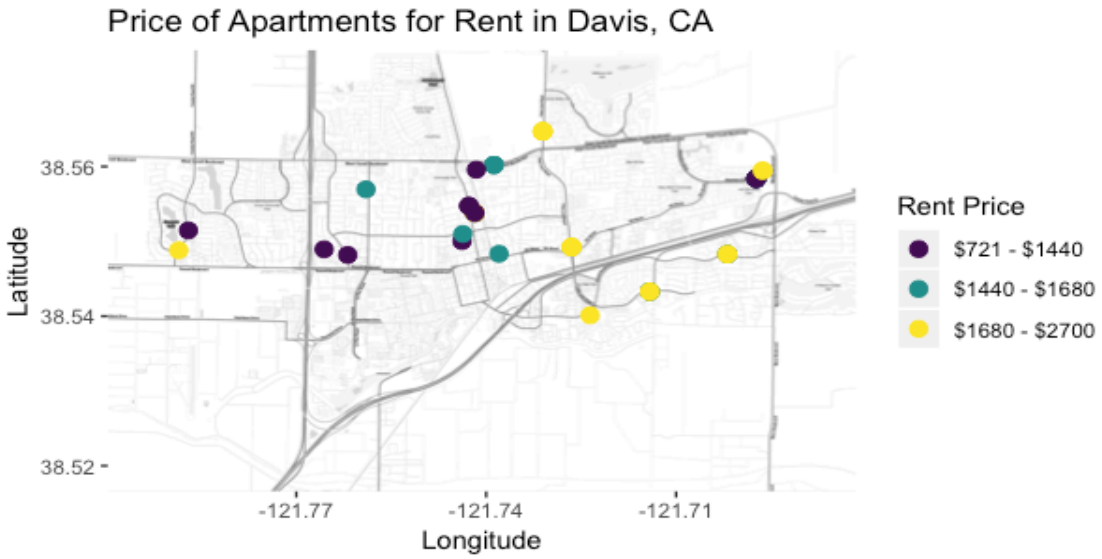
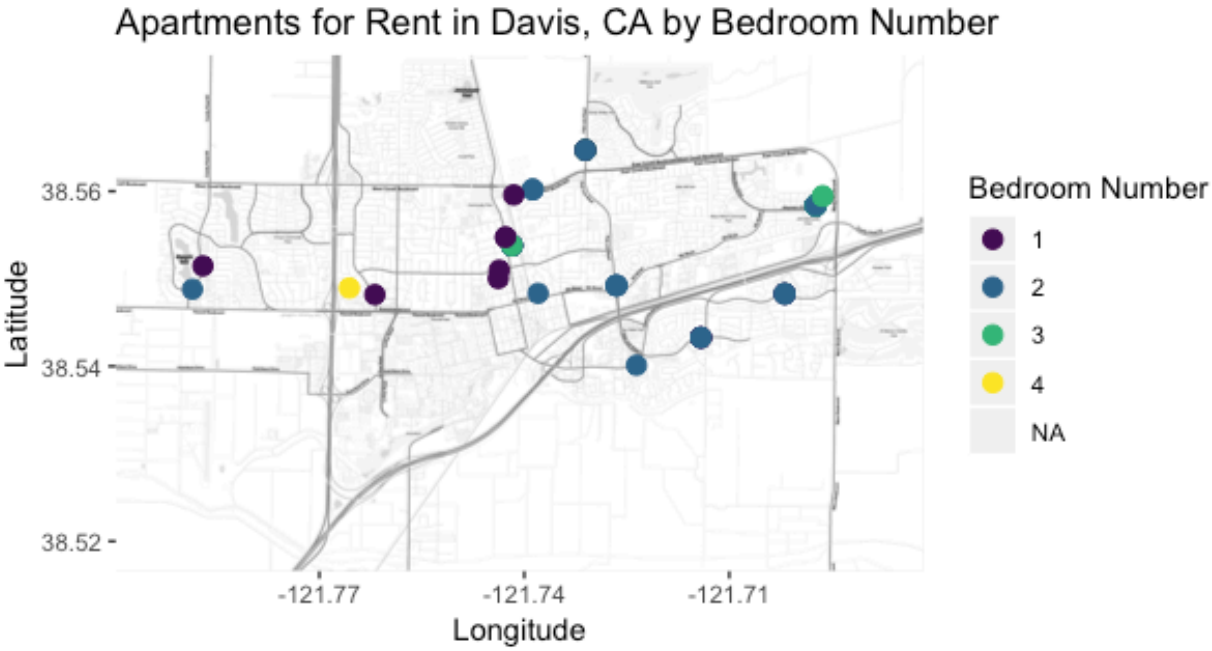
**Craigslist Posts in Davis**



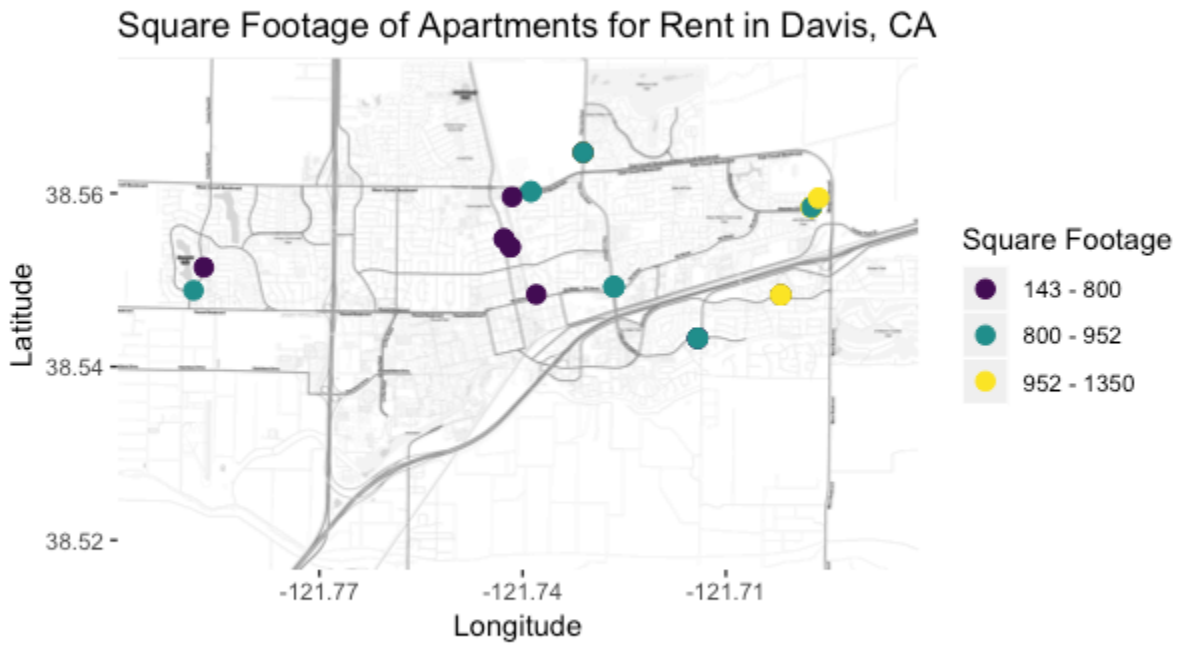

Of the apartments for rent in Davis which could be mapped based on the presence of latitude and longitude data, more expensive apartments tended to be located on the outskirts of town, farther away from campus. Cheaper apartments tended to be located in central Davis.



