Commuting Distance from Campus & Student Success

The Route to Geo-spatial Student Analysis

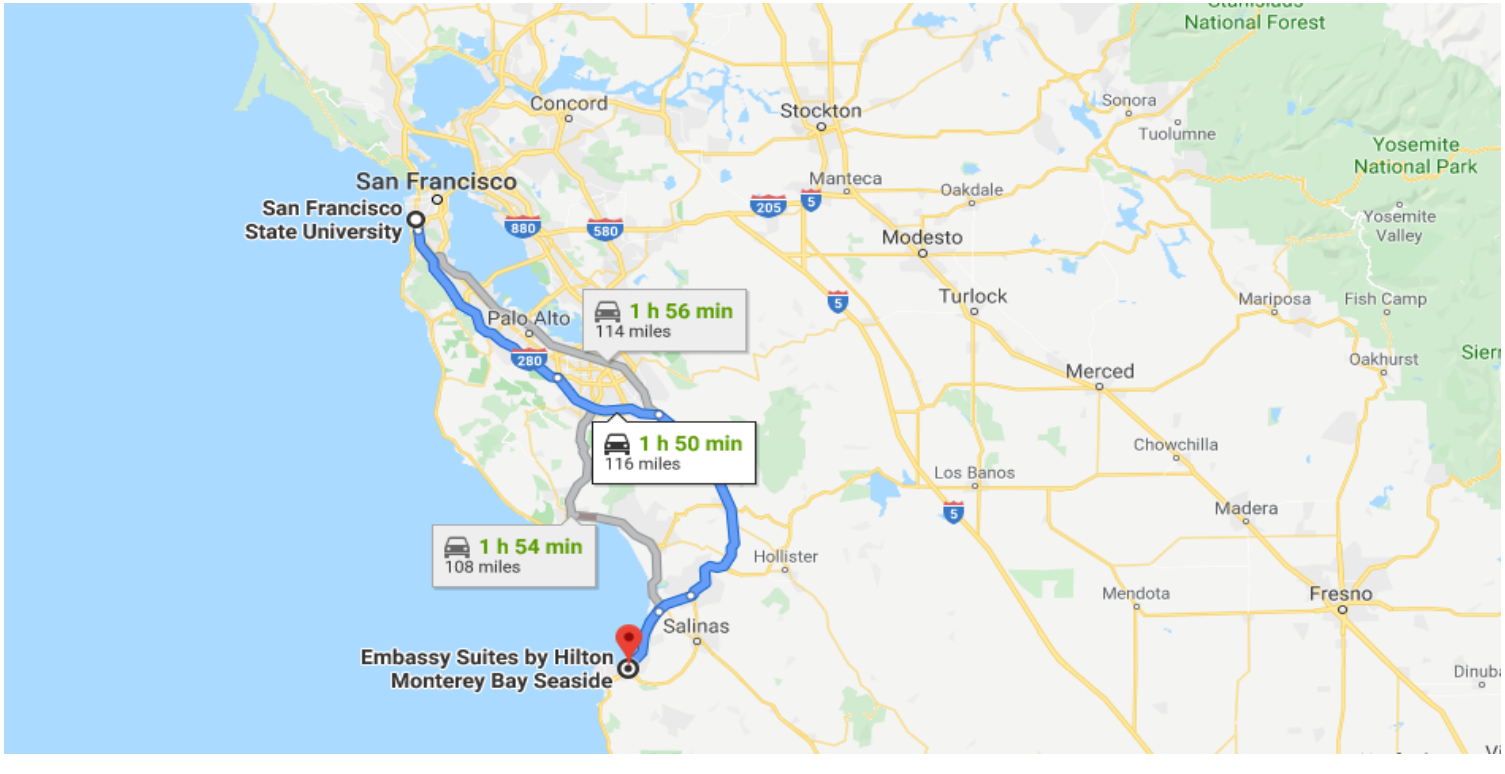
Greg Dubrow & David Ranzolin

November 8, 2019

About Us

- Greg
- Director of IR at SFSU
- Data nerd
- Map nerd

- David
- Senior Analyst at SFSU
- R Partisan
- GIS Enthusiast



Today's Presentation

- Road map to building geo-spatial student analysis
  - Rationale
  - Tools
  - Steps
  - Recommendations

Why do this? What do you need?

