STEVENS
INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY

Predicting Facebook Check-ins

Divya Rathore

Shruti Tripathi

Instructor: Khasha Dehnad

CS-513A

Knowledge Discovery in Databases

STEVENS
INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®

Introduction

Our goal is to predict which place a person would likely check-in.

Facebook created the artificial data consisting of more than 100,000 places located in a 10 km by 10 km square.
This data was fabricated to resemble location signals coming from mobile devices, giving us a flavor of what it takes to work with real data complicated by inaccurate and noisy values

For a given set of coordinates, our goal is to find out the most likely checked in places.

Data

From Kaggle.com

29 million rows and 5 columns

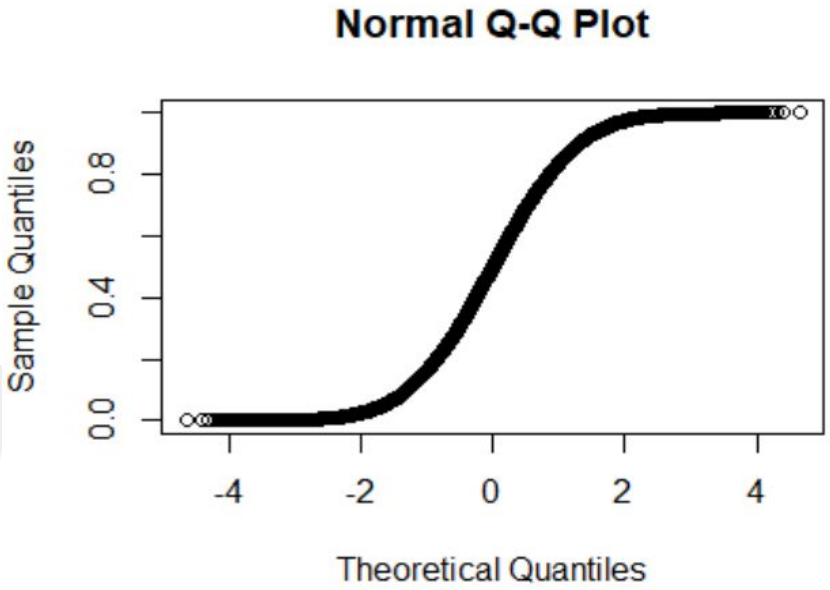
- *x & y: coordinates*
  - X&Y are bounded between the range from 0 to 10
- *accuracy: location accuracy*
- *time: timestamp*
  - Timestamp of check-in's in minutes
- *place_id: id of the business, this is the target we are predicting.*
  - Place Id are identifiers for approx 100,000 uniques places

Data Preprocessing

We sliced the data for top six Place IDs with most frequent check ins from all the places given in the dataset.

```
> summary(fbData)
     row_id              x                  y                accuracy          time
 Min.   :       0   Min.   : 0.00000   Min.   : 0.000000   Min.   :   1.00000   Min.   :     1.0
 1st Qu.: 7279505   1st Qu.: 2.53470   1st Qu.: 2.496700   1st Qu.:  27.00000   1st Qu.:203057.0
 Median :14559010   Median : 5.00910   Median : 4.988300   Median :  62.00000   Median :433922.0
 Mean   :14559010   Mean   : 4.99977   Mean   : 5.001814   Mean   :  82.84912   Mean   :417010.4
 3rd Qu.:21838515   3rd Qu.: 7.46140   3rd Qu.: 7.510300   3rd Qu.:  75.00000   3rd Qu.:620491.0
 Max.   :29118020   Max.   :10.(                                                Max.   :239.0

       place_id
 8772469670:    1849
 1623394281:    1802
 1308450003:    1757
 4823777529:    1738
 9586338177:    1718
 9129780742:    1716
 (Other)   :29107441
> sum(is.na(fbData))
[1] 0
```

