Data exploration & graphics with R Studio

In this exercise, we will work with data on Canadian occupational prestige from 1971, in relation to education, income and % women for 102 occupational categories. The data is in the car package, as Prestige. We will also learn to use R Studio somewhat more. You should be familiar with the layout of panels and what they do.

For this data set, we would be interested in predicting prestige from the other variables. The focus here is on using graphics to explore variable distributions and relations among them.

Start R Studio from the desktop icon or the Start menu.

RStudio File Edit Code View Plots Session Build Debug Profile Tools Help 🍳 📲 🚭 📲 🔝 🔒 🚺 🍝 Go to file/function 🛛 😨 🔹 Addins 🔹 Reviect: (None) Inv-ex1.R × DataExplorationR.R × Environment History 1 #' --- 📄 🔐 Import Dataset 👻 🞻 Sour 🗏 List 🗕 🥵 2 #' title: "Data exploration with R"
#' author: "Wichael Friendly"
#' date: "06 oct 2015"
#' ----Global Environment 👻 values Ofit1 List of 30 Ofit2 List of 30 library(car) Script panels(s) data(Prestige) str(Prestige) summary(Prestige) Environment: active objects History: previous commands boxplot(prestige ~ type, data=Prestige, ylab="prestige")
#' prestige should be considered an ordered factor 13 14 * prestige should be considered an ordered factor PrestigeStype <- ordered(PrestigeStype, levels=c("bc", "wc", "prof boxplot(prestige ~ type, data=Prestige, ylab="prestige") 14 #' prestig 15 prestige\$t 16 boxplot(pr 17 18 19 20 #' boxplot 21 * boxplot 21 * boxplot 21 * boxplot 21 * boxplot Files Plots Packages Help Viewer #' boxplots for symmetry 😳 New Folder 🛛 🝳 Delete 👍 Rename 🛛 🔮 More 🗣 $\fbox{$\square > C: > Dropbox > Documents > 6140 > tutorials}$ R Script ‡ Name Size Modified t -----Console ~/ DataExplorationR.R 1.7 KB Sep 28, 2016, 1:01 PM R version 3.2.5 (2016-04-14) -- "Very, Very Secure Dishes" Copyright (C) 2016 The R Foundation for Statistical Computing Platform: x86_64-w64-mingw32/x64 (64-bit) Sep 22, 2016, 11:21 PM Sep 22, 2016, 12:50 PM matrixR.html Files: file browset, 2016, 9:30 PM R is free software and comes with ABSOLUTELY NO WARWYCONSOLE You are welcome to redistribute it under certain controls Sole Type 'license()' or 'licence()' for distribution details. ■ ¹/_{matrixR.R} Plots: grap²⁷/_{MS}s Sep 21, 2016, 9:30 PM Sep 21, 2016, 9:23 PM Sep 21, 2016, 9:23 PM R is a collaborative project with many contributors. Type 'contributors()' for more information and 'citation()' on how to cite R or R packages in publications. □ ThroR.pdf Help introR.R 1.6 KB Sep 6, 2016, 3:30 PM 136.7 KB Sep 6, 2016, 3:19 PM Type 'demo()' for some demos, 'help()' for on-line help, or 'help.start()' for an HTML browser interface to help. Type 'q()' to quit R. Sep 6, 2016, 3:19 PM 🔲 👻 IntroR.docx 127.2 KB Sep 6, 2016, 2:57 PM introR.R.bak 1.6 KB Aug 22, 2016, 1:42 PM 🔲 🔁 R_install_guide.pdf 43.6 KB [Workspace loaded from ~/.RData] MixedModels.pdf 47.2 KB Apr 5, 2016, 9:53 AM > | MixedModels.docx 30.3 KB Apr 5, 2016, 9:53 AM 🔲 🖭 psych9-sem.R Mar 29, 2016, 9:58 AM 1.1 KB 🔲 🔁 CFA.pdf 30.7 KB Mar 29, 2016, 9:57 AM CFA.docx 27.3 KB Mar 29, 2016, 9:56 AM psych9-sem.R.bak 1.1 KB Mar 29. 2016. 9:41 AM

In this exercise, you should create a new R script to save your work (File -> New file -> R script, or Ctrl+shift+N). Change the working directory to your home directory (X:) and save there.

1. Load the car package, and get a quick summary of the variables in the Prestige data set.

```
library(car)
data(Prestige)
str(Prestige)
summary(Prestige)
```

2. Boxplots are useful for showing the distributions of variables, often stratified by a factor. Note the difference between the two plots below.

boxplot(prestige ~ type, data=Prestige, ylab="prestige")
#' prestige should be considered an ordered factor
Prestige\$type <- ordered(Prestige\$type, levels=c("bc", "wc", "prof"))
boxplot(prestige ~ type, data=Prestige, ylab="prestige")</pre>

3. Other uses of boxplots include comparing the shapes of a distribution with various transformations, usually to make them more symmetric.

```
#' boxplots for symmetry
symbox(~ education, data=Prestige)
symbox(~ income, data=Prestige)
```

4. Examine the bivariate relations among the variables with a scatterplot matrix. The basic plot in R would be plot(Prestige), but car::scatterplotMatrix() has many more options.

```
scatterplotMatrix(~prestige + education + income + women, data=Prestige)
scatterplotMatrix(~prestige + education + income + women | type,
data=Prestige)
```

5. You can create new variables or transform/recode existing ones in many ways. The simplest way is just to assign new variable names in the data frame.

```
Prestige$educ2 <- Prestige$education^2
Prestige$loginc <- log10(Prestige$income)</pre>
```

6. Let's fit a model predicting prestige from the original income and education (educ). The R function is lm(). When you have fit a model, there are lots of functions to explore it

```
mod1 <- lm(prestige ~ education + income, data=Prestige)
summary(mod1)
plot(mod1)</pre>
```

Try fitting another regression model, this time using loginc and educ2 as predictors.

7. The car package has a bunch of other functions that produces nicer versions of these plots.

```
qqPlot(mod1, id.n=2)
influencePlot(mod1, id.n=2)
```

The script for this exercise is on the N: drive, N:\psy6140\tutorials\DataExploration.R. You can also find it on the web, <u>http://friendly.apps01.yorku.ca/psy6140/tutorials/DataExplorationR.R</u>. It also includes sections on 3D plots and multivariate outliers.