Visualizations SF Crime in the Summer of 2014

Jovita De Loatch

January 28, 2016

Report Description:

San Francisco, 'The City' like all cities, has an interest in using data to understand social patterns, such as crime. Crimes of particular interest to both visitors and residence alike are those against persons or their property. This report seeks to visualize when and where crimes in the City are more likely happens against persons or the theft their of property. It uses data from the summer of 2014 giving time and location crime reports from across all of San Francisco's neighborhoods. The goal is to visually expose how incidents vary by time of day, day of the week or month and what is suggested broadly about the differences in these general patterns in San Francisco. The findings are summarized at the end of the report.

The Data

The data set contains 28993 records June 1, 2014 to August 31, 2014. Data was sourced from SFPD Incidents from the old system. San Francisco police have implemented a new system for tracking crime. Also note that new Police Station boundaries are not reflected in the dataset and cannot compare data from July 19, 2015 onward. Each row represent a criminal record including variables composed of the following data elements:

- Date – date
- Time - timestamp
- Category – The type of crime, Larceny, etc.
- Descript – A more detailed description of the crime.
- DayOfWeek – Day of crime: Monday, Tuesday, etc.
- PdDistrict - Police department district.
- Resolution - What was the outcome, Arrest, Unfounded, None, etc.
- Address – Street address of crime.
- X and Y – GPS coordinates of crime.

Reading the Data

First, the data was read. Data was processed for the three month period provided using a Pareto chart (a barplot where the categories are ordered in non increasing order, adding a line to show the cumulative sum) from the Quality Control Charts package(qcc). Below represents the cumulative summation of over 85% of the crimes in The City.

- Category | Count | CumSum | Percentage | CumPercent
- LARCENY/THEFT | 9466 | 9466 | 32.649260166 | 32.64926
- OTHER OFFENSES | 3567 | 13033 | 12.302969682 | 44.95223
- NON-CRIMINAL | 3023 | 16056 | 10.426654710 | 55.37888
- ASSAULT | 2882 | 18938 | 9.940330425 | 65.31921
- VEHICLE THEFT | 1966 | 20904 | 6.780947125 | 72.10016
- WARRANTS | 1782 | 22686 | 6.146311179 | 78.24647
- DRUG/NARCOTIC | 1345 | 24031 | 4.639050805 | 82.88552
- SUSPICIOUS OCC | 1300 | 25331 | 4.483840927 | 87.36937

Of the 37 categories of crime in this dataset 30 categories represent less than 18% of total crimes reported. Further, both 'Other Offenses' and 'Non-Criminal' categories (over 22% of total) appeared too broadly applied in my view to extract meaningful patterns. For example, 'Other Offenses' includes items from MISCELLANEOUS INVESTIGATION to TRAFFIC VIOLATION to OBSCENE PHONE CALLS(S). Also, 'Non-Criminal' included curious descriptions of 'INVESTIGATIVE DETENTION.' Without clarification I was unsure what offense was being tracked. I selected 3 of the 6 left to focus on crimes, representing almost half of the total crimes reported: 'Larceny/Theft,' 'Assault,' and 'Vehicle Theft' offenses.

- $32.649260166 + 9.940330425 + 6.780947125 = 49.370537716$ %

Data Visualization

The goal is to analyze criminal incident data from San Francisco to visualize patterns of activity for a small set of data for the summer of 2014. To visualize when and where crime in the City happens against persons or their property a few visualizations that can suggest a general pattern of activity. The first spacial plots use contour maps to graphically represent the data, where each value contained in the map are marked in colors. These maps are used to explore the relationships among location or time and category.

