Cleaning: Percent Vacant Variable

for ECONorthwest

Salma Huque

04/01/20

Load Packages + Read in Data

```
# Load Packages
library(tidyverse)
library(sf)
library(sp)
library(stringr)
library(data.table)
library(magrittr)
library(readxl)
library(here)

#-----------------------
# GDB sent 10/10 via USB
gdb2 <- here::here("DATA", "data1.gdb")

taxlots <- st_read(gdb2, "taxlots_20191010")
footprints <- st_read(gdb2, "building_footprints_20191010")
```

Percent Vacant Properties within 200ft

This chunk takes the taxlot level data and uses the packages sp and sf (similar to ArcGIS) to create 200ft buffers around each of the taxlots. For each taxlots, it then calculates the percentage of properties within that buffer that are classfied as "vacant". The end product is a column where every observation is a taxlot and every value is a percentage. The variable was then joined to the full dataset.

```
# explicitly define the missing values in a factor level
taxlots$STATE_ID <- fct_explicit_na(taxlots$STATE_ID)

# selects variables
# df of unique obs -- 197,131 obs
vacancy_pruned <- taxlots %>%
  dplyr::select(STATE_ID, PRPCD_DESC, Shape) %>%
  group_by(STATE_ID) %>%
  mutate(n = n()) %>%
  filter(n == 1)

# create vacant column
```

```r
vacancy_pruned$VACANT <- vacancy_pruned$PRPCD_DESC == "VACANT LAND"

# creates a buffer called buffer_dist
# adds this buffer as an additional geometry to the main dataset
# 1 ft = 0.3048 meters
conv <- 0.3048
ft <- 200
buffer_dist <- ft * conv

# creates buffer of size `buffer_dist` around buffy
buffy <- taxlots_pruned %>%
  rename(ShapeBuffy = Shape) %>%
  st_buffer(buffer_dist)

# join buffy and vacancy_pruned
vacant_join_buffy <- st_join(buffy, vacancy_pruned, left = TRUE)

# calculates percent vacant houses in buffer
# taking out na's when calcing percent vacant
vacant_var <- vacant_join_buffy %>%
  arrange(STATE_ID.x) %>%
  group_by(STATE_ID.x) %>%
  mutate(n = n()) %>%
  mutate(VACANT = VACANT*1) %>%
  filter(!is.na(VACANT)) %>%
  summarize(percent_vacant = sum(VACANT)/n[1]) %>%
  rename(STATE_ID = STATE_ID.x) %>%
  st_drop_geometry()

# join to dataframe
full_data <- left_join(fugly, vacant_var, by = "STATE_ID")
print(paste0("vacant Join: ", as.character(sum(data.frame(table(fugly$STATE_ID.x))$Freq > 1))))
```