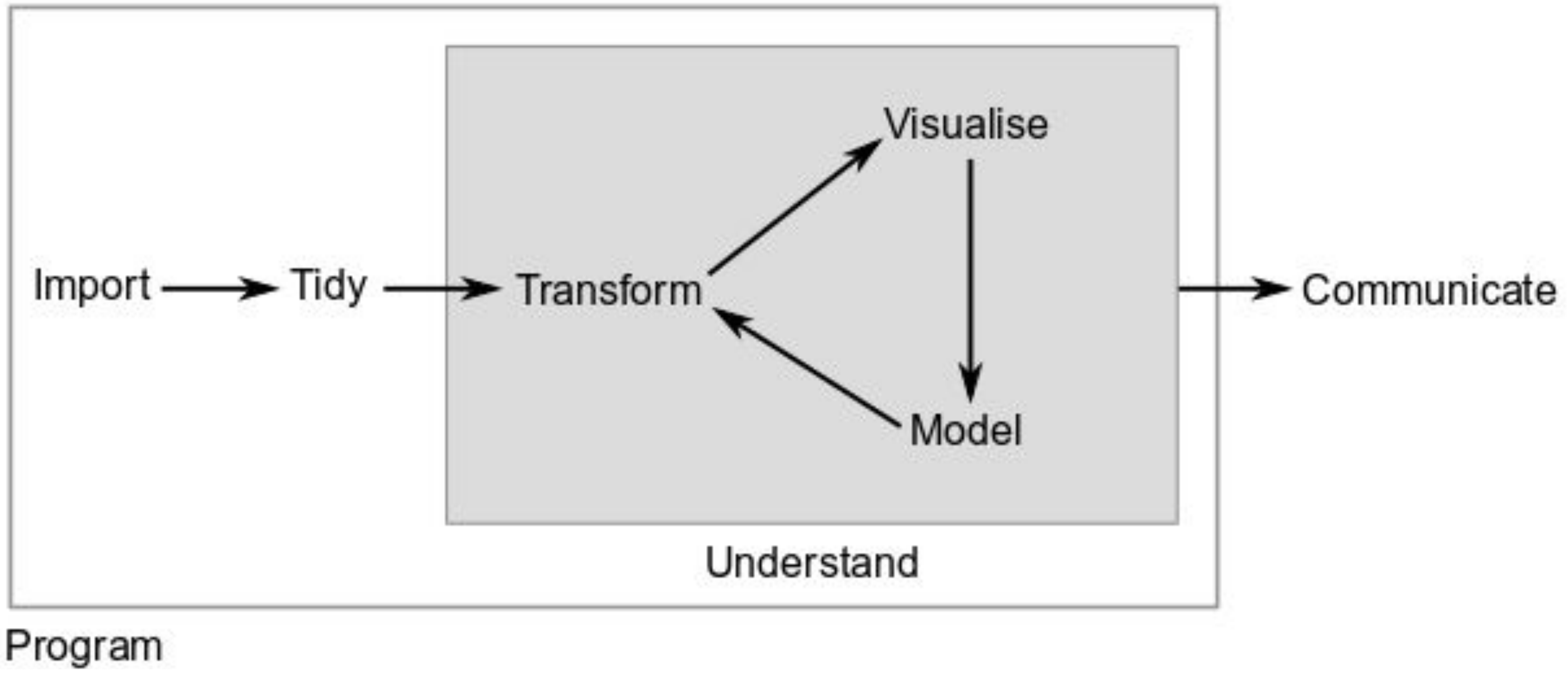
Practice and Philosophy of Data Cleaning

Mentee: Joy Li
Mentor: Ellen Graham

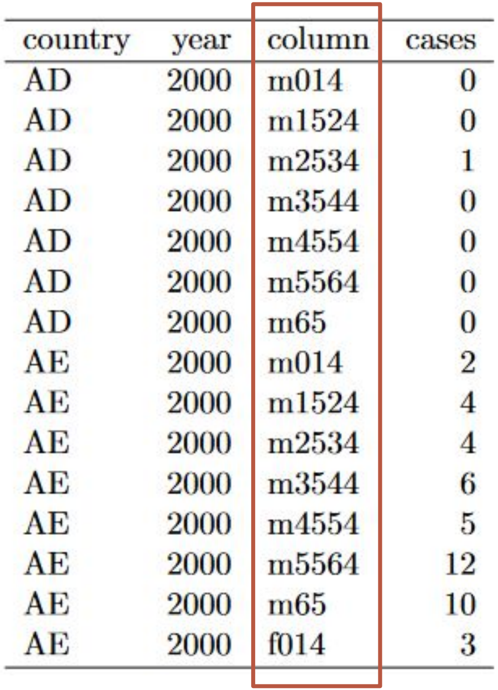

Import
Tidy
Transform
Visualise
Model
Understand
Communicate
Program

“80% of data analysis is spent on the process of cleaning and preparing the data”

- Hadley Wickham, *Tidy Data*

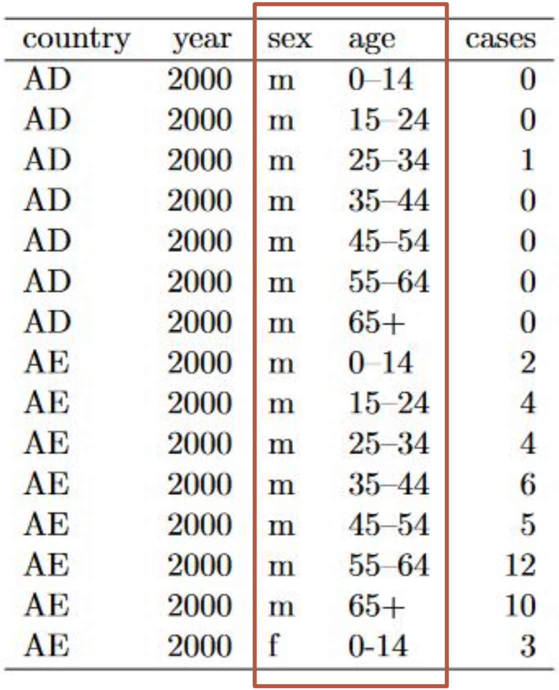
What is data cleaning?