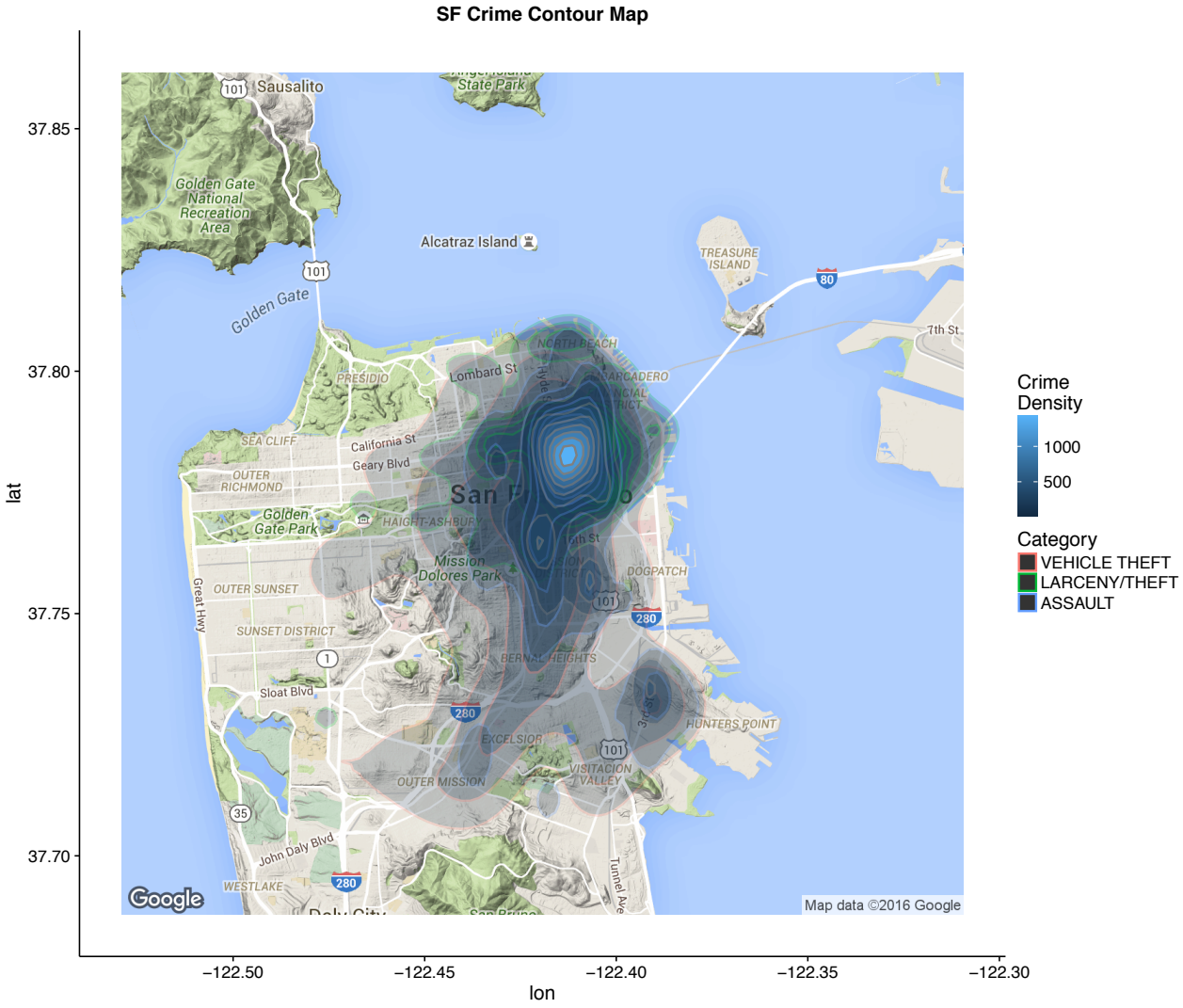
The following visualizations are offered for review.

Plot 1 Spacial map with ggmap This plot uses ggmap to create a contour map overview of the locations of the 3 activities. Note, the northeastern corner of the City and the Mission are the most densely populated area are also the area that has the most criminal activity. (See Appendix A) The central downtown including the Tenderloin district appears to be the epicenter of activity where the contours peak, represented by the highest density contours for all categories.

```
############################################################

# Focus on major crimes of concern
SF_crimes <- subset(SFCrime,
                          Category != "OTHER OFFENSES" &
                            Category != "NON-CRIMINAL"
)

# rank SF crimes
SF_crimes$Category <-
  factor(SF_crimes$Category,levels = c( "VEHICLE THEFT", "LARCENY/THEFT","ASSAULT")
  )

############################################################
# Plot 1 with ggmap
############################################################

# get a color map)
SFMap <- get_map("San Francisco", zoom = 12)
```

```
## Map from URL : http://maps.googleapis.com/maps/api/staticmap?center=San+Francisco&zoom=12&size=640x6

## Information from URL : http://maps.googleapis.com/maps/api/geocode/json?address=San%20Francisco&sens
```

```
SFMap <- ggmap(SFMap, extent = "normal", legend = "right")

# a filled contour plot...
SFMap +
  stat_density2d(aes(x = X, y = Y, colour = Category, fill = ..level.., alpha = ..level..), size = .75,
  scale_fill_gradient("Crime\nDensity") +
  ggtitle("SF Crime Contour Map") +
  scale_alpha(range = c(.2, 1), guide = FALSE) +
  guides(fill = guide_colorbar(barwidth = 1, barheight = 5))
```



