Data At Scale Assignment 6 Week 1

Ford

November 29, 2015

For this assignment, I wanted to look at a couple different visualizations. For the first, I used the ggmap library to plot a map of the city of the Seattle, then overlay the distribution of crimes on the map. This way, I could see where crimes actually occured within the city.

For the second, I wanted to look at how, if at all, the occurence of crimes differs by the hour of the day.

Below is the code I used to read the data and perform any necessary manipulations. I created a few new variables. First, I converted the datetime variable, which was read as a character variable, into a datetime variable. From this, I could create both an hour variable and a day of week variable.

I also created a couple functions to create 2 additional variables, The first, theft_func, returns a 1 if the input is equal to either "BURGLARY" or "ROBBERY". This function allows me to create a "theft" variable that is equal to 1 when the crime is considered a theft.

The second function, weekend_func, is similar to the first function, but returns a 1 if the input is equal to either "Saturday" or "Sunday". This function allows me to create a "weekend" variable that is equal to 1 when the crime occured on the weekend.

```
library(ggplot2)
library(ggmap)

setwd("~/git/data-at-scale/assignment6/")
d <- read.csv("seattle_incidents_summer_2014.csv", stringsAsFactors = FALSE)

# date/time stuff
d$Date.Reported <- strptime(d$Date.Reported, format="%m/%d/%Y %I:%M:%S %p")
d$hour <- d$Date.Reported$hour
d$dow <- weekdays(d$Date.Reported)

theft_func <- function(x) {
  if(x %in% c("BURGLARY", "ROBBERY")) {
    return(1)
  } else {
    return(0)
  }
}

weekend_func <- function(x) {
  if(x %in% c("Sunday", "Saturday")) {
    return(1)
  } else {
    return(0)
  }
}

# create the theft and weekend variables
d$theft <- as.numeric(lapply(d$Summarized.Offense.Description, theft_func))
d$weekend <- as.numeric(lapply(d$dow, weekend_func))
```

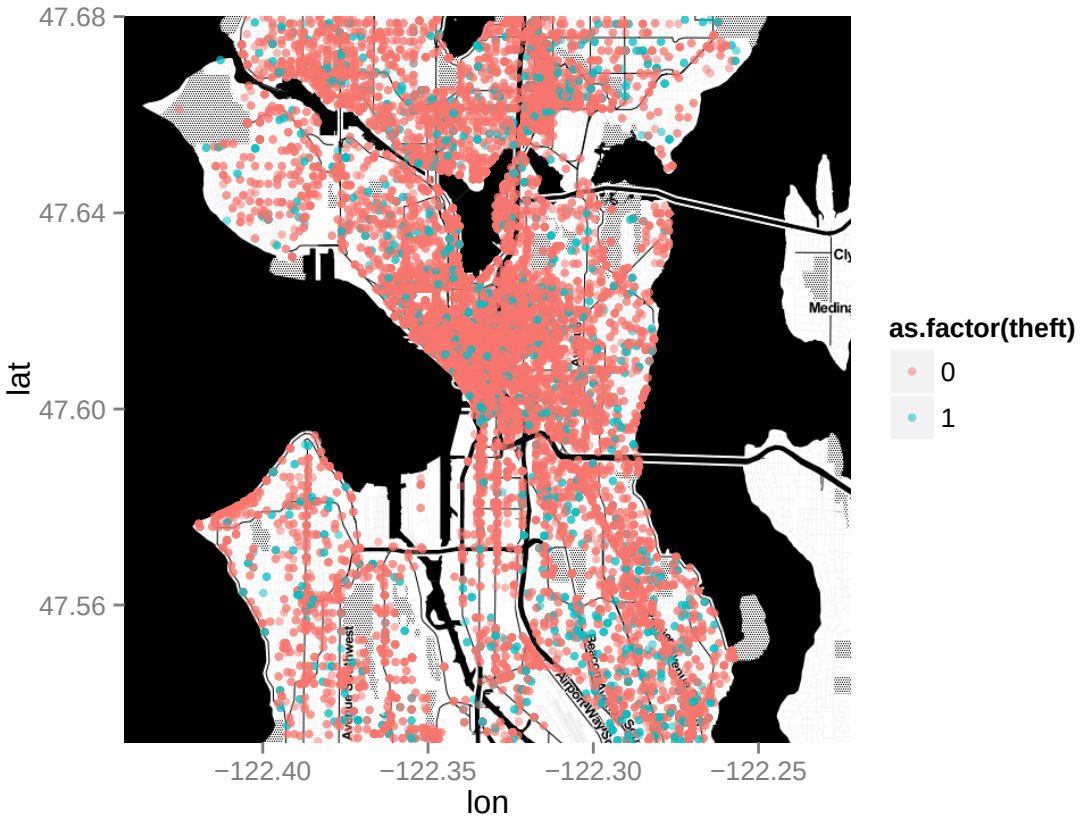
In order to get a map of Seattle, I use the ggmap library. I can use the qmap function, below, to save a plot of the map of Seattle.

```
## Map from URL : http://maps.googleapis.com/maps/api/staticmap?center=seattle&zoom=12&size=640x640&map
## Information from URL : http://maps.googleapis.com/maps/api/geocode/json?address=seattle&sensor=false
## Map from URL : http://tile.stamen.com/toner/12/654/1429.png
## Map from URL : http://tile.stamen.com/toner/12/655/1429.png
## Map from URL : http://tile.stamen.com/toner/12/656/1429.png
## Map from URL : http://tile.stamen.com/toner/12/657/1429.png
## Map from URL : http://tile.stamen.com/toner/12/654/1430.png
## Map from URL : http://tile.stamen.com/toner/12/655/1430.png
## Map from URL : http://tile.stamen.com/toner/12/656/1430.png
## Map from URL : http://tile.stamen.com/toner/12/657/1430.png
## Map from URL : http://tile.stamen.com/toner/12/654/1431.png
## Map from URL : http://tile.stamen.com/toner/12/655/1431.png
## Map from URL : http://tile.stamen.com/toner/12/656/1431.png
## Map from URL : http://tile.stamen.com/toner/12/657/1431.png
```

