Assignment 2

1. Plotly - San Francisco crime analysis

URL: https://plot.ly/dashboard/panjwani.h:18

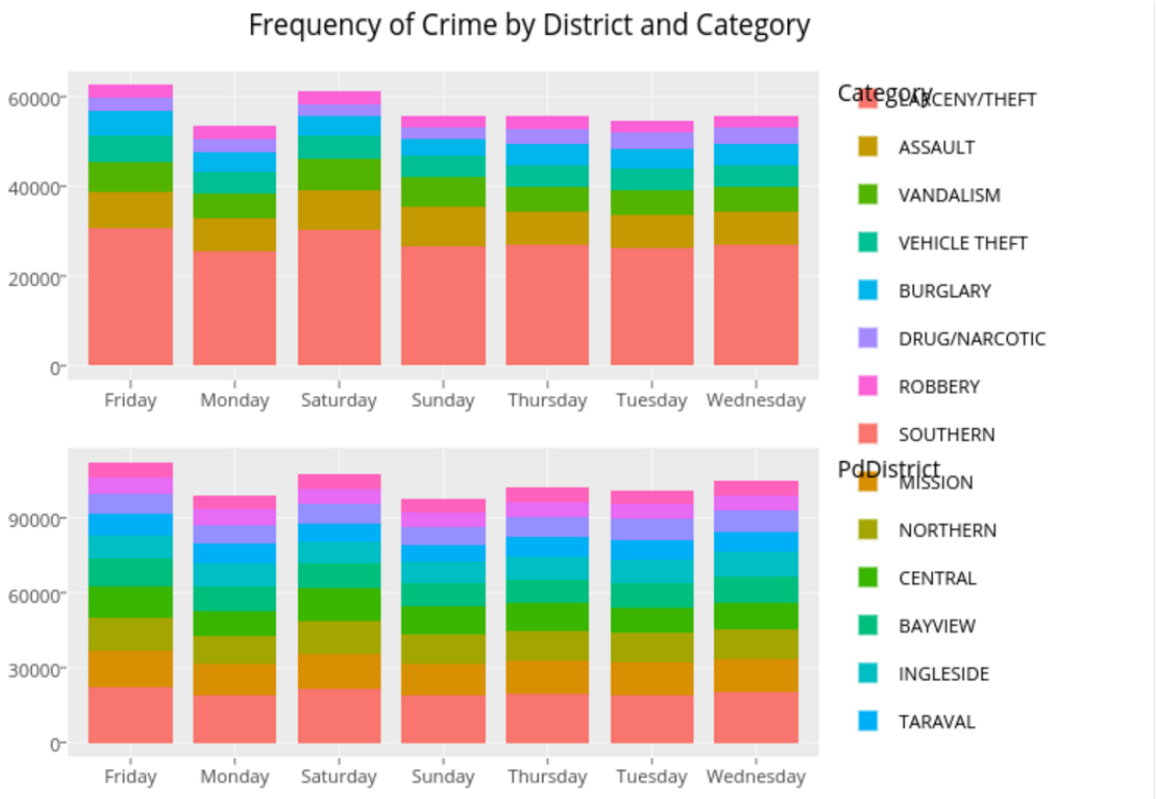
Description

San Francisco is currently the cultural, commercial and financial center of Northern California. Today the city is known more for its tech scene but it has a massive criminal past. The sudden growth in the population has brought an inequality in terms of living, housing shortages leading to no scarcity of crime in the city by the bay.