Unusable data



country	year	column	cases
AD	2000	m014	0
AD	2000	m1524	0
AD	2000	m2534	1
AD	2000	m3544	0
AD	2000	m4554	0
AD	2000	m5564	0
AD	2000	m65	0
AE	2000	m014	2
AE	2000	m1524	4
AE	2000	m2534	4
AE	2000	m3544	6
AE	2000	m4554	5
AE	2000	m5564	12
AE	2000	m65	10
AE	2000	f014	3

Clean data



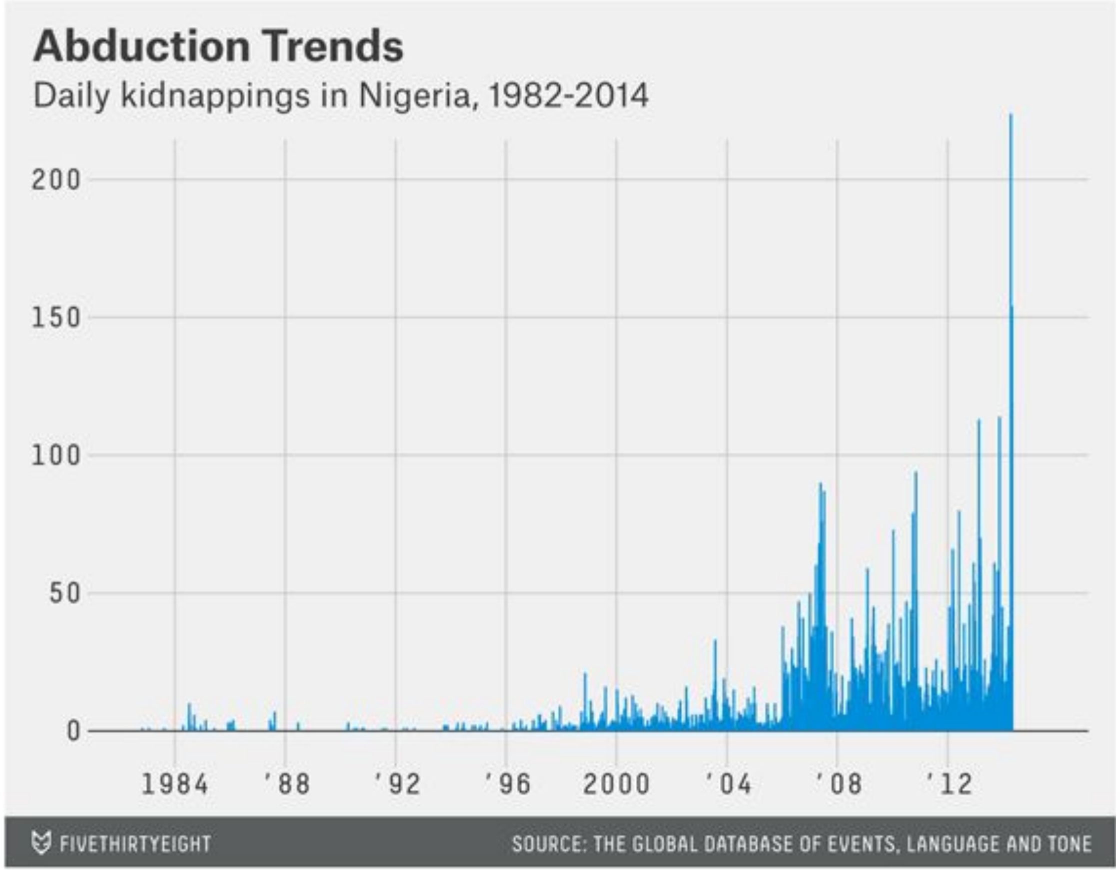
country	year	sex	age	cases
AD	2000	m	0–14	0
AD	2000	m	15–24	0
AD	2000	m	25–34	1
AD	2000	m	35–44	0
AD	2000	m	45–54	0
AD	2000	m	55–64	0
AD	2000	m	65+	0
AE	2000	m	0–14	2
AE	2000	m	15–24	4
AE	2000	m	25–34	4
AE	2000	m	35–44	6
AE	2000	m	45–54	5
AE	2000	m	55–64	12
AE	2000	m	65+	10
AE	2000	f	0-14	3

Definition

Data cleaning is any steps taken to transform data from the form it arrives in to one where it can readily be used for analysis

“the reality of women’s lives is simply not captured in quantitative statistics”
–Valerie Hudson

Case study:



