Basic Data Analysis and Visualization in R

Joe Brillantes

R Users Group – Philippines

Joe Brillantes

Management Scientist: Improve decision-making through rigorous analytical methods

PH Data & Analytics Lead, FrieslandCampina (Alaska Milk)

Led a team of 7 composing of 2 data engineers, 4 business analytics specialists, and 1 data operations engineer that enabled PHP326M (USD5.8M), and PHP507M (USD9M) in business value in 2021 and 2022, respectively

Operations Advanced Analytics Manager, Dyson

Implemented 25K individual control charts, and detection of violations to Nelson Rules in BigQuery and R (backend), and Tableau (frontend) for a Statistical Process Control implementation, which could decrease the cost of production by GBP17M

Analyst - Business Analytics, Lattice Semiconductor

Created statistical models of demand for 300 product lines in R, automated five-quarter demand forecast production in under 8 hours previously from 48 hours, and decreased quarterly Plan of Record variation by 25%

Business Analyst, Maersk

Mathematically programmed South China Cluster's Export Customer Portfolio to minimize variation of returns for a target return, which had the potential to improve profits by USD2.6M in a quarter

ATENEO

M. Mathematics (Business Track)

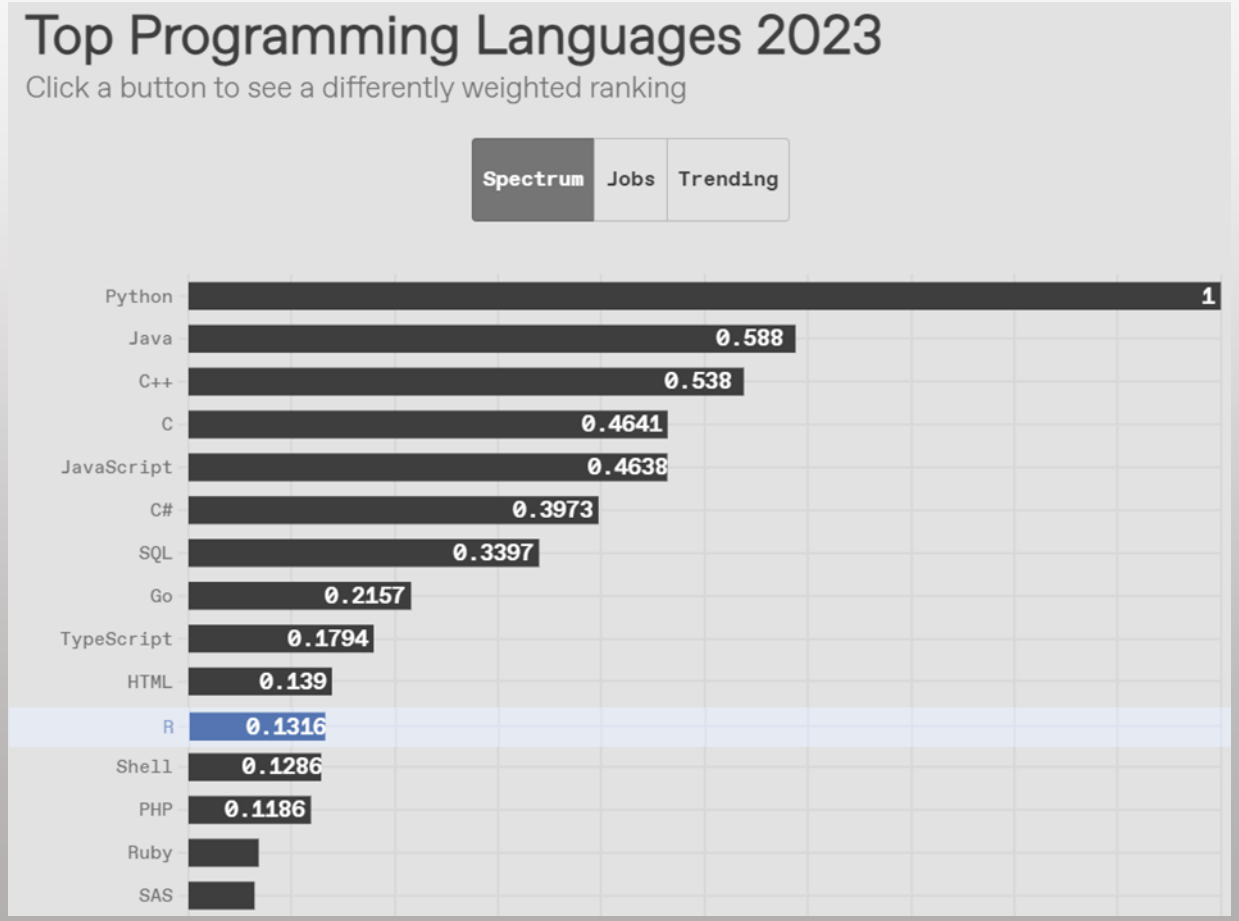
BYU HAWAII

B.S. Mathematics Education

Summary

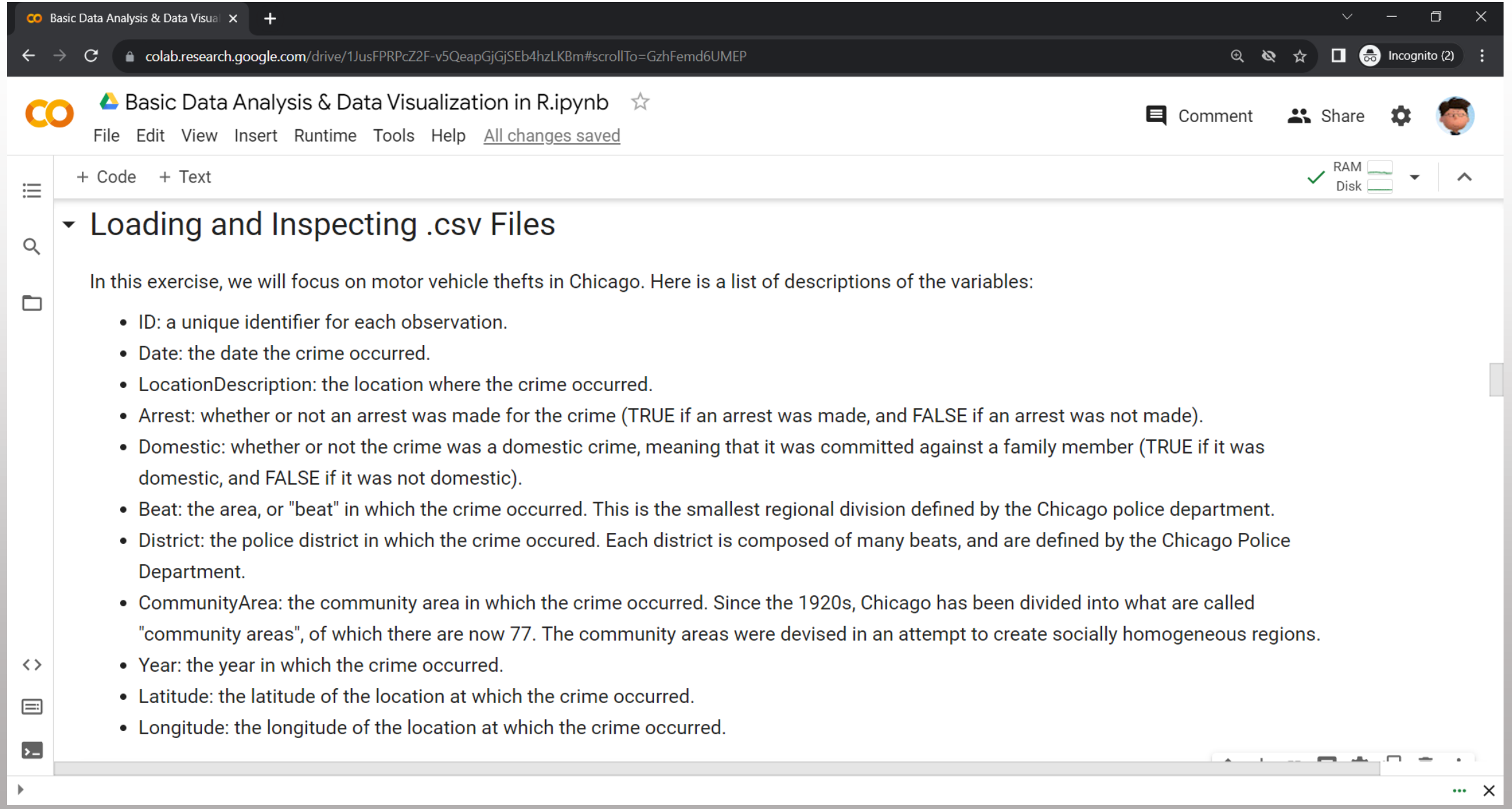
15 years of business analytics experience; Proficient in Basic Statistics, Econometrics, Elastic-Net Regularized Generalized Linear Models, Decision Trees, Random Forest, Gradient Boosting Machines, K-means and Hierarchical Clustering, Principal Components Analysis, Linear Programming, Modern Portfolio Theory, R, SQL, BigQuery, AzureML and PySpark

