Data Wrangling

Lindsay Poirier

Extended Example: NYC Stop, Question, and Frisk 2011

Getting Started

Load Packages

```
library(tidyverse)
```

Load Data

```
sqf_url <- "https://www1.nyc.gov/assets/nypd/downloads/zip/analysis_and_planning/stop-qu
temp <- tempfile()
download.file(sqf_url, temp)
sqf_zip <- unzip(temp, "2011.csv")
sqf_2011 <- read.csv(sqf_zip, stringsAsFactors = FALSE)
sqf_2011_race_cat <- read.csv("https://raw.githubusercontent.com/lindsaypoirier/STS-101/


rm(sqf_url)
rm(temp)
rm(sqf_zip)
```

Add Race Categories to Dataset

The original dataset codes each race with a single letter. This adds a column for race_cat that writes out each racial category in accordance with the data documentation.

```
sqf_2011 <-
  sqf_2011 %>%
  left_join(sqf_2011_race_cat, by = "race")
```

```
rm(sqf_2011_race_cat)
```

Convert "Y" and "N" to 1 and 0

The replaces every instance of "Y" in the dataset with 1 and every instance of "N" with 0. This will allow us to sum the Yes's in the dataset.

```
for(i in c(13:15, 17, 21:73, 76:78)){
  sqf_2011[,i]<-ifelse(sqf_2011[,i] == "Y", 1, 0)
}
```

Add new columns indicating 1) whether a weapon was found or 2) an arrest/summons was made

The original dataset had separate variables for indicating whether a pistol, rifle, assault weapon, knife, machine gun, or other weapon was found on a suspect. We create a variable equal to 1 if any of these weapons were found on the suspect.

The original dataset had separate variables for indicating whether a stop resulted in an arrest made or summons issued. We create a variable equal to 1 if either occurred.

```
sqf_2011 <-
  sqf_2011 %>%
  #Add a variable for weapon found
  mutate(wpnfound = ifelse((pistol == 1 |
                              riflshot == 1 |
                              asltweap == 1 |
                              knifcuti== 1 |
                              machgun == 1 |
                              othrweap == 1), 1, 0))


sqf_2011 <-
  sqf_2011 %>%
  #Add a variable for arrest made or summons issued
  mutate(arrestsumm = ifelse((sumissue == 1 |
                                arstmade == 1), 1, 0))
```

Create Data Frame with Relevant Variables

The original dataset has 112 variables - one for every entry on the UF-250 form. We are only interested in 6 variables. How do we create a data frame with those variables?

```
sqf_2011_sub <-
  sqf_2011 %>%
  ---_____(pct, arrestsumm, _____, wpnfound, race_cat, _____)
```

Analysis

How many stops were there in 2011?

Remember from previous lectures that nrow() calculates the number of rows in the dataset. Since every row in this dataset is a stop, calculating the number of rows will tell us how many stops were reported in 2011.

```
total_stops <-
  nrow(sqf_2011_sub)


total_stops
```

How many stops did not result in an arrest or summons in 2011? What percentage of stops did not result in an arrest or summons?

```
sqf_2011_sub %>%
  #Subset to rows where suspect innocent
  ---_____(arrestsumm ---_____ 0) %>%
  #Calculate number of observations
  ---_____(total_innocent = n(),
            percent_innocent = ---_____ / total_stops * 100)
```

In how many stops were the individuals aged 14-24? In what percentage of stops were the individuals aged 14-24?

```
sqf_2011_sub %>%
  #Subset to rows where suspect age 14-24
  ---_____(age ---_____ 14 & age ---_____ 24) %>%
```

```
  #Calculate number of observations and percentage of observations
  ---------(total_14_24 = ---------,
            percent_14_24 = n() / total_stops * 100)
```

Why doesn't this match the values we see on the NYCLU website?

Note the following from the NYCLU's 2011 report on Stop, Question, and Frisk data:

> "In a negligible number of cases, race and age information is not recorded in the database. Throughout this report, percentages of race and age are percentages of those cases where race and age are recorded, not of all stops."

```
total_stops_age_recorded <-
  sqf_2011_sub %>%
  #Subset to rows where age is not 999
  ---------(age --------- 999) %>%
  nrow()


sqf_2011_sub %>%
  filter(age >= 14 & age <= 24) %>%
  summarize(total_14_24 = n(),
            percent_14_24 = n() / total_stops_age_recorded * 100)
```

This still doesn't match the values we see on the website, but it does match the values we see in the NYCLU's 2011 report on Stop, Question, and Frisk data. This is typically when I would reach out to a representative at the NYCLU to inquire about the discrepancy.

How many stops were there per race in 2011? What percentage of stops per race in 2011? Arrange by number of stops in descending order.

```
total_stops_race_recorded <-
  sqf_2011_sub %>%
  #Subset to rows where race_cat is not NA or "OTHER"
  filter(---------(race_cat) & race_cat --------- "OTHER") %>%
  nrow()


sqf_2011_sub %>%
```

```
  #Subset to rows where race_cat is not NA or "OTHER"
  ---------(---------(race_cat) & race_cat ---------- "OTHER") %>%
  #Group by race
  ---------(race_cat) %>%
  #Calculate number of observations
  ---------(stops = n(),
            percent_stops = n() / total_stops_race_recorded * 100) %>%
  #Sort by stops in descending order
  ---------(----------(stops))
```

Note how this dataset categorizes race. Many different government datasets categorize race in many different ways. This, for instance, is not how the US Census categorizes race. How we categorize race matters for how we can talk about discrimination and racial profiling. Imagine if WHITE was not categorized separately from WHITE-HISPANIC. The values in this dataset would appear very differently! The NYCLU chose to aggregate two racial categories in this dataset into the one category - Latino - in order to advance certain claims regarding discrimination. What we should remember is that these racial categories are not reported by those stopped; they are recorded by officers stopping individuals. They may not reflect how individuals identify themselves.

```
sqf_2011_sub %>%
  #Subset to rows where race_cat is "WHITE-HISPANIC" or "BLACK-HISPANIC"
  ---------(race_cat ---------- c("WHITE-HISPANIC", "BLACK-HISPANIC")) %>%
  #Calculate number of observations
  ---------(stops_Latinx = ----------,
            percent_Latinx = n() / total_stops_race_recorded * 100)
```

What percentage of stops in 2011 resulted in a frisk per race?

```
sqf_2011_sub %>%
  filter(!is.na(race_cat) & race_cat != "OTHER") %>%
  group_by(race_cat) %>%
  summarize(stops = n(),
            percent_stops = n() / total_stops_race_recorded * 100,
            #Calculate total frisked
            percent_frisked = ----------(frisked) / n() * 100) %>%
```

```
  arrange(desc(stops))
```

What percentage of stops in 2011 resulted in a weapon found per race? What percentage of stops in 2011 resulted in an arrest or summons per race?

```
sqf_2011_sub %>%
  filter(!is.na(race_cat) & race_cat != "OTHER") %>%
  group_by(race_cat) %>%
  summarize(stops = n(),
            percent_stops = n() / total_stops_race_recorded * 100,
            percent_frisked = sum(frisked) / n() * 100,
            percent_wpnfound = sum(wpnfound) / n() * 100,
            percent_arrestsumm = sum(arrestsumm) / n() * 100) %>%
  arrange(desc(stops))
```