Autumn 2022 UW Statistics and Probability Association Directed Reading Program

Mentee: Joy Li, Mentor: Ellen Graham

December 16, 2022

Practice and Philosophy of Data Cleaning

Introduction:

During the course of this quarter, I worked with my wonderful mentor, Ellen Graham, to learn about data cleaning, first from a theoretical standpoint, then from an applied perspective with a dataset. Through readings, I learned about the complexities data is fraught with at all stages, which one must consider when they clean a dataset; I was exposed to philosophies on classification systems from varied perspectives; and I began to critically consider the motivations that a data scientist must keep in mind when they set out to clean data.

Theoretical Approach:

In the grand scheme of data analysis, the beginning step of cleaning or tidying data may not seem especially significant. However, in reality, “80% of data analysis is spent on the process of cleaning and preparing the data” (Wickham). Data will rarely be workable in its raw form, and a plethora of operations may have to be applied to it for it to be usable. Data cleaning is that process transforming a raw dataset into one that can readily be used for analysis.

Yet behind the surface, data cleaning is not as simple as righting formatting issues in a .csv file—data is more than just numbers and characters because it is knowledge that is situated in a context that may be deeply influential on what that data can represent. For example, “the reality of women’s lives is simply not captured in quantitative statistics,” the founder of WomanStats, a project that explores whether women’s status had a relationship to state security, explains (Hudson, qtd. in Kanarinka and Klein). Certain statistics, like those representing domestic abuse, cannot be viewed as an accurate proxy for underlying rates, because they rely so heavily on reported cases, an unreliable and small sample. Data is not neutral; it cannot be divorced from its roots.

We should strive to find or collect data that matches, as well as possible, the question we intend to investigate, because otherwise “data matched to the question may be surrogates for covariates that measure the underlying data phenomena” (McGowan et al.). In that case, the relationship between the surrogate data and underlying phenomena will have to be further explored.

Application:

For the second half of my project, I switched to a practical, hands-on application of data cleaning. I chose to work with the Public Life Data from the Seattle Department of Transportation. The study measures how people utilize public spaces; in it, human observers recorded details about all the people that move through or stay in specific locations in the city during designated collection periods.

The Public Life study provided observational and people-centered data. However, it is important to note that its data is not objective. While collecting data, observers had to make split second judgements about traits regarding the people that passed through their location, such as their perceived gender, race, and age range. Along with this, plenty of human error and other limitations exist. “It is important to note that this data does most likely not mirror all public life activity at any given time” (SDOT).

Thus, my goal with working with this dataset was not to draw meaningful statistical conclusions, but rather to gain more experience with R, especially by working with geographic data, and make high level visualizations.

The study comes in 5 interconnected data frames:

```
geography <- read_csv("data/Public_Life_Data_-_Geography.csv")
locations <- read_csv("data/Public_Life_Data_-_Locations.csv")
study <- read_csv("data/Public_Life_Data_-_Study.csv")
people_moving <- read_csv("data/Public_Life_Data_-_People_Moving.csv")
people_staying <- read_csv("data/Public_Life_Data_-_People_Staying.csv")
```