Abduction Trends

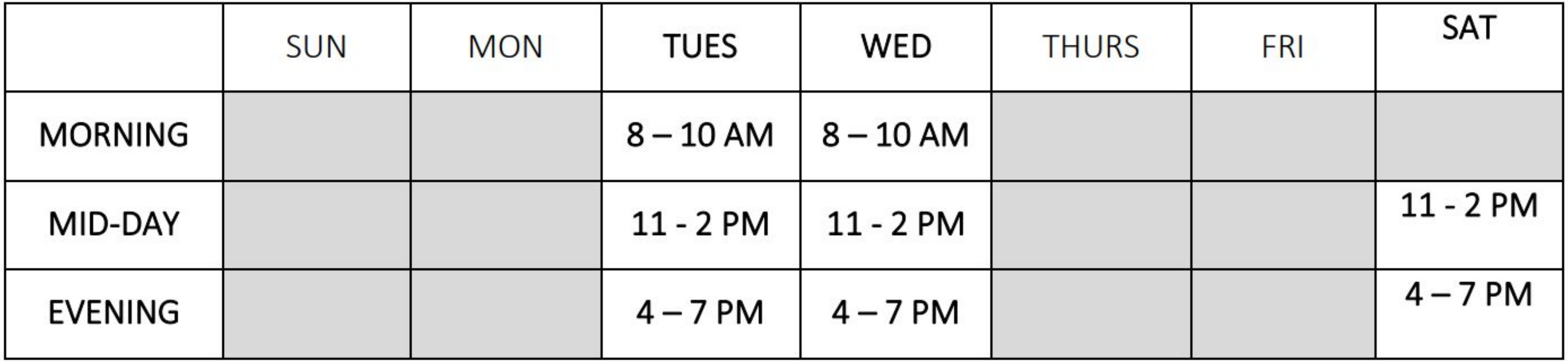
Daily kidnappings in Nigeria, 1982-2014

SOURCE: THE GLOBAL DATABASE OF EVENTS, LANGUAGE AND TONE

Public Life Data

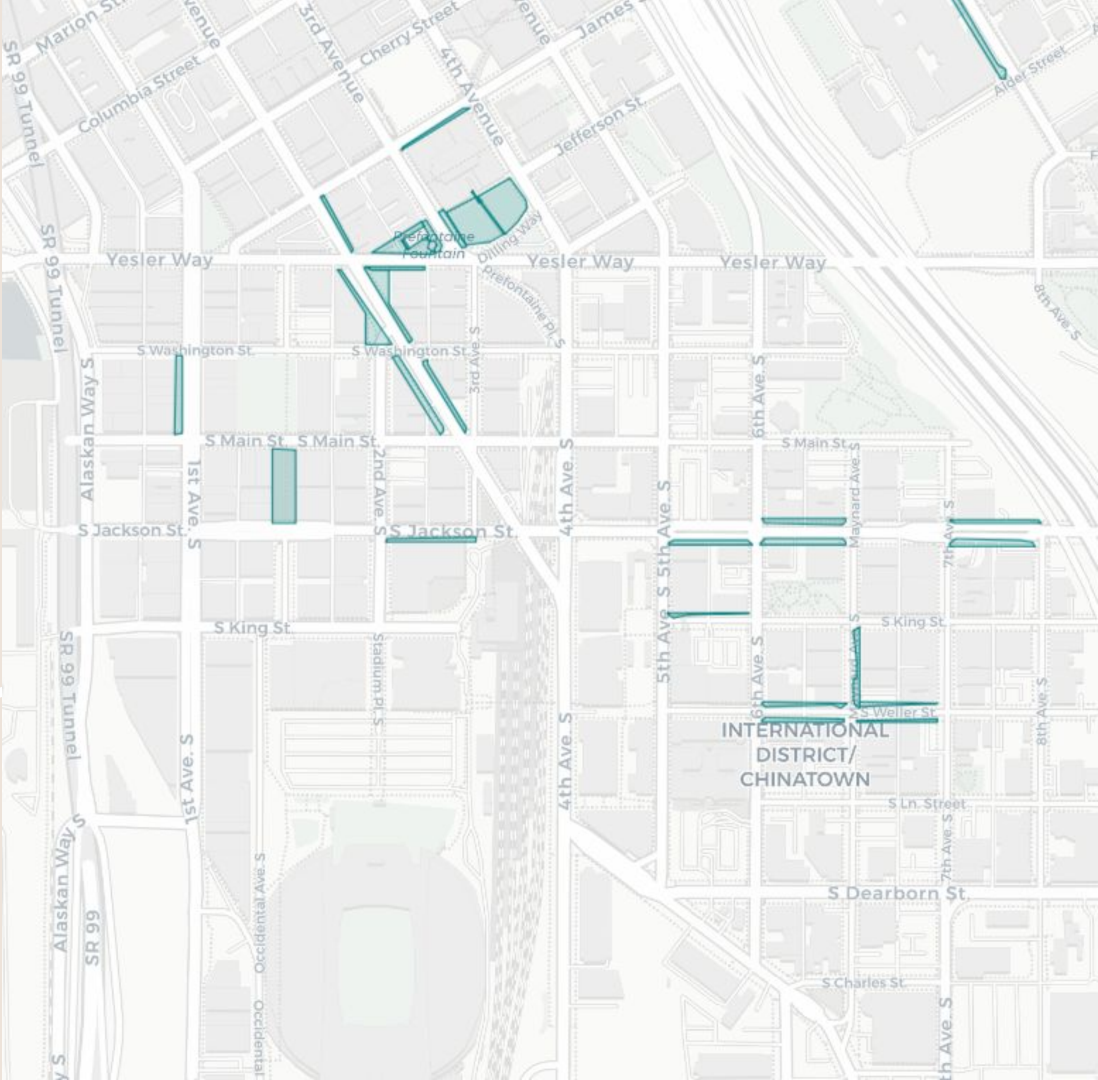
Seattle
Department of
Transportation

Typical collection schedule:



	SUN	MON	TUES	WED	THURS	FRI	SAT
MORNING			8 – 10 AM	8 – 10 AM			
MID-DAY			11 - 2 PM	11 - 2 PM			11 - 2 PM
EVENING			4 – 7 PM	4 – 7 PM			4 – 7 PM