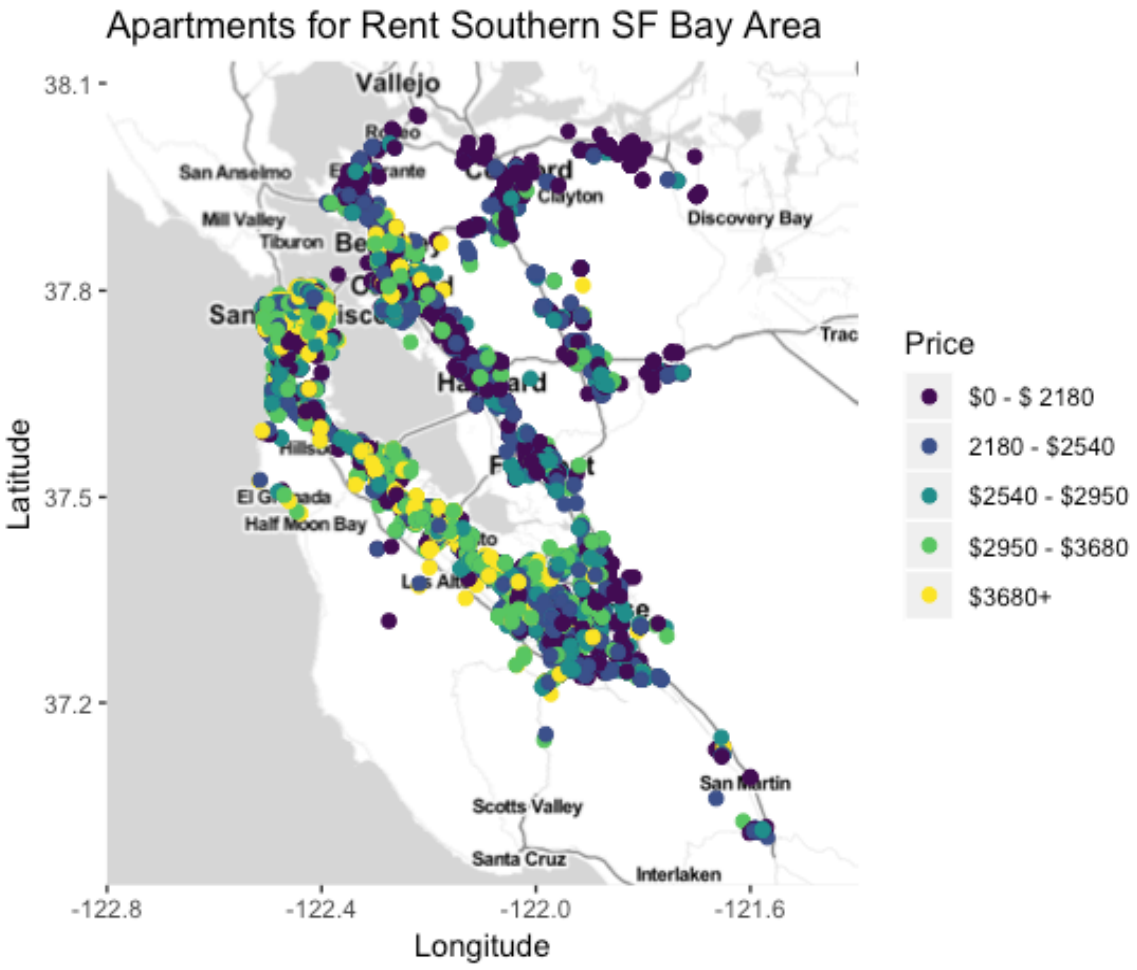

Apartments in Davis don't follow any major geographic pattern in regards to number of bedrooms. Most of the two bedroom apartments are located in the eastern half of Davis, while the one bedroom apartments are located in central and west Davis. However, because there are

so few data points, we cannot reasonably conclude that this trend applies to the entire population of apartments in Davis. This is one limitation of the craigslist data set I am using – for Davis, so few apartments have latitude and longitude data available, it is difficult to form a correct conclusion about Davis apartments.

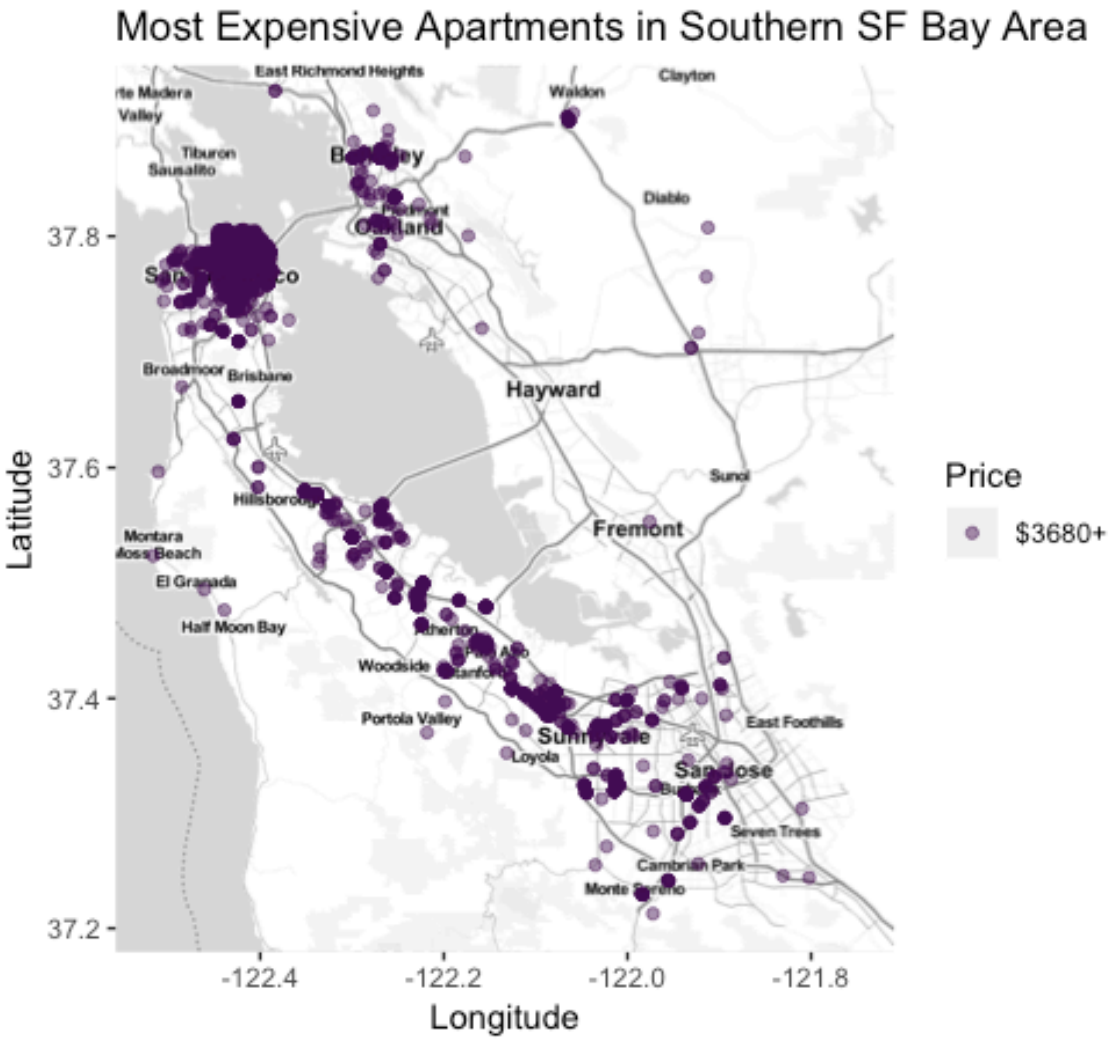
I also examined whether there was a geographical relationship with square footage in Davis. Of the apartments with data available, the largest apartments with 952- 1350 square feet were located on the edge of town in East Davis, while apartments with 143-800 and 800-952 square feet were equally distributed throughout Davis. Again, we need to be careful in assuming this trend extends to all the apartments in Davis, due to the small sample size.