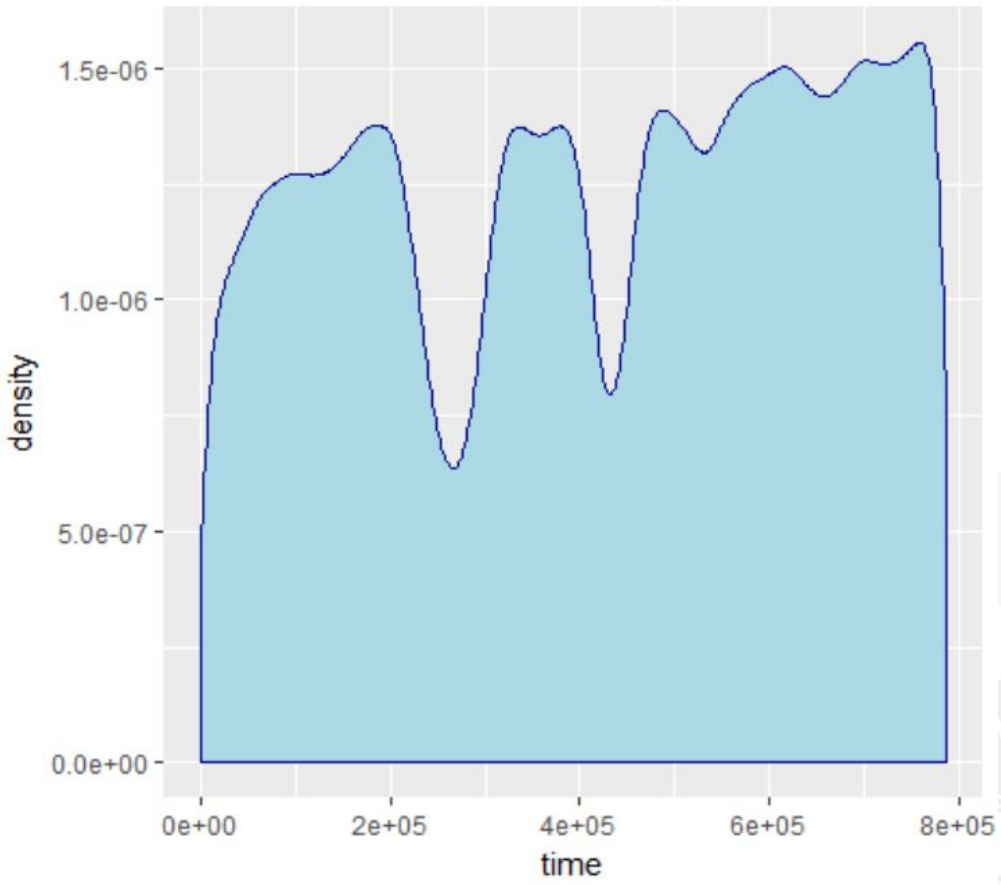



```
minmax <- function(x){(x-min(x))/(max(x)-min(x))}

fbCheckinData$trans_X <- minmax(fbCheckinData$x)
fbCheckinData$trans_Y <- minmax(fbCheckinData$y)
fbCheckinData$trans_accuracy <- minmax(fbCheckinData$accuracy)
fbCheckinData$trans_time <- minmax(fbCheckinData$time)
```

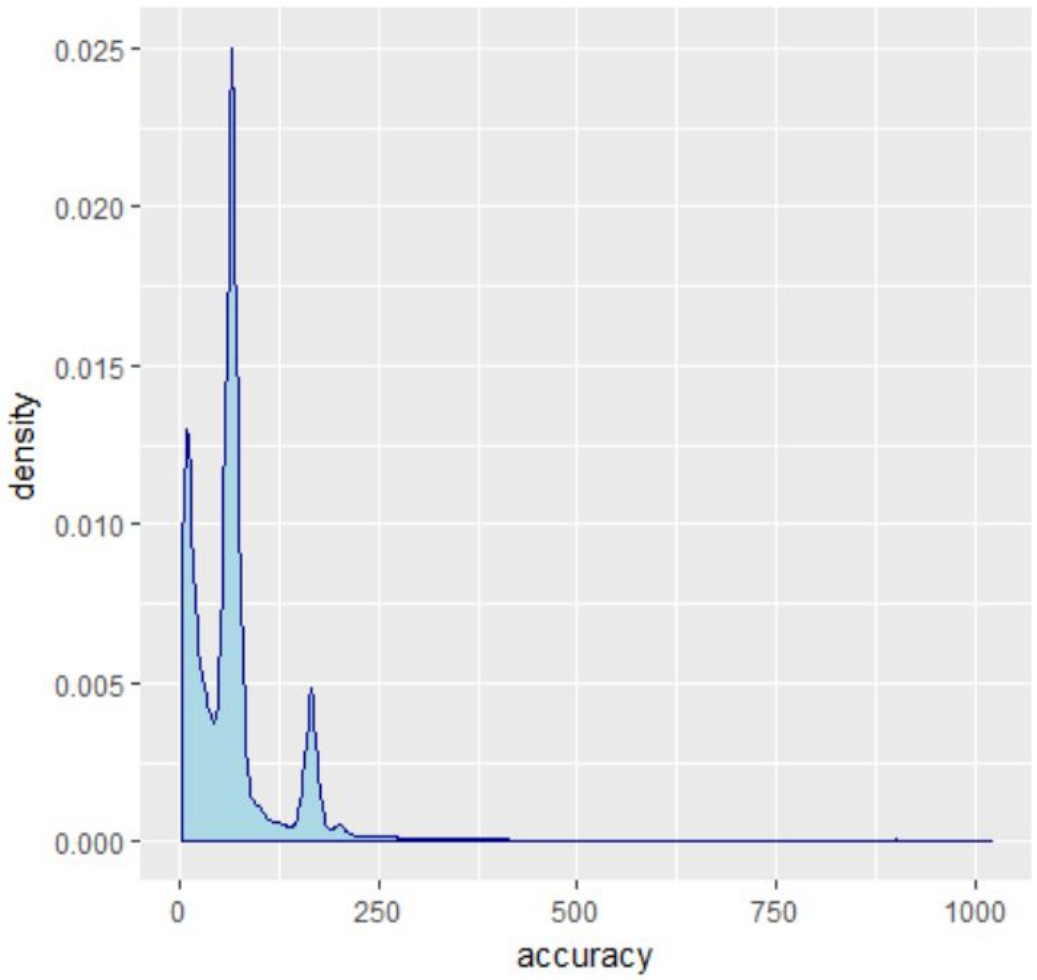
Time-Density plot

- Interestingly time variable is evenly dispersed.
- There are two big dips in the plot.
- We breakdown time variable to hours, days, weekdays, months and year.



