Data Management in R

Sam Marcotte
Neil Williams

Objectives

- Recoding
- Common problems/mistakes/solutions
- Missingness
- Dealing with real data
- General Tips

The Tidyverse

```
install.packages("tidyverse")
install.packages("dplyr")
```

Data types

- `df`, `data.frame`
- `tibble`
- `matrix`
- `list()`

Variable types

- `numeric`
- `factor`
- `character`

Name columns in data

- Change variable names in base r

```
names(data) <- c("new_name", "another_new_name")
colnames(data)[colnames(data)=="old_name"] <- "new_name"
```

- Tidyverse

```
rename(data, new_name = old_name)
select(data, variable = starts_with('S')) #rename
#columns in a group
```

Recoding

- In R

```
data$Variable[data$Variable== 1] <- "Value"
data$Variable[data$Variable > 1] <- "Value2"

recode(Variable, `1` = "Value", .default = "Value2") #dplyr
```

- In Stata

```
replace Variable = "Value" if Variable == 1
replace Variable = "Value2" if Variable > 1
```

Missing data

- In R

```
data$Variable[data$Variable < 3 &
              data$Variable2 == "True"] <- NA
```

Missing data

- In R

```
# create new dataset without missing data
newdata <- na.omit(data)

# list rows of data that have missing values
mydata[!complete.cases(mydata),]
```

- Handling missing data in Stata

Creating variables

- In R

```
data$Var_log <- log(data$Variable)
```

- In Stata

```
gen Var_log = log(Variable)
```

Removing variables

- In R

```
newdata <- mydata[!mpg]
```

- In Stata

```
drop if mpg > 21
drop mpg
```

Applications

- Comparative Study of Electoral Systems (CSES) Integrated Dataset

```
library(haven)
cses_work <- read_dta("cses_work.dta")
```

```
ncol(cses_work)
```

```
## [1] 137
```

CSES Codebook

```
))) CSES IMD VARIABLES: DEMOGRAPHIC DATA

    IMD2001_1    >>>         AGE OF RESPONDENT (IN YEARS)
    IMD2001_2    >>>         AGE OF RESPONDENT (IN CATEGORIES)
    IMD2002      >>>         GENDER
    IMD2003      >>>         EDUCATION
    IMD2004      >>>         MARITAL STATUS
    IMD2005      >>>         RELIGIOUS DENOMINATION
    IMD2006      >>>         HOUSEHOLD INCOME
    IMD2007      >>>         RURAL OR URBAN RESIDENCE


))) CSES IMD VARIABLES: MICRO-LEVEL (SURVEY) DATA

    IMD3001      >>>         TURNOUT - MAIN ELECTION
    IMD3001_PR_1 >>>         TURNOUT - CURRENT PRESIDENTIAL ELECTION -
                             ROUND 1
    IMD3001_PR_2 >>>         TURNOUT - CURRENT PRESIDENTIAL ELECTION -
                             ROUND 2
    IMD3001_LH   >>>         TURNOUT - CURRENT LOWER HOUSE ELECTION
    IMD3001_UH   >>>         TURNOUT - CURRENT UPPER HOUSE ELECTION
    IMD3002_PR_1 >>>         CURRENT PRESIDENTIAL ELECTION: VOTE CHOICE -
                             ROUND 1
    IMD3002_PR_2 >>>         CURRENT PRESIDENTIAL ELECTION: VOTE CHOICE -
                             ROUND 2
    IMD3002_LH_PL >>>        CURRENT LOWER HOUSE ELECTION: VOTE CHOICE -
                             PARTY LIST
    IMD3002_LH_DC >>>        CURRENT LOWER HOUSE ELECTION: VOTE CHOICE -
                             DISTRICT CANDIDATE
    IMD3002_UH_PL >>>        CURRENT UPPER HOUSE ELECTION: VOTE CHOICE -
                             PARTY LIST
    IMD3002_UH_DC_1 >>>      CURRENT UPPER HOUSE ELECTION: VOTE CHOICE -
                             DISTRICT CANDIDATE 1
    IMD3002_UH_DC_2 >>>      CURRENT UPPER HOUSE ELECTION: VOTE CHOICE -
                             DISTRICT CANDIDATE 2
    IMD3002_UH_DC_3 >>>      CURRENT UPPER HOUSE ELECTION: VOTE CHOICE -
                             DISTRICT CANDIDATE 3
    IMD3002_UH_DC_4 >>>      CURRENT UPPER HOUSE ELECTION: VOTE CHOICE -
                             DISTRICT CANDIDATE 4
    IMD3002_OUTGOV  >>>      CURRENT MAIN ELECTION: VOTE CHOICE - OUTGOING
                             GOVERNMENT (INCUMBENT)
    IMD3003_PR_1 >>>         TURNOUT - PREVIOUS PRESIDENTIAL ELECTION -
                             ROUND 1
    IMD3003_PR_2 >>>         TURNOUT - PREVIOUS PRESIDENTIAL ELECTION -
                             ROUND 2
```

