Cleaning\_Crime\_Complete

Jorge Olmos

24/7/2020

1. Set up: libraries, directory and dataset
2. Subset excerpts of specific crime events (crimeID == 1)

crime\_00 <- filter(exc\_clean, crimeID == 1 & speaker == 1 & crime == 1)  
dim(crime\_00) # 1,989 obs of 55 var

## [1] 1989 53

1. Create index of unique crime events (uniq\_crime) and index of excerpts within each unique crime (uniq\_crime\_exc)

crime\_01 <- crime\_00 %>%   
 group\_by(participant, crimeID\_num) %>%   
 mutate(uniq\_crime = cur\_group\_id(), .before = crimeID)  
length(unique(crime\_01$uniq\_crime)) # there are 722 unique crime episodes

## [1] 722

crime\_01 <- crime\_01 %>%   
 group\_by(uniq\_crime) %>%   
 mutate(uniq\_crime\_exc = row\_number(), .after = uniq\_crime)  
crime\_01 <- setorder(crime\_01, uniq\_crime, uniq\_crime\_exc) # sorting by uniq\_crime an uniq\_crime\_exc  
max(crime\_01$uniq\_crime) # 722 unique crime events

## [1] 722

dim(crime\_01)

## [1] 1989 55

1. Create dummy variables aggregated by unique crime event.

## Attitudes  
crime\_02 <- crime\_01 %>% group\_by(uniq\_crime) %>% mutate(justicia\_uc = rowsum(justicia, uniq\_crime), .after = justicia)  
crime\_02 <- crime\_02 %>% group\_by(uniq\_crime) %>% mutate(state\_uc = rowsum(state, uniq\_crime), .after = state)  
  
# Dichotomize  
crime\_02$justicia\_uc <- ifelse(crime\_02$justicia\_uc > 0, 1, 0)   
crime\_02$state\_uc <- ifelse(crime\_02$state\_uc > 0, 1, 0)  
  
## Measures  
crime\_02 <- crime\_02 %>% group\_by(uniq\_crime) %>% mutate(commu\_illegal\_uc = rowsum(commu\_illegal, uniq\_crime), .after = commu\_illegal)  
crime\_02 <- crime\_02 %>% group\_by(uniq\_crime) %>% mutate(commu\_legal\_uc = rowsum(commu\_legal, uniq\_crime), .after = commu\_legal)  
crime\_02 <- crime\_02 %>% group\_by(uniq\_crime) %>% mutate(state\_illegal\_uc = rowsum(state\_illegal, uniq\_crime), .after = state\_illegal)  
crime\_02 <- crime\_02 %>% group\_by(uniq\_crime) %>% mutate(state\_legal\_uc = rowsum(state\_legal, uniq\_crime), .after = state\_legal)  
  
# Dichotomize  
crime\_02$commu\_illegal\_uc <- ifelse(crime\_02$commu\_illegal\_uc > 0, 1, 0)  
crime\_02$commu\_legal\_uc <- ifelse(crime\_02$commu\_legal\_uc > 0, 1, 0)  
crime\_02$state\_illegal\_uc <- ifelse(crime\_02$state\_illegal\_uc > 0, 1, 0)  
crime\_02$state\_legal\_uc <- ifelse(crime\_02$state\_legal\_uc > 0, 1, 0)  
  
## Moral Reasonings  
crime\_02 <- crime\_02 %>% group\_by(uniq\_crime) %>% mutate(conseq\_uc = rowsum(conseq, uniq\_crime), .after = conseq)  
crime\_02 <- crime\_02 %>% group\_by(uniq\_crime) %>% mutate(dehuman\_uc = rowsum(dehuman, uniq\_crime), .after = dehuman)  
crime\_02 <- crime\_02 %>% group\_by(uniq\_crime) %>% mutate(deonto\_uc = rowsum(deonto, uniq\_crime), .after = deonto)  
crime\_02 <- crime\_02 %>% group\_by(uniq\_crime) %>% mutate(human\_uc = rowsum(human, uniq\_crime), .after = human)  
  
# Dichotomize  
crime\_02$conseq\_uc <- ifelse(crime\_02$conseq\_uc > 0, 1, 0)  
crime\_02$dehuman\_uc <- ifelse(crime\_02$dehuman\_uc > 0, 1, 0)  
crime\_02$deonto\_uc <- ifelse(crime\_02$deonto\_uc > 0, 1, 0)  
crime\_02$human\_uc <- ifelse(crime\_02$human\_uc > 0, 1, 0)  
  