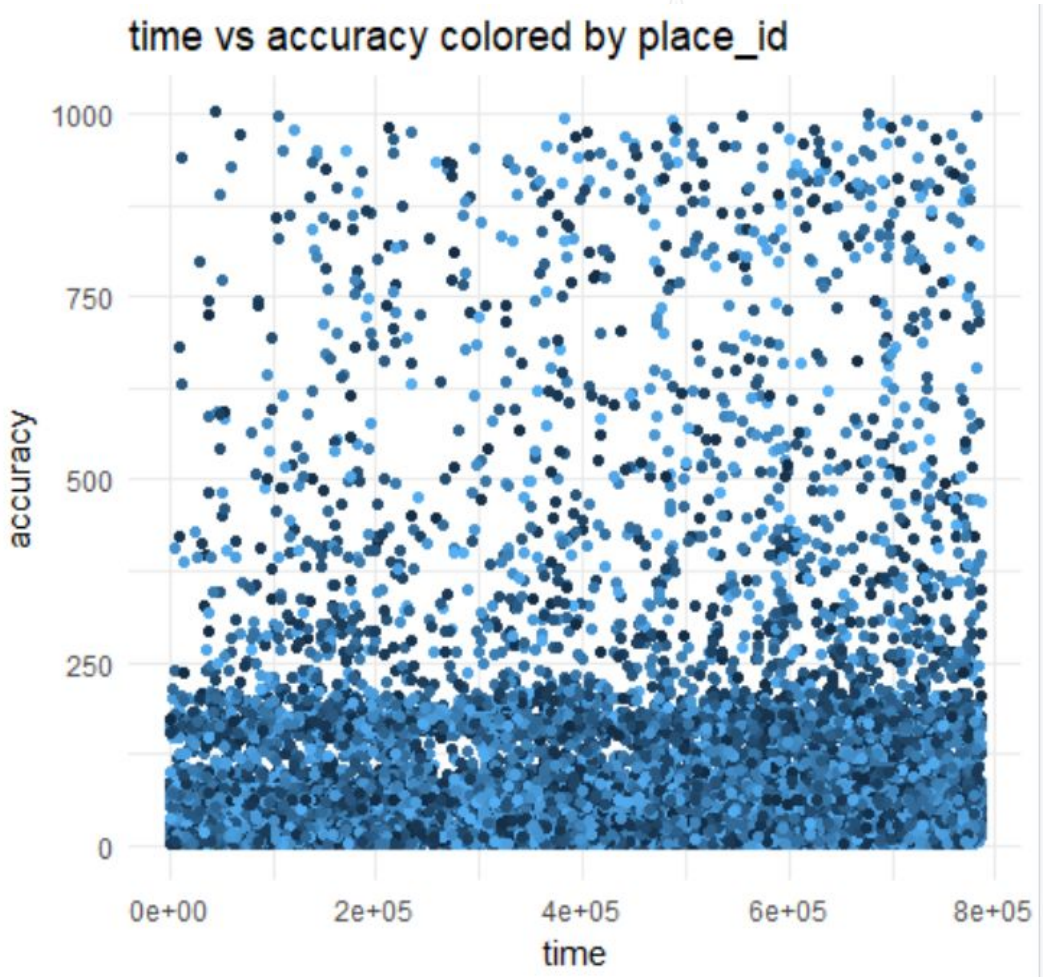
Accuracy Density Plot

- Most check-ins have accuracy between the rage from 0 to 200.
- There are 3 peaks in the plot.
- We might infer that we have different accuracy at different locations.



