markdown1

Orpaz Goldstein

April 7, 2016

Ex2_a

San Francisco Crime Classification

Downloaded from a Kaggle competition. (https://www.kaggle.com/c/sf-crime/data?test.csv.zip)

The idea of this competition is to predict the category of crimes commited in San Fransisco. But for this drill i will start with analyzing the most dangerous area and time for a crime to happen, and use these results for the competition.

```
# install.packages("RCurl")
#library(RCurl)

#URL <- "https://www.kaggle.com/c/sf-crime/download/test.csv.zip"
#x <- getURL(URL)
## Or
#x <- getURL(URL, ssl.verifypeer = FALSE)
#data <- read.csv(text = x)

#if Already downloaded
data <- read.csv("/Users/orpaz/Downloads/test.csv", header=T, sep=",")


head(data)
```

```
##   Id               Dates DayOfWeek PdDistrict                   Address
## 1  0 2015-05-10 23:59:00    Sunday    BAYVIEW  2000 Block of THOMAS AV
## 2  1 2015-05-10 23:51:00    Sunday    BAYVIEW       3RD ST / REVERE AV
## 3  2 2015-05-10 23:50:00    Sunday   NORTHERN   2000 Block of GOUGH ST
## 4  3 2015-05-10 23:45:00    Sunday  INGLESIDE 4700 Block of MISSION ST
## 5  4 2015-05-10 23:45:00    Sunday  INGLESIDE 4700 Block of MISSION ST
## 6  5 2015-05-10 23:40:00    Sunday    TARAVAL    BROAD ST / CAPITOL AV
##           X        Y
## 1 -122.3996 37.73505
## 2 -122.3915 37.73243
## 3 -122.4260 37.79221
## 4 -122.4374 37.72141
## 5 -122.4374 37.72141
## 6 -122.4590 37.71317
```

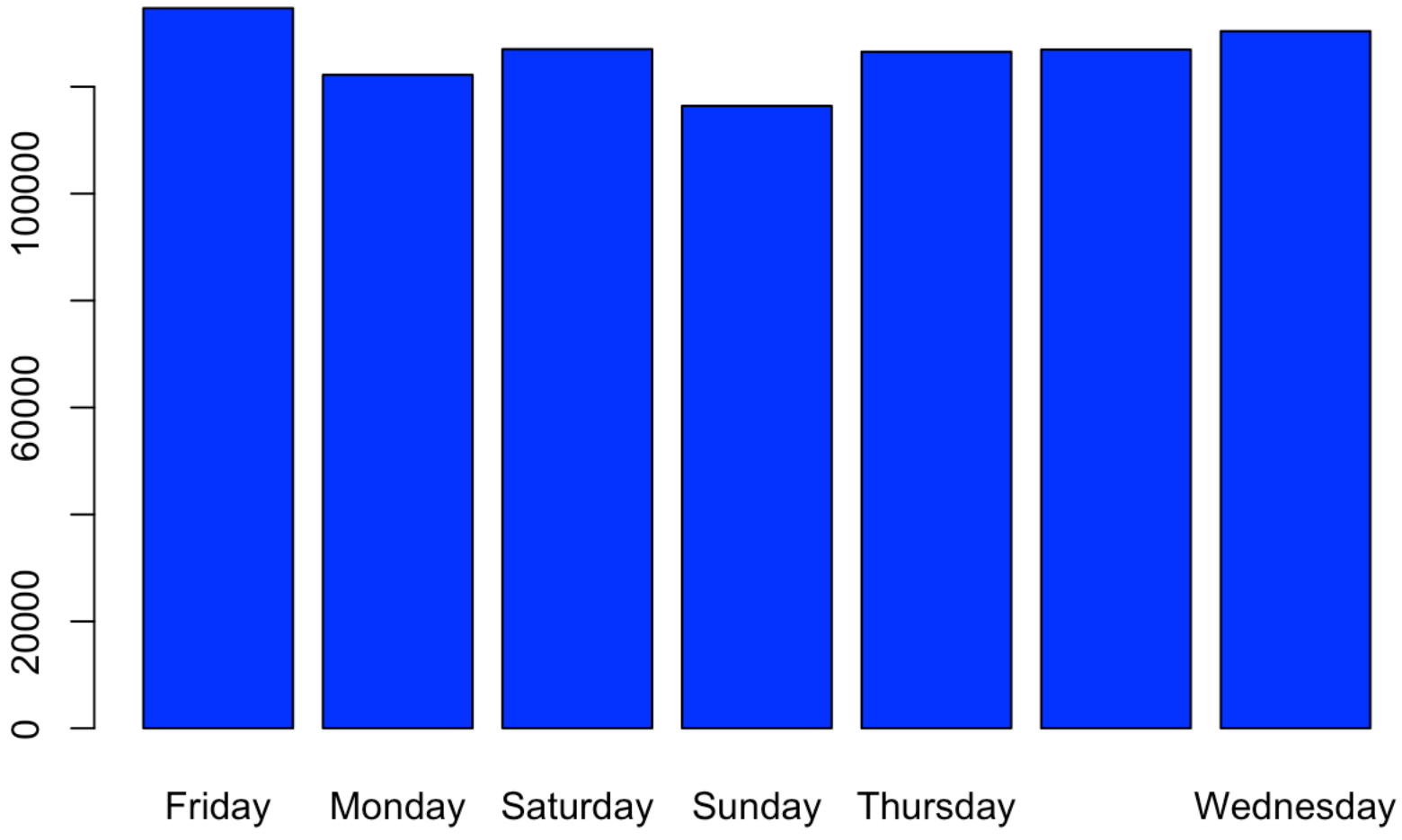
The data set contains an id, time stamp, day of week, district in SF, address, geolocation. I filtered and added an 'hour' collumn so i can more effectively filter the time of the crime.

```
library(lubridate)
data$hour <- c(hour(data$Dates))
```