1. Study
2. Location
3. Geography
4. People Moving
5. People Staying



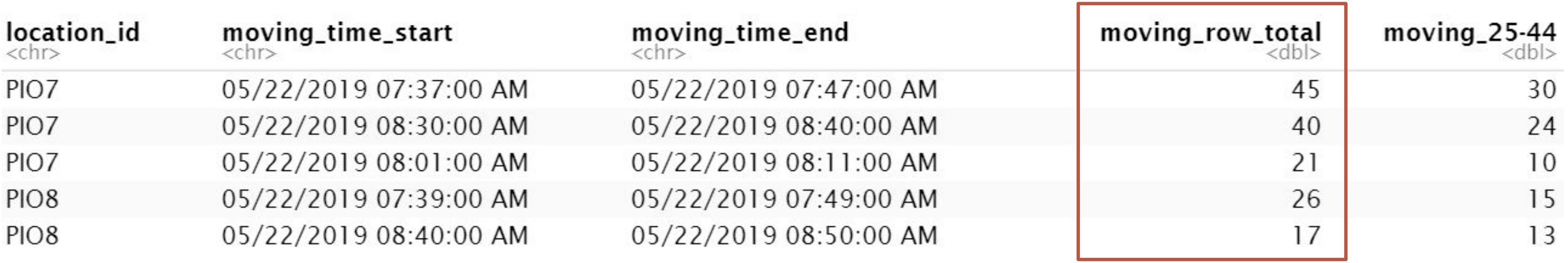
Background and limitations

"It is important to note that this data does most likely not mirror all public life activity at any given time" (SDOT).

Goal:

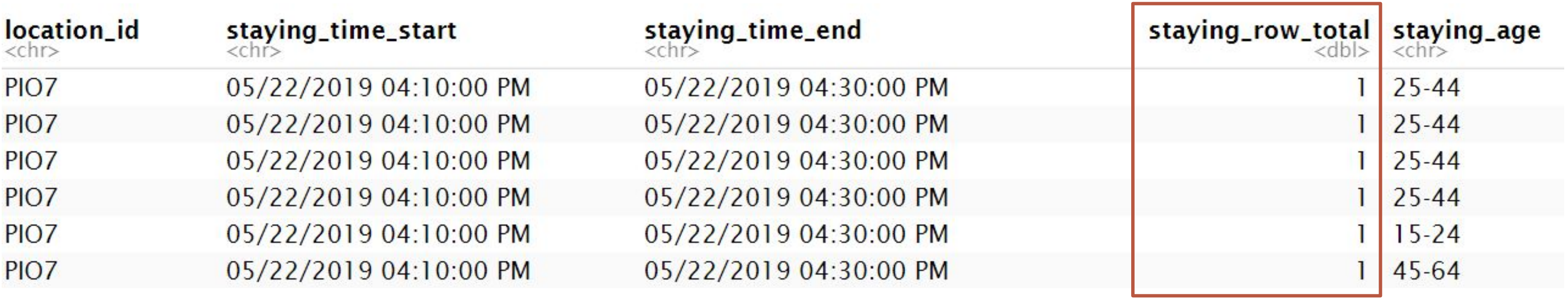
- Practice data cleaning in R
  - leaflet, tidycensus, ggplot
- High level visualizations
- Compare to other metrics of demographic information