Figure 1: A peek at the CSES Codebook

Organizing loaded data

```
cses_org <- cses_work %>%
  dplyr::select(Country = IMD1006_NAM,
                Country_year = IMD1004,
                Year  = IMD1008_YEAR,
                Age = IMD2001_1, #in years
                Female = IMD2002, #1 is male, 2 is female
                Education = IMD2003,
                Income = IMD2006,
                Urban = IMD2007,
                Turnout = IMD3001,
                Ideology = IMD3006,
                Identifier =  IMD3005_1,
  )%>%
  drop_na(Turnout, Ideology, Age)
```

Summarize data

- `head()` for visualizing the top rows

```
head(cses_org)
```

Country	Country_year	Year	Age	Female	Education
Albania	ALB_2005	2005	41	2	3
Albania	ALB_2005	2005	66	1	1
Albania	ALB_2005	2005	43	2	2
Albania	ALB_2005	2005	19	2	2
Albania	ALB_2005	2005	43	2	3
Albania	ALB_2005	2005	61	1	2

Summarize data

- And use `tail()` for the bottom rows

```
tail(cses_org)
```

Country	Country_year	Year	Age	Female	Education
South Africa	ZAF_2014	2014	37	2	1
South Africa	ZAF_2014	2014	45	2	1
South Africa	ZAF_2014	2014	56	2	1
South Africa	ZAF_2014	2014	36	1	2
South Africa	ZAF_2014	2014	23	1	1
South Africa	ZAF_2014	2014	22	1	2

Summarize data

- `str()` or the "structure" command is useful to get a sense of your data frame

```
str(cses_org)
```

- `summary()` also gives descriptive statistics of data frame

```
summary(cses_org)
```

Subsetting Filtering data

- Base R

```
cses_new <- cses_org[which(cses_org$Female==1
& cses_org$Age > 25), -which(names(cses_org) %in%
                                c("Ideology", "Identifier"))]
```

- subset() is also a good option

```
cses_new <- subset(cses_org, Female==1 & Age > 25,
                   select=Country:Turnout))
```

Filtering data

- The `filter()` function is very useful!!
- Can also use this with other functions within the pipe

```
cses_org <- cses_org %>%
  filter(Female <= 2)%>%
  mutate(Female = (Female - 1))%>%
  filter(Year > 2005)%>%
  filter(!grepl('South', Country))
```

```
unique(cses_work$IMD2002)
```

```
## [1] 2 1 9
```

```
unique(cses_org$Female)
```

```
## [1] 1 0
```

Look at our data again

Country	Country_year	Year	Age	Female	Education
Argentina	ARG_2015	2015	41	1	2
Argentina	ARG_2015	2015	61	0	1
Argentina	ARG_2015	2015	50	1	2
Argentina	ARG_2015	2015	58	1	1
Argentina	ARG_2015	2015	29	1	3
Argentina	ARG_2015	2015	35	0	2

Look at our data again

Country	Country_year	Year	Age	Female
United States of America	USA_2012	2012	30	0
United States of America	USA_2012	2012	68	0
United States of America	USA_2012	2012	19	1
United States of America	USA_2012	2012	52	1
United States of America	USA_2012	2012	63	0
United States of America	USA_2012	2012	18	0

What if we want to add another data source?