The data set contains an id, time stamp, day of week, district in SF, address, geolocation. I filtered and added an 'hour' collumn so i can more effectively filter the time of the crime.

- First i checked what day of the week crime accures.

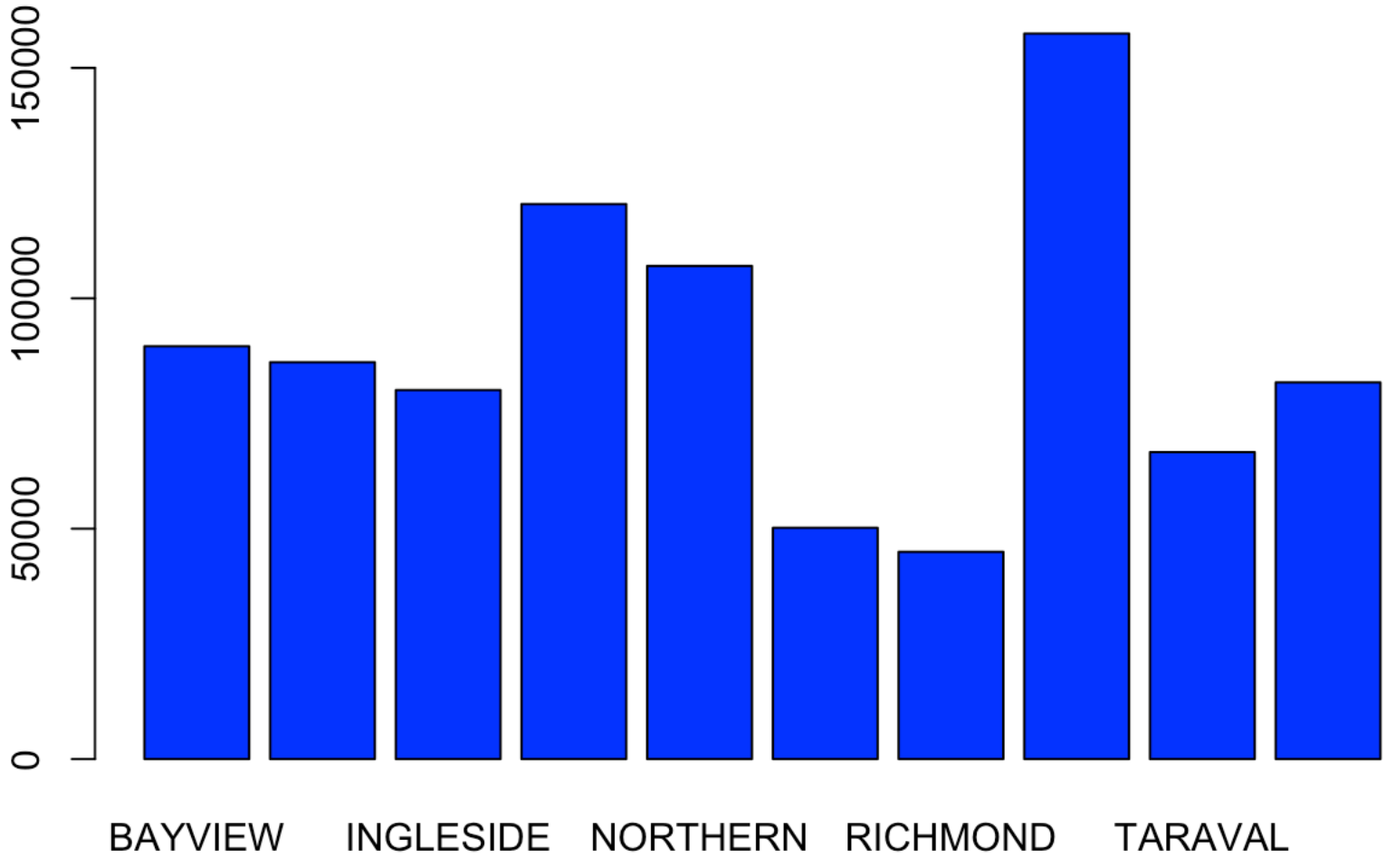
```
barplot(table(data$DayOfWeek), col='blue')
```



Seemed that they crime accures about the same on every day.

- Next i checked if a certain area has higher crime rate then others.

