Exploratory Data Analysis in R (Handout 2)

Exploratory data analysis involves summarizing sample data and generating visual displays of the variables of interest.

1. Summary statistics and graphing for a single variable

A. Summary statistics & graphing for single **categorical** variable

```
library(descr)
# freq() function gives univariate output in the form of a frequency table and also generates a univariate
# graph in the form of a bar chart
freq(as.ordered(my_data$CategVar), main="plot title", names=c("category1","category2"),
y.axis="percent", ylab="y-axis label",xlab="x-axis label",col="color")

# add the argument y.axis="percent" to show percentages instead of counts on the y-axis

# alternatively a bar chart can be generated with the plot() function, main= etc… still apply
plot(my_data$CategVar)

# Note: Including \n inserts a hard return into a character string
# cex.names = 0.75 will reduce the font size of the category labels to 75%.
# Try different values but don't make the font size so small that it can't be read easily
```

B. Summary statistics & graphing for single **quantitative** variable

```
summary(my_data$QuantVar)
mean(my_data$QuantVar, na.rm = TRUE)
sd(my_data$QuantVar, na.rm = TRUE)

# the logical value na.rm tells R to remove NA values before calculating summary statistics

hist(my_data$QuantVar, main="plot title", xlab="x-axis label", ylab="y-axis label", col="color")
```

2. Summary statistics and graphing for an association between two variables

A. **C®C**

```
# Summary statistics is a table of proportions
tab1 <- table (my_data$CategResponseVar, my_data$CategExplanatoryVar)
tab1                                                    # table of counts

#use only if unused factors are showing for one of the variables
tab1 <- table (droplevels(my_data$CategResponseVar), droplevels(my_data$CategExplanatoryVar))
tab1

tab1_colProp <-prop.table(tab1, 2)          # column proportions indicated by the 2
round(tab1_colProp, 3)                      # display table of proportions rounded to 3 decimal places

# Graph is a bar plot where each column is an explanatory category and the height of the columns
# are the sample proportions for ONE response category indicated by row_num
barplot (tab1_colProp [row_num, ], ylim = c(0, 0.5))

# Note: use ylim to set the lower and upper limits for the vertical axis, main = etc. still apply
# indexing tables is by [row,col], if col is blank as above, then all columns in the row will be plotted
```

B. **C®Q**

```
# Summary statistics are the sample means and standard deviations for each category of the
# explanatory variable. Also, the length function gives the sample size for each category.
# Use either the tapply() or by() functions
 tapply(my_data$QuantResponseVar, my_data$CategExplanatoryVar, mean, na.rm = TRUE)
 tapply(my_data$QuantResponseVar, my_data$CategExplanatoryVar, sd, na.rm = TRUE)
 tapply(my_data$QuantResponseVar, my_data$CategExplanatoryVar, length)

 by(my_data$QuantResponseVar, my_data$CategExplanatoryVar, mean, na.rm = TRUE)
 by(my_data$QuantResponseVar, my_data$CategExplanatoryVar, sd, na.rm = TRUE)
 by(my_data$QuantResponseVar, my_data$CategExplanatoryVar, length)

 #Graph of C → Q is either a box plot for each category or a bar plot of category means.
 #box plot:
 plot(QuantResponseVar ~ factor(CategExplanatoryVar), data=my_data, names=c("cat1", "cat2",
 "cat3", "cat4"), main="title")

 #bar plot:
 groupMeans1 <- by(my_data$QuantResponseVar, my_data$CategExplanatoryVar, mean,
 na.rm=TRUE)
 barplot(groupMeans1)          # main = and all other graphics parameters apply

 # If the by() function isn't working you can create a your own vector that lists the group means and
 # then becomes the first argument of the barplot() function.
```

C. **Q®Q**

```
# summary statistics are the mean and standard deviation for the response variable and the mean
# and standard deviation of the explanatory variable.
# Graph for Q → Q is a scatterplot with a line of best fit.
plot(QuantResponseVar ~ QuantExplanatoryVar, data = my_data)
abline(lm(my_data$QuantResponseVar ~ my_data$QuantExplanatoryVar))
```