## Types of Response  
crime\_02 <- crime\_02 %>% group\_by(uniq\_crime) %>% mutate(arrest\_uc = rowsum(arrest, uniq\_crime), .after = arrest)  
crime\_02 <- crime\_02 %>% group\_by(uniq\_crime) %>% mutate(death\_uc = rowsum(death, uniq\_crime), .after = death)  
crime\_02 <- crime\_02 %>% group\_by(uniq\_crime) %>% mutate(commu\_serv\_uc = rowsum(commu\_serv, uniq\_crime), .after = commu\_serv)  
crime\_02 <- crime\_02 %>% group\_by(uniq\_crime) %>% mutate(life\_uc = rowsum(life, uniq\_crime), .after = life)  
crime\_02 <- crime\_02 %>% group\_by(uniq\_crime) %>% mutate(missing\_uc = rowsum(missing, uniq\_crime), .after = missing)  
crime\_02 <- crime\_02 %>% group\_by(uniq\_crime) %>% mutate(no\_resp\_uc = rowsum(no\_resp, uniq\_crime), .after = no\_resp)  
crime\_02 <- crime\_02 %>% group\_by(uniq\_crime) %>% mutate(physical\_uc = rowsum(physical, uniq\_crime), .after = physical)  
crime\_02 <- crime\_02 %>% group\_by(uniq\_crime) %>% mutate(prevention\_uc = rowsum(prevention, uniq\_crime), .after = prevention)  
crime\_02 <- crime\_02 %>% group\_by(uniq\_crime) %>% mutate(reparations\_uc = rowsum(reparations, uniq\_crime), .after = reparations)  
crime\_02 <- crime\_02 %>% group\_by(uniq\_crime) %>% mutate(reintegration\_uc = rowsum(reintegration, uniq\_crime), .after = reintegration)  
crime\_02 <- crime\_02 %>% group\_by(uniq\_crime) %>% mutate(jail\_uc = rowsum(jail, uniq\_crime), .after = jail)  
  
# Dichotomize  
crime\_02$arrest\_uc <- ifelse(crime\_02$arrest\_uc > 0, 1, 0)  
crime\_02$death\_uc <- ifelse(crime\_02$death\_uc > 0, 1, 0)  
crime\_02$commu\_serv\_uc <- ifelse(crime\_02$commu\_serv\_uc > 0, 1, 0)  
crime\_02$life\_uc <- ifelse(crime\_02$life\_uc > 0, 1, 0)  
crime\_02$missing\_uc <- ifelse(crime\_02$missing\_uc > 0, 1, 0)  
crime\_02$no\_resp\_uc <- ifelse(crime\_02$no\_resp\_uc > 0, 1, 0)  
crime\_02$physical\_uc <- ifelse(crime\_02$physical\_uc > 0, 1, 0)  
crime\_02$prevention\_uc <- ifelse(crime\_02$prevention\_uc > 0, 1, 0)  
crime\_02$reparations\_uc <- ifelse(crime\_02$reparations\_uc > 0, 1, 0)  
crime\_02$reintegration\_uc <- ifelse(crime\_02$reintegration\_uc > 0, 1, 0)  
crime\_02$jail\_uc <- ifelse(crime\_02$jail\_uc > 0, 1, 0)  
  
dim(crime\_02)

## [1] 1989 76

1. Subset unique crime variables. Add dummy variables which identify the presence of any measures (measure\_uc), moral reasonings (moral\_uc), and types of responses (response\_uc) by unique crime. Add variables which count different measures (measure\_count\_uc) (0-4), moral reasonings (moral\_count\_uc) (0-4), and types of response (response\_count\_uc) (0-11) by unique crime.

crime\_03 <- dplyr::select(crime\_02,   
 uniq\_crime, participant, exc\_range, exc\_start, exc\_end, exc\_creator, exc\_date, exc\_raw,  
 codes, justicia:state, speaker:hypothetical,  
 crime:verbal, measure:state\_legal, moral:human, response:jail,  
 crime\_count, measure\_count, moral\_count, response\_count,  
 commu\_illegal\_uc:state\_legal\_uc, conseq\_uc:human\_uc,  
 arrest\_uc:jail\_uc)  
  