R is the most popular programming language for statistical computing.



https://spectrum.ieee.org/the-top-programming-languages-2023

"R [...] came to prominence with the rise of big data several years ago. Although powerful, it's not easy to learn, with enigmatic syntax and functions typically being performed on entire vectors, lists, and other high-level data structures. But although there are Python libraries that provide similar analytic and graphical functionality, R has remained popular, likely precisely because of its peculiarities. They make R scripts hard to port, a significant issue given the enormous body of statistical analysis and academic research built on R. Entire fields of researchers and analysts would have to learn a new language and rebuild their work. (Side note: We use R to crunch the numbers for the TPL.)"


Basic Data Analysis & Data Visual
colab.research.google.com/drive/1JusFPRPcZ2F-v5QeapGjGjSEb4hzLKBm#scrollTo=GzhFemd6UMEP
Incognito (2)
Basic Data Analysis & Data Visualization in R.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comment
Share
+ Code + Text
RAM
Disk
Loading and Inspecting .csv Files
In this exercise, we will focus on motor vehicle thefts in Chicago. Here is a list of descriptions of the variables:
ID: a unique identifier for each observation.
Date: the date the crime occurred.
LocationDescription: the location where the crime occurred.
Arrest: whether or not an arrest was made for the crime (TRUE if an arrest was made, and FALSE if an arrest was not made).
Domestic: whether or not the crime was a domestic crime, meaning that it was committed against a family member (TRUE if it was domestic, and FALSE if it was not domestic).
Beat: the area, or "beat" in which the crime occurred. This is the smallest regional division defined by the Chicago police department.
District: the police district in which the crime occured. Each district is composed of many beats, and are defined by the Chicago Police Department.
CommunityArea: the community area in which the crime occurred. Since the 1920s, Chicago has been divided into what are called "community areas", of which there are now 77. The community areas were devised in an attempt to create socially homogeneous regions.
Year: the year in which the crime occurred.
Latitude: the latitude of the location at which the crime occurred.
Longitude: the longitude of the location at which the crime occurred.

Basic Data Analysis & Data Visual

colab.research.google.com/drive/1JusFPRPcZ2F-v5QeapGjGjSEb4hzLKBm#scrollTo=GzhFemd6UMEP

Incognito (2)

Basic Data Analysis & Data Visualization in R.ipynb

File Edit View Insert Runtime Tools Help All changes saved

Comment Share

+ Code + Text

RAM Disk

```
mvt <- read.csv('https://ocw.mit.edu/courses/15-071-the-analytics-edge-spring-2017/123f9aa4885b259db7f3aef5153835de_mvtWeek1.csv',
                header = TRUE, sep = ',', stringsAsFactors = FALSE)
head(mvt)
str(mvt)
summary(mvt)
```