- Why study student commutes?
  - Commuter students more the norm in higher ed, especially our sector
  - Time spent commuting is not spent in academic and co-curricular activities
  - Relevant to CSU efforts to increase retention and graduation rates

- What do you need?
  - Student addresses and student success data
  - Access to geo-coding for large number of addresses
  - Some skills in R or other data analysis/viz software with robust geo-spatial features (e.g. Python, QGIS, ArcGIS, etc.)
  - A curious prediliction towards, and patience for, extensive cleaning and tidying of messy address files

Addresses can be messy - example 1

> The good news - there was a query in CS that gave us plenty of address info. The not-so-good news - it looked like this:

ID	Address_1	Address_2	City	State	Postal	Addr_Type	Phone	Type_phone	Preferred_phone	Addr_Type	Type_email	Email	Preferred_email
8888888		DORMINIT 1503-2-D	San Francisco	CA	94132-4036	DORM	415-555-1212	CELL	N	DORM	OCMP	notrealstudent@ma	Y
8888888		DORMINIT 1503-2-D	San Francisco	CA	94132-4036	DORM	415-555-1212	CELL	N	DORM	OTHR	madeupperson@gm	N
8888888	1212 Homeaddress St		San Jose	CA	95127	DIPL	415-555-1212	CELL	N	DIPL	OCMP	notrealstudent@ma	Y
8888888	1212 Homeaddress St		San Jose	CA	95127	DIPL	415-555-1212	CELL	N	DIPL	OTHR	madeupperson@gm	N
8888888	1212 Homeaddress St		San Jose	CA	95127	MAIL	415-555-1212	CELL	N	MAIL	OCMP	notrealstudent@ma	Y
8888888	1212 Homeaddress St		San Jose	CA	95127	MAIL	415-555-1212	CELL	N	MAIL	OTHR	madeupperson@gm	N
8888888	500 Permaddress Ave		Milpitas	CA	95035	PERM	415-555-1212	CELL	N	PERM	OCMP	notrealstudent@ma	Y
8888888	500 Permaddress Ave		Milpitas	CA	95035	PERM	415-555-1212	CELL	N	PERM	OTHR	madeupperson@gm	N
8888888		DORMINIT 1503-2-D	San Francisco	CA	94132-4036	DORM	408-555-1212	MAIN	Y	DORM	OCMP	notrealstudent@ma	Y
8888888		DORMINIT 1503-2-D	San Francisco	CA	94132-4036	DORM	408-555-1212	MAIN	Y	DORM	OTHR	madeupperson@gm	N
8888888	1212 Homeaddress St		San Jose	CA	95127	DIPL	408-555-1212	MAIN	Y	DIPL	OCMP	notrealstudent@ma	Y
8888888	1212 Homeaddress St		San Jose	CA	95127	DIPL	408-555-1212	MAIN	Y	DIPL	OTHR	madeupperson@gm	N
8888888	1212 Homeaddress St		San Jose	CA	95127	MAIL	408-555-1212	MAIN	Y	MAIL	OCMP	notrealstudent@ma	Y
8888888	1212 Homeaddress St		San Jose	CA	95127	MAIL	408-555-1212	MAIN	Y	MAIL	OTHR	madeupperson@gm	N
8888888	500 Permaddress Ave		Milpitas	CA	95035	PERM	408-555-1212	MAIN	Y	PERM	OCMP	notrealstudent@ma	Y
8888888	500 Permaddress Ave		Milpitas	CA	95035	PERM	408-555-1212	MAIN	Y	PERM	OTHR	madeupperson@gm	N

Addresses can be messy - example 2

> There are 50 Eskimo words for snow, and 20-some ways to spell "San Francisco". Regex can be your friend.