dim(crime\_03) # 1,989 obs of 46 var

## [1] 1989 72

## Dummy variables  
measures <- c("commu\_illegal", "commu\_legal", "state\_illegal", "state\_legal")  
crime\_03 <- crime\_03 %>% add\_column(measure\_uc = NA, .before = 'measure')  
crime\_03$measure\_uc <- ifelse(rowSums(crime\_03[measures]) > 0, 1, 0)  
  
morals <- c("conseq", 'deonto', 'dehuman', 'human')  
crime\_03 <- crime\_03 %>% add\_column(moral\_uc = NA, .before = 'moral')  
crime\_03$moral\_uc <- ifelse(rowSums(crime\_03[morals]) > 0, 1, 0)  
  
responses <- c('arrest', 'death', 'commu\_serv', 'life', 'no\_resp', 'physical', 'prevention', 'reparations', 'reintegration', 'jail')  
crime\_03 <- crime\_03 %>% add\_column(response\_uc = NA, .before = 'response')  
crime\_03$response\_uc <- ifelse(rowSums(crime\_03[responses]) > 0, 1, 0)  
  
## Count variables  
crime\_03 <- crime\_03 %>% mutate(measure\_count\_uc = commu\_illegal\_uc + commu\_legal\_uc + state\_illegal\_uc + state\_legal\_uc, .before = moral\_uc)  
  
crime\_03 <- crime\_03 %>% mutate(moral\_count\_uc = conseq\_uc + dehuman\_uc + deonto\_uc + human\_uc, .before = response\_uc)  
  
crime\_03 <- crime\_03 %>% mutate(response\_count\_uc = arrest\_uc + death\_uc + commu\_serv\_uc + life\_uc + missing\_uc + no\_resp\_uc + physical\_uc + prevention\_uc + reparations\_uc + reintegration\_uc + jail\_uc)  
  
dim(crime\_03) # 1,989 obs of 50 variables

## [1] 1989 78

1. For each unique crime, identify excerpts that are contained and overlapping.

## Remove perfect duplicates:  
crime\_04 <- crime\_03[!duplicated(crime\_03[,c("uniq\_crime", "exc\_start", "exc\_end")]),] # 1,988 obs (only 1 obs deleted)  
  
## Identify longest, first and last excerpts:  
crime\_04$exc\_length <- crime\_04$exc\_end - crime\_04$exc\_start  
crime\_04 <- crime\_04 %>%   
 group\_by(uniq\_crime, participant) %>%   
 mutate(exc\_start\_min = min(exc\_start), exc\_end\_max = max(exc\_end), exc\_length\_max = max(exc\_length))   
crime\_04$exc\_longest <- ifelse(crime\_04$exc\_length == crime\_04$exc\_length\_max, 1, 0)  
crime\_04$exc\_first <- ifelse(crime\_04$exc\_start==crime\_04$exc\_start\_min, 1, 0)  
crime\_04$exc\_last <- ifelse(crime\_04$exc\_end==crime\_04$exc\_end\_max, 1, 0)  
crime\_04 <- crime\_04 %>%  
 mutate(exc\_longest\_start = exc\_start[exc\_longest==1],  
 exc\_longest\_end = exc\_end[exc\_longest==1])  
  
## Identify excerpts completely contained in longest (plus 5 char margin):  
crime\_04$contained <- ifelse(crime\_04$exc\_start >= crime\_04$exc\_longest\_start - 5 &   
 crime\_04$exc\_end <= crime\_04$exc\_longest\_end + 5, 1, 0)  
table(crime\_04$contained)

##   
## 0 1   
## 280 1708

## Identify excerpts that are totally unique from the longest:  
crime\_04$unique <- ifelse(crime\_04$exc\_end < crime\_04$exc\_longest\_start | crime\_04$exc\_start > crime\_04$exc\_longest\_end, 1, 0)  
table(crime\_04$unique)

##   
## 0 1   
## 1712 276

## Identify excerpts that partially overlap with longest:  
crime\_04$overlap <- ifelse(crime\_04$unique==0 & crime\_04$contained==0, 1, 0)  
table(crime\_04$overlap) ## 4 partially overlapping excerpts that need ot be appended

##   
## 0 1   
## 1984 4

1. Append text from overlapping excerpts.

crime\_05 <- crime\_04[crime\_04$exc\_longest == 1,]  
unique <- crime\_04[crime\_04$unique==1,]  
ovlp <- crime\_04[crime\_04$overlap == 1,]  
  