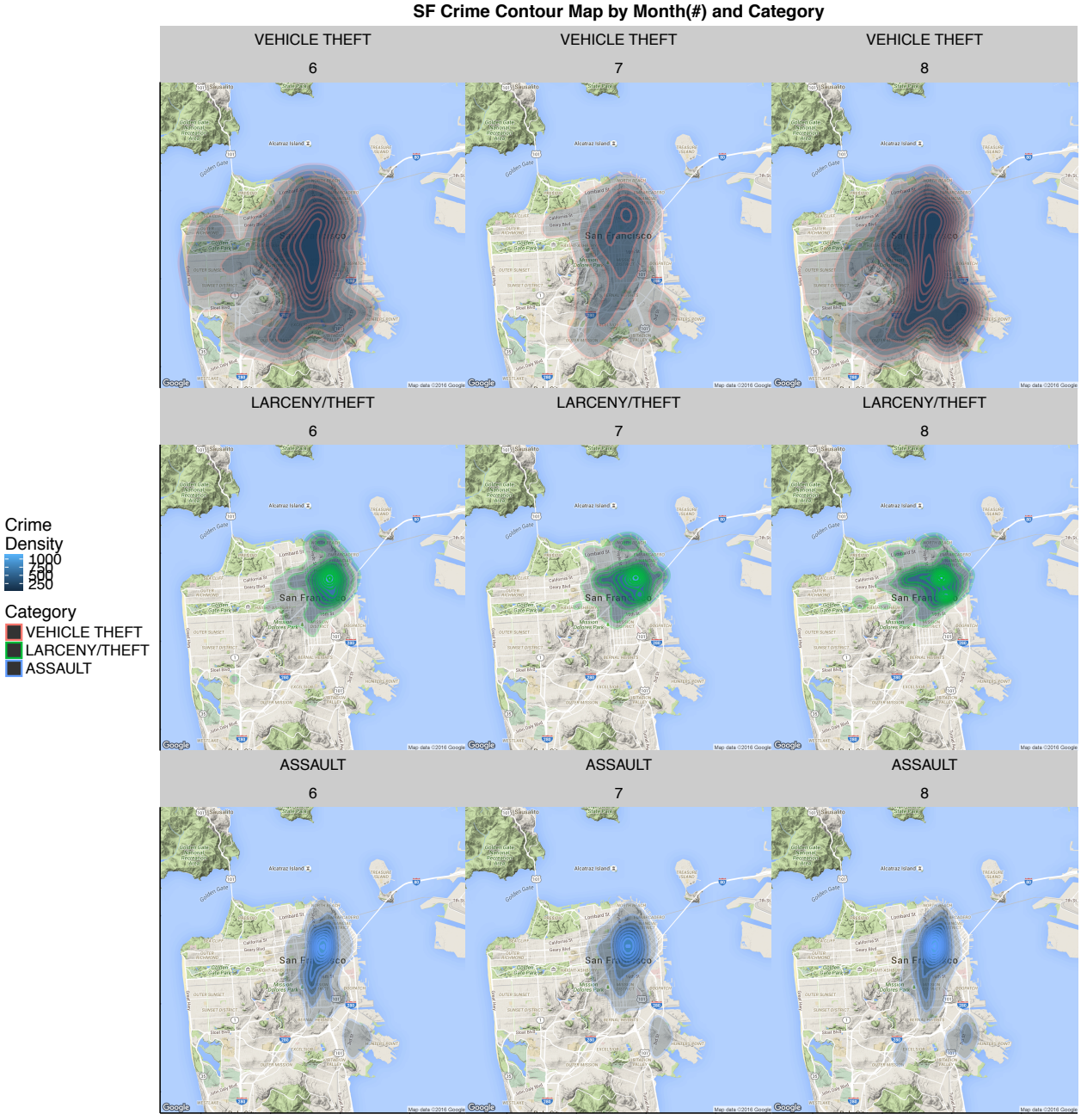
Plot 2a Temporal/Spacial map with ggmap The vehicle theft contour lines are shown a to be wide spread at low level of activity throughout the City, contrast with an unusually high activity through the Mission District and a finger of increased activity to the southern neighborhoods to the Outer Mission. Vehicle theft is generally less dense during July and seems to abandon a good part of the western side of the City. This is typically when the Bay Area surf is down and the City and Ocean Beach at its westend, in particular, are foggy.