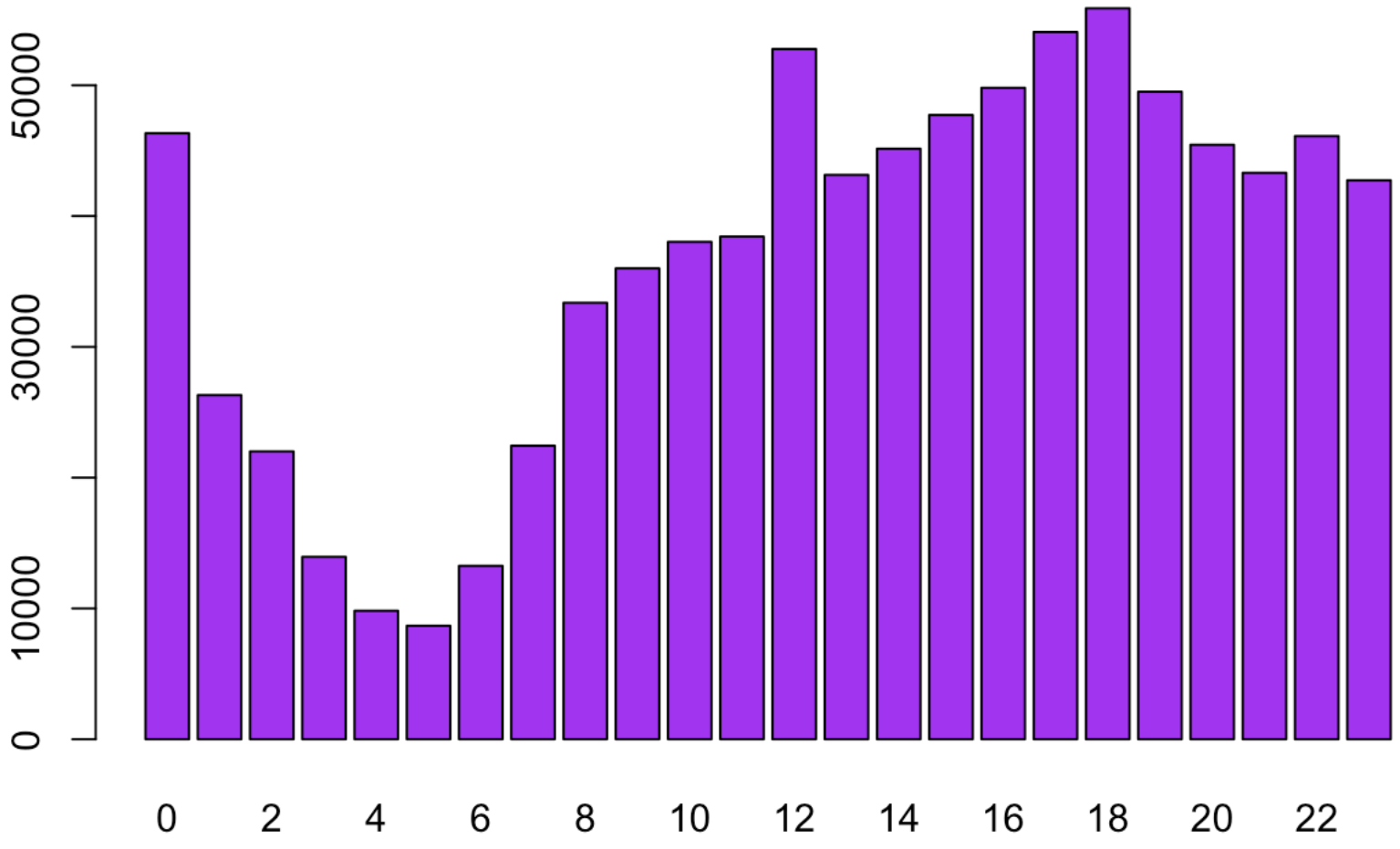
```
barplot(table(data$PdDistrict), col='blue')
```



Looks like Southern leads the crime rate in SF!

- Now i want to see what at what time criminals like working

```
barplot(table(data$hour), col='purple')
```



Crime peaks after-noon from 16-19, also at 12 and midnight…

- Since 18 looks like the most dangerous hour to be outside, i checked what the crime looks like at 18 on the south side of San fransisco
  -# getting the map

```
#display results on a map
library(ggmap)
```

```
## Loading required package: ggplot2
```

```
# Subset for South Side
data1 <- subset(data, PdDistrict == 'SOUTHERN')

# Get The Map
map <- get_map(location = c(lon = mean(data1$X), lat = mean(data1$Y)), zoom = 15,
               maptype = "satellite", scale = 2)
```

```
## Map from URL : http://maps.googleapis.com/maps/api/staticmap?center=37.784812,-122
.4051&zoom=15&size=640x640&scale=2&maptype=satellite&language=en-EN&sensor=false
```

```
# plotting the map with some points on it
ggmap(map) +
  geom_point(data = data1, aes(x = X, y = Y, colour = ifelse(hour==18,F,T), alpha = 0
.1), size = 2, shape = 21) +  guides(fill=FALSE, alpha=FALSE, size=FALSE)
```

```
## Warning: Removed 26690 rows containing missing values (geom_point).
```