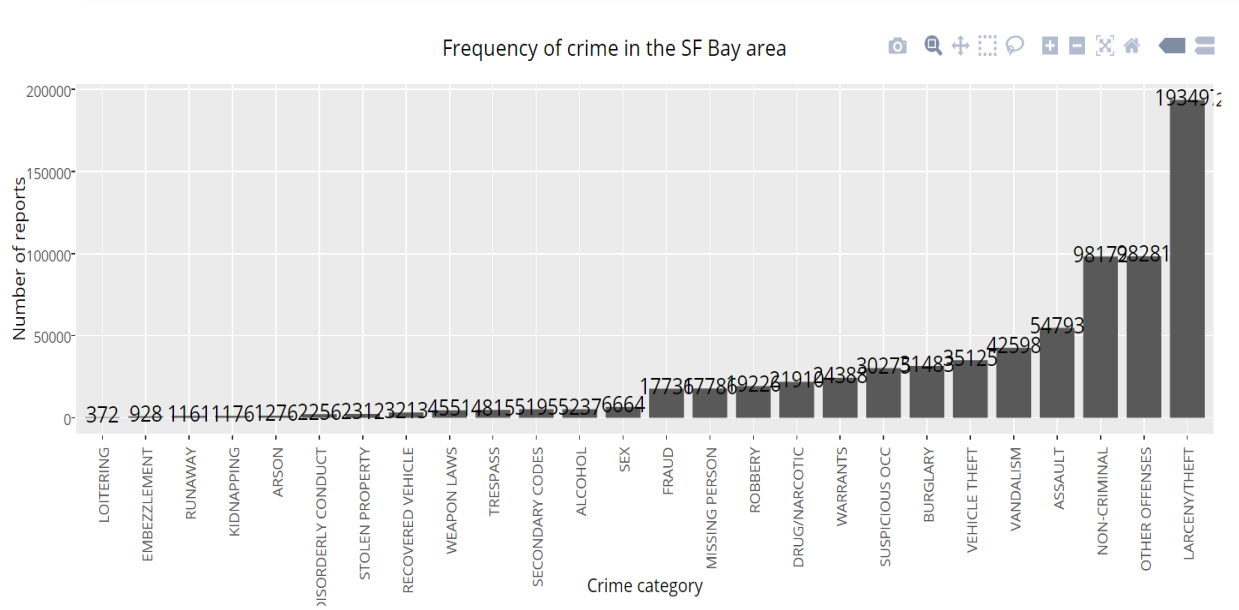
Business Case:

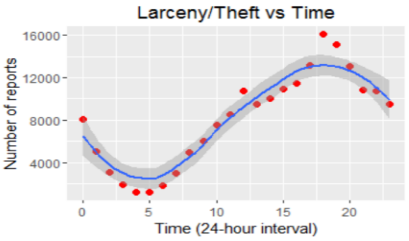
Our analysis could help the police department to get an overall view on the category of crime occurring in a particular area. Based on our analysis the police department could set up extra patrolling/ checks in notorious areas to avoid criminal activities in the city of San Francisco.

