Assignment 6

Janice Luong
STA 141B
March 19, 2017

In this assignment, you'll analyze a collection of data sets from the San Francisco Open Data Portal (http://data.sfgov.org/) and Zillow (https://www.zillow.com/). The data sets have been stored in the SQLite database `sf_data.sqlite`, which you can download here (http://anson.ucdavis.edu/~nulle/sf_data.sqlite). The database contains the following tables:

Table	Description
`crime`	Crime reports dating back to 2010.
`mobile_food_locations`	List of all locations where mobile food vendors sell.
`mobile_food_permits`	List of all mobile food vendor permits. More details here (https://data.sfgov.org/api/views/rqzj-sfat/files/8g2f5RV4PEk0_b24iJEtgEet9gnh_eA27GlqoOjjK4k?download=true&filename=DPW_DataDictionary_Mobile-Food-Facility-Permit.pdf).
`mobile_food_schedule`	Schedules for mobile food vendors.
`noise`	Noise complaints dating back to August 2015.
`parking`	List of all parking lots.
`parks`	List of all parks.
`schools`	List of all schools.
`zillow`	Zillow rent and housing statistics dating back to 1996. More details here (https://www.zillow.com/research/data/).

The `mobile_food_` tables are explicitly connected through the `locationid` and `permit` columns. The other tables are not connected, but you may be able to connect them using dates, latitude/longitude, or postal codes.

Shapefiles for US postal codes are available here (https://www.census.gov/geo/maps-data/data/cbf/cbf_zcta.html). These may be useful for converting latitude/longitude to postal codes.

Shapefiles for San Francisco Neighborhoods are available here (https://data.sfgov.org/Geographic-Locations-and-Boundaries/SF-Find-Neighborhoods/pty2-tcw4).

I worked with Richard Safran, Chad Pickering and Edie Espejo.

Exercise 1.1. Which mobile food vendor(s) sells at the most locations?

In [1]:
```
import sqlite3
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import shapely.geometry as geom
import geopandas as gpd
import matplotlib.patches as mpatches
import warnings
warnings.filterwarnings('ignore')

# Make Jupyter to automatically display matplotlib plots.
%matplotlib inline

from mpl_toolkits.basemap import Basemap
import numpy as np
from math import log
```

In [2]:
```
# Reads database
db = sqlite3.connect("sf_data.sqlite")
```

In [3]:
```
pd.read_sql("SELECT * FROM sqlite_master", db)
```

Out[3]:

	type	name	tbl_name	rootpage	sql
0	table	crime	crime	2	CREATE TABLE "crime" (\n"IncidntNum" INTEGER,\...
1	table	noise	noise	35775	CREATE TABLE "noise" (\n"CaseID" INTEGER,\n "...
2	table	parking	parking	35921	CREATE TABLE "parking" (\n"Owner" TEXT,\n "Ad...
3	table	schools	schools	35944	CREATE TABLE "schools" (\n"Name" TEXT,\n "Ent...
4	table	parks	parks	35961	CREATE TABLE "parks" (\n"Name" TEXT,\n "Type"...
5	table	zillow	zillow	35967	CREATE TABLE "zillow" (\n"RegionName" INTEGER,...
6	table	mobile_food_permits	mobile_food_permits	36050	CREATE TABLE "mobile_food_permits" (\n"permit"...
7	table	mobile_food_locations	mobile_food_locations	36060	CREATE TABLE "mobile_food_locations" (\n"locat...
8	table	mobile_food_schedule	mobile_food_schedule	36079	CREATE TABLE "mobile_food_schedule" (\n"locati...

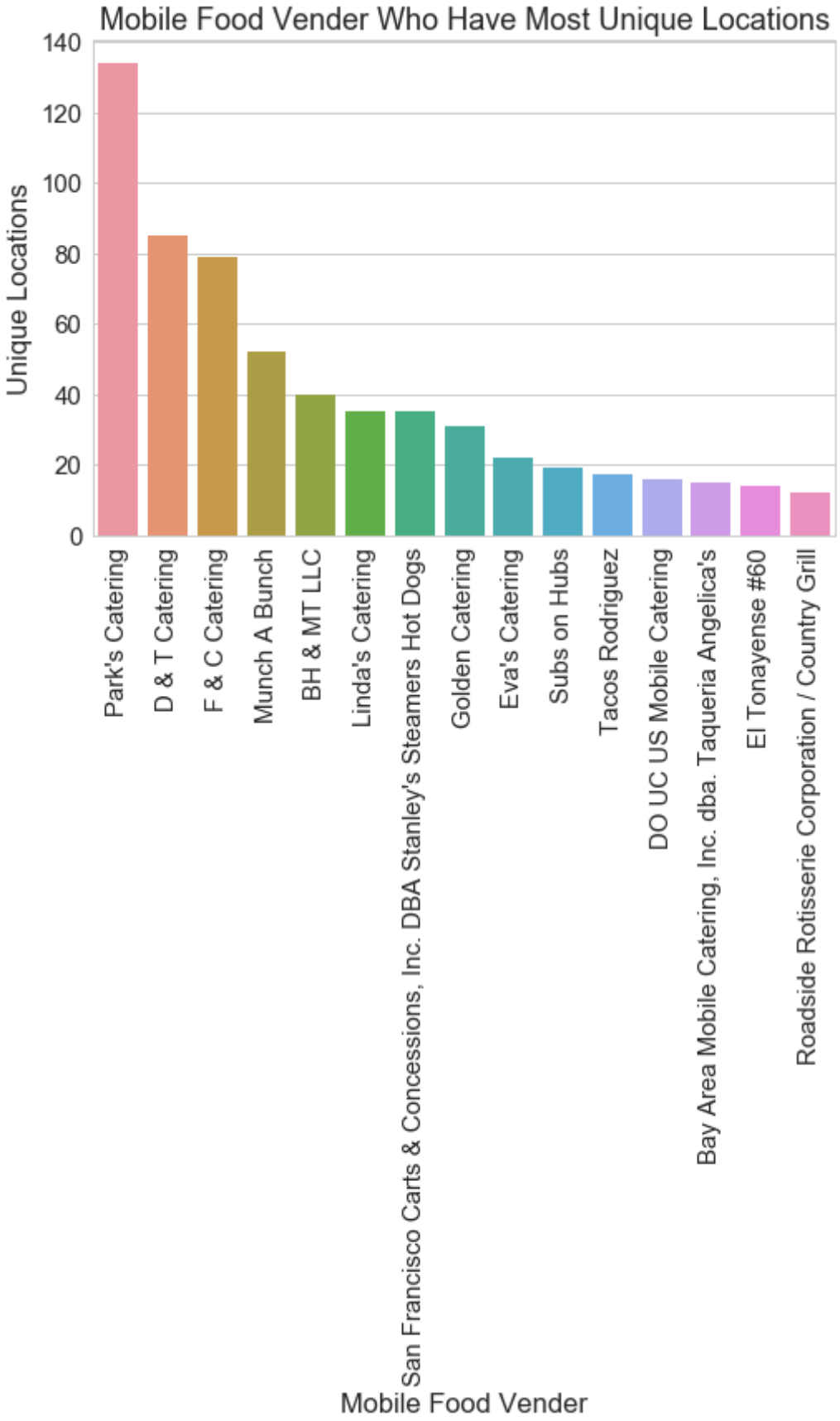
In [129]:
```
most_location_mobile = pd.read_sql("SELECT p.Applicant, \
                                    COUNT(s.locationid) as 'Unique Locations',
 Status \
                                    FROM mobile_food_permits p \
                                    INNER JOIN mobile_food_schedule s ON p.perm
it = s.permit \
                                    WHERE p.Status = 'APPROVED' \
                                    GROUP BY Applicant \
                                    ORDER BY COUNT(s.locationid) \
                                    DESC \
                                    LIMIT 15", db)
most_location_mobile
```