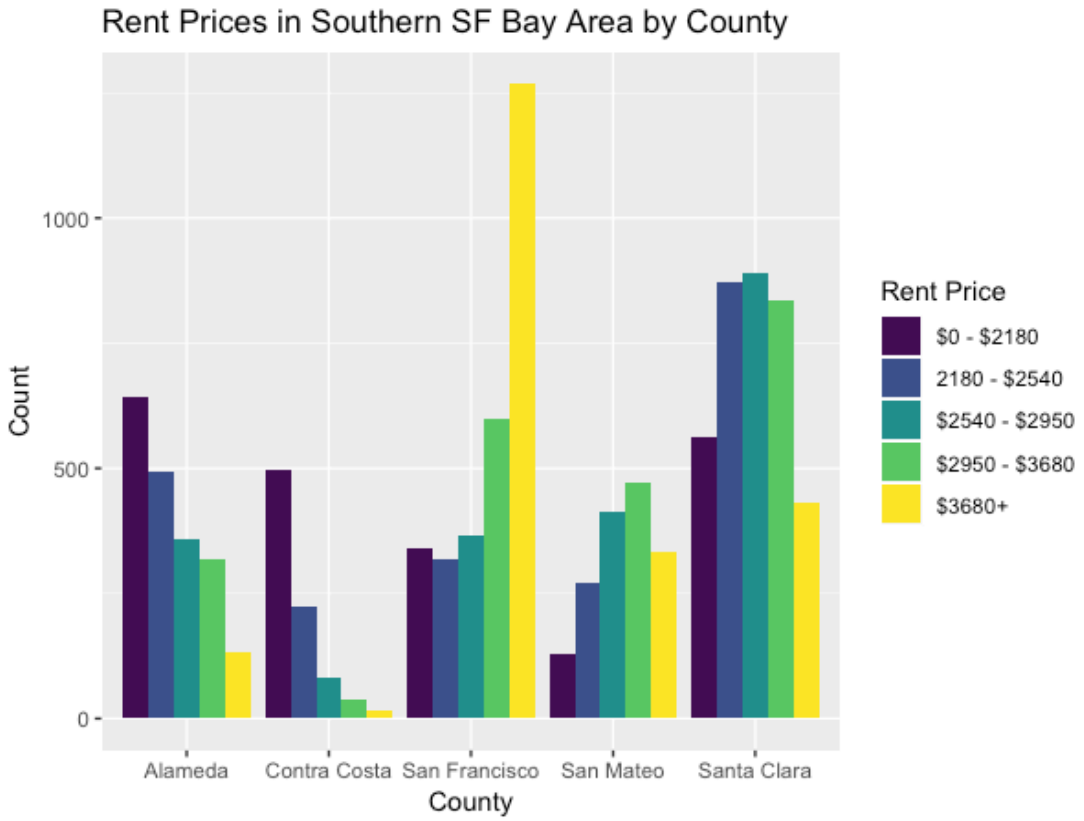
**Craigslist in SF Bay Area**



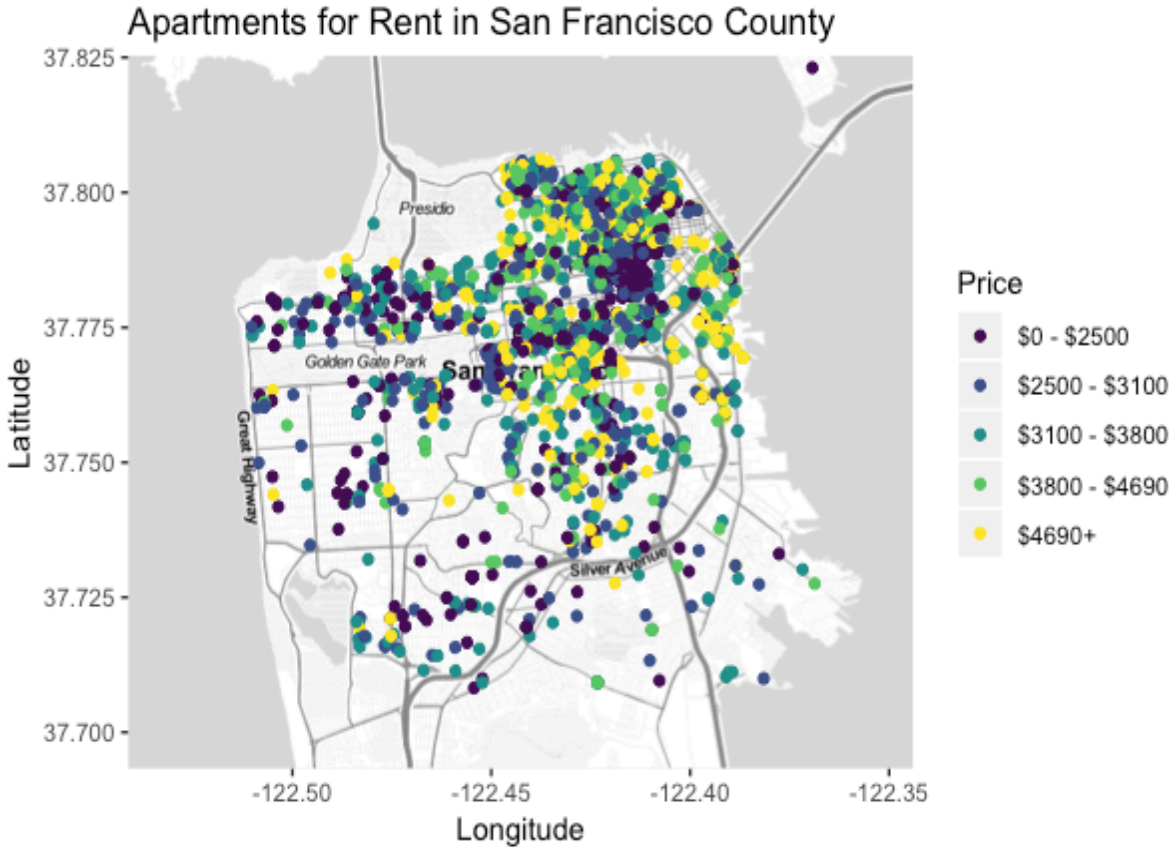

In examining apartments in the SF Bay Area, I first looked at how rent prices were distributed geographically across the Bay. Dividing rent price into five discrete categories, it appeared from the map above that more expensive apartments were located in San Francisco, Oakland, and along the Peninsula. However, this map contains a large amount of over-plotting. To help alleviate the over-plotting, I generated a graph displaying the location of all the apartments in the most expensive range, $3600+. This graph also confirms that the most expensive apartments are located on the Peninsula and in San Francisco.