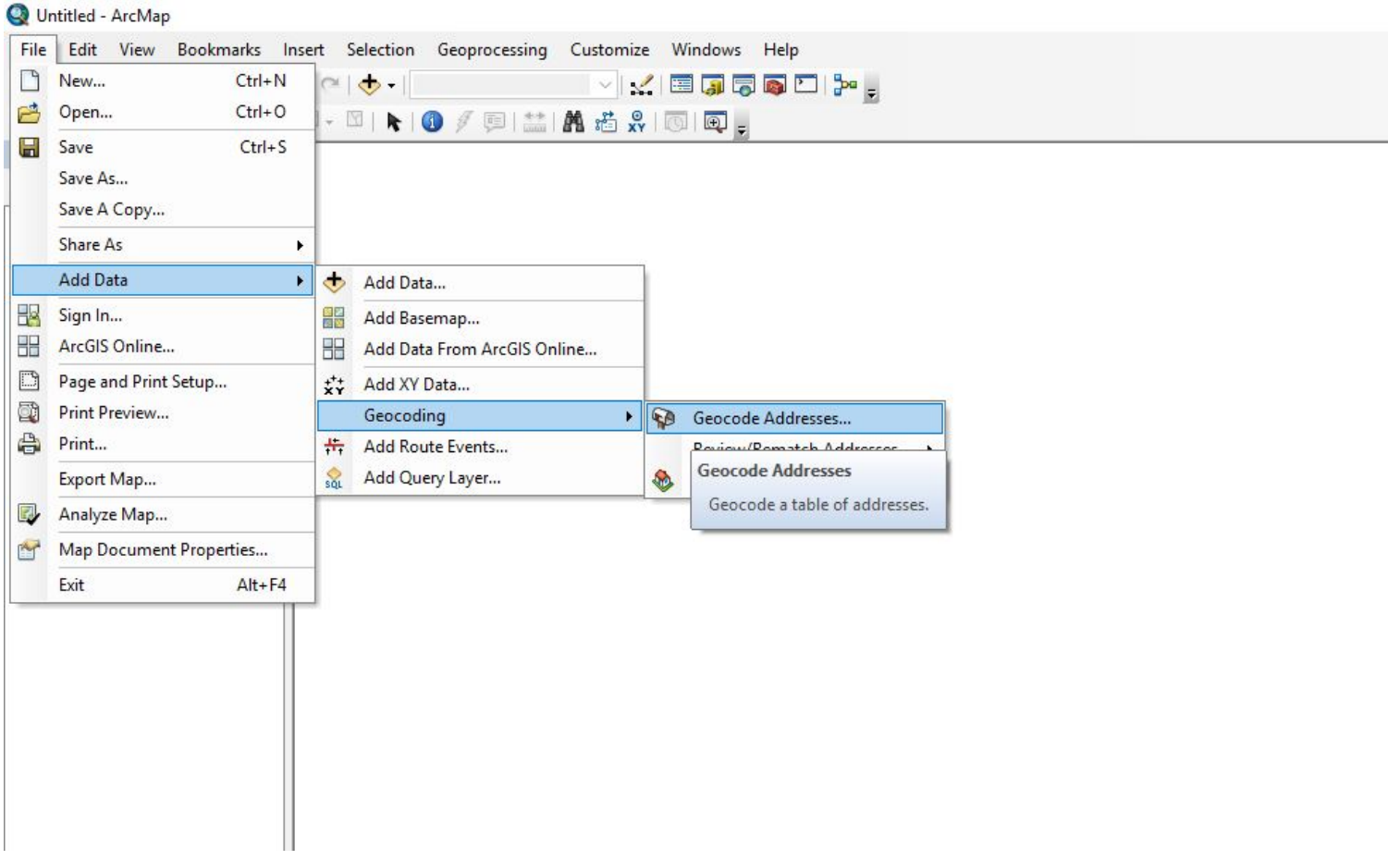
```
library(rsfsu)
library(tidyverse)
library(tidylog)
library(janitor)

## read in address file

address %>%
  mutate(City = ifelse(City == "Sf" | City == "San  Francisco" | City == "Sanfrancisco"
                       | City == "San Fancisco", 'San Francisco', City)) %>%
  mutate(City = ifelse(grepl("^San Fr.*", City),'San Francisco', City)) %>%

  mutate(City = ifelse(grepl("^South San Fr.*", City),'South San Francisco', City)) %>%
  mutate(City = ifelse(grepl("^So. San Francisco", City),'South San Francisco', City)) %>%
  mutate(City = ifelse(grepl("^S. San Francisco", City),'South San Francisco', City)) %>%
  mutate(City = ifelse(grepl("^South  San Francisco", City),'South San Francisco', City)) %>%
  mutate(City = ifelse(City == "Ssf", 'South San Francisco', City)) %>%
  mutate(City = ifelse(City == "South Sanfrancisco", 'South San Francisco', City))
```