Out[129]:

	Applicant	Unique Locations	Status
0	Park's Catering	134	APPROVED
1	D & T Catering	85	APPROVED
2	F & C Catering	79	APPROVED
3	Munch A Bunch	52	APPROVED
4	BH & MT LLC	40	APPROVED
5	Linda's Catering	35	APPROVED
6	San Francisco Carts & Concessions, Inc. DBA St...	35	APPROVED
7	Golden Catering	31	APPROVED
8	Eva's Catering	22	APPROVED
9	Subs on Hubs	19	APPROVED
10	Tacos Rodriguez	17	APPROVED
11	DO UC US Mobile Catering	16	APPROVED
12	Bay Area Mobile Catering, Inc. dba. Taqueria A...	15	APPROVED
13	El Tonayense #60	14	APPROVED
14	Roadside Rotisserie Corporation / Country Grill	12	APPROVED

For my querie, I did an inner join on the tables mobile_food_permits and mobile_food_schedule and I counted the locationid's while also making sure that the Applicant recieved and APPROVED for their Status because if they were note approved, I do not want to count it since if they are not approved, they will not be at that location. This gives me the unique counts of each mobile food vender in case they have two of the same venders in one location.

In [130]:
```
# Set the size of plot or else its too small to read
plt.rcParams['figure.figsize'] = (10, 10)
sns.set(font_scale = 1.5)

sns.set_style("whitegrid")
mobile_locat = sns.barplot(x = "Applicant", y = "Unique Locations", data = mos
t_location_mobile)
for item in mobile_locat.get_xticklabels():
    item.set_rotation(90)
sns.plt.title('Mobile Food Vender Who Have Most Unique Locations')
mobile_locat.set(ylabel = 'Unique Locations', xlabel = 'Mobile Food Vender')
sns.plt.show()
```



The food venders that sell at the most locations are Park's Catering, D & T Catering, F & C Catering, Munch A Bunch, BH & MT LLC, Linda's Catering, and San Francisco Carts & Concessions, Inc.

Exercise 1.2. Ask and use the database to analyze 5 questions about San Francisco. For each question, write at least 150 words and support your answer with plots. Make a map for at least 2 of the 5 questions.

You should try to come up with some questions on your own, but these are examples of reasonable questions:

- Which parts of the city are the most and least expensive?
- Which parts of the city are the most dangerous (and at what times)?
- Are noise complaints and mobile food vendors related?
- What are the best times and places to find food trucks?
- Is there a relationship between housing prices and any of the other tables?

Please make sure to clearly state each of your questions in your submission.

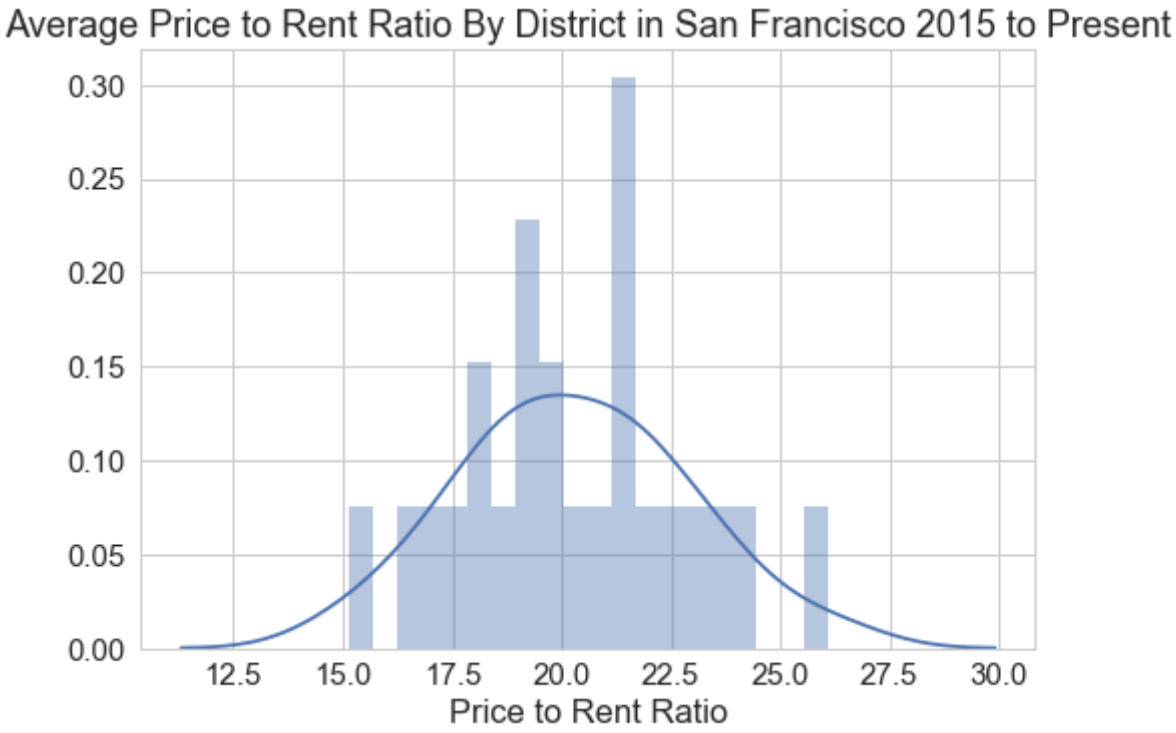
1. Which parts of the city are the most and least expensive?

In [131]:
```
ratio_rent_zip = pd.read_sql("SELECT AVG(PriceToRentRatio_AllHomes) as 'Price
 to Rent Ratio', RegionName as 'Zipcode' \
                              FROM zillow \
                              WHERE PriceToRentRatio_AllHomes IS NOT NULL \
                              GROUP BY RegionName \
                              ORDER BY AVG(PriceToRentRatio_AllHomes) \
                              DESC", db)
ratio_rent_zip.head()
```

Out[131]:

	Price to Rent Ratio	Zipcode
0	26.086711	94123
1	23.930658	94114
2	23.869211	94127
3	22.997500	94121
4	22.725921	94131

In [132]:
```
sns.distplot(ratio_rent_zip['Price to Rent Ratio'], bins=20)
sns.plt.title('Average Price to Rent Ratio By District in San Francisco 2015 t
o Present')
plt.show()
```



For my querie, I selected the PriceToRentRation_AllHomes column because the PriceToRentRation_AllHomes is calculated as the ratio of home prices to annual rental rates. So, for example, in a real estate market where, on average, a home worth $200,000 could rent for $1000 a month, the price-rent ratio is 16.67. That's determined using the formula: $\$200,000 \div (12 \times \$1,000)$. So this column would give me a general idea how expensive or cheap an area in San Francisco are because the more a home is worth, the home's price-rent ratio would be higher.

