summary stats of data cleaning

Devraj Kori

3/14/2020

```
#set the working directory to the main git directory
setwd("..")
file_names<-list.files(path=paste0(getwd(),"/data preparation"),pattern="*.txt")
#use lapply to read in each file (using the read_tsv function) and store them in a list
data_list<-lapply(file_names,function(x) read_tsv(paste0(getwd(),"/data preparation/",x)))
```

```
## Parsed with column specification:
## cols(
##   HA = col_character(),
##   CLIENT_ID = col_double(),
##   GENDER = col_character(),
##   RACE = col_character(),
##   RELATIONSHIP = col_character(),
##   MOVEINDATE = col_character(),
##   MOVEOUTDATE = col_character(),
##   PRIMARYSTREET = col_character(),
##   SECONDARYSTREET = col_character(),
##   CITY = col_character(),
##   STATE = col_double()
## )
## Parsed with column specification:
## cols(
##   HA = col_character(),
##   CLIENT_ID = col_double(),
##   GENDER = col_character(),
##   RACE = col_character(),
##   RELATIONSHIP = col_character(),
##   MOVEINDATE = col_character(),
##   MOVEOUTDATE = col_character(),
##   PRIMARYSTREET = col_character(),
##   SECONDARYSTREET = col_character(),
##   CITY = col_character(),
##   STATE = col_double()
## )
## Parsed with column specification:
## cols(
##   HA = col_character(),
##   CLIENT_ID = col_double(),
##   GENDER = col_character(),
##   RACE = col_character(),
##   RELATIONSHIP = col_character(),
##   MOVEINDATE = col_character(),
##   MOVEOUTDATE = col_character(),
```

Cleaning Summary Statistics

Unique Physical Addresses:	
Before Cleaning:	32,558
After Cleaning:	22,860
Rows associated with heads of household:	
Before Cleaning:	51,585
After Cleaning:	48,424
Missing Move-out Dates:	
Before Cleaning:	16,399
After Cleaning:	12,937

```
##   PRIMARYSTREET = col_character(),
##   SECONDARYSTREET = col_character(),
##   CITY = col_character(),
##   STATE = col_double()
## )
```

```
#use do.call to rbind the three files
dat<-do.call(rbind,data_list)

#load cleaned data
load("cleaned and geocoded data 01-April.Rdata")
```

```
starting_addresses<-length(unique(dat[dat$RELATIONSHIP=="Head",]$PRIMARYSTREET))%>%comma()
final_addresses<-length(unique(cleaned_and_geocoded$PRIMARYSTREET))%>%comma()

original_rows<-nrow(dat[dat$RELATIONSHIP=="Head",])%>%comma()
final_rows<-nrow(cleaned_and_geocoded)%>%comma()

missing_moveout_original<-nrow(dat[dat$RELATIONSHIP=="Head" & is.na(dat$MOVEOUTDATE),])%>%comma()
missing_moveout_final<-nrow(cleaned_and_geocoded[is.na(cleaned_and_geocoded$MOVEOUTDATE),])%>%comma()

#put these into a kable table
cleaning_summary<-data.frame(` `=c("Unique Physical Addresses:",
                               "  Before Cleaning:",
                               "  After Cleaning:",
                               "Rows associated with heads of household:",
                               "  Before Cleaning:",
                               "  After Cleaning:",
                               "Missing Move-out Dates:",
                               "  Before Cleaning:",
                               "  After Cleaning:"),
                             `  `=c(" ",starting_addresses,final_addresses,
                                    " ", original_rows, final_rows,
                                    " ", missing_moveout_original,missing_moveout_final),
                             check.names=FALSE)
kable(cleaning_summary,format="latex",booktabs=TRUE,caption="Cleaning Summary Statistics")%>%
  row_spec(c(1,4,7),bold=TRUE)
```

```
#summarize by race
by_race<-cleaned_and_geocoded%>%
```

Summary Statistics: Race

RACE	clients	rows
American Indian/Alaska Native	102	157
Asian	73	104
Black/African American	19,508	37,157
Multi-Racial	90	108
Native Hawaiian/Other Pacific Islander	40	52
White	7,223	10,846

Summary Statistics: Gender

GENDER	clients	rows
Female	21,856	41,204
Male	5,020	7,220

```r
  group_by(RACE)%>%
  summarise(clients = n_distinct(CLIENT_ID),
            rows = n())%>%
  ungroup()%>%
  mutate(clients=comma(clients),
         rows=comma(rows))%>%
  kable(format="latex",booktabs=TRUE,caption="Summary Statistics: Race")
by_race
```

```r
by_gender<-cleaned_and_geocoded%>%
  group_by(GENDER)%>%
  summarise(clients = n_distinct(CLIENT_ID),
            rows = n())%>%
  ungroup()%>%
  mutate(clients=comma(clients),
         rows=comma(rows))%>%
  kable(format="latex",booktabs=TRUE, caption="Summary Statistics: Gender")
by_gender
```