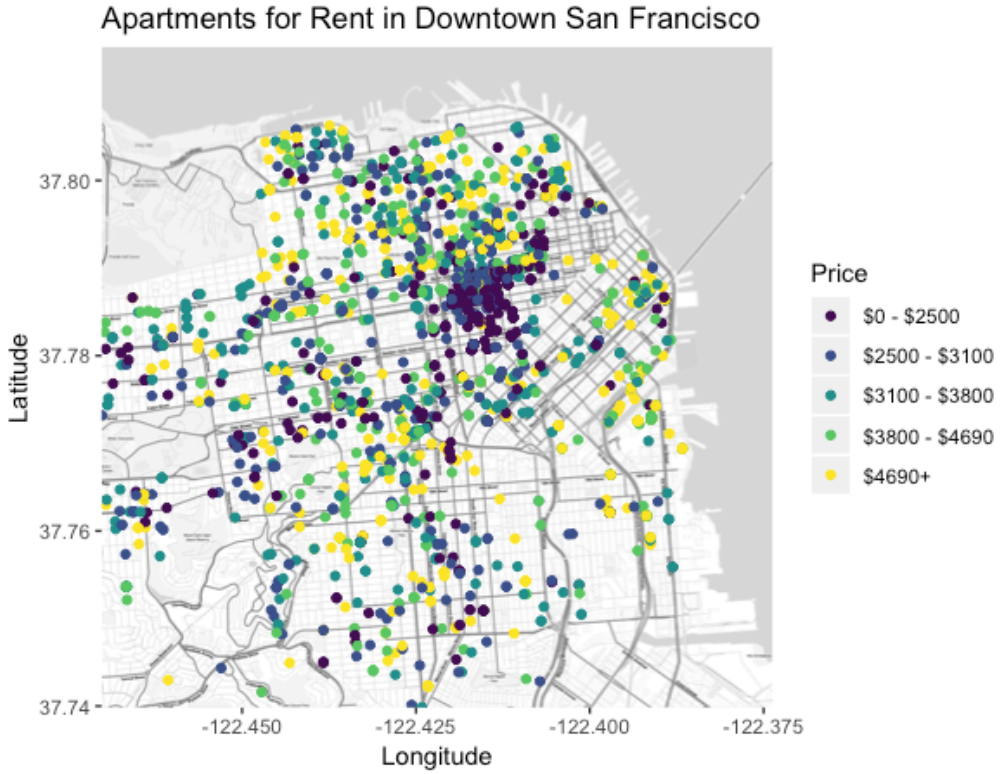

To further confirm my observation about rent price distribution, and more specifically compare price between the five counties in the Southern Bay Area, I generated the bar chart on the following page which confirms that the most expensive apartments are found in San Francisco county. San Mateo and Santa Clara county, which are located on the Peninsula, also have a higher proportion of expensive apartments ($2950- $3680, and $3680+) located in them than Alameda and Contra Costa county do. For Alameda and Contra Costa, a greater percentage of the apartments located in them are towards the less expensive range of all Bay Area apartments.




Next, I was curious about how rent prices were distributed within the most expensive region of the SF Bay, San Francisco county itself. According to the map, it appears as if rent