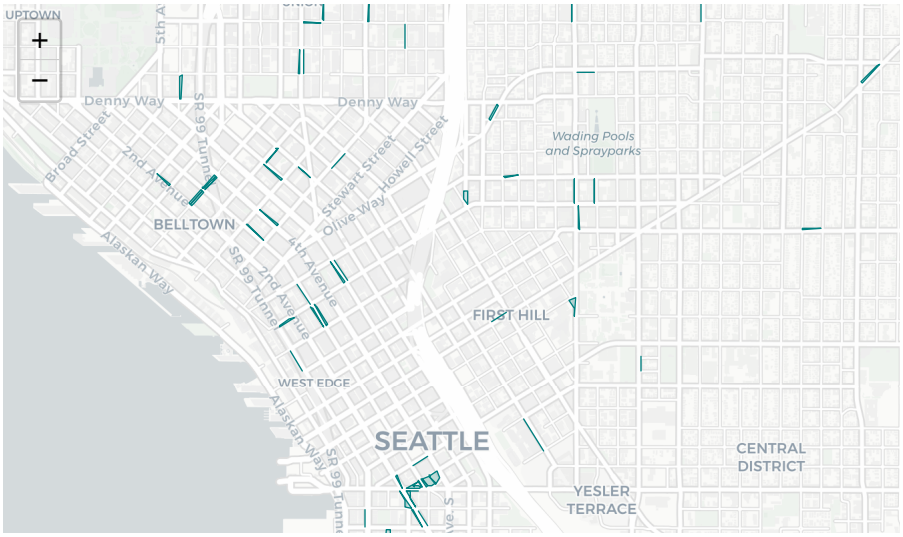

Figure 1: All locations viewed using leaflet.

I want to compare the people_moving and people_staying datasets, but their units of observation are not compatible. Each row for people_staying is a single individual who is staying in a specific location for a given observational period. For people moving, however, each row represents all the individuals that move through a location during an observational period. These formats are not compatible. Thus the goal is to convert these datasets into a format that can be universally compared: rate of people moving or staying per hour for each location. The rates for demographics will also be isolated in order to compare with census data.
To create a dataset that shows distinct age groups, I need to make new categorizations for ages. There are redundant names for ages and incorrect labels (5 converted to may), so this will be sorted into 4 buckets.

```
## # A tibble: 11 × 2
##    staying_age                        n
##    <chr>                          <int>
##  1 0-4                               95
##  2 14-May                           176
##  3 15-24                           1006
##  4 25-44                           5669
##  5 25-64                            416
##  6 45-64                           2864
##  7 65+                              607
##  8 Baby - Lying / Crawling            1
##  9 Toddler - Walking unconfidently    2
## 10 Up to 5 - Walking confidently      5
## 11 <NA>                             361
```

Figure 2: categories of age in the raw data

The product of this cleaning is a data frame with total staying and moving per hour for each location, as well as separated by age groups.

Now that the data is in a form where it can be compared, relations in the data can be explored. Here, the rates of staying and moving for locations seem to have some correlation.

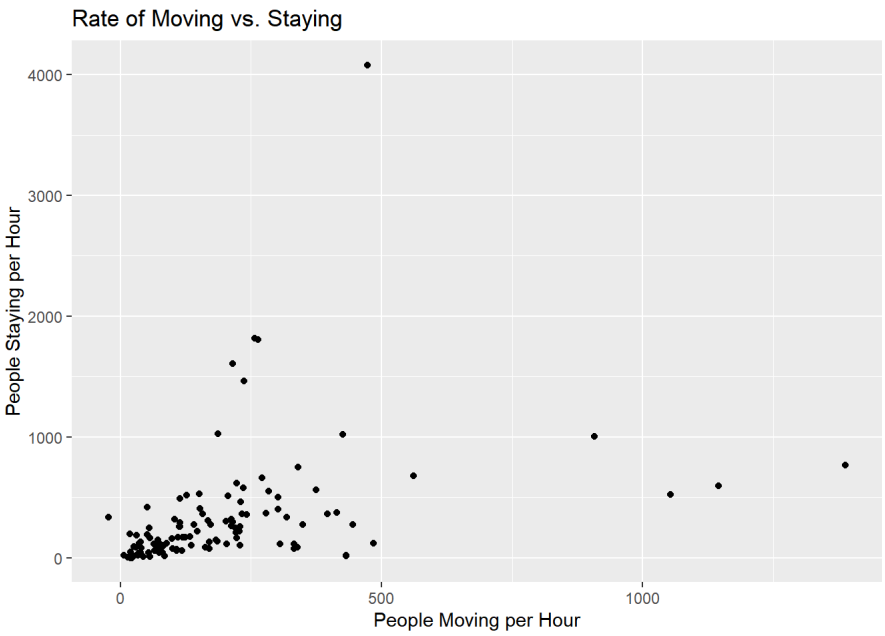

Figure 3: relationship between rates of staying and moving

How does this data compare to demographic data from the census? Is the actual population density in a location reflected in the observed public life there?