Addresses can be messy - example 3

> People don't always fill out the address fields as you'd like

```
address %>%
  # if apartment/unit infor in address 1, moves to new field
  mutate(apttxt = ifelse(str_detect(Address_1, "Apt"), str_extract(Address_1, "(Apt).*"),
                         ifelse(str_detect(Address_1, "Apartment"),
                                str_extract(Address_1, "(Apartment).*"),
                                ifelse(str_detect(Address_1, "Unit"),
                                       str_extract(Address_1, "(Unit).*"),
                                       ifelse(str_detect(Address_1, "#"),
                                              str_extract(Address_1, "(#).*"), NA))))) %>%
  mutate(Address_1a = ifelse(str_detect(Address_1, "Apt"), str_remove(Address_1, "(Apt).*"),
                             ifelse(str_detect(Address_1, "Apartment"),
                                    str_remove(Address_1, "(Apartment).*"),
                                    ifelse(str_detect(Address_1, "Unit"),
                                           str_remove(Address_1, "(Unit).*"),
                                           ifelse(str_detect(Address_1, "#"),
                                                  str_remove(Address_1, "(#).*"),
                                                  Address_1))))) %>%
  mutate(Address_1a = str_remove(Address_1a, ",")) %>%
  mutate(Address_2 = ifelse(is.na(Address_2) & !is.na(apttxt), apttxt, Address_2)) %>%
```