Time vs Accuracy

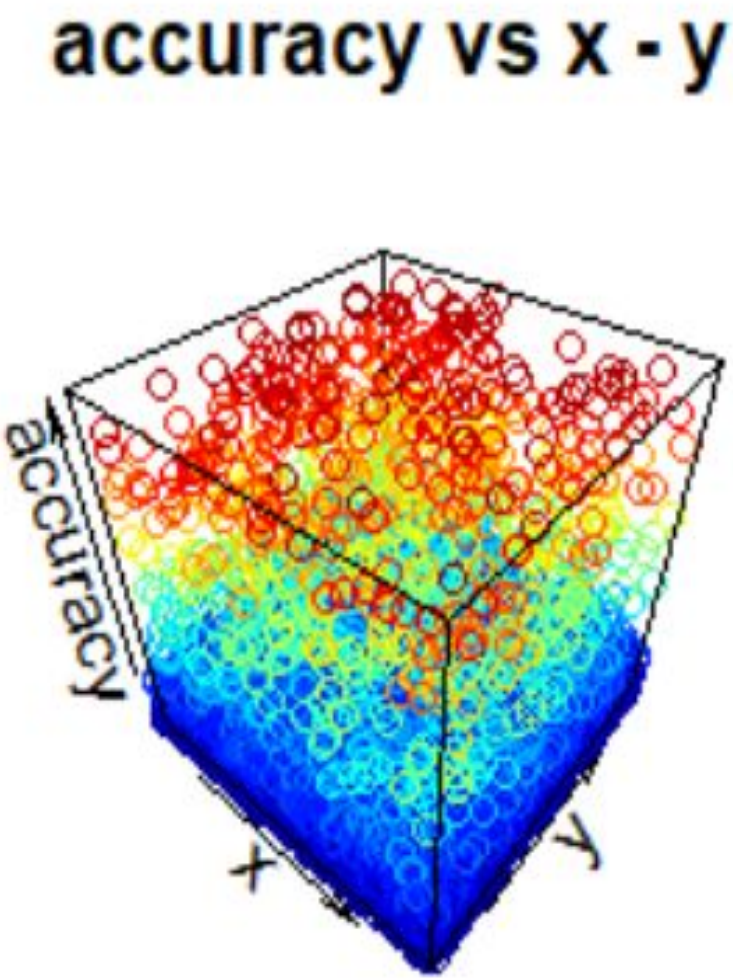
- No visible relation between time and accuracy.
- Distribution of accuracy is uniform throughout time range
- Irrespective of time, trend of accuracy stays the same.



STEVENS
INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY

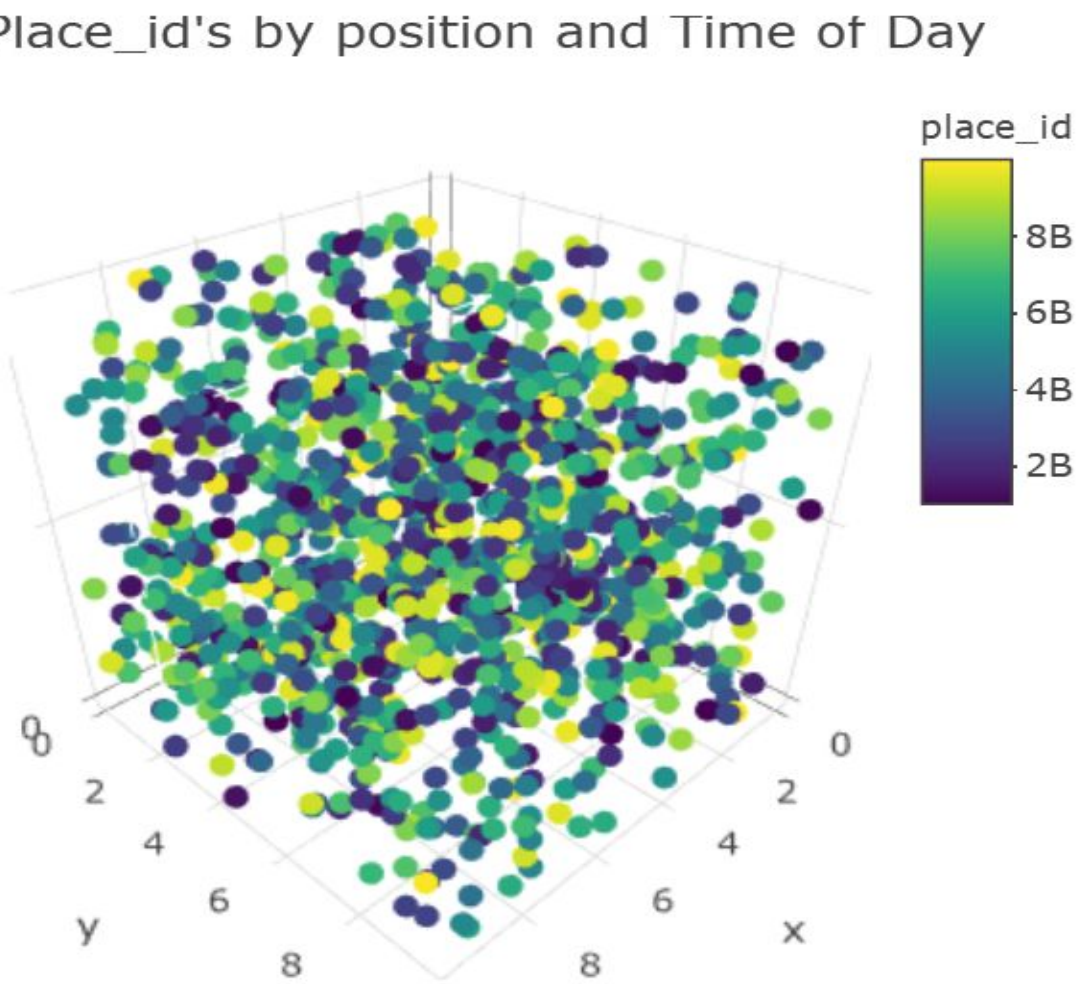
3D Plot

- Accuracy is different for different places.
- Most of the Place IDs have low accuracy, only a selected few have higher accuracies.