For each zipcode, I took the average of the price to rent ratio so I could get an idea of how high or low the rent in San Francisco is. Since the price to rent ratio is derived from the house's real estate market, we know that if the price to rent ratio is higher for certain areas, then that means the house has a higher value. And vice versa for areas with a low price to rent ratio.

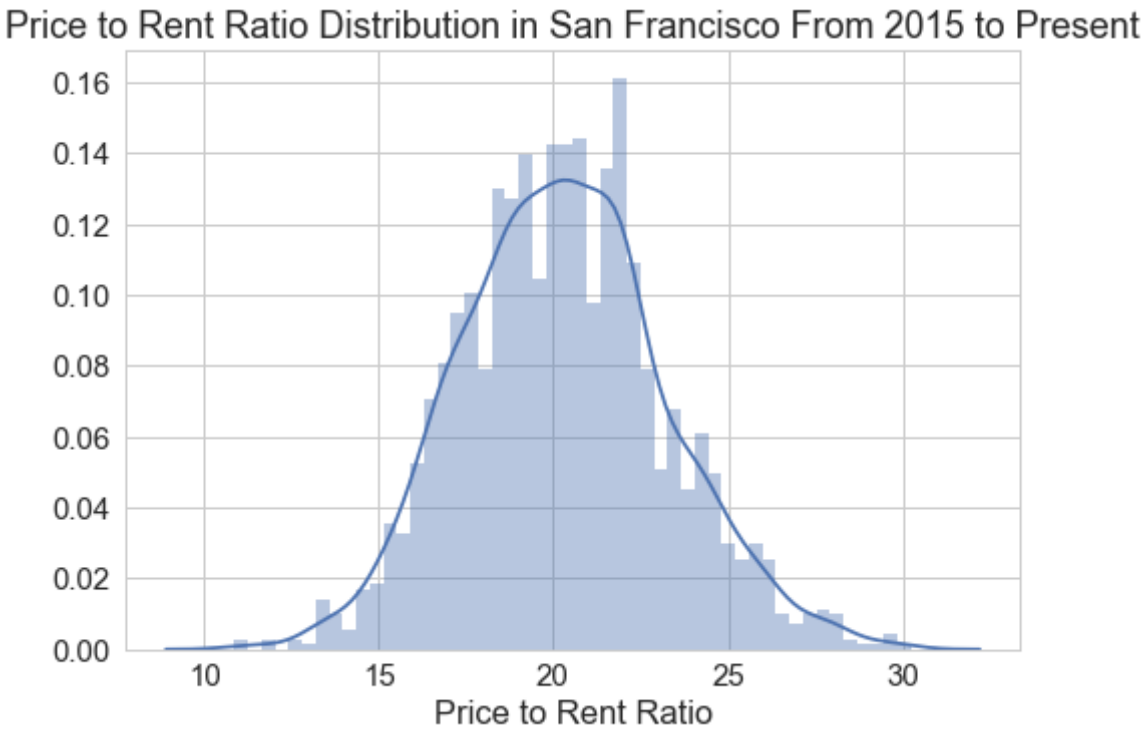
The most expensive areas are the Marina District, the Castro, North Beach, and Pacific Heights, which have a price range of $3,200 to $4,000. The least expensive areas are Hunders Point, Visitation Valley, Excelsior, Outter Mission, Oceanview, Crocker-Amazon and Ingleside, which have a price range of $1,900 to $2,600.

However, on issue I did notice with the price-rent ratio is that areas such as Soma, Financial District and South Beach do not appear at the top of the list for most expensive areas to live in San Francisco because most of the residents there live in apartments and apartments are generally cheaper to rent than compared to rending a house in the Marina District. However the areas such as the Marina the Castro, North Beach, and Pacific Heights are still areas of San Francisco that are more expensive to rent.

In [133]:
```
rent_zip = pd.read_sql("SELECT PriceToRentRatio_AllHomes as 'Price to Rent Rat
io', RegionName as 'Zipcode' \
                       FROM zillow \
                       WHERE PriceToRentRatio_AllHomes IS NOT NULL", db)
```

There were some Price to rent ratios that were empty, I removed those because they would not add anything meaningful to my histogram.

In [134]:
```
sns.distplot(rent_zip['Price to Rent Ratio'], bins=50)
sns.plt.title('Price to Rent Ratio Distribution in San Francisco From 2015 to
 Present')
plt.show()
```



The histogram plot above of the price to rent ratio is not centered around 0. If it was a normal distribution, it would be skewed left. Therefore, most places in San Francisco generally is an expensive area to rent/live in. Most of the prices are between a price to rent ratio of $15 to $25 a month because ~68% of the price to rent ratio are within 1 standard deviation from the mean $20. Overall, San Francisco is not a cheap place to rent. If there are cheap places to rent, the neighborhood is generally not a safe palce to live (Hunter's Point, Visitacion Valley, etc.)

2. Which parts of the city are the most dangerous (and at what times)?

In [135]:
```
pd.read_sql("SELECT Lon, Lat, PdDistrict, strftime('%H', Datetime) as 'Hour of
 the Day', COUNT(*) as 'PdDistCount' \
            FROM crime \
            WHERE PdDistrict IS NOT NULL \
            GROUP BY PdDistrict, Datetime \
            ORDER BY COUNT(*) \
            DESC \
            LIMIT 10"
                  , db)
```

Out[135]:

	Lon	Lat	PdDistrict	Hour of the Day	PdDistCount
0	-122.402672	37.756423	BAYVIEW	00	48
1	-122.404286	37.796142	CENTRAL	11	35
2	-122.422063	37.789920	NORTHERN	12	34
3	-122.476057	37.762657	TARAVAL	12	34
4	-122.408761	37.715900	INGLESIDE	00	26
5	-122.401817	37.788441	SOUTHERN	12	26
6	-122.403405	37.775421	SOUTHERN	00	23
7	-122.397251	37.778980	SOUTHERN	00	22
8	-122.474308	37.716643	TARAVAL	12	22
9	-122.419059	37.759423	MISSION	00	21

For my query to get the most dangerous location and times of the city, I took the columns that contained the Police Department's District because if a particular Police Department received a lot of crimes, then this means the surrounding area of the Police Department has a lot of crimes because the Police Stations are located in the area that they are supposed to serve and patrol. So then I did a count of how many times a PdDistrict appears for each hour of the day. This will show us which neighborhoods have the most amount of crimes and between what hours the most crime occurs. I did it by hour because there is no real connection between what day of the week crimes occur. Most cries occur during hours where they know a house is empty or when they notice security is on break, which is why I was more interested in the hour than the date.

Most of the crime tends to happen at midnight or at noon. This may be because most people are drunk at night (hence 10pm to 12am). Also most crimes may occur during the afternoon because the suspect knows most places go off on lunch around then, so the victim (store or company) may have less security up due to them taking their lunch break. The parts of the city that are most dangerous are Bayview, Tenderloin, Mission Bay (Southern) and Ingleside. These are areas in San Francisco where then is generally cheaper, so there are high amounts of low-income families living there.