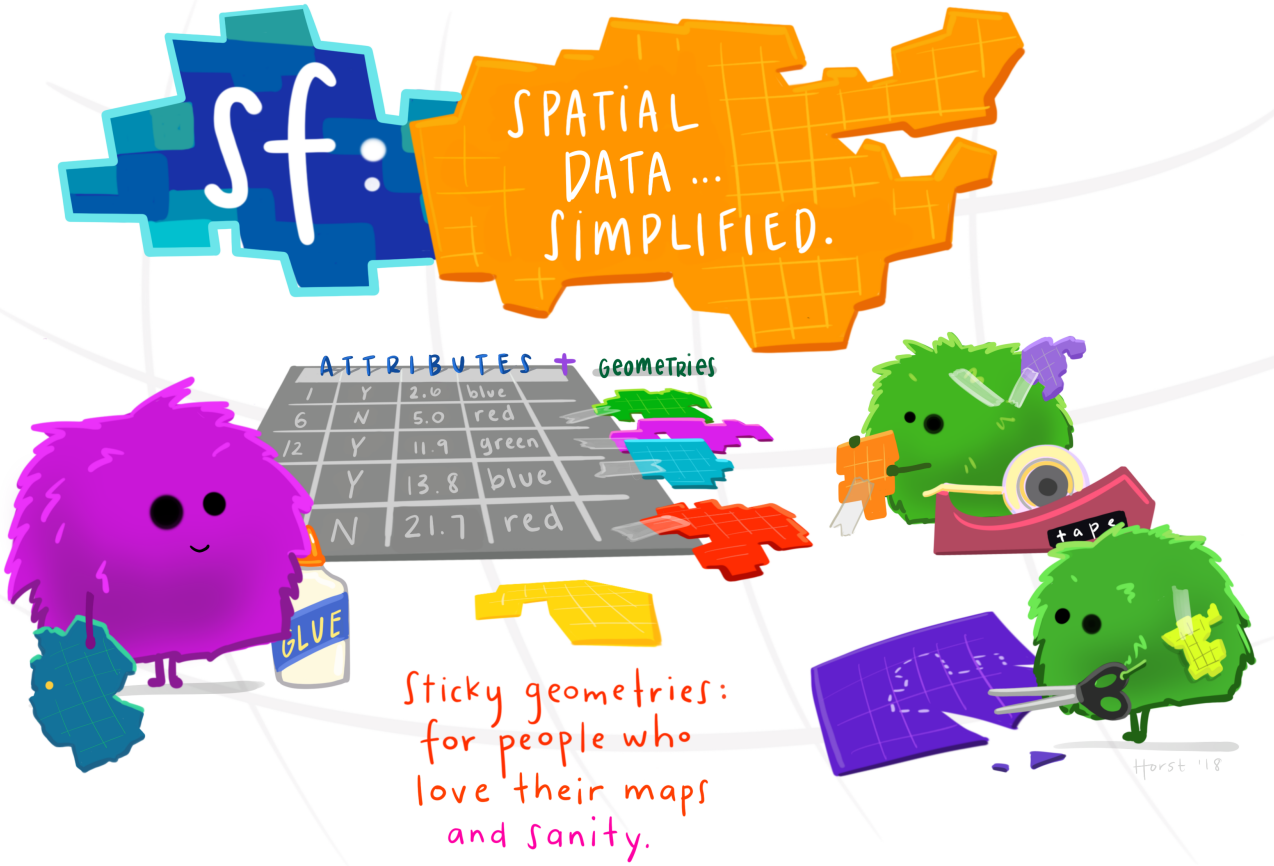
Time to get longitude & latitude



Build student outcomes, merge with addresses

```
# connect to analytics datamart and get variables
stuoutcomes <- dm_enr %>%
  select(a bunch of student outcomes variables)

# read in geocoded addresses
geocoded <- read.csv(file="geocodedaddress.csv",
                     header=TRUE, sep=",",
                     stringsAsFactors = FALSE)


finaldata1 <- merge(stuoutcomes, geocoded, by = "SID")
```

Geospatial Fun



GIS Things to Know

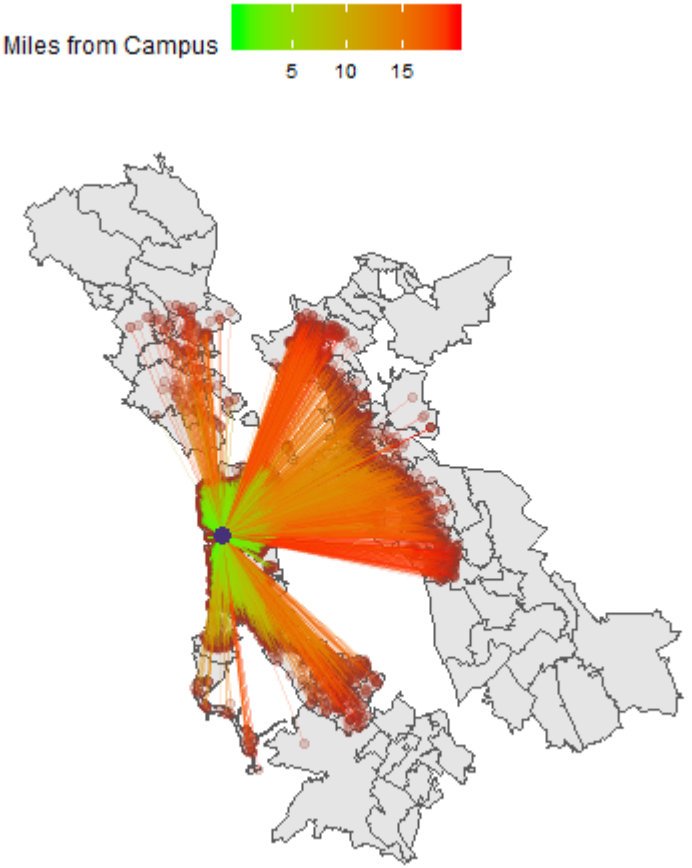
- Geographic data
  - Vector data
  - Rasters
- Coordinate Reference Systems and Projections
- File formats
  - .shp
  - .tif
  - .json, .geojson, .topojson
- Geometric operations
- Tools
  - ArcMap / ArcGIS Pro
  - QGIS
  - R
  - Python