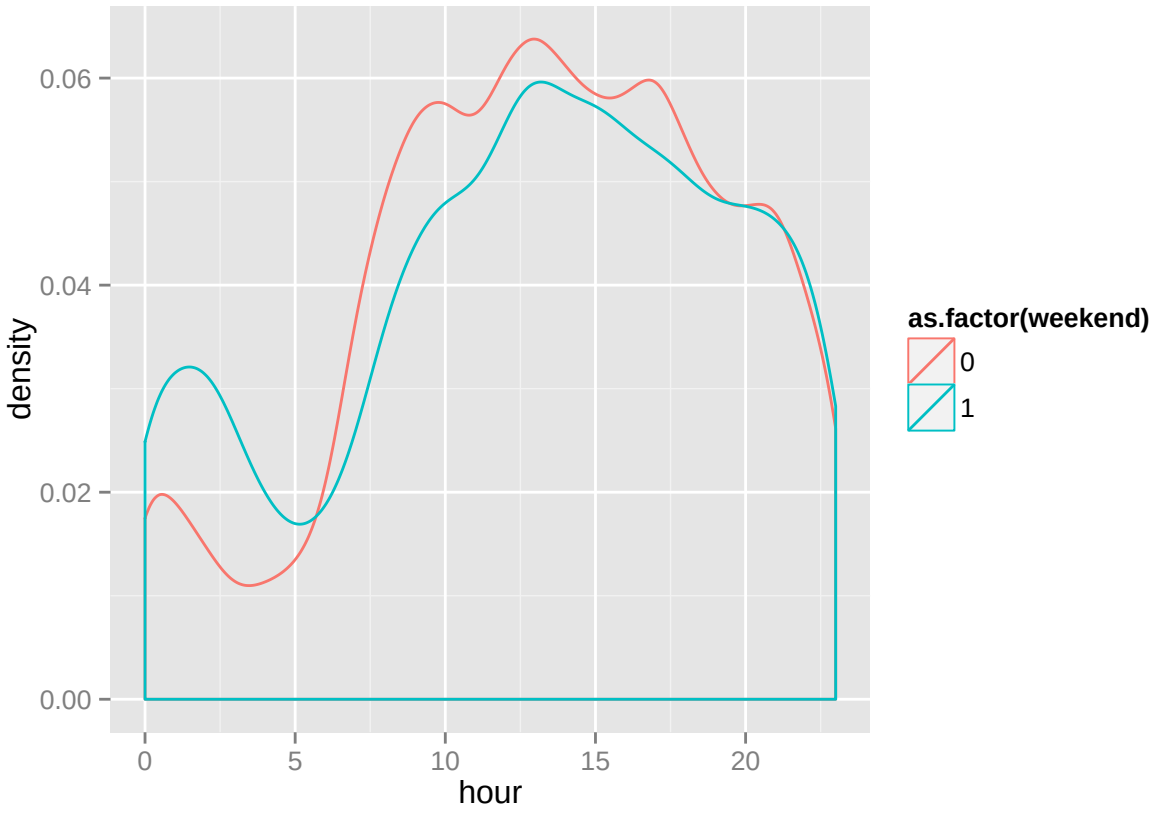
After the map of Seattle is read into R, I can overlay a scatter plot of the crimes, using both Latitude and Longitude. I set the alpha to .5 so that the plot is a bit easier to read. Additionally, I wanted to create the visualization a bit more interesting so I added a third dimension of visualization to the plot. I included the theft variable as a third dimension, which is visualized as the blue points. It is interesting to see that the vast majority of crimes that occur in the downtown area are not actually thefts. Instead, in the southeast corner of the map there seem to be fewer overall crimes. However, a very high proportion of the crimes are burgulary or thefts.

```
seattle_map + geom_point(data=d, aes(x=Longitude, y=Latitude, color=as.factor(theft)), alpha=.5, size=1
```

```
## Warning: Removed 9970 rows containing missing values (geom_point).
```



For the second visualization, I wanted to look at when crimes are occuring. Again, I thought it would be more interesting to add another dimension to the visualization. For this reason I added the weekend variable that I created earlier. This graph shows the density of crimes by hour for both weekdays and weekends. The blue curve is for weekends, while the red is for weekdays. It is interesting to see that the curves are very similar. Both seem to peak around 1PM. An interesting difference is that on weekends, the bump in crime density around 2AM is much more pronounced. This seems to make sense; people are awake later on weekends and thus we'd expect a higher density of crimes.


0.06
0.04
0.02
0.00
density
0
5
10
15
20
hour
as.factor(weekend)
0
1