- Varieties of Democracy Project (Wave 8)
- Information on country-level indicators about democracy

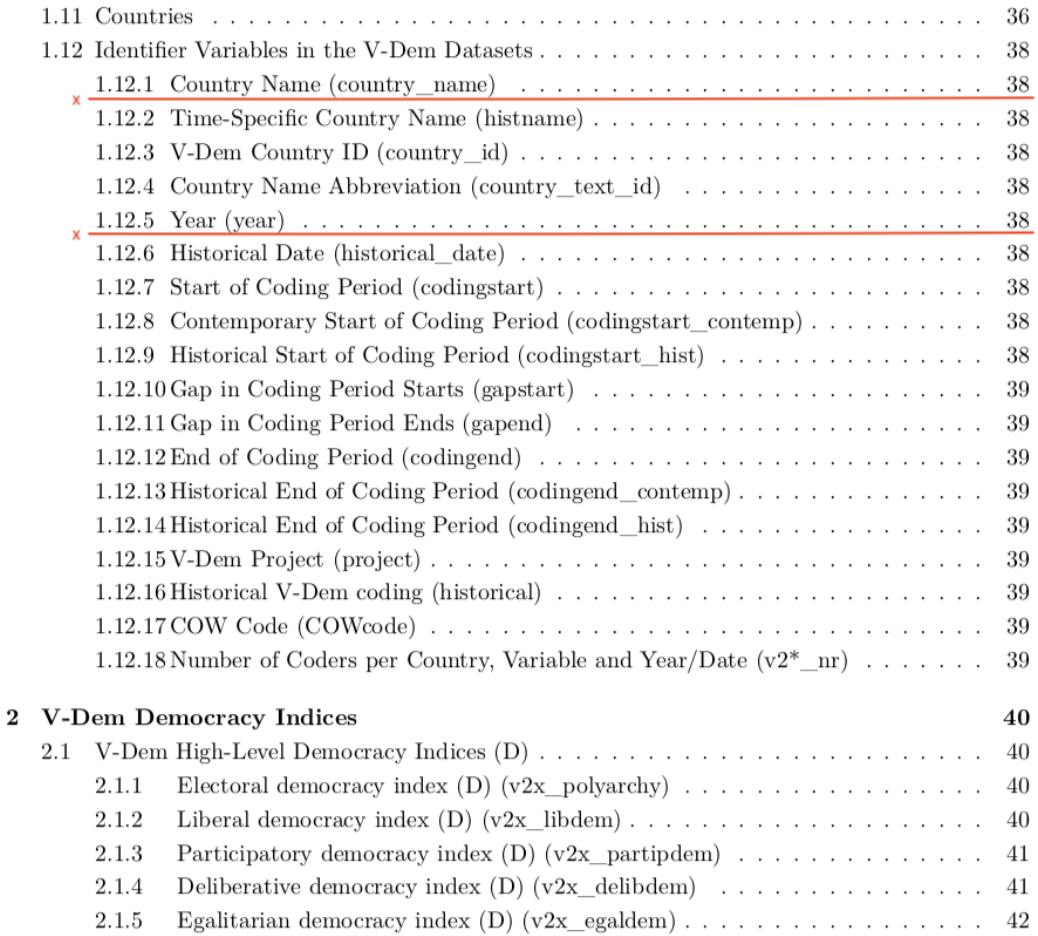
```
vdem_work <- read_dta("vdem_work.dta")
```

Combining data

- `rbind()` (same number of columns) and `cbind()` (same number of rows)
- `paste()`
  - combines values of separate vectors into one vector and pushes the separate values into one
- `coalesce()`
  - returns non-null elements when combining vectors or rows.
- We can combine datasets at different units of analysis by matching columns using the `merge()` command.

Merging

- Search for columns of interest in common between the two datasets



1.11 Countries 36
1.12 Identifier Variables in the V-Dem Datasets 38
1.12.1 Country Name (country_name) 38
1.12.2 Time-Specific Country Name (histname) 38
1.12.3 V-Dem Country ID (country_id) 38
1.12.4 Country Name Abbreviation (country_text_id) 38
1.12.5 Year (year) 38
1.12.6 Historical Date (historical_date) 38
1.12.7 Start of Coding Period (codingstart) 38
1.12.8 Contemporary Start of Coding Period (codingstart_contemp) 38
1.12.9 Historical Start of Coding Period (codingstart_hist) 38
1.12.10 Gap in Coding Period Starts (gapstart) 39
1.12.11 Gap in Coding Period Ends (gapend) 39
1.12.12 End of Coding Period (codingend) 39
1.12.13 Historical End of Coding Period (codingend_contemp) 39
1.12.14 Historical End of Coding Period (codingend_hist) 39
1.12.15 V-Dem Project (project) 39
1.12.16 Historical V-Dem coding (historical) 39
1.12.17 COW Code (COWcode) 39
1.12.18 Number of Coders per Country, Variable and Year/Date (v2*_nr) 39