A data.frame: 6 × 11

	ID	Date	LocationDescription	Arrest	Domestic	Beat	District	CommunityArea	Year	Latitude	Longitude
	<int>	<chr>	<chr>	<lgl>	<lgl>	<int>	<int>	<int>	<int>	<dbl>	<dbl>
1	8951354	12/31/12 23:15	STREET	FALSE	FALSE	623	6	69	2012	41.75628	-87.62164
2	8951141	12/31/12 22:00	STREET	FALSE	FALSE	1213	12	24	2012	41.89879	-87.66130
3	8952745	12/31/12 22:00	RESIDENTIAL YARD (FRONT/BACK)	FALSE	FALSE	1622	16	11	2012	41.96919	-87.76767
4	8952223	12/31/12 22:00	STREET	FALSE	FALSE	724	7	67	2012	41.76933	-87.65773
5	8951608	12/31/12 21:30	STREET	FALSE	FALSE	211	2	35	2012	41.83757	-87.62176
6	8950793	12/31/12 20:30	STREET	TRUE	FALSE	2521	25	19	2012	41.92856	-87.75400

```
'data.frame':   191641 obs. of  11 variables:
 $ ID                  : int  8951354 8951141 8952745 8952223 8951608 8950793 8950760 8951611 8951802 8950706 ...
 $ Date                : chr  "12/31/12 23:15" "12/31/12 22:00" "12/31/12 22:00" "12/31/12 22:00" ...
```

Basic Data Analysis & Data Visual

colab.research.google.com/drive/1JusFPRPcZ2F-v5QeapGjGjSEb4hzLKBm#scrollTo=GzhFemd6UMEP

Incognito (2)

Basic Data Analysis & Data Visualization in R.ipynb

File Edit View Insert Runtime Tools Help All changes saved

Comment Share

+ Code + Text

RAM Disk

1s

5	8951608	12/31/12 21:30	STREET	FALSE	FALSE	211	2	35	2012	41.83757	-87.62176
6	8950793	12/31/12 20:30	STREET	TRUE	FALSE	2521	25	19	2012	41.92856	-87.75400

```
'data.frame':   191641 obs. of  11 variables:
$ ID                : int  8951354 8951141 8952745 8952223 8951608 8950793 8950760 8951611 8951802 8950706 ...
$ Date              : chr  "12/31/12 23:15" "12/31/12 22:00" "12/31/12 22:00" "12/31/12 22:00" ...
$ LocationDescription: chr  "STREET" "STREET" "RESIDENTIAL YARD (FRONT/BACK)" "STREET" ...
$ Arrest            : logi  FALSE FALSE FALSE FALSE FALSE TRUE ...
$ Domestic          : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
$ Beat              : int  623 1213 1622 724 211 2521 423 231 1021 1215 ...
$ District          : int  6 12 16 7 2 25 4 2 10 12 ...
$ CommunityArea     : int  69 24 11 67 35 19 48 40 29 24 ...
$ Year              : int  2012 2012 2012 2012 2012 2012 2012 2012 2012 2012 ...
$ Latitude          : num  41.8 41.9 42 41.8 41.8 ...
$ Longitude         : num  -87.6 -87.7 -87.8 -87.7 -87.6 ...
      ID              Date           LocationDescription   Arrest
Min.   :1310022   Length:191641      Length:191641       Mode :logical
1st Qu.:2832144   Class :character   Class :character    FALSE:176105
Median :4762956   Mode  :character   Mode  :character    TRUE :15536
Mean   :4968629
3rd Qu.:7201878
Max.   :9181151

 Domestic           Beat          District     CommunityArea        Year
Mode :logical   Min.   : 111   Min.   : 1.00   Min.   : 0      Min.   :2001
```

Basic Data Analysis & Data Visual

colab.research.google.com/drive/1JusFPRPcZ2F-v5QeapGjGjSEb4hzLKBm#scrollTo=OOoQvT2NMGws

Incognito (2)