In [137]:
```
crime_location = pd.read_sql("SELECT Lon, Lat, PdDistrict, COUNT(*) as 'PdDist
Count' \
                              FROM crime \
                              WHERE PdDistrict IS NOT NULL \
                              GROUP BY PdDistrict, Lon \
                              ORDER BY COUNT(*) \
                              DESC \
                              LIMIT 2000"
                              , db)
crime_location.head()
```

Out[137]:

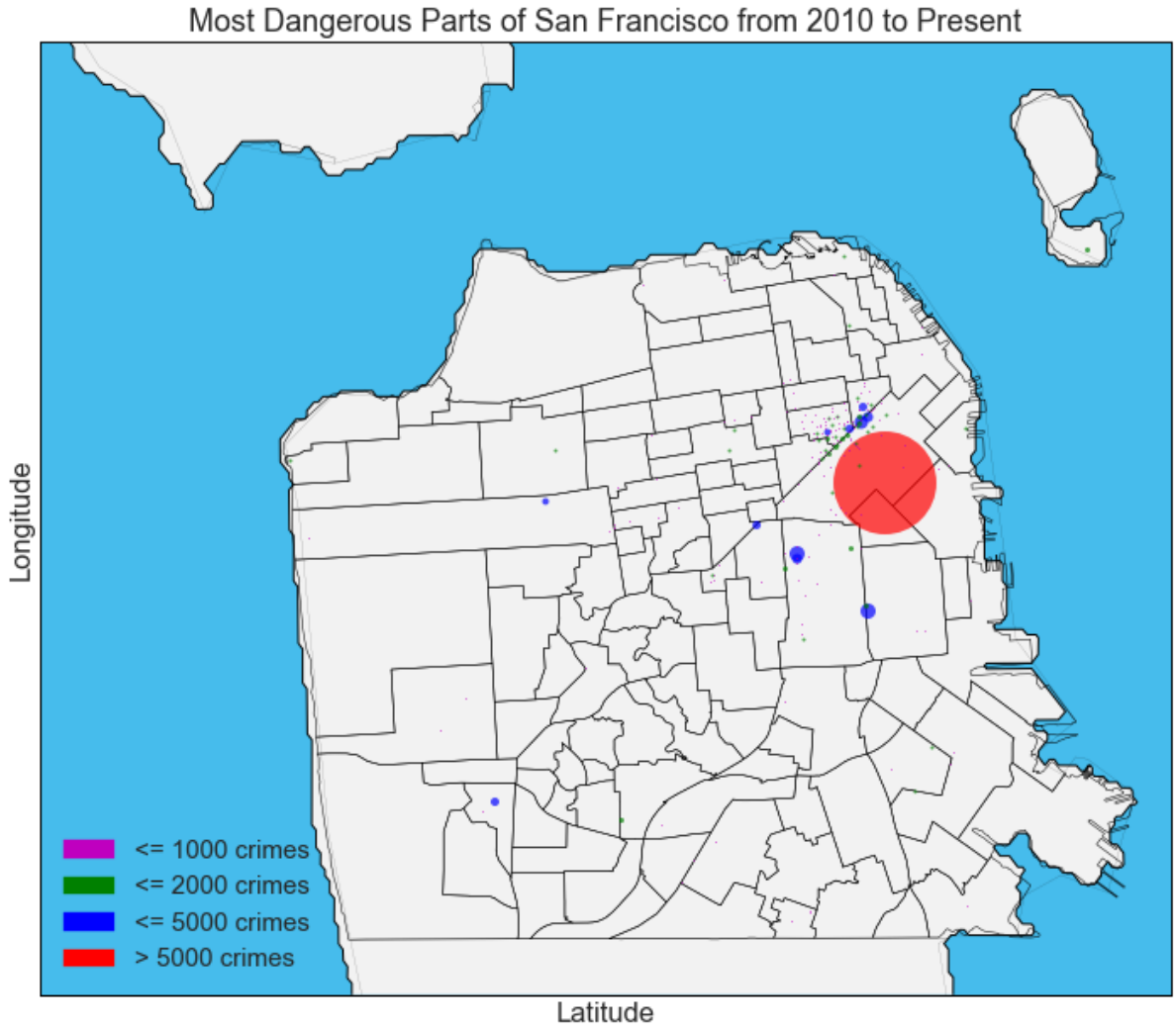
	Lon	Lat	PdDistrict	PdDistCount
0	-122.403405	37.775421	SOUTHERN	30673
1	-122.406539	37.756486	MISSION	4631
2	-122.419672	37.765050	MISSION	4580
3	-122.407634	37.784189	SOUTHERN	4410
4	-122.419658	37.764221	MISSION	3497

I will use this query to plot what parts of the city are most dangerous. I grouped by police district so I could get all the number of crimes that police station has gotten since 2010. Then I grouped by the longitide (could have also done latitude) so that way I can get the unique locations of where all crimes have happeneed. I limited to only 2000 rows because there are many crimes and if I were to plot all of them, it would have taken too long. I also notcied around row 1970, the number of crimes in the particular location are <100 so by that point, it would ve a very small dot on the map. Any smaller, it might not even be visiable on the map.

In [138]:
```
crime_location.columns = crime_location.columns.str.lower()
```

In [182]:
```
# Make plots larger.
plt.rcParams['figure.figsize'] = (12, 12)

#llcrnrlon lower left corner, urcrnrlon upper right corner
my_map = Basemap(llcrnrlon=-122.56, llcrnrlat=37.7, urcrnrlon=-122.35, urcrnrl
at=37.84, resolution="f", projection="merc")
my_map.drawcoastlines()
my_map.drawmapboundary(fill_color='#46bcec')
my_map.fillcontinents(color='#f2f2f2',lake_color='#46bcec')
my_map.drawcounties()

# read in the shape file, where sfneighborhoods is what I will name it
my_map.readshapefile("geo_export_8c430ac2-fab6-412a-8efe-6d7d4ad53402", "sfnei
ghborhoods")

for longi, lata, dist in zip(crime_location.lon, crime_location.lat, crime_loc
ation.pddistcount):
    x, y = my_map(longi, lata)
    dot_size = dist / 500
    # >5000 are red, 5000 - 2001 are blue, 2000 - 10001 are green, <1000 are y
ellow
    if dist <= 1000:
        color = "m"
    elif dist <= 2000:
        color = "g"
    elif dist <= 5000:
        color = "b"
    else:
        color = "r"
    # print x, y, dist, dot_size, color, alpha = transparentcy
    my_map.plot(x, y, color + "o", markersize = dot_size, alpha = 0.7)

m_patch = mpatches.Patch(color='m', label='<= 1000 crimes')
g_patch = mpatches.Patch(color='g', label='<= 2000 crimes')
b_patch = mpatches.Patch(color='b', label='<= 5000 crimes')
r_patch = mpatches.Patch(color='r', label='> 5000 crimes')
sns.plt.title('Most Dangerous Parts of San Francisco from 2010 to Present')
plt.legend(handles=[m_patch, g_patch, b_patch, r_patch], loc = 'lower left')
plt.xlabel("Latitude")
plt.ylabel("Longitude")
```

```
Out[182]: <matplotlib.text.Text at 0x1a2ccb70>
```



According the map, the largest amount of crimes are in the Mission Bay (Southern Police Department). Lots of small amounts of crime around the Tenderloin and Mission District (Downtown San Francisco). There are a few large blue dots, which are also part of Mission and Mission Bay of San Francisco. It is also interesting to note that all these areas are areas in San Francisco where rent is cheaper and have high amounts of low-income families living there.