2 V-Dem Democracy Indices 40
2.1 V-Dem High-Level Democracy Indices (D) 40
2.1.1 Electoral democracy index (D) (v2x_polyarchy) 40
2.1.2 Liberal democracy index (D) (v2x_libdem) 40
2.1.3 Participatory democracy index (D) (v2x_partipdem) 41
2.1.4 Deliberative democracy index (D) (v2x_delibdem) 41
2.1.5 Egalitarian democracy index (D) (v2x_egaldem) 42

Figure 2: Varieties of Democracy codebook

Merging

```
cses_merge <- merge(cses_org, vdem_work,
                    by = c("Country", "Year"),
                    type = "left", match = "all")
```

Country	Year	Age	Female	Polity_score	GDP_growth_capita
Argentina	2015	61	0	9	0.0166316
Argentina	2015	38	1	9	0.0166316
Argentina	2015	46	0	9	0.0166316
Argentina	2015	41	1	9	0.0166316
Argentina	2015	28	1	9	0.0166316
Argentina	2015	50	1	9	0.0166316

Summarizing this data by group

- What if you want to see variations by Country or Year?
- Could use the base R `aggregate()` function
- or `group_by()` and `summarize_at()`

Summarizing Turnout and Gender

```
unique(cses_merge$Turnout)
```

```
## [1]       1 9999997       0 9999999 9999998 9999993
```

```
cses_merge$Turnout[cses_merge$Turnout > 1 ] <- NA
```

```
cses_summary <- cses_merge %>%
  group_by(Country)%>%
  summarize_at(vars(Turnout.mean = "Turnout",
                    Female.mean = "Female"),
               mean, na.rm = TRUE)
```

Summarizing Turnout and Gender

Country	Turnout.mean	Female.mean
Argentina	1.0000000	0.5263158
Australia	1.0000000	0.6111111
Austria	0.9000000	0.4000000
Belarus	0.7142857	0.6666667
Brazil	0.9230769	0.5000000
Bulgaria	0.5714286	0.5000000
Thailand	1.0000000	0.4230769
Turkey	0.9000000	0.7142857
United States of America	0.9655172	0.5937500
Uruguay	1.0000000	0.5833333

Reshaping data

- Can use `reshape` library
- `melt()` function

```
melt.data <- melt(cses_merge, id=c("Country","Year"))
```

	Country	Year	variable	value
1	Argentina	2015	Country_year	ARG_2015
5000	Switzerland	2007	Education	NA
15000	Mexico	2012	Freedom_house_status	NA

- `cast()` function

```
# cast(data, formula, function)
Country.means <- cast(melt.data, Country~variable, mean)
Year.means <- cast(melt.data, Year~variable, mean)
```

Gather and Spread

- Long vs. wide formats
- Analysis usually occurs with long
- Gather and spread can help convert between the two

Tips

- **Never save over your original data!**
- Save work and altered data frequently
- Document your code for both **you** and for **others**
- Create an organized folder structure on your computer for data analysis
    - Check out the Tier Protocol for examples of this

Tips

- Know what you are trying to do
- Be familiar with codebooks!
- Use google, stackexchange, helpsites, stata forums (for stata users)
- **Ask for help!**

Other Tools

- RStudio
  - R window and workflow manager (**Requires base R to be installed**)
- RMarkdown (**Requires RStudio**)
  - PDFs
  - HTML
  - Presentations (Beamer, Slidy, Powerpoint)
- Sweave
  - Latex within RStudio (**Requires Latex compiler to be installed on computer**)

Other Resources

- Wickham, Hadley, and Garrett Grolemund. R for data science: import, tidy, transform, visualize, and model data. " O'Reilly Media, Inc.", 2016.
- DPLYR cheatsheet
- Workflow of Data Analysis Using Stata