## UNIQUE, NON-LONGEST EXCERPTS  
  
## Identify longest, first and last excerpts among secondary unique excerpts:  
unique$exc\_length <- unique$exc\_end - unique$exc\_start  
unique <- unique %>%   
 group\_by(uniq\_crime, participant) %>%   
 mutate(exc\_start\_min = min(exc\_start), exc\_end\_max = max(exc\_end), exc\_length\_max = max(exc\_length))   
unique$exc\_longest <- ifelse(unique$exc\_length == unique$exc\_length\_max, 1, 0)  
unique$exc\_first <- ifelse(unique$exc\_start==unique$exc\_start\_min, 1, 0)  
unique$exc\_last <- ifelse(unique$exc\_end==unique$exc\_end\_max, 1, 0)  
unique <- unique %>%  
 mutate(exc\_longest\_start = exc\_start[exc\_longest==1],  
 exc\_longest\_end = exc\_end[exc\_longest==1])  
  
## Identify excerpts completely contained in longest (plus 5 char margin):  
unique$contained <- ifelse(unique$exc\_start >= unique$exc\_longest\_start - 5 &   
 unique$exc\_end <= unique$exc\_longest\_end + 5,  
 1, 0)  
table(unique$contained)

##   
## 0 1   
## 69 207

## Drop contained text  
unique <- unique %>%   
 subset(contained==0)  
  
## Hand code a few  
unique$contained[unique$uniq\_crime==57 & unique$exc\_range=="42528-42693"] <- 1  
unique$contained[unique$uniq\_crime==155 & unique$exc\_range=="14911-15017"] <- 1  
unique$contained[unique$uniq\_crime==173 & unique$exc\_range=="45797-46423"] <- 1  
unique$contained[unique$uniq\_crime==249 & unique$exc\_range=="56512-56842"] <- 1  
unique$contained[unique$uniq\_crime==249 & unique$exc\_range=="56843-57749"] <- 1  
unique$contained[unique$uniq\_crime==267 & unique$exc\_range=="81420-81973"] <- 1  
unique$contained[unique$uniq\_crime==297 & unique$exc\_range=="32533-34459"] <- 1  
unique$contained[unique$uniq\_crime==519 & unique$exc\_range=="1301-2116"] <- 1  
  
## Drop contained text  
unique <- unique %>%   
 subset(contained==0)  
  
## Concatenate unique second excerpts  
unique <- unique %>%  
 group\_by(uniq\_crime, participant) %>%   
 summarize(exc\_append = concatenate(exc\_raw, collapse = " \*\*BREAK\*\* "))

## `summarise()` has grouped output by 'uniq\_crime'. You can override using the  
## `.groups` argument.

## Append unique excerpts  
crime\_05$appended <- ifelse(crime\_05$uniq\_crime %in% unique$uniq\_crime, 1, 0)  
crime\_05 <- merge(crime\_05, unique, by = c('participant', 'uniq\_crime'), all.x = T)  
crime\_05$exc\_raw\_append <- ifelse(crime\_05$appended==1,   
 paste(crime\_05$exc\_raw, crime\_05$exc\_append, sep = " \*\*BREAK\*\* "),  
 crime\_05$exc\_raw)  
crime\_05$exc\_append=NULL

1. Calculate word counts.

crime\_05 <- crime\_05 %>%   
 group\_by(uniq\_crime, participant) %>%  
 mutate(char\_count = nchar(exc\_raw\_append))  
  
crime\_05 <- crime\_05 %>%   
 group\_by(uniq\_crime, participant) %>%  
 mutate(word\_count = wordcount(exc\_raw\_append))

1. Calculate some crime variables

## Severity of violence  
# Physical violence crime  
crime\_05 <- crime\_05 %>% dplyr::mutate(phy\_cri = ifelse(assault == 1 | kidnap == 1 | dom\_vio == 1 | sexual == 1 | murder == 1, 1, 0))  
# Serious crime  
crime\_05 <- crime\_05 %>% dplyr::mutate(grave\_cri = ifelse(murder == 1 | kidnap == 1 | sexual == 1, 1, 0))  
# High impact crime  
crime\_05 <- crime\_05 %>% dplyr::mutate(alto\_cri = ifelse(rob == 1 | kidnap == 1 | extort == 1 | sexual == 1 | murder == 1 | drug == 1, 1, 0))  
  