Moving

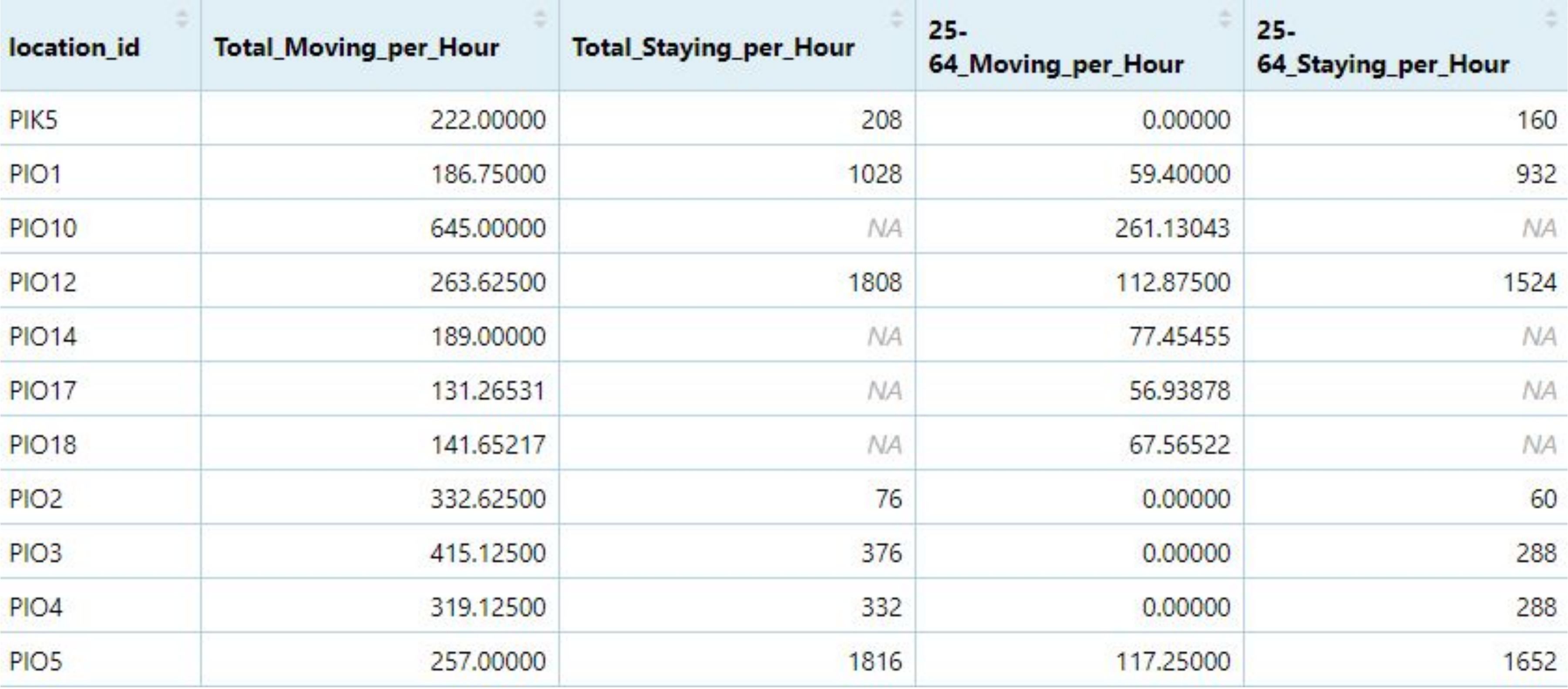


location_id <chr>	moving_time_start <chr>	moving_time_end <chr>	moving_row_total <dbl>	moving_25-44 <dbl>
PIO7	05/22/2019 07:37:00 AM	05/22/2019 07:47:00 AM	45	30
PIO7	05/22/2019 08:30:00 AM	05/22/2019 08:40:00 AM	40	24
PIO7	05/22/2019 08:01:00 AM	05/22/2019 08:11:00 AM	21	10
PIO8	05/22/2019 07:39:00 AM	05/22/2019 07:49:00 AM	26	15
PIO8	05/22/2019 08:40:00 AM	05/22/2019 08:50:00 AM	17	13