The sf package

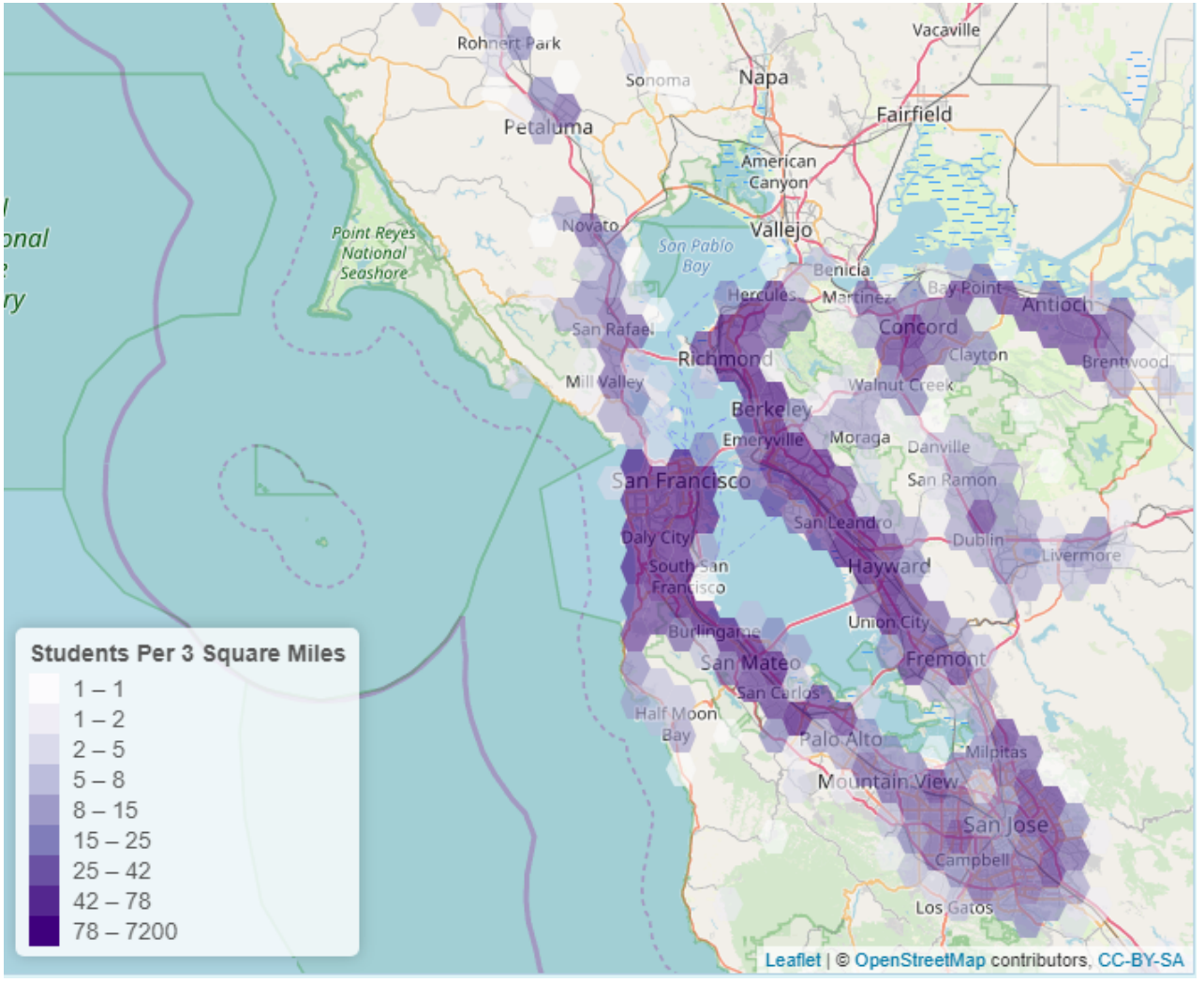
```
library(sf)

addresses <- finaldata1 # dataset with addresses & student outcomes
addresses <- st_as_sf(addresses, coords = c("X", "Y"))
addresses_sfc <- st_sfc(st_geometry(addresses), crs = 4326)

sfsu <- st_read(".../Shapefiles/sfsu_point")
sfsu_sfc <- st_sfc(st_geometry(sfsu), crs = 4326)

distances <- st_distance(sfsu_sfc, addresses_sfc, by_element = TRUE)
str(distances)
```

```
'units' num [1:26504] 28076 6971 18026 26968 6448 ...
- attr(*, "units")=List of 2
 ..$ numerator  : chr "m"
 ..$ denominator: chr(0)
 ..- attr(*, "class")= chr "symbolic_units"
```