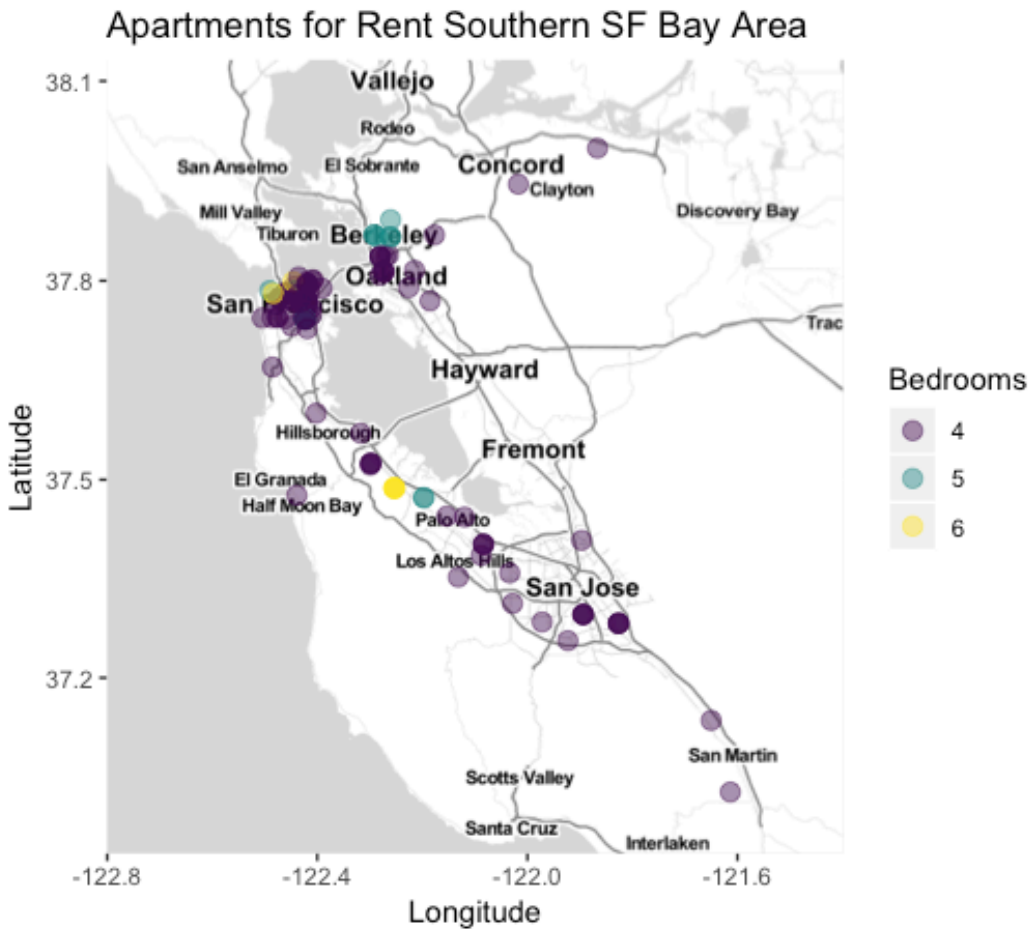


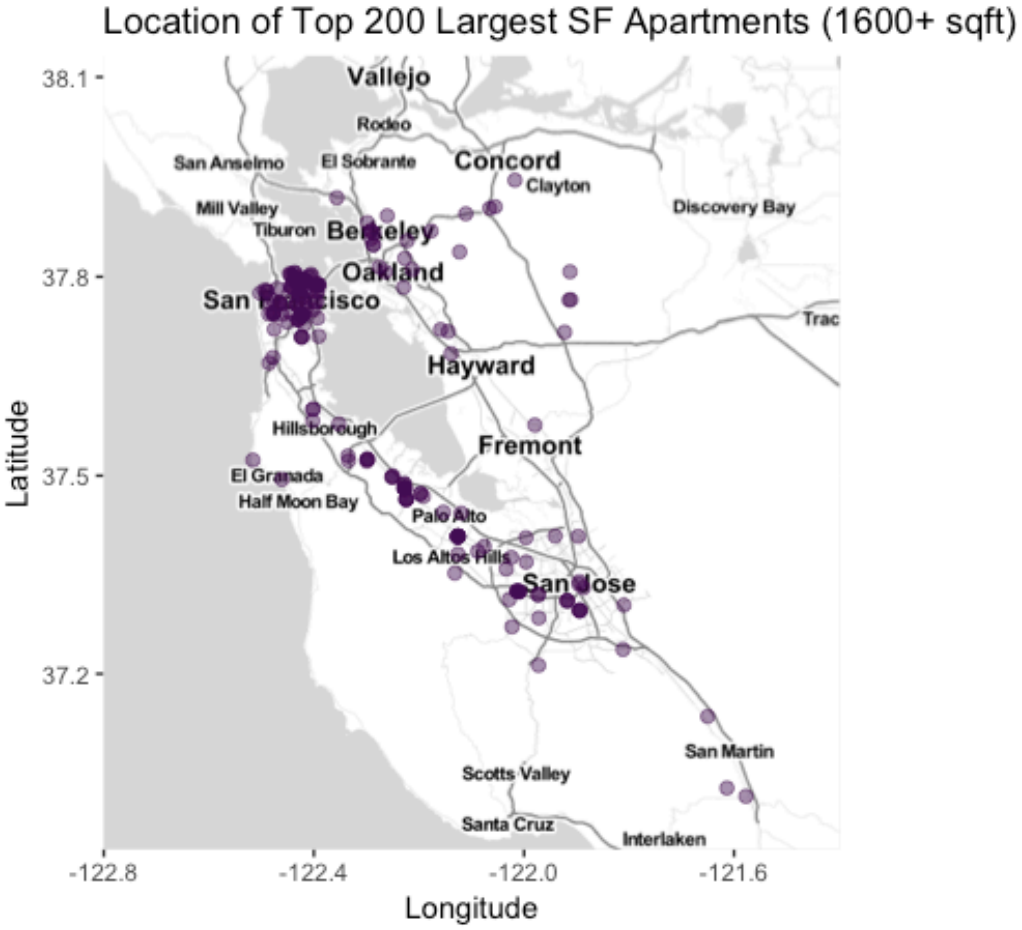
price is fairly equally distributed across the city, with apartments at both ends of the price range found all over the city. To address the high amount of clustering in the downtown SF area, I generated another plot to examine any geographical trends in price in that specific region. Apartments in downtown San Francisco similarly followed no noticeable trend in price.




I was also curious about where in the Bay Area the apartments with the most bedrooms were located, and whether there was any relationship to geography. I found that the four, five and six bedroom apartments were mainly located on the Peninsula, in San Francisco and in Oakland, mirroring the location trend of the most expensive apartments previously examined in this report. This makes sense because apartments with more bedrooms would be assumed to cost more. The top 200 largest apartments in terms of square footage similarly follow this geographic trend.

The results for San Francisco Bay Area are trustworthy, because there is such a large sample size of apartments that we can reasonably assume these trends are representative of the whole population of San Francisco and Bay Area apartments. This is why we can make stronger conclusions about apartment trends in Davis. There are a much higher number of apartments with latitude and longitude data in the Bay Area than in Davis, so we can assume trends aren't just due to random chance.

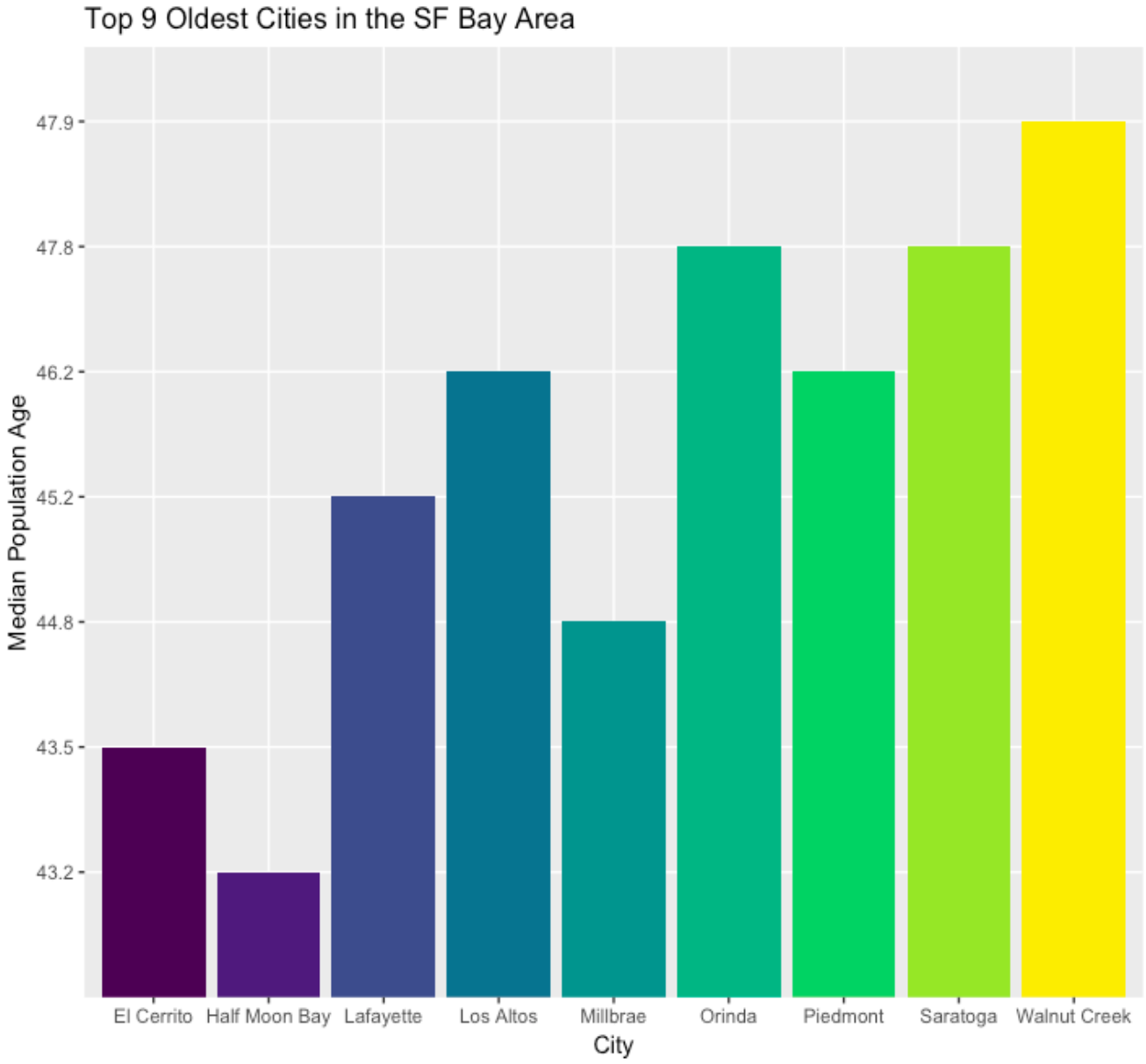

Apartments for Rent Southern SF Bay Area
Vallejo
Rodeo
San Anselmo
El Sobrante
Concord
Clayton
Discovery Bay
Mill Valley
Tiburon
Oakland
Hayward
Hillsborough
Fremont
El Granada
Half Moon Bay
Palo Alto
Los Altos Hills
San Jose
San Martin
Scotts Valley
Santa Cruz
Interlaken
38.1
37.8
37.5
37.2
-122.8
-122.4
-122.0
-121.6
Latitude
Longitude
Bedrooms
4
5
6