## Punishment preferences  
# Death  
crime\_05 <- crime\_05 %>% dplyr::mutate(lethal\_pun = ifelse(death\_uc == 1, 1, 0))   
# Any physical punishment  
crime\_05 <- crime\_05 %>% dplyr::mutate(let\_phy\_pun = ifelse(death\_uc == 1 | physical\_uc == 1, 1, 0))   
# Categorical  
crime\_05 <- crime\_05 %>%  
 dplyr::mutate(mult\_pun = ifelse(death\_uc == 1, 4,  
 ifelse(death\_uc == 0 & physical\_uc == 1, 3,  
 ifelse(death\_uc ==0 & physical\_uc == 0 & life\_uc == 1, 2,  
 ifelse(death\_uc ==0 & physical\_uc == 0 & life\_uc == 0 & jail\_uc == 1, 1, 0)))))

1. Merge crime events dataset with individual dataset

# Read in clean ind data  
ind <- readit("../02\_Data/02\_Clean/individual\_clean.xlsx")

## File guessed to be xls/xlsx (Excel) ("../02\_Data/02\_Clean/individual\_clean.xlsx")

# Modifying 'participant' var so datasets can match  
crime\_06 <- crime\_05  
crime\_06$participant <- str\_sub(crime\_06$participant, -6, -4) # 722 obs of 44 var  
crime\_06$participant <- gsub("\_", "", crime\_06$participant)  
  
# Merge datasets  
crind <- merge(crime\_06, ind, by = 'participant', all.x = T) # Interviews 7T and 3M not matched with ind data

1. Fix a few codings manually

crind$deonto\_uc[crind$uniq\_crime==121] <- 1  
crind$deonto\_uc[crind$uniq\_crime==154] <- 0  
crind$deonto\_uc[crind$uniq\_crime==238] <- 0  
crind$deonto\_uc[crind$uniq\_crime==390] <- 0  
crind$deonto\_uc[crind$uniq\_crime==394] <- 0  
crind$deonto\_uc[crind$uniq\_crime==405] <- 0  
crind$deonto\_uc[crind$uniq\_crime==407] <- 0  
crind$deonto\_uc[crind$uniq\_crime==444] <- 0  
crind$deonto\_uc[crind$uniq\_crime==445] <- 0  
crind$deonto\_uc[crind$uniq\_crime==556] <- 1  
crind$deonto\_uc[crind$uniq\_crime==560] <- 0  
crind$deonto\_uc[crind$uniq\_crime==563] <- 0  
crind$deonto\_uc[crind$uniq\_crime==580] <- 0  
crind$deonto\_uc[crind$uniq\_crime==654] <- 0  
crind$deonto\_uc[crind$uniq\_crime==716] <- 0  
crind$deonto\_uc[crind$uniq\_crime==78] <- 0  
crind$deonto\_uc[crind$uniq\_crime==81] <- 0  
  
crind$human\_uc[crind$uniq\_crime==237] <- 1  
crind$human\_uc[crind$uniq\_crime==24] <- 0  
crind$human\_uc[crind$uniq\_crime==272] <- 0  
crind$human\_uc[crind$uniq\_crime==394] <- 1  
crind$human\_uc[crind$uniq\_crime==36] <- 0  
crind$human\_uc[crind$uniq\_crime==38] <- 0  
crind$human\_uc[crind$uniq\_crime==32] <- 0  
crind$human\_uc[crind$uniq\_crime==34] <- 0  
crind$human\_uc[crind$uniq\_crime==272] <- 0  
crind$human\_uc[crind$uniq\_crime==567] <- 0  
crind$human\_uc[crind$uniq\_crime==115] <- 0  
crind$human\_uc[crind$uniq\_crime==118] <- 0  
crind$human\_uc[crind$uniq\_crime==688] <- 0  
crind$human\_uc[crind$uniq\_crime==466] <- 0  
crind$human\_uc[crind$uniq\_crime==311] <- 0  
crind$human\_uc[crind$uniq\_crime==496] <- 0  
crind$human\_uc[crind$uniq\_crime==498] <- 0  
crind$human\_uc[crind$uniq\_crime==385] <- 0  
crind$human\_uc[crind$uniq\_crime==13] <- 0  
crind$human\_uc[crind$uniq\_crime==245] <- 0

1. Write cleaned and merged dataset that includes individual-level characteristics. Unit of analysis: unique crime events (rows)

write\_xlsx(crind, '../02\_Data/02\_Clean/crime\_set1\_clean.xlsx')

1. Write anonymized clean dataset (no text)

crind <- crind %>%   
 select(-exc\_raw, -exc\_raw\_append)  
write\_xlsx(crind, '../02\_Data/02\_Clean/crime\_set1\_clean+anon.xlsx')