3. Multivariate Output and Graphing

A. Multivariate Output & Graphing for Categorical Explanatory Variable, Quantitative Response Variable, Categorical 3rd Variable: **C+C®Q**

```
#library(stats) needed for the ftable() function

tbl <- ftable(by(my_data$QuantResponseVar,
        list(my_data$CategThirdVar, my_data$CategExpVar), mean, na.rm = TRUE))

barplot(tbl, beside = TRUE, main =, names.arg=c(), ylab= ,xlab= ,col=c("col1", "col2"))
legend("topleft", c(), fill=c("col1","col2"))

# names.arg= lists the labels for the categorical explanatory variable
# legend() lists the names of the categorical third variable, the location can be "topleft", "topright" or
# other options listed in R help.
```

B. Multivariate Output & Graphing for Categorical Explanatory Variable, Categorical Response Variable, Categorical 3rd Variable: **C+C®C**

```
# The code below works best if both the categorical response variable (row var) and categorical third
# moderating variable have only two categories. The categorical explanatory variable (col var) can have
# more than two categories.

tbl <- ftable(my_data$colVar ~ my_data$rowVar + my_data$thirdModVar)
tbl
p_tbl <- prop.table(tbl, 2)        # column proportions
p_tbl                              # widen the console screen and study the table carefully to understand
                                   # the output, identify which two rows you want to plot side-by-side.

barplot(p_tbl[3:4, ], beside = TRUE, main =, names.arg = c(), ylab=, xlab=, col=c())
legend("topleft", c(), fill=c("col1","col2"))

# The example above is plotting rows 3 and 4 of the prop table side-by-side.
# Study the prop table to identify which two rows make sense for your situation.
# names.arg= lists the labels for the categorical explanatory variable (horizontal axis)
# legend() lists the names of the categorical third variable, the location can be “topleft”, “topright” or
# other options listed in R help.

# An alternative approach to graphing C to C with moderation is to use the lattice graphics package
library(lattice)                   # may need to install the package the first time

histogram(~ CategResponseVar | CategExpVar, data=my_data)                    # C®C
histogram(~ CategResponseVar | CategExpVar + CategThirdVar, data=my_data)  # C+C®C

# Instead of requiring the $ symbol, this function takes the data frame name as the second argument and
# then variable names can be used directly.

# The plot will be very difficult to interpret unless you rename response codes with understandable labels
levels(my_data$VAR1)
levels(my_data$VAR1) <- c("value0label", "value1label", "value2label", "value3label")

# make sure you reassign the levels in the correct order!
```

C. Multivariate Output & Graphing for Quantitative Explanatory Variable, Quantitative Response Variable, Categorical 3rd Variable: **Q+C®Q**

```
#The example below assumes the categorical moderating variable has two levels coded 1 and 2.

plot(my_data$QuantResponsVar[my_data$CategThirdVar == 1] ~
       my_data$QuantExpVar[my_data$CatThirdVar == 1]), ylab = " ", xlab = " ", col = "col1")

abline(lm(my_data$QuantResponsVar[my_data$CategThirdVar == 1] ~
       my_data$QuantExpVar[my_data$CatThirdVar == 1]), col = "col1")

points(my_data$QuantResponsVar[my_data$CategThirdVar == 2] ~
       my_data$QuantExpVar[my_data$CatThirdVar == 2], col = "col2")

abline(lm(my_data$QuantResponsVar[my_data$CategThirdVar == 1] ~
       my_data$QuantExpVar[my_data$CatThirdVar == 1]), col = "col")
```

```
legend("topright", c("Category 1", "Category 2"), fill = c("col1","col2"))
```

4. Sources of help for data visualization in R:
   A. To get a complete list of graphical parameters (ie. how to set character widths, fonts, colors…) type ***?par()*** into the console. Type ***?legend()*** to see all of the settings for the legend in a side-by-side bar chart.
   B. Watch the following videos on bivariate graphing:

      From passion driven statistics:

      https://www.youtube.com/watch?v=Yf121fdtjDA&list=PLDEF0B9CBD27AD37E&index=55

      This video covers a lot of the graphing parameters, as well as a strategy for moderation by gender of a scatterplot: Q to Q by Gender:

      https://www.youtube.com/watch?v=lPOSwfxMd3c

      More about bar charts:

      https://www.youtube.com/watch?v=rl1tB9p3FLg&index=19&list=PLqzoL9-eJTNBDdKgJgJzaQcY6OXmsXAHU