The ACS data is cleaned so that for each census tract, the density (people per kilometer squared) is represented

```
acs_data_joined <-
  acs_data %>%
  st_as_sf(crs = "4326") %>%
  pivot_wider(names_from = "variable", values_from = c("estimate", "moe")) %>%
  mutate(across(.cols = contains("estimate"), .fns = list(density = ~(.x*1000000/ALAND))))

public_life_and_acs <- by_ages %>%
  st_as_sf(crs = "4326") %>%
  st_join(acs_data_joined, join = st_intersects)
```

Figure 4: Visualization of the density with option to toggle between layers

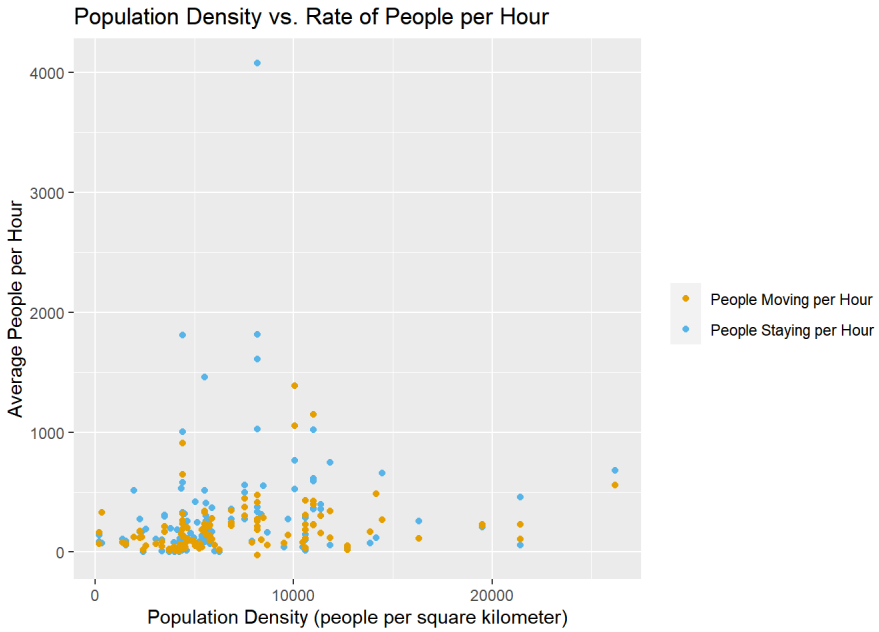

Figure 5: Visualization of the relationship between the total population density and the amount of people staying in a location

Conclusion:

Through this project, I got a thorough introduction to both the philosophical and practical approaches to data cleaning. People that work with data often say, “garbage in, garbage out,” and what I learned exemplifies that. If the proper considerations are not taken while cleaning data, the rest of the data analysis that follows it will essentially be useless.

Works Cited:

Bowker, Geoffrey C, and Susan Leigh Star. Sorting Things out: Classification and Its Consequences. Cambridge, Massachusetts, Mit Press, 2000.
Kanarinka, and Lauren F Klein. Data Feminism. Cambridge, Massachusetts, The Mit Press, 2020.
McGowan, Lucy D’Agostino, et al. “Design Principles for Data Analysis.” Journal of Computational and Graphical Statistics, 19 Sept. 2022, pp. 1–8,

10.1080/10618600.2022.2104290.
Rawson, Katie, and Trevor Muñoz. “Against Cleaning.” Curatingmenus.org, 6 July 2016, curatingmenus.org/articles/against-cleaning/.
Seattle Department of Transportation. “Public Life Data - Study.” City of Seattle Open Data Portal, 12 Feb. 2021, data.seattle.gov/Transportation/Public-Life-Data-Study/7qru-sdcp.
Wickham, Hadley. “Tidy Data.” Journal of Statistical Software, vol. 59, no. 10, 2014, 10.18637/jss.v059.i10.
Wickham, Hadley, and Garrett Grolemund. R for Data Science. O’Reilly Media, 12 Dec. 2016.