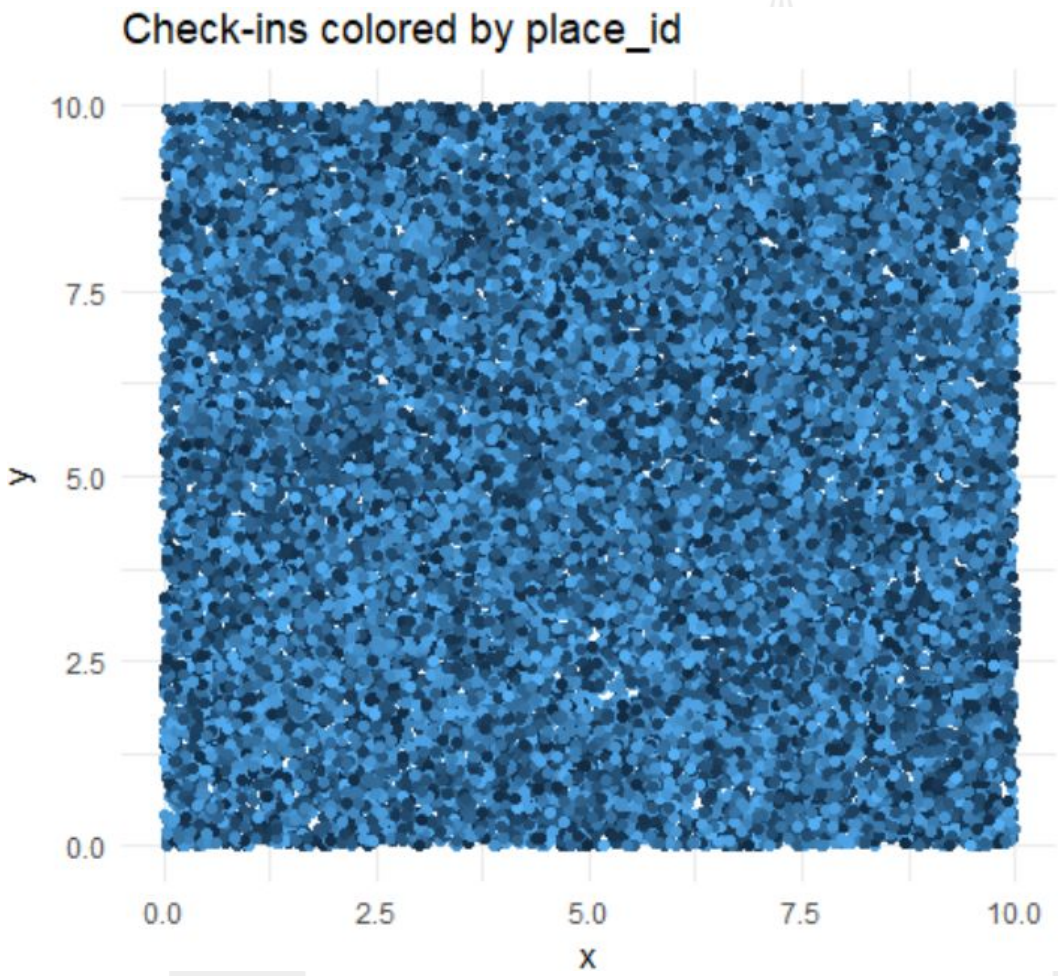
3D Plot *X-Y with Hour*

- The Check-ins appears to be evenly dispersed when we are considering a small chunk of time.