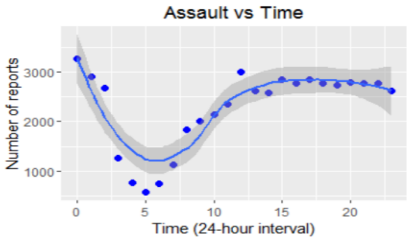
Crime Analysis

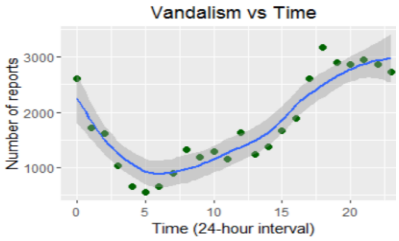


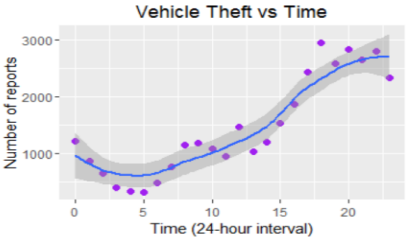
Assignment 2

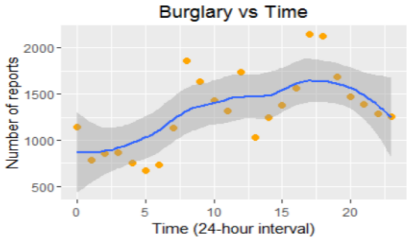


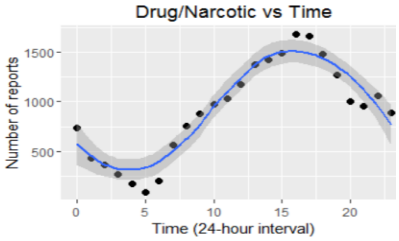


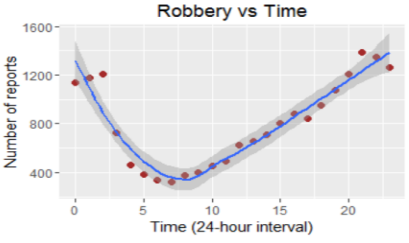




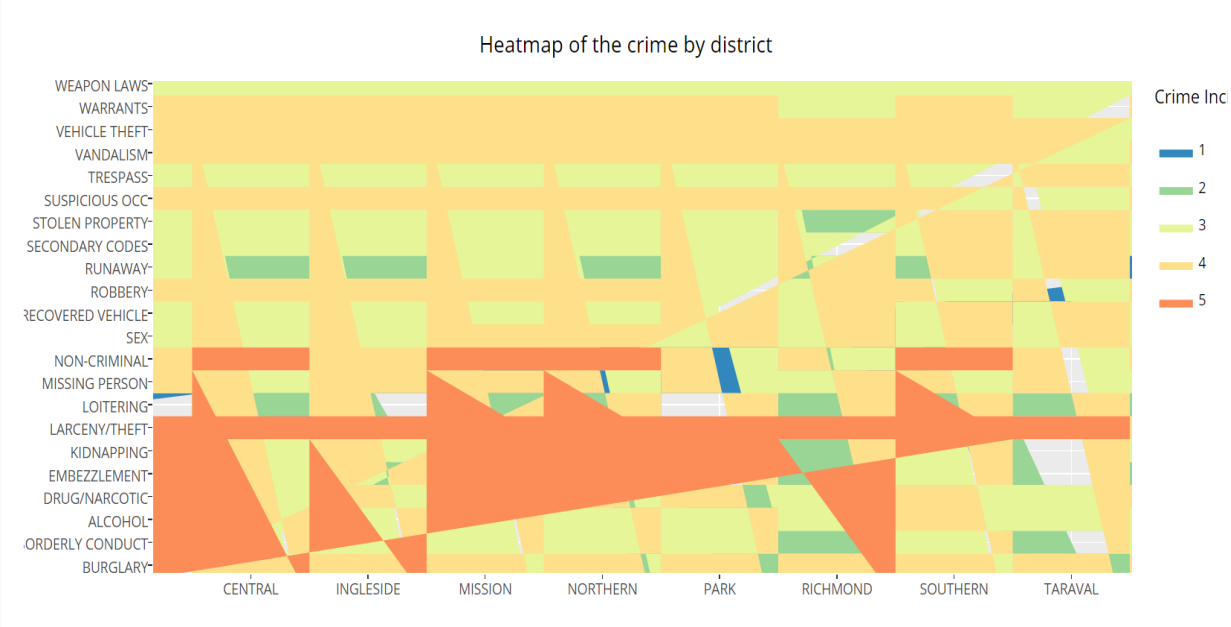








Assignment 2



2. R Lab Code

a. 0-Intro.R

Exercise: select 2x2 subsection from the "bottom left" of matrix mm
Solution:

mm[c(3,4), c(1,2)]

Output

```
> ## exercise: select 2x2 subsection from the "bottom left" of matrix mm
> ## [your code here]
> mm[c(3,4),c(1,2)]
     [,1] [,2]
[1,]    3    7
[2,]    4    8
```

b. 1-data.R

Exercise:
obtain this data view from "df":

	X	Grad.Rate
1	James Madison University	98
2	Incarnate Word College	95
3	Johns Hopkins University	90
4	John Carroll University	89
5	Kenyon College	88
6	King's College	87

Assignment 2

Solution:

arrange(select(df, X, Grad.Rate), desc(Grad.Rate))

```
> arrange(select(df, X, Grad.Rate), desc(Grad.Rate))
                                          X Grad.Rate
1                  James Madison University         98
2                    Incarnate Word College         95
3                  Johns Hopkins University         90
4                   John Carroll University         89
5                            Kenyon College         88
6                            King's College         87
7                        La Salle University         84
8              Illinois Wesleyan University         83
9                           Juniata College         80
10                             Knox College         79
11                          Ithaca College         75
12                    John Brown University         75
13                       Immaculata College         69
14 Indiana University at Bloomington         68
15         Kansas Wesleyan University         68
16                           Iona College         66
17                    Iowa State University         65
18                           King College         65
19                 Kentucky Wesleyan College         62
```

Exercise: extract distinct (unique) rows

select(df, S.F.Ratio)

Solution:

unique(select(df, S.F.Ratio))

```
> unique(select(df, S.F.Ratio))
    S.F.Ratio
270      21.0
271      12.9
272      11.2
273      11.4
274      16.6
275      21.3
276      39.8
277      16.0
278      19.2
279      11.5
280      17.9
281      17.0
282      14.9
283      13.3
284      14.5
285       3.3
286      14.4
287      10.6
288      12.7
289      18.5
290      12.4
291      16.1
292      11.3
```

Assignment 2

Exercise: find max and min tuition ("Outstate") grouped by private/public school, in dataset 'df' and 'college'

DF:

Private max min

1 No 9766 3946

2 Yes 19240 6398

college:

Private max min

1 No 15732 2580

2 Yes 21700 2340

Solution:

```
dfx <- group_by(df, Private)
dfx <- summarise(dfx, max=max(Outstate), min = min(Outstate))
dfx
```

```
> dfx <- group_by(df, Private)
> dfx <- summarise(dfx, max=max(Outstate), min = min(Outstate))
> dfx
# A tibble: 2 x 3
  Private   max   min
   <fctr> <int> <int>
1      No  9766  3946
2     Yes 19240  6398
```

```
collegex <- group_by(college, Private)
collegex <- summarise(collegex, max=max(Outstate), min = min(Outstate))
college
```

```
> collegex <- group_by(college, Private)
> collegex <- summarise(collegex, max=max(Outstate), min = min(Outstate))
> collegex
# A tibble: 2 x 3
  Private   max   min
   <fctr> <int> <int>
1      No 15732  2580
2     Yes 21700  2340
```