Looking at the contour map of both Larceny/Theft and Assault are very consentrated in the Downtown area

with modest fluctuations of Larceny/Theft in the smaller North Beach tourist area and constant Assault activity at the southeastern end of the City at Bayview/Hunters Point.

```
## Map from URL : http://maps.googleapis.com/maps/api/staticmap?center=San+Francisco&zoom=12&size=640x6
```

```
## Information from URL : http://maps.googleapis.com/maps/api/geocode/json?address=San%20Francisco&sens
```




```
##    Friday    Monday  Saturday    Sunday  Thursday   Tuesday Wednesday
##      3436      3069      3357      3340      3049      2990      3162
```

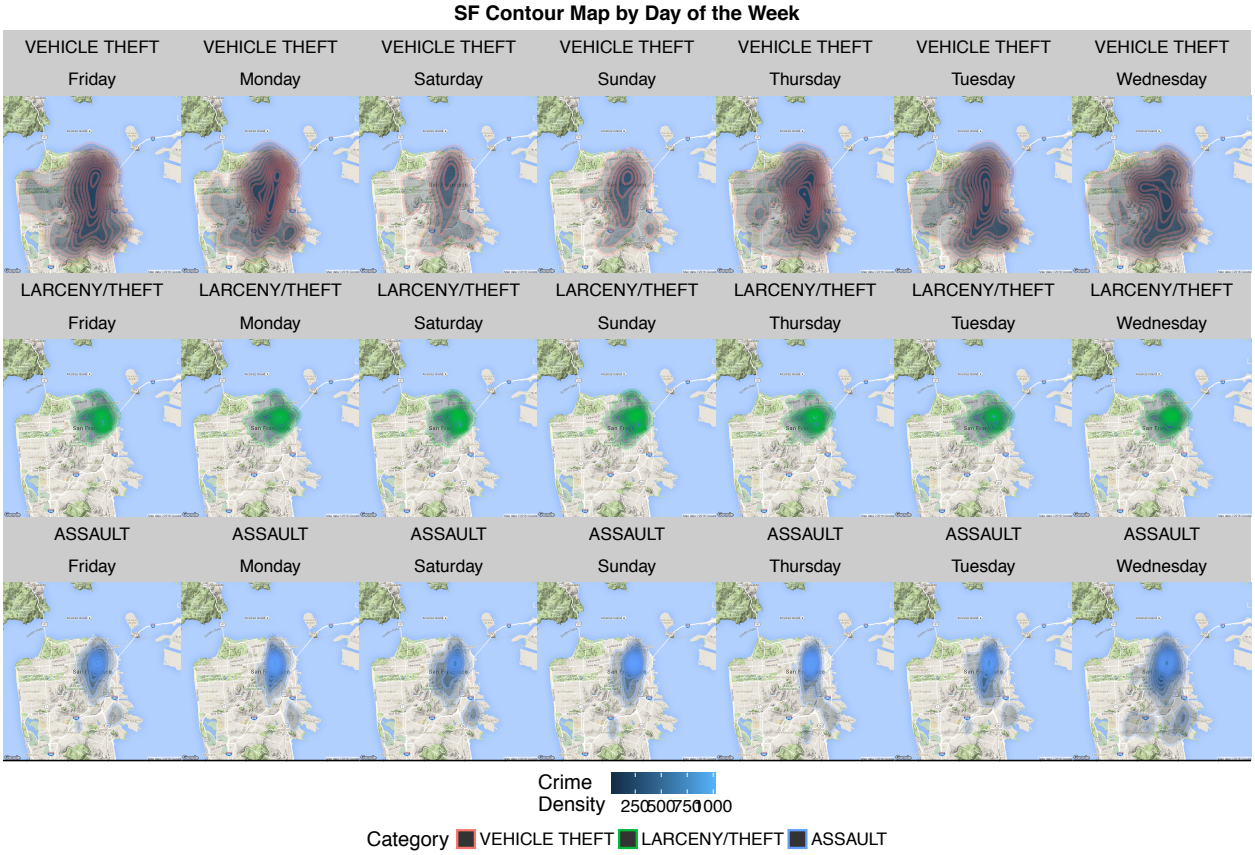
```
## [1] 2990 3436
```

```
## [1] 0.1298021
```