3. Are noise complaints and mobile food vendors related?

In [4]:
```
# get only mobile_food_facility, entertainment, amplified_sound_electronics
noise_unique = pd.read_sql("SELECT Type \
                           FROM noise \
                           WHERE Lat IS NOT NULL \
                           AND Lon > -123 \
                           GROUP BY Type"
                           , db)

noise_complaints = pd.read_sql("SELECT Neighborhood, Lat, Lon, Type \
                               FROM noise \
                               WHERE Lat IS NOT NULL \
                               AND Lon > -123 \
                               AND (Type = 'mobile_food_facility' OR Type = 'e
ntertainment' OR Type = 'amplified_sound_electronics')"
                               , db)

noise_complaints.head()
```

Out[4]:

	Neighborhood	Lat	Lon	Type
0	Mission Dolores	37.769148	-122.424475	amplified_sound_electronics
1	Outer Richmond	37.774201	-122.484448	amplified_sound_electronics
2	South of Market	37.773292	-122.411780	amplified_sound_electronics
3	Lower Pacific Heights	37.785036	-122.441537	amplified_sound_electronics
4	South of Market	37.779203	-122.402444	amplified_sound_electronics

In [5]:
```
food_truck_location = pd.read_sql("SELECT l.LocationDescription, l.Latitude,
 l.Longitude, p.Status \
                                  FROM mobile_food_locations l \
                                  LEFT JOIN mobile_food_schedule s ON l.locati
onid = s.locationid \
                                  LEFT JOIN mobile_food_permits p ON s.permit
 = p.permit\
                                  WHERE l.Latitude IS NOT NULL \
                                  AND l.Latitude IS NOT 0.0 \
                                  AND p.Status = 'APPROVED'"
                                  , db)
food_truck_location.head()
```

Out[5]:

	LocationDescription	Latitude	Longitude	Status
0	TOWNSEND ST: 05TH ST to 06TH ST (400 - 499)	37.774871	-122.398532	APPROVED
1	TOWNSEND ST: 05TH ST to 06TH ST (400 - 499)	37.774871	-122.398532	APPROVED
2	TOWNSEND ST: 05TH ST to 06TH ST (400 - 499)	37.774871	-122.398532	APPROVED
3	TOWNSEND ST: 05TH ST to 06TH ST (400 - 499)	37.774871	-122.398532	APPROVED
4	TOWNSEND ST: 05TH ST to 06TH ST (400 - 499)	37.774871	-122.398532	APPROVED

For my query, I first selected the latitude and longitide columns from the tables noise and mobile_food_locations.

I grabbed only the columns where the latitude was not null because some latitude columns were empty. I also had to filter it so that the latitude did not contain the number 0.0 because a latitude with that coordinate is out side the boundaries of San Francisco or else it would being plotting points that were very far from the City and County of San Francisco.

For the noise complaints, I filtered them based on the "Type" that I believed were related to food venders. For example, I knew that complaints about construction or major event venue would not be related to food venders because construction is about buildings and vensue music are usually concerts. For the types of noise complaints, I picked mobile_food_facility, entertainment, amplified_sound_electronics. I picked ones that were related to music and entertainment because sometimes food venders do have music playing because they are at a local fair.

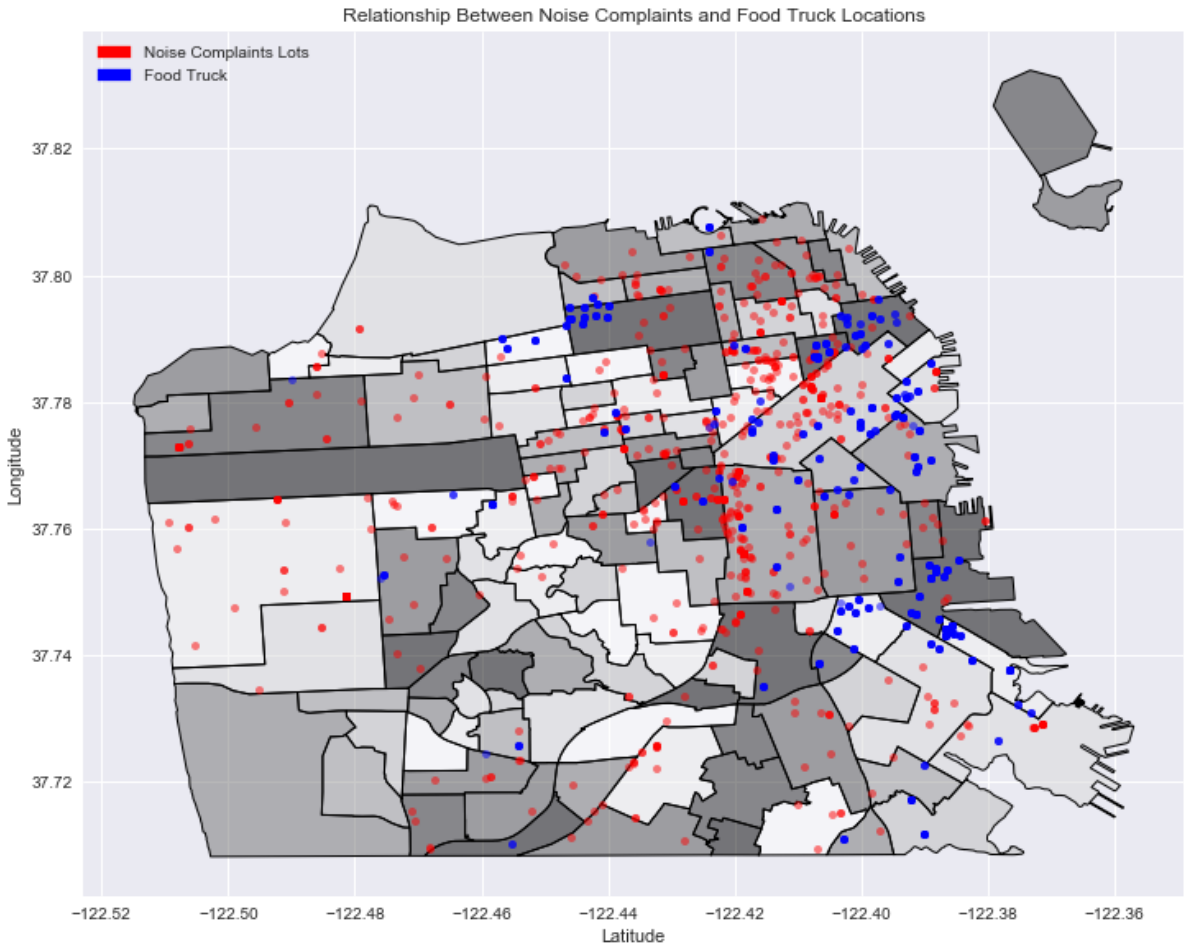
For the food truck locations, I only took the locations were the food truck had the Status of APPROVED because if they are approved then they will have permission from the City and Count of San Francisco to park their food truck and sell food. I chose not to pick the unique locations because that way when I plot the food truck location, if the dot is darker, that means there are many food trucks there (overlapping so color darkens). More food truck could mean more noise because more people will be going there to try out all the different food trucks. Of course the noise level of the area would increase if more people are gathered there.