Location of Top 200 Largest SF Apartments (1600+ sqft)
Vallejo
Rodeo
San Anselmo
El Sobrante
Concord
Clayton
Discovery Bay
Mill Valley
Tiburon
Oakland
Hayward
Hillsborough
Fremont
El Granada
Half Moon Bay
Palo Alto
Los Altos Hills
San Jose
San Martin
Scotts Valley
Santa Cruz
Interlaken
38.1
37.8
37.5
37.2
-122.8
-122.4
-122.0
-121.6
Latitude
Longitude

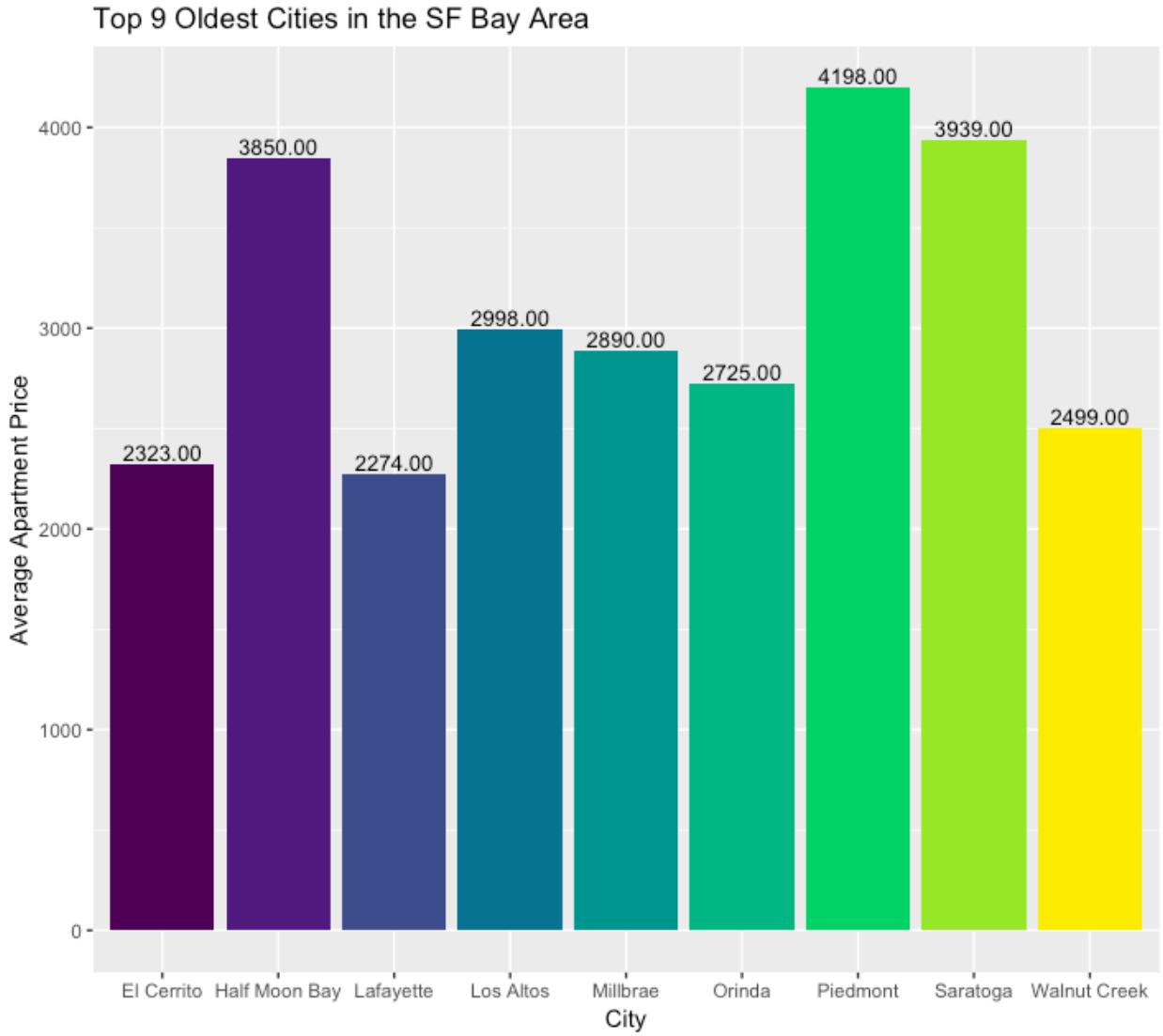
Oldest populations in the Southern San Francisco Bay Area

I decided to answer the question of which cities in the SF Bay Area contained the oldest populations, and how it related to the rental market.

Using the data from the 2010 US Census, I extracted the cities with the highest median age for their total population. These cities were El Cerrito, Half Moon Bay, Lafayette, Los Altos, Millbrae, Orinda, Piedmont, Saratoga, and Walnut Creek.



I was curious about whether higher average age of inhabitants of these cities would affect the prices of apartments for rent. I generated the following bar graph examining average rental price for each city.




In my previous analysis, I looked at the distribution of rent price across the entire Bay Area, separating rent price into five discrete, evenly spaced categories: ($0-$2180, $2180-$2540, $2540 - $2950, $2950-$3680, and $3680). Seven out of the top nine cities with the oldest populations also fell into the top three categories for highest apartment price range. These cities are Half Moon Bay, Los Altos, Millbrae, Orinda, Piedmont, Saratoga, and Walnut Creek. It appears there is a modest correlation between cities with older populations and higher rent prices. However, this could be a result simply of the location in the Bay Area of these cities, given that I have already demonstrated that higher rent prices are found in San Francisco, the Peninsula and Oakland. Many of these cities are located in those regions.

Sources:
https://stackoverflow.com/questions/21812625/ggmap-package-plot-filled-polygon-on-map

