Exploring the Structure of Data in RStudio

Data

During the 1996 election between Bill Clinton & Bob Dole, the American National Election Study collected data on the party affiliations, demographics, and voting behavior of a sample of Americans. (Data downloaded from ICPSR.)

Cases & Variables

variables
attributes of the cases - these can *vary* from case to case

cases
individual objects in the sample

PID	age	educ	income	vote	popul	TVnews
Republican	36	HS	1.5	Dole	0	7
Democrat	20	College	1.5	Clinton	190	1
Democrat	24	College	1.5	Clinton	31	7
Democrat	28	College	1.5	Clinton	83	4
Democrat	68	College	1.5	Clinton	640	7
Democrat	21	College	1.5	Clinton	110	3
Democrat	77	College	1.5	Clinton	100	7
Independent	21	College	1.5	Clinton	31	1
Independent	31	College	1.5	Clinton	180	7

Quantitative vs Categorical

PID	age	educ	income	vote	popul	TVnews
Republican	36	HS	1.5	Dole	0	7
Democrat	20	College	1.5	Clinton	190	1
Democrat	24	College	1.5	Clinton	31	7
Democrat	28	College	1.5	Clinton	83	4
Democrat	68	College	1.5	Clinton	640	7
Democrat	21	College	1.5	Clinton	110	3
Democrat	77	College	1.5	Clinton	100	7
Independent	21	College	1.5	Clinton	31	1
Independent	31	College	1.5	Clinton	180	7

Import Data into RStudio

How? It depends on the *structure* and *storage location* of the data. For example, I've stored the election data as a *csv* file and posted it *online* at

http://www.macalester.edu/~ajohns24/data/partyID.csv
(http://www.macalester.edu/~ajohns24/data/partyID.csv)

In this case, we can import the data into RStudio using the `read.csv` function. We'll store this data under the name `election`:

```
election = read.csv("http://www.macalester.edu/~ajohns24/data/partyID.csv")
```

NOTE:

We can name and store (thus later retrieve) data, numbers, etc in RStudio. The names we use *cannot contain any spaces*. For example, trying to store the data as `election data` would return an error message.

Import Data into RStudio

Let's confirm that this actually worked and get a sense of what the data look like in RStudio. The `View` function will print the data in the upper left hand window of RStudio:

```
View(election)
```

Typing the name of the data alone will print the whole mess in the console:

```
election
```

For simplicity, we can just check out the first 6 cases:

```
head(election)
```

```
##          PID age    educ income    vote popul TVnews
## 1 Republican  36      HS    1.5    Dole     0      7
## 2   Democrat  20 College    1.5 Clinton   190      1
## 3   Democrat  24 College    1.5 Clinton    31      7
## 4   Democrat  28 College    1.5 Clinton    83      4
## 5   Democrat  68 College    1.5 Clinton   640      7
## 6   Democrat  21 College    1.5 Clinton   110      3
```

Explore the Data Structure

Before doing any analysis, we must understand the key features of our data. To this end, we'll discuss the following functions. NOTE: RStudio ignores any content after the pound sign #. Thus we use this to 'comment' and organize our code.

```
head(election)        #the first 6 rows
dim(election)         #dimensions = number of cases & variables
names(election)       #labels/names of variables
summary(election)     #summary statistics of each variable
levels(election$PID)  #labels of levels in a categorical variable
```

We can also *subset* our data:

```
election$PID                                      #just the PID variable
indy = subset(election, PID == "Independent")     #just the independent cases
```

Explore the Data Structure

Data dimensions (the # of cases & variables, respectively)

```
dim(election)
```

```
## [1] 944   7
```

Variable names

```
names(election)
```

```
## [1] "PID"    "age"    "educ"   "income" "vote"   "popul"  "TVnews"
```

Explore the Data Structure

Quick summary statistics of each variable

```
summary(election)
```

```
##          PID           age             educ         income      
##  Democrat   :380   Min.   :19.00   College:631   Min.   :  1.50  
##  Independent:239   1st Qu.:34.00   HS     :313   1st Qu.: 23.50  
##  Republican :325   Median :44.00                 Median : 37.50  
##                    Mean   :47.04                 Mean   : 46.58  
##                    3rd Qu.:58.00                 3rd Qu.: 67.50  
##                    Max.   :91.00                 Max.   :115.00  
##       vote          popul            TVnews      
##  Clinton:551   Min.   :   0.0   Min.   :0.000  
##  Dole   :393   1st Qu.:   1.0   1st Qu.:1.000  
##                Median :  22.0   Median :3.000  
##                Mean   : 306.4   Mean   :3.728  
##                3rd Qu.: 110.0   3rd Qu.:7.000  
##                Max.   :7300.0   Max.   :7.000
```

Working with Subsets: Variables

Recall that the 7 variables in our data have the following labels:

```
names(election)
```

```
## [1] "PID"    "age"    "educ"   "income" "vote"   "popul"  "TVnews"
```

Suppose we're only interested in a *subset* of these variables. We can isolate the data on a single variable say, `PID`, using the `$` notation:

```
election$PID
```

Working with Subsets: Cases

Or we might only be interested in certain *cases*, for example voters that identified as "independents." First, let's check out the 'coding' of the PID variable to determine how independents are labeled in our data:

```
levels(election$PID)
```

```
## [1] "Democrat"    "Independent" "Republican"
```

Next, let's create a smaller data set with only the `Independent` voters:

```
indy = subset(election, PID == "Independent")
dim(indy)
```

```
## [1] 239   7
```

Quiz

On the previous slide, we created the subset `indy` of only independent voters:

```
indy = subset(election, PID == "Independent")
```

Consider two more attempts to create this subset, `indy1` and `indy2`. What errors are we making and how does RStudio respond? (See answers on next slides.)

```
indy1 = subset(election, PID = "Independent")
indy2 = subset(election, PID == "independent")
```

Quiz Answers

With `indy1` we typed `PID =` instead of `PID ==`. This is a dangerous mistake because RStudio didn't return an error message. Rather, it defined `indy1` to be the original data set (not just independents). Always check your data to make sure it is as you intended.

```
indy1 = subset(election, PID = "Independent")
head(indy1)
```

```
##          PID age    educ income    vote popul TVnews
## 1 Republican  36      HS    1.5    Dole     0      7
## 2   Democrat  20 College    1.5 Clinton   190      1
## 3   Democrat  24 College    1.5 Clinton    31      7
## 4   Democrat  28 College    1.5 Clinton    83      4
## 5   Democrat  68 College    1.5 Clinton   640      7
## 6   Democrat  21 College    1.5 Clinton   110      3
```

Quiz Answers

With `indy2` we typed `PID == "independent"` instead of `PID == "Independent"`. This is a dangerous mistake because RStudio didn't return an error message. Rather, it defined `indy2` to be an *empty* data set. Always check your data to make sure it is as you intended.

```
indy2 = subset(election, PID == "independent")
head(indy2)
```

```
## [1] PID    age    educ   income vote   popul  TVnews
## <0 rows> (or 0-length row.names)
```