In [6]:
```
noise_complaints.columns = noise_complaints.columns.str.lower()
food_truck_location.columns = food_truck_location.columns.str.lower()
```

In [7]:
```
zips = gpd.read_file("geo_export_8c430ac2-fab6-412a-8efe-6d7d4ad53402.shp")
```

In [8]:
```
plt.rcParams['figure.figsize'] = (12, 12)

lonlat = [geom.Point(lon, lat) for lon, lat in zip(noise_complaints.lon, noise
_complaints.lat)]
lonlat2 = [geom.Point(lon, lat) for lon, lat in zip(food_truck_location.longit
ude, food_truck_location.latitude)]

noise_complaints = gpd.GeoDataFrame(noise_complaints, geometry = lonlat, crs =
 {'init' :'epsg:4326'})
food_truck_location = gpd.GeoDataFrame(food_truck_location, geometry =
lonlat2, crs = {'init' :'epsg:4326'})
base = zips.plot()

noise_complaints.plot(ax = base, marker = "o", color = "red", markersize = 5,
alpha = 0.5)
food_truck_location.plot(ax = base, marker = "o", color = "blue", markersize =
 5, alpha = 0.5)
red_patch = mpatches.Patch(color='red', label='Noise Complaints Lots')
blue_patch = mpatches.Patch(color='blue', label='Food Truck')
sns.plt.title('Relationship Between Noise Complaints and Food Truck
Locations')
plt.legend(handles=[red_patch, blue_patch])
plt.xlabel("Latitude")
plt.ylabel("Longitude")
```

Out[8]: <matplotlib.text.Text at 0xc3c0400>



It appears that there is no relationship between food truck locations and noise because most of the noise complaints are not near a food truck location. However, I do notice that most of the noise complaints are around North Beach, Downtown and Market areas of San Francisco, but the complaints are not near a food truck location. Those are the areas where most of the night life are (bars, clubs, concerts). the noise complaints are not near food trucks possibly because most foods trucks are open during day time hours and not night time. A noise complaint in San Francisco is only vaild if it is past 11pm. Again, I noticed that most of the food trucks are also located in Downtown and Market areas of San Francisco.

4. What days of the week do crimes happen the most and what days of the week do crimes happen the least? Are these different across districts/neighborhoods?

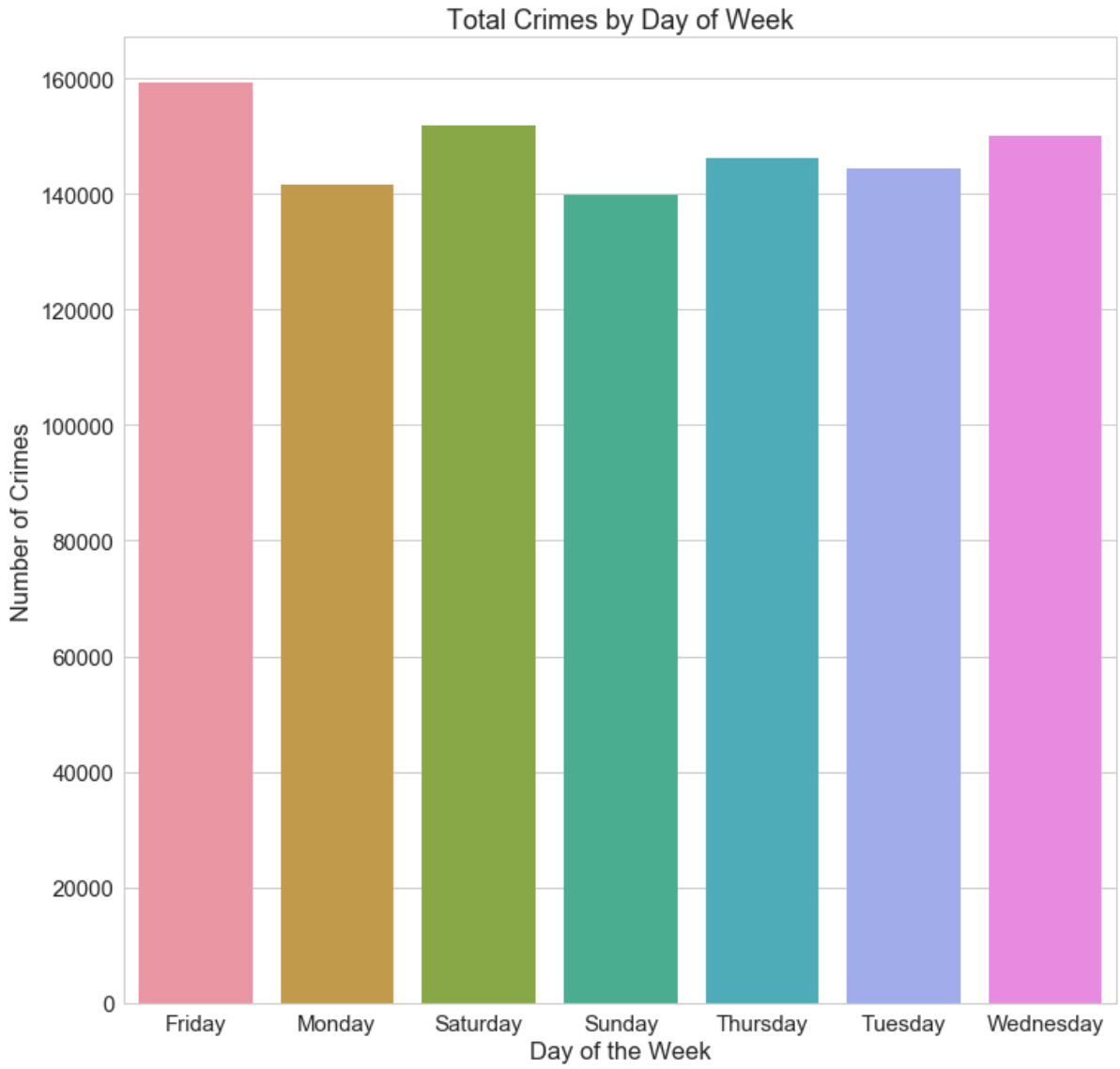
In [151]:
```
crime_by_day = pd.read_sql("SELECT DayOfWeek, COUNT(*) as 'CrimeCount' \
                           FROM crime \
                           GROUP BY DayOfWeek"
                           , db)
crime_by_day
```

Out[151]:

	DayOfWeek	CrimeCount
0	Friday	159182
1	Monday	141683
2	Saturday	151941
3	Sunday	139915
4	Thursday	146130
5	Tuesday	144353
6	Wednesday	150197

For my query, I selected the days of the week and made a count of how many crimes appear on certain days of the week by doing COUNT(*). This will give me a table with 7 rows for each day of the week with the count of how many crimes have occures on a paticular day of the week.

In [152]:
```
crime_day = sns.barplot(y=crime_by_day["CrimeCount"], x=crime_by_day["DayOfWeek"], data=crime_by_day)
plt.title("Total Crimes by Day of Week")
crime_day.set(ylabel = 'Number of Crimes', xlabel = 'Day of the Week')
plt.show()
```



In the plot above, we can see that the two most popular days to commit a crime are Fridays and Saturdays. This makes intuitive sense because Friday and Saturdays are generall not work or school days so most people go out and most minor crimes in San Francisco are usually people getting drunk and often times this results in people not being able to control themselves. Fridays and Saturdays are also very popular times of the week for people to go out clubbing etc, which may also lead to crimes such as violence.