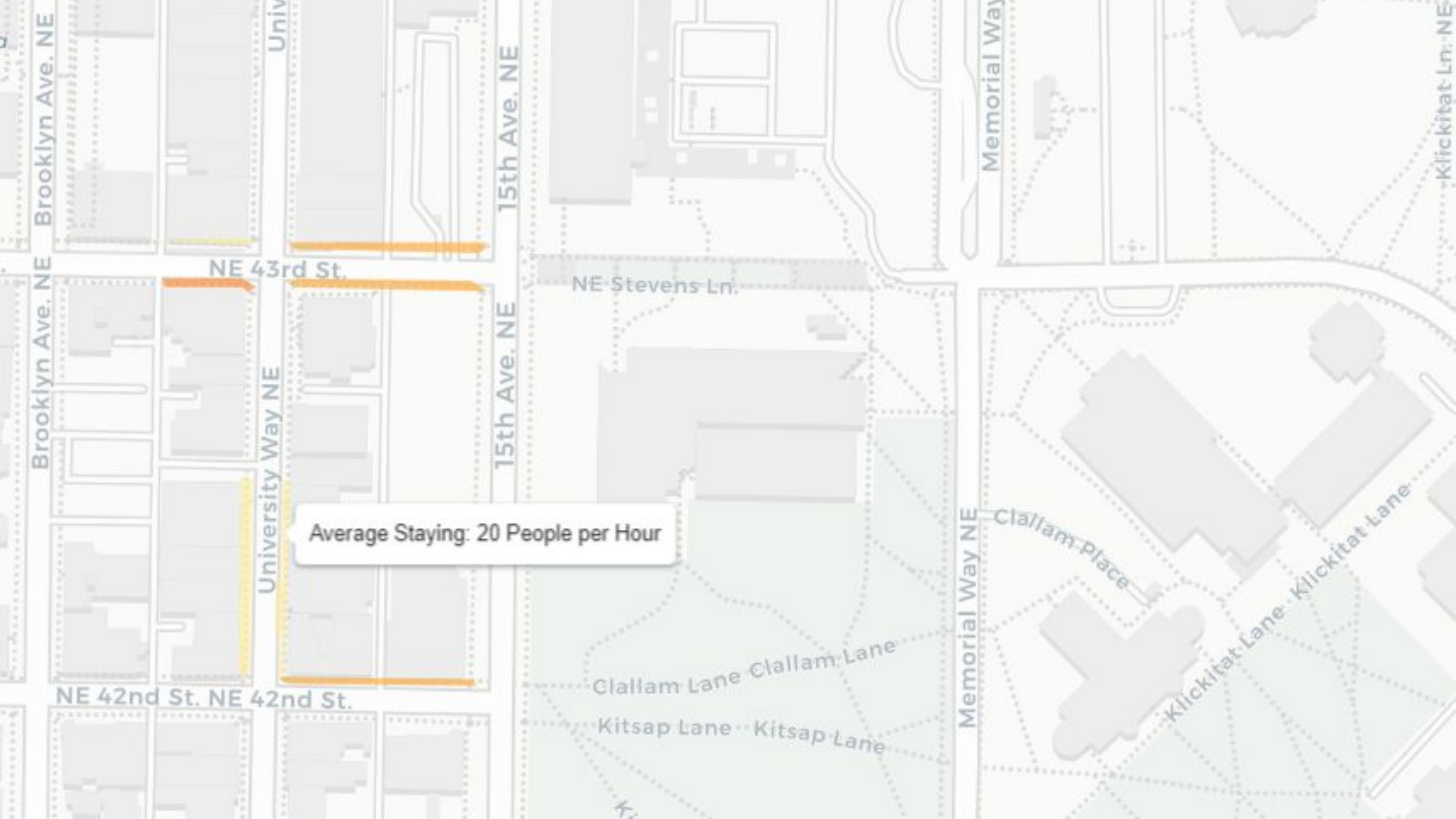
Staying



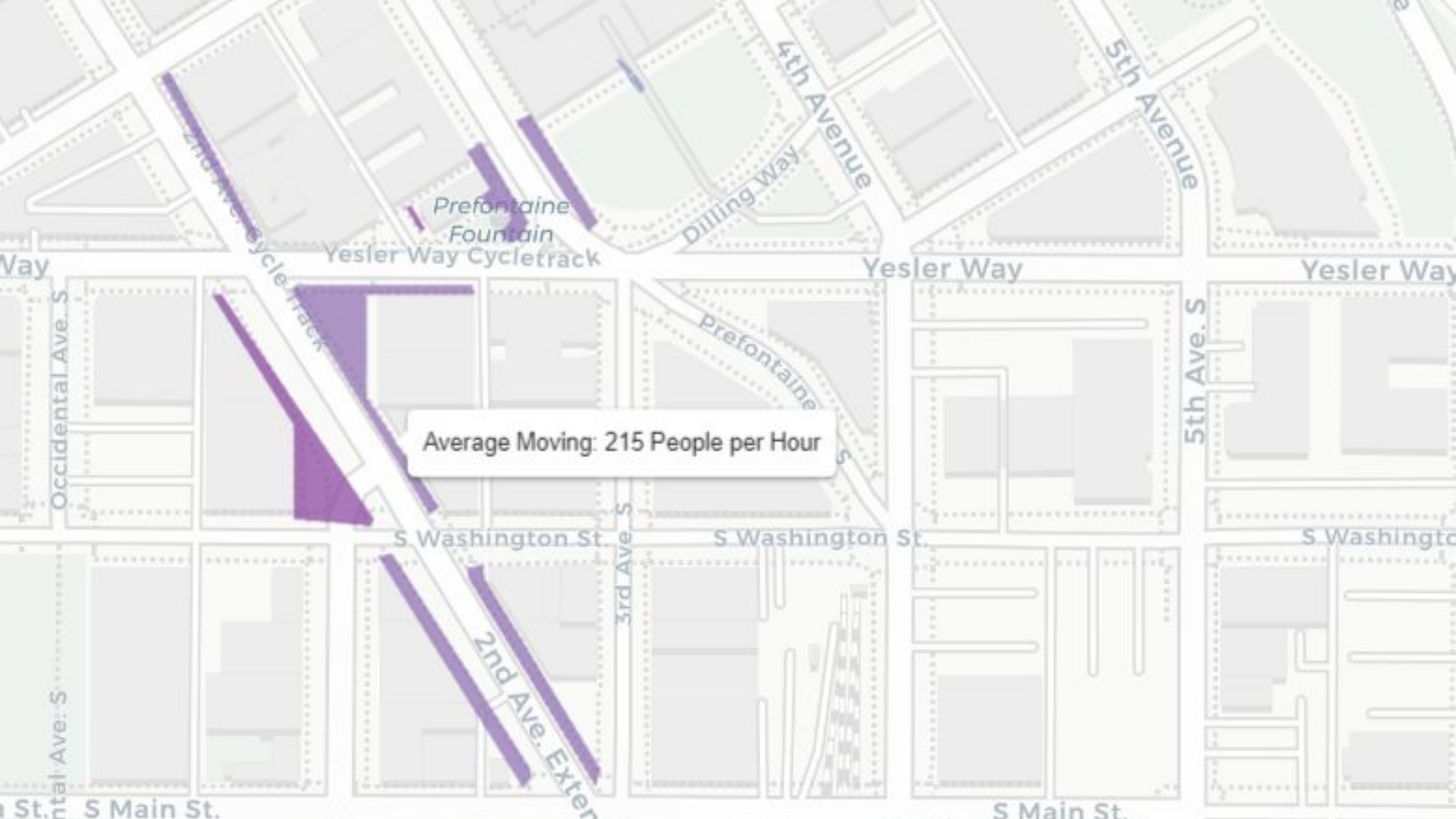
location_id <chr>	staying_time_start <chr>	staying_time_end <chr>	staying_row_total <dbl>	staying_age <chr>
PIO7	05/22/2019 04:10:00 PM	05/22/2019 04:30:00 PM	1	25-44
PIO7	05/22/2019 04:10:00 PM	05/22/2019 04:30:00 PM	1	25-44
PIO7	05/22/2019 04:10:00 PM	05/22/2019 04:30:00 PM	1	25-44
PIO7	05/22/2019 04:10:00 PM	05/22/2019 04:30:00 PM	1	25-44
PIO7	05/22/2019 04:10:00 PM	05/22/2019 04:30:00 PM	1	15-24
PIO7	05/22/2019 04:10:00 PM	05/22/2019 04:30:00 PM	1	45-64