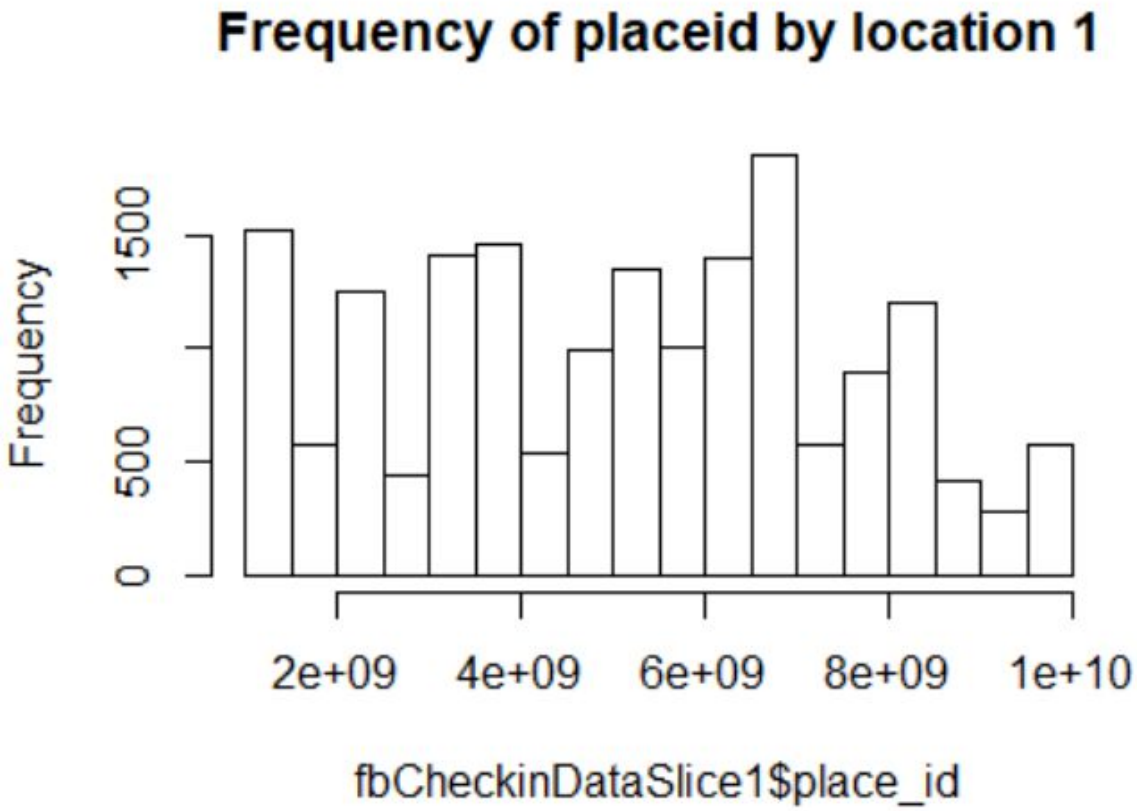
X-Y by PlaceID

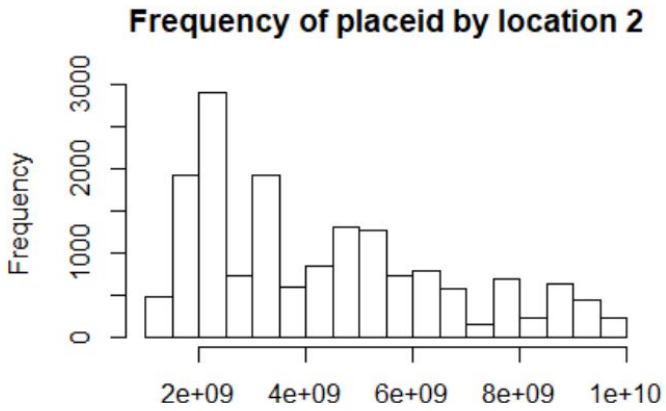
Place Ids are homogeneously distributed across coordinates.

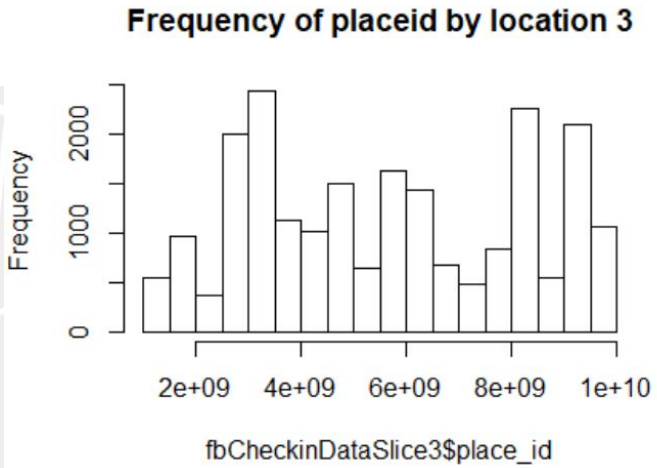


STEVENS
INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY

Getting a closer look:

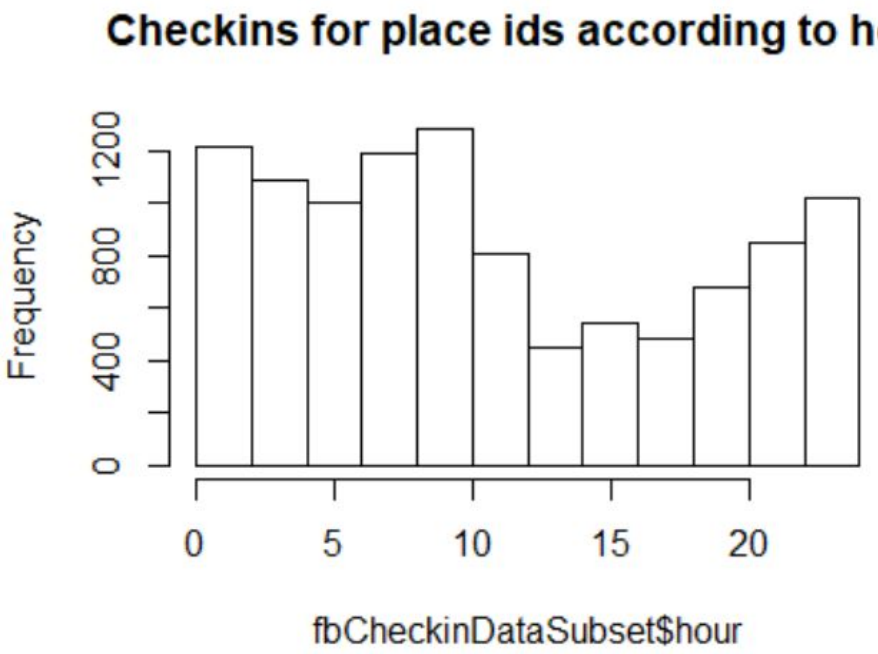


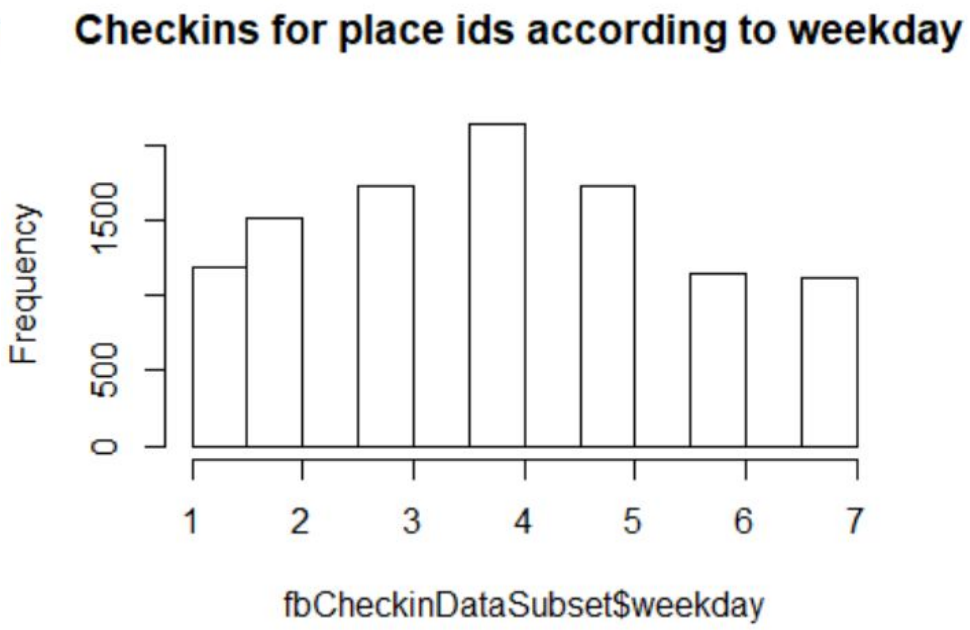




Check-Ins with Hour & Weekdays

Analysing hourly and weekly trends.





STEVENS
INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®

Classification Models

k nearest Neighbor

Naive Bayes

ANN

Random Forest

STEVENS
INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY

k-Nearest Neighbor

K nearest neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions).

Error with this model taking k=3 is 5.7
Accuracy Rate= 94.3%

```
> table(knn_fb,place_id=testing$place_id)
            place_id
knn_fb       1308450003 1623394281 4823777529 8772469670 9129780742 9586338177
  1308450003        491          0          1          0          0          0
  1623394281          0        537          0          0          0          0
  4823777529          1          0        525          0          0          6
  8772469670          0          0          0        477         76          0
  9129780742          0          0          0         94        448          0
  9586338177          1          0          4          1          0        512
> knn_error_rate
[1] 0.05797101449
>
```

Naive Bayes

The probability of a place given its features can be expressed in the form
p(place | x, y, time, accuracy) ∝ p(x, y, time, accuracy | place)p(place)
We reason that the time, location and accuracy features are independent given the place.