EDA 1 - Euclidian Distance

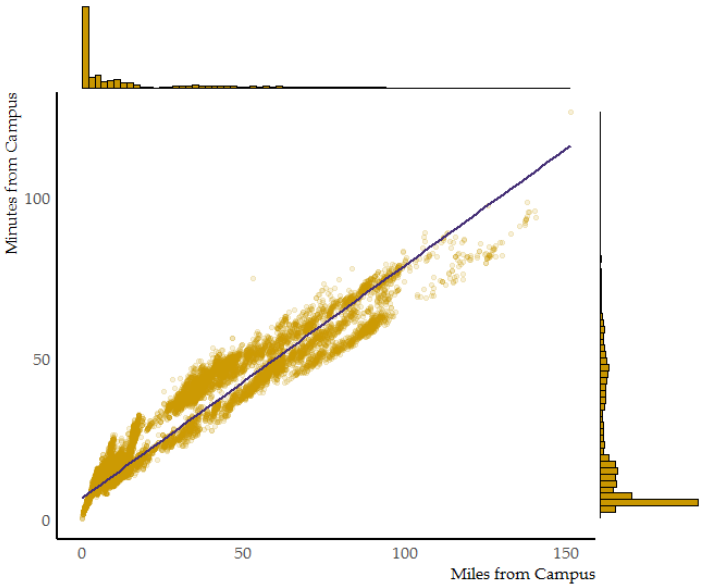


EDA 2 - Hexbins



EDA 3 - OSRM

- OSRM is a routing service based on OpenStreetMap data. The tool computes distances (travel time and kilometric distance) between points and travel time matrices.



EDA 5 - Term GPAs

Fall 2018 Term GPA by Entering Type

	GPA	students
Freshmen	2.83	15,107
Transfers	2.98	10,337

Fall 2018 Term GPA by Enroll Status

	GPA	students
New Freshman	2.76	4,285
New Transfer	2.97	3,398
Continuing/Return	2.90	17,761

EDA 6 - Addresses can be messy, example 4

Fall 2018 Distance to Campus by Entry Type

	Avg Dist	Median Dist	Max Dist
Freshmen	23.4	0.5	2,579
Transfers	40.0	9.2	2,434

EDA 6 - Addresses can be messy, example 4

Fall 2018 Distance to Campus by Entry Type

	Avg Dist	Median Dist	Max Dist
Freshmen	23.4	0.5	2,579
Transfers	40.0	9.2	2,434

Fall 2018 Distance to Campus by Entry Type

Filtered to CA addresses only

	Avg Dist	Median Dist	Max Dist
Freshmen	22.1	0.5	525
Transfers	37.8	9.1	527

EDA 6 - Addresses can be messy, example 4

Fall 2018 Distance to Campus by Entry Type

	Avg Dist	Median Dist	Max Dist
Freshmen	23.4	0.5	2,579
Transfers	40.0	9.2	2,434

Fall 2018 Distance to Campus by Entry Type

Filtered to CA addresses only

	Avg Dist	Median Dist	Max Dist
Freshmen	22.1	0.5	525
Transfers	37.8	9.1	527

Fall 2018 Distance to Campus by Enrollment Group

Filtered to addresses within 100 miles

	Avg Dist	Median Dist	Max Dist
Freshmen	6.7	0.1	100
Transfers	15.3	5.8	100

EDA 7 - Enrollment context matters

Fall 2018 Distance to Campus by Enrollment Group

Filtered to addresses within 100 miles

	Avg Dist	Median Dist
New Freshman	6.5	0.1
New Transfer	16.3	7.5
Continuing/Return	9.9	2.5