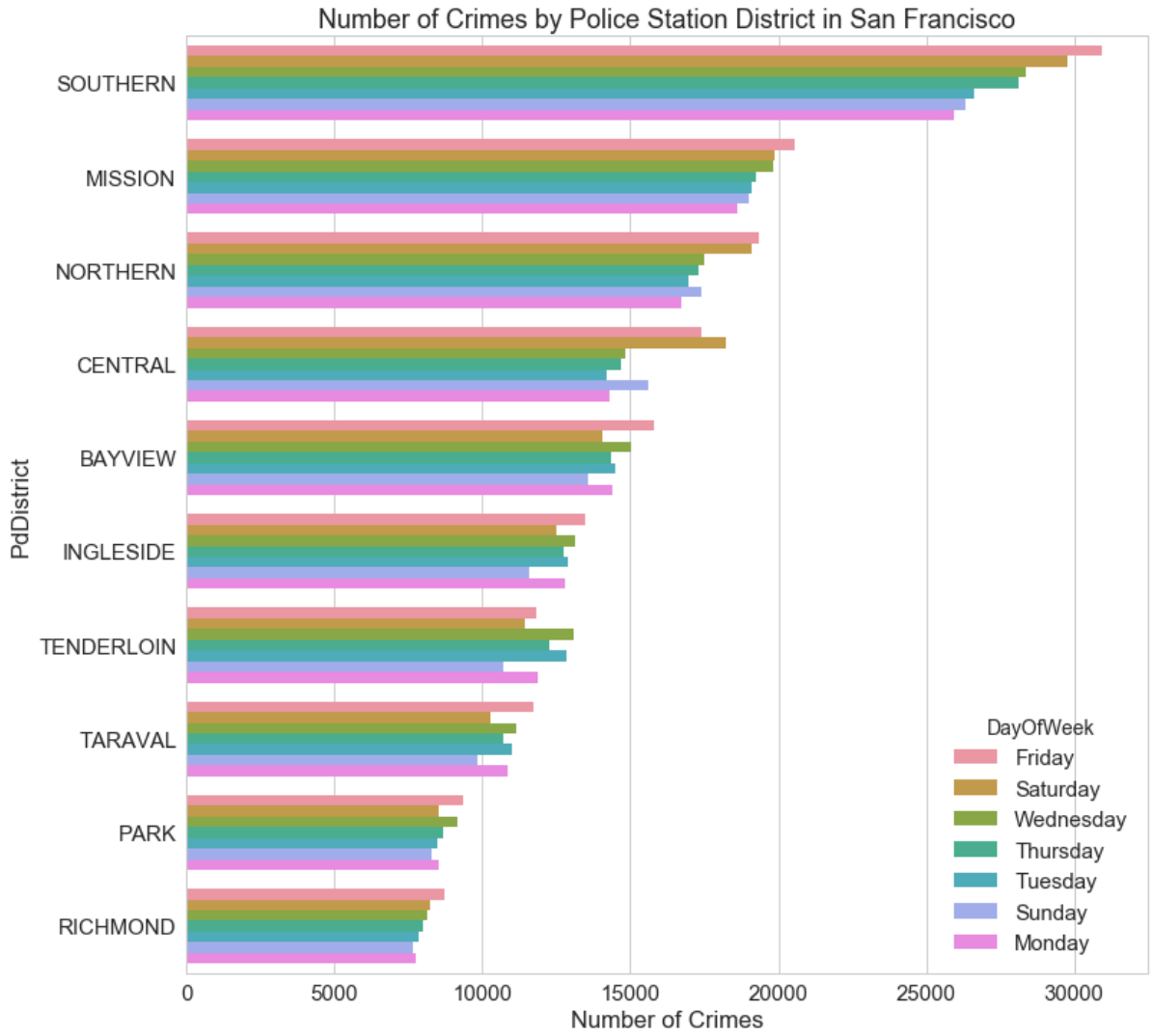
In [153]:
```
crime_by_neigh = pd.read_sql("SELECT DayOfWeek, COUNT(*) as 'CrimeCount', PdDi
strict \
                              FROM crime \
                              GROUP BY PdDistrict, DayOfWeek \
                              ORDER BY COUNT(*) \
                              DESC"
                              , db)
crime_by_neigh.head()
```

Out[153]:

	DayOfWeek	CrimeCount	PdDistrict
0	Friday	30952	SOUTHERN
1	Saturday	29790	SOUTHERN
2	Wednesday	28394	SOUTHERN
3	Thursday	28150	SOUTHERN
4	Tuesday	26614	SOUTHERN

In the query above, in order to get the number of crimes by day of the week and by neighborhood, I grouped by both day of the week and Police Department District. The Police Department District is the same as the neighborhood of San Francsico because the City and County of San Francisco has one police department for each neighborhood. So a police department would have documentation of crimes of only the ones that occur within their neighborhood/near by area around that police station. So the more crimes a police station has, then the more crimes occur within that neighborhood/district.

In [154]:
```
plot = sns.barplot(x="CrimeCount", y="PdDistrict", hue="DayOfWeek", data=crime
_by_neigh)
plt.title("Number of Crimes by Police Station District in San Francisco")
plt.xlabel("Number of Crimes")
plt.show()
```



In the plot above, we can clearly see that Southern Police Department, which is located in the Mission Bay, has the most number of crimes, and they are usually on Friday, Saturday and Wednesday. The Mission District has a lot of crimes on Fridays and Saturdays. However, as we go down the plot, we begin to notice that there is no real pattern on what days of the week are most popular to commit a crime when we split up the crimes by neighborhood. Some say the number of crimes generally increase during the holidays (residential theft) and that crimes by states also have no real pattern on why certain days of the week more crimes are commited.

I noticed that neighborhoods such as the Richmond and the Sunset (Taravel) have lower crime rates, but they alos have a higher price to rent ration. It appears that high rent means less crimes. This makes sense as people are generally willing to pay more to live in a safer area.

In San Francisco it remains relatively clear that weekends are when more crimes are commited. But these crimes are likely related to partying/clubbing.

5. Where are most of the schools in San Francisco located? Does location affect what type of schools there are in the neighborhood? Are there more schools in certain neighborhoods?

In [155]:
```
public_schools = pd.read_sql("SELECT Name, GradeRange, Category, Entity, Lat,
 Lon \
                             FROM schools \
                             WHERE GradeRange = '9-12' AND (Entity = 'Community
 College District' OR Entity = 'SFUSD')"
                             , db)
public_schools.head()
```

Out[155]:

	Name	GradeRange	Category	Entity	Lat	Lon
0	Asawa, Ruth Asawa San Francisco School Of The ...	9-12	USD Grades 9-12	SFUSD	37.745316	-122.448830
1	Balboa High School	9-12	USD Grades 9-12	SFUSD	37.721142	-122.441399
2	Burton, Phillip And Sala Burton High School	9-12	USD Grades 9-12	SFUSD	37.721546	-122.406555
3	City Arts And Tech High School	9-12	USD Grades 9-12	SFUSD	37.718784	-122.424667
4	Downtown High School	9-12	USD Grades 9-12	SFUSD	37.761398	-122.403702

For public schools in San Francisco, children must go to the elementary and middle school that is nearest/closest to them. This means children will likely attend a elementary and middle school that is within the same neighborhood. There is no choice on what elementary or middle school they attend. This is why i filter out the schools by GradeRange of 9-12 (high school level in USA) because once the students get to high school, the students will be assigned a high school based on their choice.