Plot 2b Temporal/Spacial map with ggmap Temporally, the view of the day of the week in decending order of frequency show the most assault and larceny activity on Friday. Saturday and Sunday follow respectively. It seems a bit surprising that Monday would be a more active day than the other weekend days. If I were to speculate, the data may be skewed to Monday because some records may reflect the time discovered after the weekend away and not within an hour of the crime as is assumed here.

```
## Map from URL : http://maps.googleapis.com/maps/api/staticmap?center=San+Francisco&zoom=12&size=640x6
```

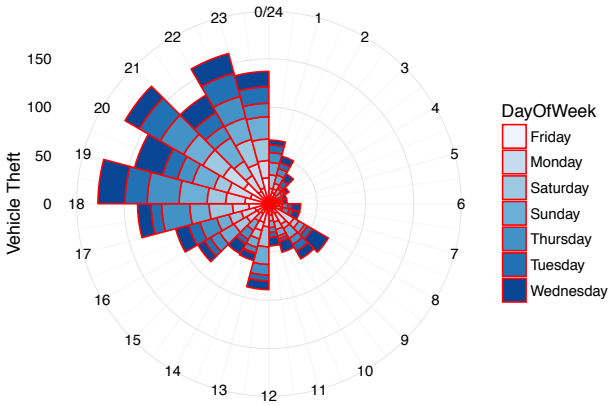
```
## Information from URL : http://maps.googleapis.com/maps/api/geocode/json?address=San%20Francisco&sens
```

