggregate = mean, na.rm = TRUE)

1. Create a boxplot of the data above, showing average RT for the Congruent and Incongruent conditions.

temp <- dcast(Mydata\_1, SubNo ~ Congruence, value.var = "RT", fun.aggregate = mean, na.rm = TRUE)

boxplot(temp$Cong, temp$Incong)

**Session 3: 1\_3 Descriptive Statistics TTests Correlation**

1. Generate the average RT for each Item from the Homophone\_RT\_Long data set – use ddply and summarise  
     
   ddply(Homophone\_RT\_Long, "Item", summarise,

MeanRT = mean(RT, na.rm=TRUE))

1. Generate the average RT for each Item from the Homophone\_RT\_Long data set – use the psych package and describeBy  
     
   describeBy(Homophone\_RT\_Long$RT,Homophone\_RT\_Long$Item)
2. Generate a dataset that has the average Participant RT for the two conditions in WordCondition.   
   Hint: you can use ddply and summarise.   
   Store it in ‘temp’

*students should be able to re-run code from lines 16*-17  
temp <- ddply(Homophone\_RT\_Long, c("Participant","WordCondition"), summarise,

MeanRT = mean(RT, na.rm=TRUE))

1. From Mydata, Generate a dataset that has the average Participant RT for the two different experiments in Task.  
   Store it in ‘temp’.

temp <- ddply(Mydata, c("SubNo","Task"), summarise,  
MeanRT = mean(RT, na.rm=TRUE))

1. Create a boxplot for the above data, with labelled axes and a title

boxplot(temp$MeanRT~temp$Task, title = "Flanker vs CRT RTs",

ylab = "Mean RT in ms",

xlab = "Task", names = c("CRT", "Flanker"))

1. Read in the following data / use the following line of code:

movies <- read.csv(url("http://s3.amazonaws.com/dcwoods2717/movies.csv"))

1. Create a new columns called ‘profit’. This is the column ‘gross’ minus the ‘budget’. Hint: you can use mutate from dplyr to do this  
     
   movies <- mutate(movies, profit = gross - budget)
2. Correlate the ‘ratings’ with the ‘profit’. What do you find?

cor(movies$rating,movies$profit, method="spearman", use="pairwise.complete.obs")

or

library(Hmisc)

movies<-data.matrix(movies)

rcorr(movies[,c(11:12)], type="spearman")

1. How could you visualize/create a graph for this data?

plot(movies$rating,movies$profit)

**Session 4: 1\_4 Chi Square**

1. Complete a chi square analysis for the association between sex and smoking

temp <- table(survey$Sex, survey$Smoke)

chisq.test(temp)

**Session 5: 2\_1 regression / linear models**

(1) Write some code for a regression that predicts the movie’s profits from its budget, year and cast facebook likes. Save in lm.mov.2

lm.mov.2<- lm(profit ~ budget + year + cast\_facebook\_likes, data = movies)

(2) Now write code for a regression that predicts the movie’s profits from its budget, year, cast facebook likes and genre. Save in lm.mov.3

lm.mov.3<- lm(profit ~ budget + year + cast\_facebook\_likes + genre, data = movies)

(3) Now compare lm.mov.2 with lm.mov.3  
Does including the genre improve the prediction of the movies profit?

Why?

anova(lm.mov.2,lm.mov.3)

yes = significant improvement in model fit

summary(lm.mov.3)

certain genres significantly improve profit -adventure, animation, comedy, family and horror

(4) You should now have four models (lm.mov.1, 2 3 and 4). Look at the plots for each one – do the residuals look OK / normally distributed?   
Make sure you use plot and hist(resid()).  
  
plot(lm.mov.1)  
hist(resid(lm.mov.1))

plot(lm.mov.2)  
hist(resid(lm.mov.2))

plot(lm.mov.3)  
hist(resid(lm.mov.3))

plot(lm.mov.4)  
hist(resid(lm.mov.4))

(6) Go back to the code we used for correlation. Look at the correlation between the predictor variables we are using for our movie regression.  
budget, year, cast\_facebook\_likes, genre  
Hint: use corrplot to look at many correlations at the same time  
  
library(corrplot)

names(movies)

#genre is column 2, year is column 4, budget is column 7, facebook\_likes column # facebook\_likes is column 8

# Asking for correlation between these columns of data

M <- cor(movies[,c(2,4,7,8)],method="spearman",use="pairwise.complete.obs")

#get an error as not all of these are numbers

#need to check which are numbers

str(movies)

#only budget, facebook\_likes and year

M <- cor(movies[,c(4,7,8)],method="spearman",use="pairwise.complete.obs")

#repeat those correlations, but plot with labelled axes

names(movies)

rownames(M)<-names(movies)[ c(4,7,8)]

colnames(M)<-names(movies)[ c(4,7,8)]

corrplot(M, method = "number")

(7) Are any of the predictors correlated? Why?

Yes, budget and cast\_facebook\_likes are correlated positively.

1. Plot a line graph just for the predictor Year. What do you see?

plot(eff$year)

#Later years show lower profits