Error rate with this model: 1.2
Accuracy= 98%

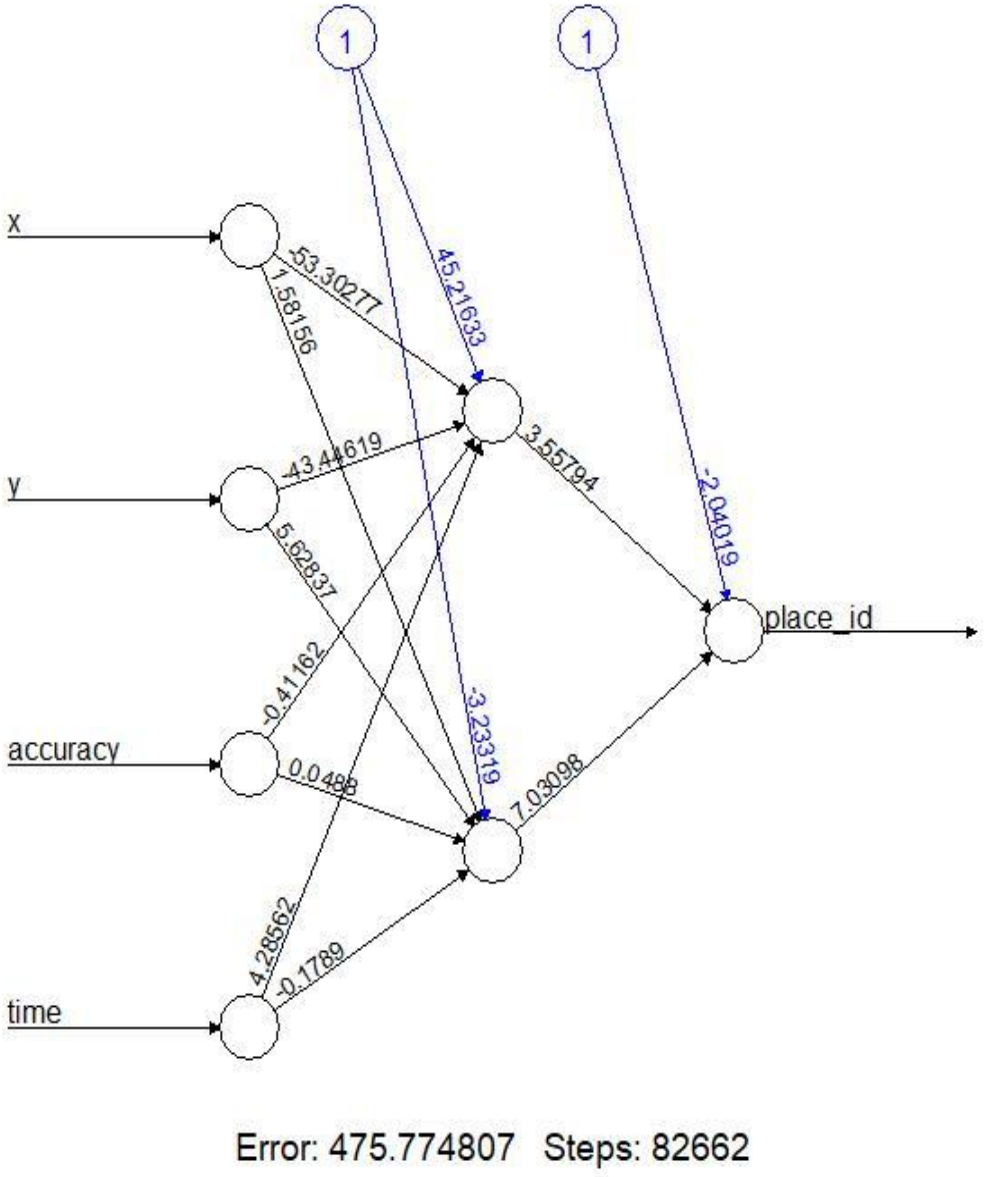
```
> table(NBayes=category,place_id=testing$place_id)
          place_id
NBayes     1308450003 1623394281 4823777529 8772469670 9129780742 9586338177
  1308450003        489          0          0          0          1          1
  1623394281          0        529          1          0          0          0
  4823777529          3          1        526          5          3          1
  8772469670          0          0          2        560          0          0
  9129780742          0          0          0          0        516          0
  9586338177          1          7          1          7          4        516
> NB_error_rate
[1] 0.01197227473
```

STEVENS
INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY

Neural Network

The neural network algorithm is the backpropagation algorithm. It is modeled loosely after the human brain and is designed to recognize patterns.

Error rate with random forest : 2.6
Accuracy = 97.4%



STEVENS
INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY

Random Forest

Random forest is a supervised learning algorithm which builds multiple decision trees and merges them together to get a more accurate and stable prediction.

Error rate with random forest : 0.2
Accuracy=99%

```
> table(actual=testing$place_id ,Prediction)
           Prediction
actual      1308450003 1623394281 4823777529 8772469670 9129780742 9586338177
  1308450003        493          0          0          0          0          0
  1623394281          0        537          0          0          0          0
  4823777529          0          0        530          0          0          0
  8772469670          0          0          0        570          2          0
  9129780742          0          0          0          3        521          0
  9586338177          0          0          3          0          0        515
> error_rate
[1] 0.002520478891
>
```

STEVENS
INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY

Conclusion

For a given set of coordinates, the places most likely checked in depend not only on location, but on time and accuracy of check in as well

STEVENS
INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®

stevens.edu

Thank You!