House Number EDA

Bo Jumrustanasan

11/4/19

House Number EDA on the Sample 1910 Census Data

1. Read and clean

This script explores the 1910 census sample data in order to gain insights about filling in missing house numbers. Furthermore, address information is only available in household records except for `microfilm.page.number.3` that may be useful for house number filling down. Furthermore, most of the columns are irrelevant. Thus, these rows and columns will not be considered during the EDA.

```
HN1 <- read.csv("/Users/panchanok/Desktop/HNYC/data/us1910m_usa_sample100k.csv") %>%
    as.data.frame() %>%
    fill(Microfilm.page.number.3, .direction = "up") %>%
    filter(Record.type=="H") %>%
    select(Microfilm.page.number.3, Enumeration.district.2, Household.serial.number.2, House.number, St
  mutate(House.number = ifelse(House.number=="", NA, as.character(House.number)))
```

The raw data has some issues that have to be cleaned. The table below displays those potential problems:

```
## subset of the sample
HN1 %>% filter(Household.serial.number.2 %in% c(15461, 15825, 20325, 15076, 15249, 15134))
```

```
##   Microfilm.page.number.3 Enumeration.district.2 Household.serial.number.2
## 1                     206                     21                      15076
## 2                     211                     21                      15134
## 3                     223                     21                      15249
## 4                     231                     21                      15461
## 5                     260                     22                      15825
## 6                     660                     45                      20325
##   House.number    Street.address.2
## 1        2 1/2        ROOSEVELT ST
## 2        14/16    ROOSEVELT STREET
## 3      9 FRONT        JAMES STREET
## 4         <NA> 191 TO 195 PARK ROW
## 5     31 TO 33       OLIVER STREET
## 6   TO 24 REAR     MULBERRY STREET
```

1. House number ranges are used instead of house numbers
2. House number ranges are put in `Street.address.2` column
3. Modifiers (e.g. "REAR") are included in `House.number` column. The modifiers may be helpful to geocoder later so we will create a new column specifically for them.
4. Some house numbers have "1/2"s. This is not informative for our project purpose and should be removed.

```
HN2 <- HN1 %>% rowwise() %>%
  mutate(House.number = gsub("\\s+[0-9]+/[0-9]+", "", House.number),           ## remove 1/2
         House.number = gsub("\\s*\\.\\s*[0-9]+", "", House.number),   ##remove decimal points (not nec
         House.number = gsub("\\s*(TO|/)\\s*", "-", House.number, ignore.case = TRUE),  ## replace "TO"
         modifier = trimws(str_extract(House.number, "[A-Za-z\\s]+")),               ## create modifie
         House.number = gsub("\\s+", " ", House.number),                        ## remove excess whit
```

```
        House.number = trimws(gsub("[A-Za-z\\s]+", "", House.number)))           ## remove words from
```

1.1 Join with cleaned street names

```
df_cleaned_mn <- read.csv("/Users/panchanok/Desktop/HNYC/data/100ksample_MN_matched.csv") %>%
    as.data.frame() %>%
    filter(Record.type=="H") %>%
    select(Household.serial.number.2, corrected_str, dscore)

HN3 <- HN2 %>% left_join(df_cleaned_mn, by = "Household.serial.number.2")
```

2. Attempt to fill in missing house numbers

Most of `House.numbers` in the sample census data are missing:

```
plot_missing <- function(HN_df){
  ggplot(data = HN_df %>% mutate(na.cnt = ifelse(is.na(House.number), "Missing", "Present")), aes(x=na.
  geom_bar() +
  xlab("counts of present and missing house numbers")
}

count_missing <- function(HN_df){
  HN_df %>% mutate(na.cnt = ifelse(is.na(House.number), "Missing", "Present")) %>% group_by(na.cnt) %>%
}

count_missing(HN3)
```

```
## # A tibble: 2 x 2
##   na.cnt  count_missing
##   <chr>           <int>
## 1 Missing         14569
## 2 Present          4868
```

2.1 Extract potential house numbers from `Street.address.2`

Some house numbers make into `Street.address.2` column (see table above). Kyi's has provided the code that eaxtracts those numbers back into `House.number`.

- numbers in a form of "100 - 120" or "100 TO 120" in `Street.address.2` is considered a house number if a current `House.number` is NA.

```
HN4 <- HN3%>% rowwise() %>%
  mutate(House.number=ifelse(is.na(House.number)&
                               !is.na(str_extract(Street.address.2,"[0-9]+\\s*(-|TO)+\\s*[0-9]+")),
                             str_extract(Street.address.2,"[0-9]+\\s*(-|TO)+\\s*[0-9]+"),
                             House.number),
         House.number = gsub("\\s*(TO)\\s*", "-", House.number, ignore.case = TRUE))  ## convert "TO" t


#gsub("\\s*.\\s*[0-9]+", "Z","34.5 he")
```

Not many missing house numbers have been recovered.

```
count_missing(HN4)
```

```
## # A tibble: 2 x 2
```

```
##   na.cnt  count_missing
##   <chr>           <int>
## 1 Missing         14560
## 2 Present          4877
```

2.2 Infer house numbers from neighbors

House number interpolation process is taken from Logan and Zhong (2018).

```
street_list <- HN4 %>% pull(corrected_str) %>% unique()
HN4 %>% filter(corrected_str==street_list[3]) %>%
  arrange(Household.serial.number.2) %>%
  pull(House.number)
```

2.2.1 Flag outliers

Logan and Zhang create a distrbution of house number difference within the same ED ans street. A house number whose difference is extreme relative to the distribution is flagged as an outlier (I am not precisely sure how they compute an outlier using `compute_cray_z_scores()`). However, this approach may not be necessary because a new ED dict will have valid house number ranges for each street and ED. Any house number that falls outside the range should be flagged.

2.2.2 Determine house numbers

[taking a filldown approach right now. Once house number sequences can be figured out, we can try to evenly assign HN to each record in a big missing value gap]

Logan and Zhong use a house number from a record that comes immediately before in a enumertaion page as a house number for the current record if it is missing under a condition that two records are likely to come from the same household(?).

Since we still working on only household records, we will compare if two records are on the same street. If so, we filldown house numbers within a enumeration page.

```
HN5 <- HN4 %>%
  group_by(Enumeration.district.2, corrected_str) %>%
  fill(House.number, .direction = "down") %>%
  ungroup() %>%
  arrange(Household.serial.number.2)

count_missing(HN5)
```

```
## # A tibble: 2 x 2
##   na.cnt  count_missing
##   <chr>           <int>
## 1 Missing           804
## 2 Present         18633
```

This kind of interpolation should recover almost all missing house numbers. However, this process can be so crude that it is inaccurate. More precise process will be explored [determining house number from house number sequence].

2.2.3 Infer a sequence from available house numbers [In Progress]

There are some characteristics we know about house number sequence that cna be used for filling in blank streets.

- An enumerator is likely to collect data from one side of a street at a time. Thus, we expect to see sequences of odd numbers and sequences of even numbers.
- There is a range for house numbers that exist on a street. Any house number outside the range can be considered as mistranscribed.

2.2.4 Evenly space out house numbers in a missing house number gap [NEXT STEP]

Note from Dan

- There is a mistake in transcription that may post a challenge in house number inferring. "3-5-7" is transcribed as "357". However, this should be less painful to handle once we have a new ED dict that comes with house number ranges.