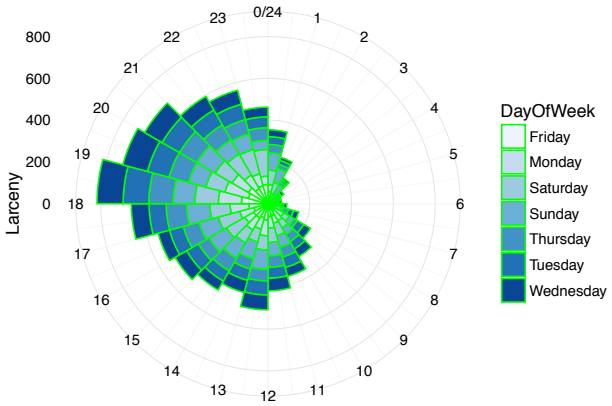


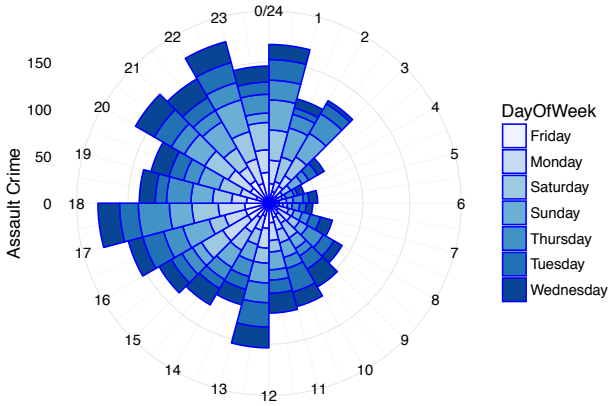
Plot 3 Time of Day with ggplot These next visualizations help to explore the trends for the time of day. The results are perhaps most surprising as they suggest that most of these crime happen not in the scary wee hours of the night but during the day and evening. As noted above, we assume that the report was made within 1 hour of the crime. If that is the case, vehicle theft trended toward the evenings in an odd pattern of a surge every other hour. There seems to be an ebb and flow.

```
#####################################################################
#Plot 3 with ggplot
#####################################################################

assaultCrime = subset(SFCrime, SFCrime$Category == "ASSAULT")
```

```
plotassaultCrime = ggplot(assaultCrime, aes(x = Hour, fill = DayOfWeek)) +
  geom_histogram(breaks = seq(0, 24), width = 1, colour = "blue") +
  coord_polar(start = 0) +
  theme_minimal() +
  scale_fill_brewer() + ylab("Assault Crime") +
  scale_x_continuous("", limits = c(0, 24), breaks = seq(0, 24),
                     labels = seq(0, 24))

theftCrime = subset(SFCrime, SFCrime$Category == "LARCENY/THEFT")
plotTheftCrime = ggplot(theftCrime, aes(x = Hour, fill = DayOfWeek)) +
  geom_histogram(breaks = seq(0, 24), width = 1, colour = "green") +
  coord_polar(start = 0) +
  theme_minimal() +
  scale_fill_brewer() + ylab("Larceny") +
  scale_x_continuous("", limits = c(0, 24), breaks = seq(0, 24),
                     labels = seq(0, 24))

carCrime = subset(SFCrime, Category == "VEHICLE THEFT")
plotCarCrime = ggplot(carCrime, aes(x = Hour, fill = DayOfWeek)) +
  geom_histogram(breaks = seq(0, 24), width = 1, colour = "red") +
  coord_polar(start = 0) +
  theme_minimal() +
  scale_fill_brewer() + ylab("Vehicle Theft") +
  scale_x_continuous("", limits = c(0, 24), breaks = seq(0, 24),
                     labels = seq(0, 24))

theme_set(theme_cowplot(font_size=1)) # reduce default font size

#plotCarCrime
#plotTheftCrime
#plotassaultCrime

w = 0.50
grid.arrange(plotCarCrime,  plotTheftCrime, plotassaultCrime,  widths=c(w,1-w), ncol=2)
```







Summary of Findings

The main findings of the report on crime in the summer of 2014 of San Francisco are as follows: - The central downtown district is the epicenter of activity - Vehicle theft has a bit of activity throughout much of the City, with an unusually high activity through the Downtown and Mission District to the southern neighborhoods. - Friday and Monday are the peak days for assault and larceny activity. Saturday and Sunday follow respectively. - There were modest Monthly fluctuations of crime in other smaller areas - A constant area of activity at the southeastern end of the City at Bayview/Hunters Point. - Vehicle theft seems to abandon a good part of the western side of the City during July. - Vehicle theft trended toward the evenings in an odd pattern of a surge every other hour.

Discussion

These visualizations suggest there are some differences in when and where crime happens. However, the question is whether the differences brought out by these visualizations actually identifies any statistically significant differences in the data given the range is often somewhat small. A second question could be asked

as to whether these deviations stray from what is expected in crime per capita. Does the pattern change with the spacial temporal changes in daily population activity?

Finally, as more complex approaches to urban scaling advance should we look beyond linear models to consider super linear models. Bettencourt LMA, Lobo J, Strumsky D, West GB (2010), have found that larger cities are disproportionally the centers of innovation, wealth as well as crime, all to approximately the same degree. They note that an obstacle to effective policy is the lack of meaningful urban metrics based on a quantitative understanding of cities. Typically, linear per capita indicators are used to characterize and rank cities. However, these implicitly ignore the fundamental role of nonlinear agglomeration are explicitly manifested by the superlinear power law scaling of most urban socioeconomic indicators with population size, all with similar exponents (footnote 1). With the density of activity in the northeastern section of San Francisco so concentrated may be better visualised potentially on an exponential scale. This suggest that exploring superlinear power law scaling in crime data may be a more fruitful approach to developing a prediction model.