Code:

```
#Assignment 4
#Mira Mastoras


install.packages("devtools")
devtools::install_github("dkahle/ggmap", ref = "tidyup")
library(ggmap)
cl = readRDS("~/sta141a/cl_apartments (1).rds")
colnames(cl)
# Craigslist in Davis

davis_apts = cl[cl$place == "Davis" | cl$city == "Davis",]

# MAPPING all the apartments in davis - only showed a handful of points
davis_bbox = c(-121.799786,38.516581,-121.681567,38.575592)
m = get_stamenmap(davis_bbox, zoom = 15, maptype = "toner-lite")
ggmap(m)

ggmap(m) + geom_point(aes(longitude, latitude), davis_apts)

# only 64 apartments in davis have latitude & longitude values
nrow(davis_apts)

shadow1 = is.na(davis_apts$latitude)
table(shadow1)

shadow2 = is.na(davis_apts$longitude)
table(shadow2)

# only 20 apartments in davis are "mappable" - they have long & lat coordinate data available
mappable_davis = davis_apts[!is.na(davis_apts$latitude) & !is.na(davis_apts$longitude),]
ncol(mappable_davis)

# Map of Davis apartments by bedroom number
ggmap(m) + geom_point(aes(x = longitude, y = latitude, color = as.character(bedrooms)),
mappable_davis, size = 3,) +
```

```
  scale_color_viridis_d() +
  labs(title = "Apartments for Rent in Davis, CA by Bedroom Number", x = "Longitude", y =
"Latitude", color = "Bedroom Number")

# looking at distribution of bedrooms in all of davis vs the ones that can be mapped
ggplot(davis_apts, aes(x = bedrooms) ) + geom_bar() + labs( title = "All Davis Apartments")

ggplot(mappable_davis, aes(x= bedrooms)) + geom_bar()

# map of davis apts by parking type

ggmap(m) + geom_point(aes(x = longitude, y = latitude, color = parking), mappable_davis, size
= 2) +
  scale_color_hue() +
  labs(title = "Apartments for Rent in Davis, CA", x = "Longitude", y = "Latitude", color = "Parking
Type")

colnames(mappable_davis)

# pets in davis
ggmap(m) + geom_point(aes(x = longitude, y = latitude, color = pets, shape = pets),
mappable_davis, size = 2) +
  scale_color_hue() +
  labs(title = "Apartments for Rent in Davis, CA", x = "Longitude", y = "Latitude", color = "Pets")

# davis apartments by prices
install.packages("arules")
library(arules)

discrete_price = discretize(mappable_davis$price)
as.data.frame(discrete_price)

levels(discrete_price) = c("$721 - $1440", "$1440 - $1680", "$1680 - $2700")

mappable_davis$discrete_price = discrete_price

ggmap(m) + geom_point(aes(x = longitude, y = latitude, color = discrete_price),
mappable_davis[!is.na(mappable_davis$discrete_price),], size = 3) +
  scale_color_hue() +
  labs(title = "Price of Apartments for Rent in Davis, CA", x = "Longitude", y = "Latitude", color =
"Rent Price") +
  scale_color_viridis_d()

# davis apartments by square feet
```

```
colnames(mappable_davis)

discrete_sqft = discretize(mappable_davis$sqft)
levels(discrete_sqft) = c("143 - 800", "800 - 952", "952 - 1350")


campus_coord = campus_coord[-2,]
campus_coord$campus_lat = "38.546168"
campus_coord
mappable_davis$latitude

colnames(new_mappable_davis)

campus_coord = as.data.frame(cbind(campus_long, campus_lat))
campus_coord = data.frame(y=c(38.528439,38.546168), x=c(-121.768034,-121.745449))

mappable_davis$discrete_sqft = discrete_sqft
ggmap(m) + geom_point(aes(x = longitude, y = latitude, color = discrete_sqft),
mappable_davis[!is.na(mappable_davis$discrete_sqft),], size = 3) +
  scale_color_hue() +
  labs(title = "Square Footage of Apartments for Rent in Davis, CA", x = "Longitude", y =
"Latitude", color = "Square Footage") +
  scale_color_viridis_d()
 # geom_polygon(mapping = aes(x = x, y = y, fill = "red", campus_coord))

# attempt to draw a polygon around davis campus
colnames(mappable_davis)

new_df =
data.frame("title","text","latitude","longitude","city_text","date_posted","date_updated","price",

"deleted","sqft","bedrooms","bathrooms","pets","laundry","parking","craigslist","place","city",
"state", "county","discrete_price")
colnames(new_df) =
c("title","text","latitude","longitude","city_text","date_posted","date_updated","price",

"deleted","sqft","bedrooms","bathrooms","pets","laundry","parking","craigslist","place","city",
"state", "county","discrete_price")
new_df[1,] = NA
new_df$longitude = "-121.745449"
new_df$latitude = "38.546168"

new_mappable_davis = rbind(mappable_davis, new_df)
```

```
new_mappable_davis

nrow(new_mappable_davis)

new_mappable_davis[65,]


colnames((mappable_davis))

# San Francisco Bay Area

#subset by SF

sf_apts = cl[cl$county == "San Francisco" | cl$county == "San Mateo" | cl$county == "Santa
Clara" | cl$county == "Alameda" | cl$county == "Contra Costa", ]

sf_bbox = c(-122.801,36.9329,-121.3983,38.1325)
sf_map = get_stamenmap(sf_bbox, zoom = 9, maptype = "toner-lite")
ggmap(sf_map)

# all apartments in SF
ggmap(sf_map) + geom_point(aes(longitude, latitude), sf_apts)

mappable_sf = sf_apts[!is.na(sf_apts$latitude) & !is.na(sf_apts$longitude),]
nrow(mappable_sf)

# apartments in entire SF area by price

discrete_sf_price = discretize(mappable_sf$price, breaks = 5, na.rm = TRUE)
levels(discrete_sf_price) = c("$0 - $2180", "2180 - $2540", "$2540 - $2950", "$2950 - $3680",
"$3680+ ")
mappable_sf$discrete_price = discrete_sf_price

ggmap(sf_map) + geom_point(aes(x = longitude, y = latitude, color =
discrete_price),mappable_sf[!is.na(mappable_sf$price),], size = 2) +
  scale_color_hue() +
  labs(title = "Apartments for Rent Southern SF Bay Area", x = "Longitude", y = "Latitude", color
= "Price") +
  scale_color_viridis_d()

ggplot(mappable_sf[!is.na(mappable_sf$price),], aes(x = county, fill = discrete_price)) +
  geom_bar(position = "dodge") +
  scale_color_viridis_d() +
```

```
  labs(title = "Rent Prices in Southern SF Bay Area by County", x = "County", y = "Count", fill =
"Rent Price") +
  scale_fill_viridis_d()


# price of apartments in just san francisco county

sf_county_bbox = c(-122.541239,37.693278,-122.343899,37.82538)
sf_county_map = get_stamenmap(sf_county_bbox, zoom = 12, maptype = "toner-lite")

ggmap(sf_county_map)

sf_county_apts = cl[cl$county == "San Francisco" ,]
mappable_sf_county = sf_county_apts[!is.na(sf_apts$latitude) & !is.na(sf_apts$longitude),]
nrow(mappable_sf_county)
discrete_sf_county_price = discretize(mappable_sf_county$price, breaks = 5, na.rm = TRUE)
levels(discrete_sf_county_price) = c("$0 - $2500", "$2500 - $3100", "$3100 - $3800", "$3800 -
$4690", "$4690+")
mappable_sf_county$discrete_price = discrete_sf_county_price


ggmap(sf_county_map) + geom_point(aes(x = longitude, y = latitude, color = discrete_price),
mappable_sf_county[!is.na(mappable_sf_county$price),], size = 1.5) +
  scale_color_hue() +
  labs(title = "Apartments for Rent in San Francisco County", x = "Longitude", y = "Latitude",
color = "Price") +
  scale_color_viridis_d()

#downtown SF district

downtown_bbox = c(-122.470211,37.739831,-122.373683,37.81526)
downtown_map = get_stamenmap(downtown_bbox, zoom = 15, maptype = "toner-lite")

ggmap(downtown_map) + geom_point(aes(x = longitude, y = latitude, color = discrete_price),