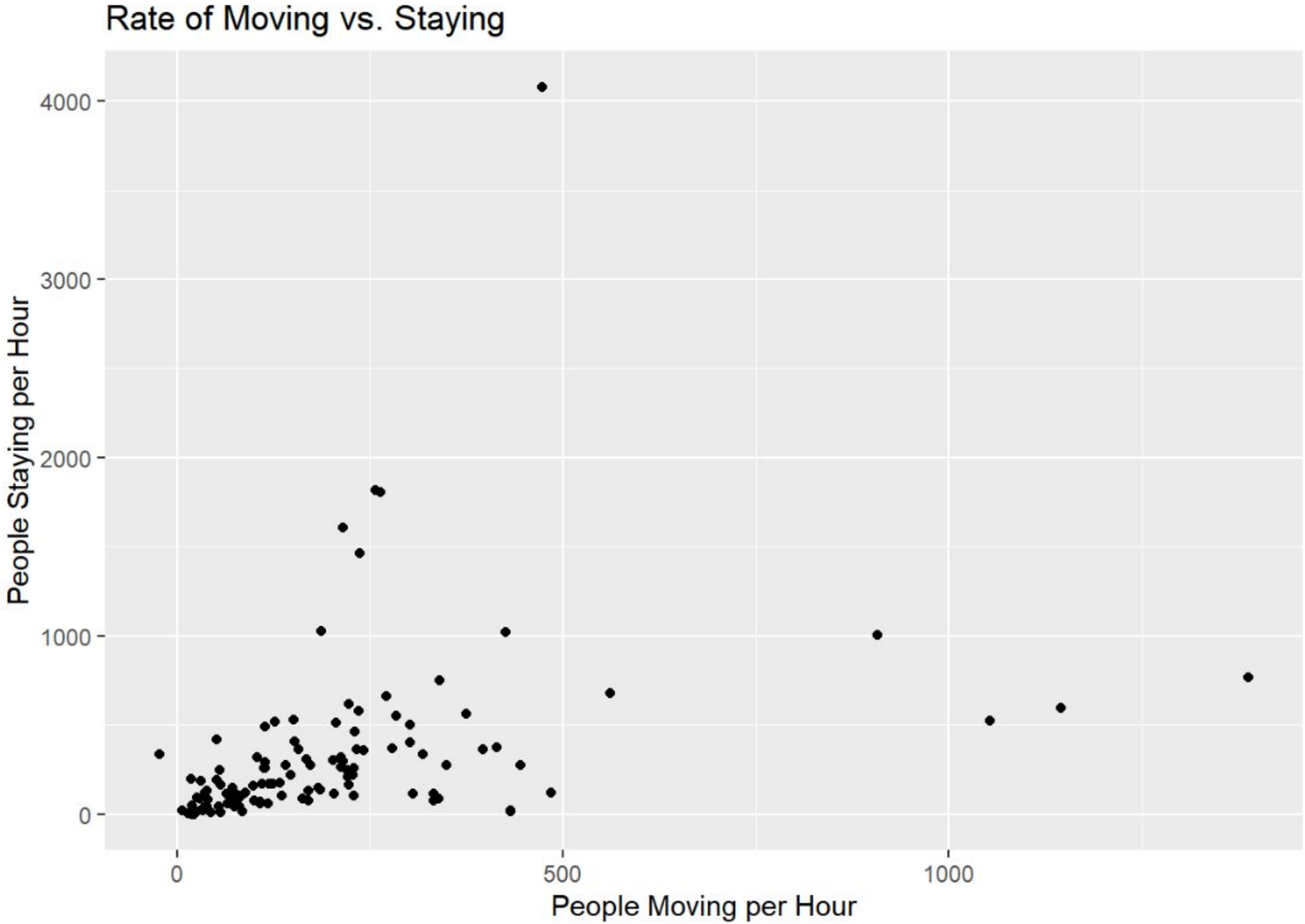
location_id	Total_Moving_per_Hour	Total_Staying_per_Hour	25-64_Moving_per_Hour	25-64_Staying_per_Hour
PIK5	222.00000	208	0.00000	160
PIO1	186.75000	1028	59.40000	932
PIO10	645.00000	NA	261.13043	NA
PIO12	263.62500	1808	112.87500	1524
PIO14	189.00000	NA	77.45455	NA
PIO17	131.26531	NA	56.93878	NA
PIO18	141.65217	NA	67.56522	NA
PIO2	332.62500	76	0.00000	60
PIO3	415.12500	376	0.00000	288
PIO4	319.12500	332	0.00000	288
PIO5	257.00000	1816	117.25000	1652