However, in order to get into their first choice they need to have a competative GPA in order to attend. Some high schools have a lower GPA requirment and some have a high GPA requirement. Some high schools, such as charter schools, do not follow the SFUSD application process, but is still a public school, which is why they are under the entity SFUSD.

For example, Lowell High School is a public charter school. They have the same application process as all SFUSD high schools, but they require students to have a completative GPA and write an essay before being considered a valid applicant and then reviewed for acceptance.

In [156]:
```
private_schools = pd.read_sql("SELECT Name, GradeRange, Category, Entity, Lat,
 Lon \
                               FROM schools \
                               WHERE Entity = 'Private'"
                               , db)
private_schools.head()
```

Out[156]:

	Name	GradeRange	Category	Entity	Lat	Lon
0	Adda Clevenger School	K-8	Independent / Private	Private	37.753738	-122.424461
1	Alt School - Alamo Square	K-5	Independent / Private	Private	37.774788	-122.430206
2	Alt School - Dogpatch 1	TK-2	Independent / Private	Private	37.761177	-122.388130
3	Alt School - Dogpatch 2	TK-2	Independent / Private	Private	37.760662	-122.387962
4	Alt School - Fort Mason	K-8	Independent / Private	Private	37.804562	-122.433907

For private schools in San Francisco, anyone can attend them as long as they can pay for the tuition (no aid or whatsoever). This is why I accepted all grade ranges, but filtered out by private entity because I want to see if the location of a private school is affected by the neighborhood.

What we probably expect to see is that neighborhoods with higher housing prices or higher rent prices would have more private schools within the same neighborhood and the lower housing/rent prices would have less private schools nearby.

In [157]:
```
public_schools.columns = public_schools.columns.str.lower()
private_schools.columns = private_schools.columns.str.lower()
```

In [158]:
```
# Read in the zipcode database
zips_of_SF = gpd.read_file("cb_2015_us_zcta510_500k.shp")
```

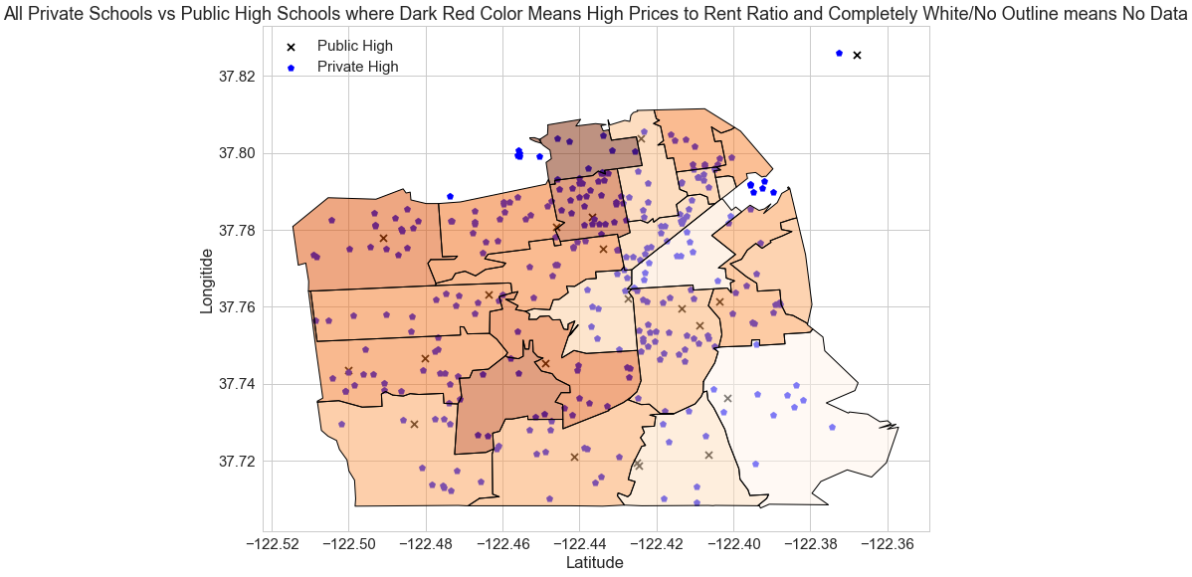
In [159]:
```
all_zipcodes = pd.read_sql("SELECT RegionName FROM zillow GROUP BY
RegionName", db)
all_zipcodes = [str(zips) for zips in all_zipcodes["RegionName"].values]
```

In [160]:
```
# Remove the zipcodes for South San Francisco
south_city_zip = ['94105', '94080']

for zipcode in south_city_zip:
    all_zipcodes.remove(zipcode)
```

Here I moved all zip codes associated with South San Francisco (AKA South City) because although South City has the words "San Francisco" it is not part of the City and County of San Francisco. It is part of San Mateo County.

In [161]:
```
#Look for the zipcodes within SF found withing the list above and take the two
 columns of interest.
zips_of_SF = zips_of_SF.loc[zips_of_SF["ZCTA5CE10"].isin(set(all_zipcodes))]
zips_of_SF = zips_of_SF.sort("ZCTA5CE10").reset_index(drop=True)
zips_of_SF = zips_of_SF[["ZCTA5CE10", "geometry"]]

# Reset the index to later merge with the other dataframe.
ratio_rent_zip = ratio_rent_zip.sort("Zipcode").reset_index(drop=True)

# Merge the two dataframes based on the sorted zipcodes
# Want to get the polygon and the price on the same data frame.
df_merged = pd.concat([zips_of_SF, ratio_rent_zip["Price to Rent Ratio"]], axi
s=1)
```

In [162]:
```
color_plot = df_merged.plot(column='Price to Rent Ratio', colormap="Oranges")
public_schools = plt.scatter(public_schools["lon"], public_schools["lat"], mar
ker='x', color = 'black')
private_schools = plt.scatter(private_schools["lon"], private_schools["lat"],m
arker='p', color = 'blue')

plt.title("All Private Schools vs Public High Schools where Dark Red Color Mea
ns High Prices to Rent Ratio and Completely White/No Outline means No Data")
plt.legend((public_schools, private_schools),
           ('Public High', 'Private High'),
           scatterpoints=1,
           loc='upper left',
           fontsize=15)
color_plot.set(ylabel = 'Longitide', xlabel = 'Latitude')
plt.show()
```



In the map above, we see two floating points, one public and one private high school. Although the map does not show, that is actually the location of Treasure Island. Treasure Island is part of the City and County of San Francisco.

As we can see on the map, there is no relationship or correlation between private school location and price to rent ratio by neighborhoods. There is no real patter to where private school locations are, but the demand for private schools have been steadly increasing since 1930's (Barrow). Since private schools do not recieve any type of funding from the Government, usually a new private school needs demand/funding from public before a new one can open. Some also believe that the growing population of private schools are to force competition between public and private schools by allowing parents to have more options in private schools than high schools. Often times private schools are said to offer a higher quality of education than publics schools because of the saying "what you pay is what you get".

We can also see that high schools appear to be scattered around San Francisco. Although students in San Francisco are not reuqired to attend a high school within the same neighborhood as them, we can see that there is about 1 high school per zipcode/neighborhood and that the price to rent ratio does not affect where a high school is placed. This could have been done on purpose, as possibly students attended high schools based on neighborhoods back then. Or the public high schools were first build by demand of a growing San Francisco.

Resources

Barrow, Lisa. Private School Location andNeighborhood Characteristics. Federal Reserve Bank of Chicago. December 2002. Web. March 2017.