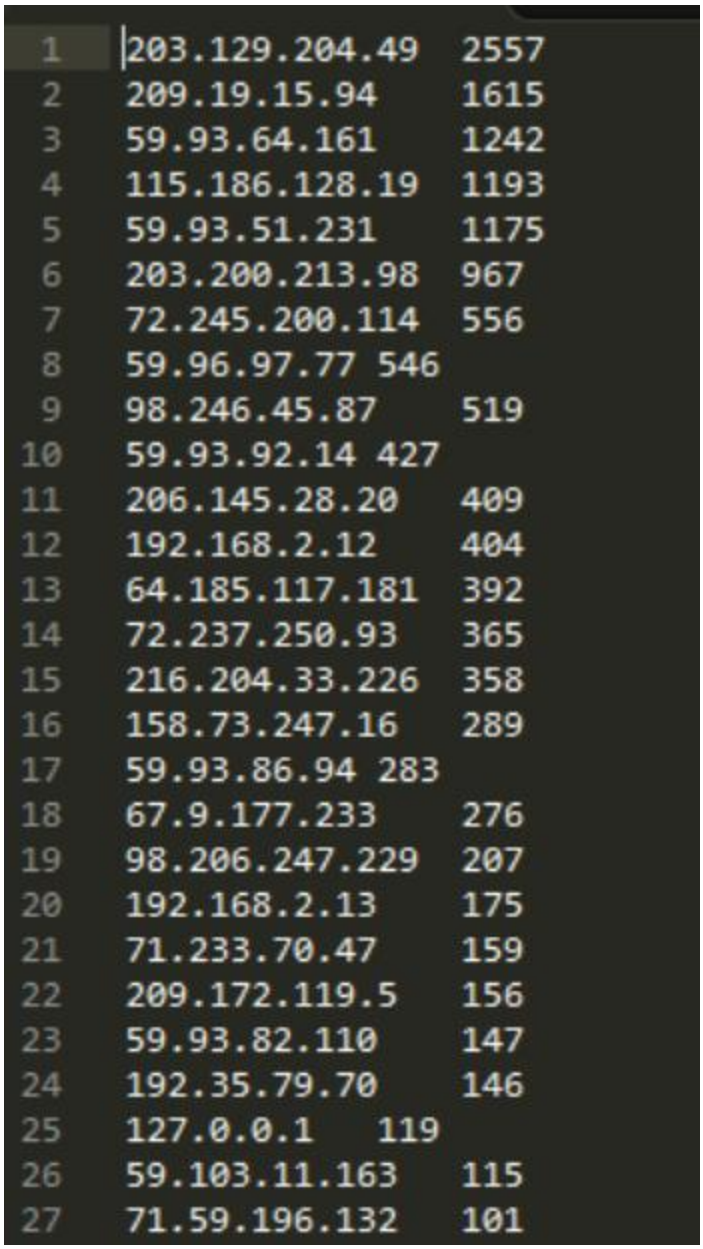
3. Hadoop Lab 3

Commands to run the code

Within the directory Hadoop Workspace/Lab 3

- `javac -classpath ../../hadoop-0.20.2/hadoop-0.20.2-core.jar -d ipcount_classes Runner.java`
- `jar cvf ipcount.jar -C ipcount_classes/ .`
- `../../hadoop-0.20.2/bin/hadoop jar ipcount.jar Cloud.ApacheLog.Runner input output`

Assignment 2



```
1   203.129.204.49  2557
2   209.19.15.94    1615
3   59.93.64.161    1242
4   115.186.128.19  1193
5   59.93.51.231    1175
6   203.200.213.98  967
7   72.245.200.114  556
8   59.96.97.77 546
9   98.246.45.87    519
10  59.93.92.14 427
11  206.145.28.20   409
12  192.168.2.12    404
13  64.185.117.181  392
14  72.237.250.93   365
15  216.204.33.226  358
16  158.73.247.16   289
17  59.93.86.94 283
18  67.9.177.233    276
19  98.206.247.229  207
20  192.168.2.13    175
21  71.233.70.47    159
22  209.172.119.5   156
23  59.93.82.110    147
24  192.35.79.70    146
25  127.0.0.1   119
26  59.103.11.163   115
27  71.59.196.132   101
```