References

Footnote 1: Bettencourt LMA, Lobo J, Strumsky D, West GB (2010) Urban Scaling and Its Deviations: Revealing the Structure of Wealth, Innovation and Crime across Cities. PLoS ONE 5(11): e13541. doi: 10.1371/journal.pone.0013541

Scrucca, L. (2004). qcc: an R package for quality control charting and statistical process control. R News 4/1, 11-17. Wetherill, G.B. and Brown, D.W. (1991) Statistical Process Control. New York: Chapman & Hall.

H. Wickham. ggplot2: elegant graphics for data analysis. Springer New York, 2009.

D. Kahle and H. Wickham. ggmap: Spatial Visualization with ggplot2. The R Journal, 5(1), 144-161. URL http://journal.r-project.org/archive/2013-1/kahle-wickham.pdf

San Francisco Demographic: D.1 Population density, Population per square mile Geographic Unit of Analysis: Census tract, City and County of San Francisco Department of Public Health Environmental Health Section http://www.sustainablecommunitiesindex.org/city_indicators/view/75

SFPD Incidents - from 1 January 2003, https://data.sfgov.org/Public-Safety/SFPD-Incidents-from-1-January-2003/tmnf-yvry

Appendix A

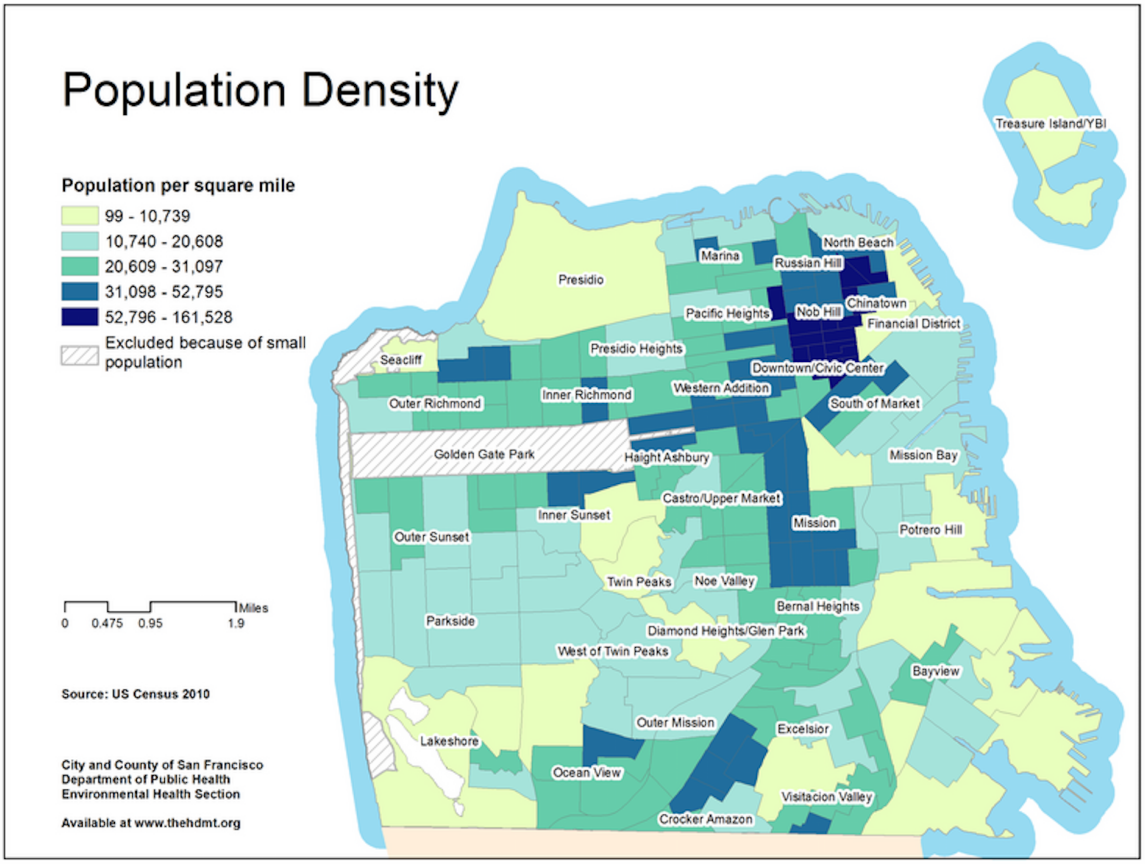


Figure 1: SF Population per square mile.