mappable_sf_county[!is.na(mappable_sf_county$price),], size = 1.5) +
  scale_color_hue() +
  labs(title = "Apartments for Rent in Downtown San Francisco", x = "Longitude", y = "Latitude",
color = "Price") +
  scale_color_viridis_d()

colnames(sf_apts)

# where are the super expensive places located from SF? answer was all over
expensive_sf = mappable_sf[mappable_sf$discrete_price == "$3680+ ",]
```

```
expensive_map
expensive_map = get_stamenmap(bbox = c(-122.5572,37.1794,-121.7092,37.948), zoom = 10,
maptype = "toner-lite")
ggmap(expensive_map) + geom_point(aes(x = longitude, y = latitude, color =
discrete_price),expensive_sf[!is.na(expensive_sf$discrete_price),], size = 1.5, alpha = 0.5) +
  scale_color_hue() +
  labs(title = "Most Expensive Apartments in Southern SF Bay Area", x = "Longitude", y =
"Latitude", color = "Price") +
  scale_color_viridis_d()


# Where are the 4, 5 and 6 bedroom apartments located?

table(sf_apts$bedrooms)

many_beds = sf_apts[sf_apts$bedrooms == "4" | sf_apts$bedrooms == "5" | sf_apts$bedrooms
== "6",]

ggmap(sf_map) + geom_point(aes(x = longitude, y = latitude, color =
as.character(bedrooms)),many_beds[!is.na(many_beds$bedrooms),], size = 3, alpha = 0.5) +
  scale_color_hue() +
  labs(title = "Apartments for Rent Southern SF Bay Area", x = "Longitude", y = "Latitude", color
= "Bedrooms") +
  scale_color_viridis_d()

# proportion of SF apartments by county

ggplot(sf_apts[!is.na(sf_apts$county),], aes(x = county, fill = county)) +
  geom_bar() +
  scale_fill_viridis_d() +
  labs(title = "Apartments for Rent in Each SF Bay County", x = "County", y = "Count") +
  guides(fill = FALSE)

#huge apartments

discrete_sf_sqft = discretize(sf_apts$sqft, breaks = 40)

sf_apts$discrete_sqft = discrete_sf_sqft
table(sf_apts$discrete_sqft)[40]
levels(sf_apts$discrete_sqft)[40] = "1600 - 8900"

large_apts = sf_apts[sf_apts$discrete_sqft == "1600 - 8900",]
table(sf_apts$discrete_sqft)
```

```
ggmap(sf_map) + geom_point(aes(x = longitude, y = latitude, color =
discrete_sqft),large_apts[!is.na(large_apts$discrete_sqft),], size = 2, alpha = 0.5) +
  scale_color_hue() +
  labs(title = "Location of Top 200 Largest SF Apartments (1600+ sqft)", x = "Longitude", y =
"Latitude", color = "Square feet") +
  scale_color_viridis_d() +
  guides(color = FALSE)


# Question 3

census_data = read.csv("~/sta141a/2010_census_data/DEC_10_SF1_SF1DP1_with_ann.csv",
stringsAsFactors = FALSE)
metadata =read.csv("~/sta141a/2010_census_data/DEC_10_SF1_SF1DP1_metadata.csv",
stringsAsFactors = FALSE, header = FALSE)

colnames(census_data)
head(metadata)
colnames(sf_apts)
colnames(census_data) = metadata$V2

colnames(census_data)
nrow(census_data)

install.packages("stringr")
library(stringr)

census_data$Geography = str_remove_all(census_data$Geography, fixed(", California"))
census_data$Geography = str_remove_all(census_data$Geography, fixed(" CDP"))
census_data$Geography = str_remove_all(census_data$Geography, fixed(" city"))

census_data$Geography
head(census_data)
sf_census = merge(census_data, sf_apts, by.y = "city", by.x = "Geography")

column_names = colnames(sf_census)

column_names
head(sf_census)

nrow(sf_census)

sf_census$Geography
```

```
# which places in SF have the oldest populations?

table(sf_census$`Number; SEX AND AGE - Total population - Under 5
years`,sf_census$Geography)

#rename column with average age for sf cities
colnames(sf_census)[42] = "total_pop_med_age_num"


sf_census[42]
sf_census$total_pop_med_age_num = as.numeric(sf_census$total_pop_med_age_num)

typeof(sf_census$total_pop_med_age_num)

discrete_med_age = discretize(sf_census$total_pop_med_age_num, breaks = 4)
as.data.frame(discrete_med_age)

levels(discrete_med_age) = c("")

sf_census$discrete_med_age = discrete_med_age

# apartments with their median location mapped - not super meaningful
ggmap(sf_map) + geom_point(aes(x = longitude, y = latitude, color =
discrete_med_age),sf_census, size = 1.5, alpha = 0.5) +
  labs(title = "Apartments for Rent in Southern SF Bay Area", x = "Longitude", y = "Latitude",
color = "Median Population Age") +
  scale_color_viridis_d()

colnames(census_data)

nrow(census_data)
bay_census_data = subset(census_data, Geography %in% sf_census$Geography)

old_cities = bay_census_data[bay_census_data$total_pop_med_age_num > 40,]

ggmap(sf_map) + geom_point(aes(x = longitude, y = latitude, color =
total_pop_med_age_num),bay_census_data, size = 1.5, alpha = 0.5) +
  labs(title = "Apartments for Rent in Southern SF Bay Area", x = "Longitude", y = "Latitude",
color = "Median Population Age") +
  scale_color_viridis_d()

# subsetting data by the oldest populations
colnames(bay_census_data)[42] = "total_pop_med_age_num"
nrow(bay_census_data)
```

```
colnames(bay_census_data)
table(bay_census_data$total_pop_med_age_num)

#histogram of median age of Bay Area populations - not meaningful
ggplot(bay_census_data, aes(x = as.numeric(total_pop_med_age_num))) +
  geom_density()

old_cities = bay_census_data[bay_census_data$total_pop_med_age_num > 43,]
old_cities_apts = subset(sf_census, Geography %in% old_cities$Geography)
table(old_cities$Geography)

table(old_cities_apts$Geography, old_cities_apts$total_pop_med_age_num)

#graph of top oldest cities in bay
ggplot(old_cities, aes(x = Geography, y = total_pop_med_age_num, fill = Geography)) +
  geom_bar(stat = "identity") +
  labs(title = "Top 9 Oldest Cities in the SF Bay Area", x = "City", y = "Median Population Age")+
  scale_fill_viridis_d() +
  guides(fill = FALSE)

# graph of apartments in oldest areas of the bay
ggmap(sf_map) + geom_point(aes(x = longitude, y = latitude, color =
as.character(total_pop_med_age_num)),old_cities_apts, size = 3, label = sprintf("%0.2f",
round(total_pop_med_age_num)))+
  labs(title = "Apartments for Rent in the Oldest Cities in the Bay Area", x = "Longitude", y =
"Latitude", color = "Median Population Age") +
  scale_color_viridis_d() +
  geom_text(position=position_dodge(width=0.9), vjust=-0.25, hjust = 0.5, size = 3.5)

#do these areas have higher prices? less apartments available? less family friendly?
sf_apts$price
#price of these cities
avg_price_old_cities = aggregate(price ~ Geography, old_cities_apts, mean)
avg_price_old_cities = as.data.frame(avg_price_old_cities)
mean(sf_apts$price, na.rm = TRUE)

ggplot(avg_price_old_cities, aes(x = Geography, y = price, fill = Geography, label =
sprintf("%0.2f", round(price)))) +
  geom_bar(stat = "identity") +
  labs(title = "Top 9 Oldest Cities in the SF Bay Area", x = "City", y = "Average Apartment
Price")+
  scale_fill_viridis_d() +
  guides(fill = FALSE) +
  geom_text(position=position_dodge(width=0.9), vjust=-0.25, hjust = 0.5, size = 3.5)
```

```
# are there less apartments in these cities?

#distribution of apartment number by city

num_apts_per_city = table(sf_apts$city)
num_apts_per_city = as.data.frame(num_apts_per_city)

ggplot(num_apts_per_city, aes(x = Var1, y = Freq )) + geom_point()

head(num_apts_per_city)
```