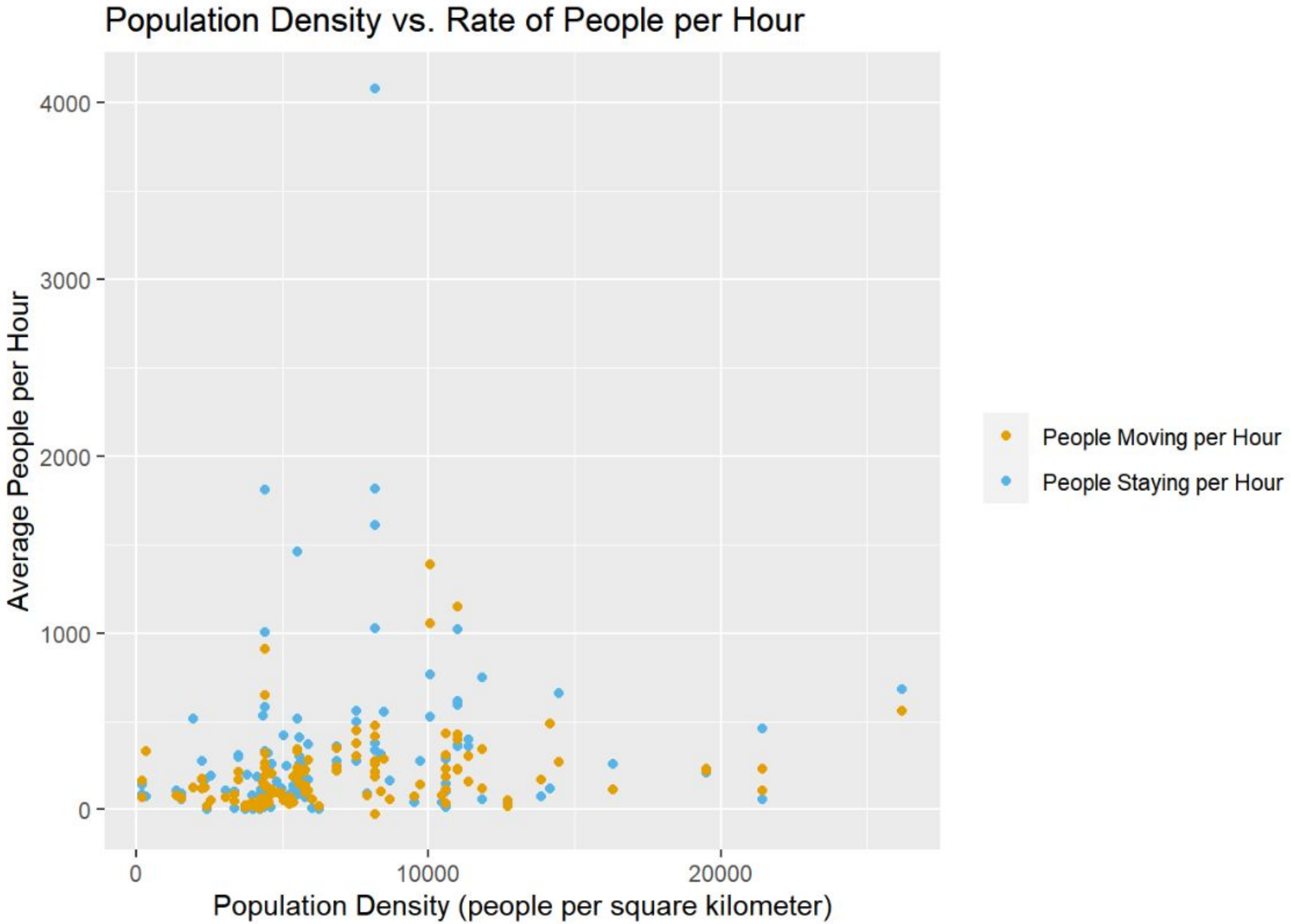


Average Staying: 20 People per Hour



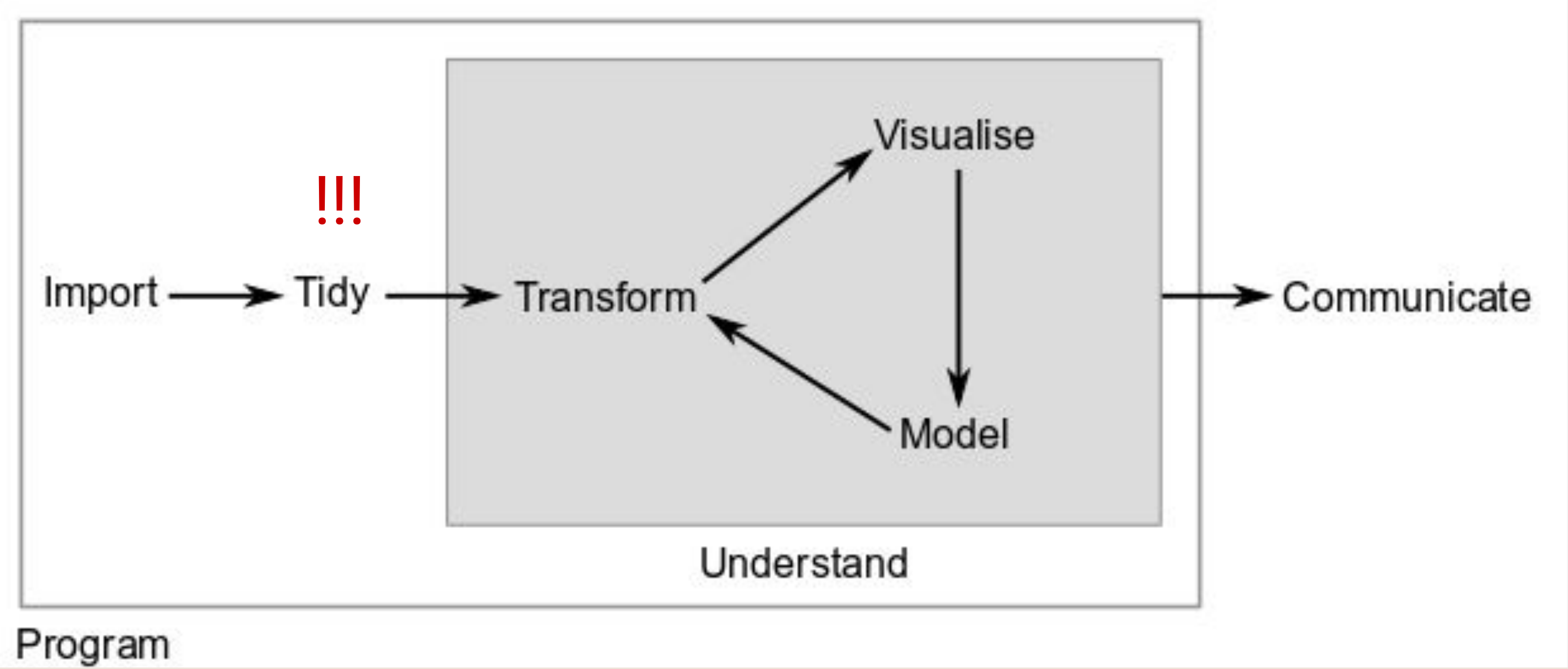
Average Moving: 215 People per Hour


Rate of Moving vs. Staying
People Staying per Hour
4000
3000
2000
1000
0
0
500
1000
People Moving per Hour



Population Density vs. Rate of People per Hour

Conclusion


!!!
Import
Tidy
Transform
Visualise
Model
Understand
Communicate
Program

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.

"Public Life Data - Study | City of Seattle Open Data Portal." *Data.seattle.gov*, 12 Feb. 2021, data.seattle.gov/Transportation/Public-Life-Data-Study/7qru-sdcp. Accessed 5 Dec. 2022.

Rawson, Katie, and Trevor Muñoz. "Against Cleaning." *Curatingmenus.org*, 6 July 2016, curatingmenus.org/articles/against-cleaning/.

Wickham, Hadley. "Tidy Data." *Journal of Statistical Software*, vol. 59, no. 10, 2014, 10.18637/jss.v059.i10. Accessed 19 Dec. 2019.

Wickham, Hadley, and Garrett Grolemund. *R for Data Science*. O'Reilly Media, 12 Dec. 2016.