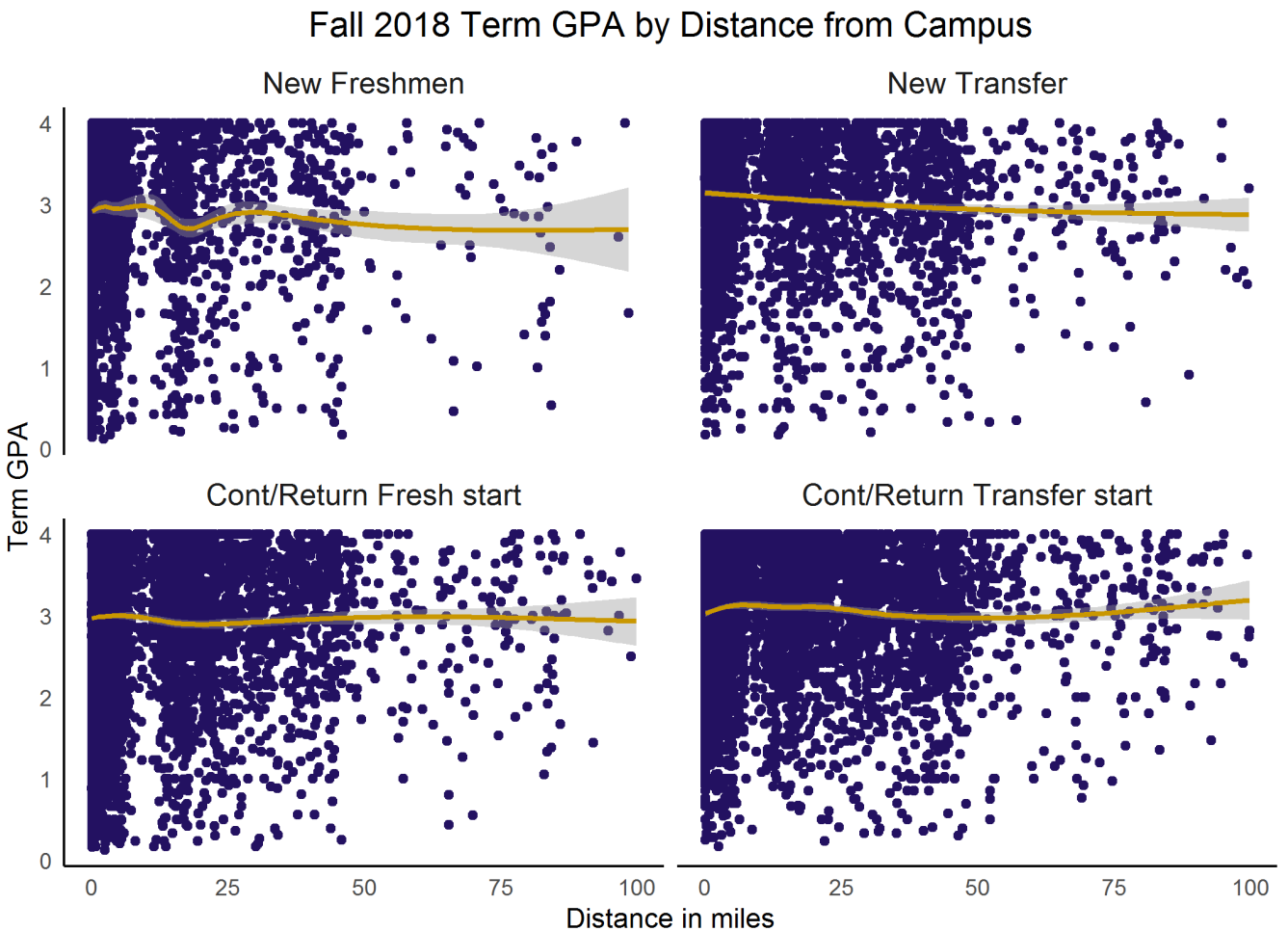
EDA 7 - Enrollment & Housing contexts

Fall 2018 Distance by Enrollment/Housing Group

Filtered to addresses within 100 miles

	Avg Dist	Median Dist
New Freshmen, on-campus	0.1	0.1
New Freshmen, off-campus	14.5	6.2
New Transfer, on-campus	0.1	0.1
New Transfer, off-campus	18.8	14.0
Freshman start, on-campus	0.1	0.1

Any actionable findings?



Recommendations

- Know your data sources
- Consider your spatial context
- Consider mass transit options
- Consider census data overlays

Resources

- R packages
  - sf - Simple features in R
  - osrm - R interface to OSRM
  - leaflet - R interface to Leaflet
  - tmap - Thematic maps in R
  - ggplot2 - Grammar of Graphics in R

Thank you!

Connect with us...

- Greg

- Twitter: @greg_dubrow
- LinkedIn: dubrowg
- GitHub: dubrowg
- Email: gdubrow@sfsu.edu

- David

- Twitter: @daranzolin
- LinkedIn: dranzolin
- GitHub: daranzolin
- Blog: daranzolin.github.io
- Email: daranzolin@sfsu.edu