Basic Data Analysis & Data Visualization in R.ipynb

File Edit View Insert Runtime Tools Help All changes saved

Comment Share

+ Code + Text

RAM Disk

Correcting Data Types

[7] 1s

```
# Nominal Variables
mvt$ID <- as.character(mvt$ID)
mvt$Beat <- as.character(mvt$Beat)
mvt$District <- as.character(mvt$District)
mvt$CommunityArea <- as.character(mvt$CommunityArea)

# Temporal Variables
mvt$Date <- strptime(mvt$Date, "%m/%d/%y %H:%M")

str(mvt)
summary(mvt)
```

```
'data.frame':   191641 obs. of  11 variables:
 $ ID                : chr  "8951354" "8951141" "8952745" "8952223" ...
 $ Date              : POSIXlt, format: "2012-12-31 23:15:00" "2012-12-31 22:00:00" ...
 $ LocationDescription: chr  "STREET" "STREET" "RESIDENTIAL YARD (FRONT/BACK)" "STREET" ...
 $ Arrest            : logi  FALSE FALSE FALSE FALSE FALSE TRUE ...
 $ Domestic          : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
 $ Beat              : chr  "623" "1213" "1622" "724" ...
 $ District          : chr  "6" "12" "16" "7" ...
 $ CommunityArea     : chr  "69" "24" "11" "67" ...
 $ Year              : int  2012 2012 2012 2012 2012 2012 2012 2012 2012 2012 ...
```

Basic Data Analysis & Data Visual

colab.research.google.com/drive/1JusFPRPcZ2F-v5QeapGjGjSEb4hzLKBm#scrollTo=OOoQvT2NMGws

Incognito (2)

Basic Data Analysis & Data Visualization in R.ipynb

File Edit View Insert Runtime Tools Help All changes saved

Comment Share

+ Code + Text

RAM Disk

[7] 1s

```
$ Latitude            : num  41.8 41.9 42 41.8 41.8 ...
$ Longitude           : num  -87.6 -87.7 -87.8 -87.7 -87.6 ...
      ID                  Date                       LocationDescription
Length:191641      Min.   :2001-01-01 00:01:00.00   Length:191641
Class :character   1st Qu.:2003-07-10 18:00:00.00   Class :character
Mode  :character   Median :2006-05-21 12:30:00.00   Mode  :character
                   Mean   :2006-08-23 13:45:08.91
                   3rd Qu.:2009-10-24 22:40:00.00
                   Max.   :2012-12-31 23:15:00.00

  Arrest          Domestic           Beat              District
Mode :logical   Mode :logical   Length:191641      Length:191641
FALSE:176105    FALSE:191226    Class :character   Class :character
TRUE :15536     TRUE :415       Mode  :character   Mode  :character



CommunityArea          Year         Latitude       Longitude
Length:191641      Min.   :2001   Min.   :41.64   Min.   :-87.93
Class :character   1st Qu.:2003   1st Qu.:41.77   1st Qu.:-87.72
Mode  :character   Median :2006   Median :41.85   Median :-87.68
                   Mean   :2006   Mean   :41.84   Mean   :-87.68
                   3rd Qu.:2009   3rd Qu.:41.92   3rd Qu.:-87.64
                   Max.   :2012   Max.   :42.02   Max.   :-87.52
                                  NA's   :2276    NA's   :2276
```

Basic Data Analysis & Data Visual

colab.research.google.com/drive/1JusFPRPcZ2F-v5QeapGjGjSEb4hzLKBm#scrollTo=ew_nNNGoBApG

Incognito (2)

Basic Data Analysis & Data Visualization in R.ipynb

File Edit View Insert Runtime Tools Help

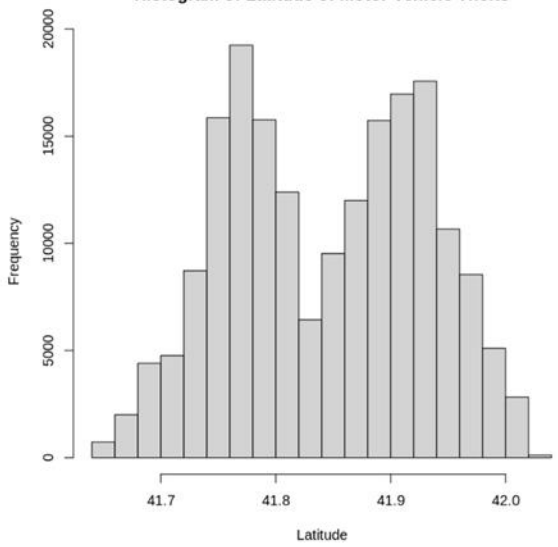
Comment Share

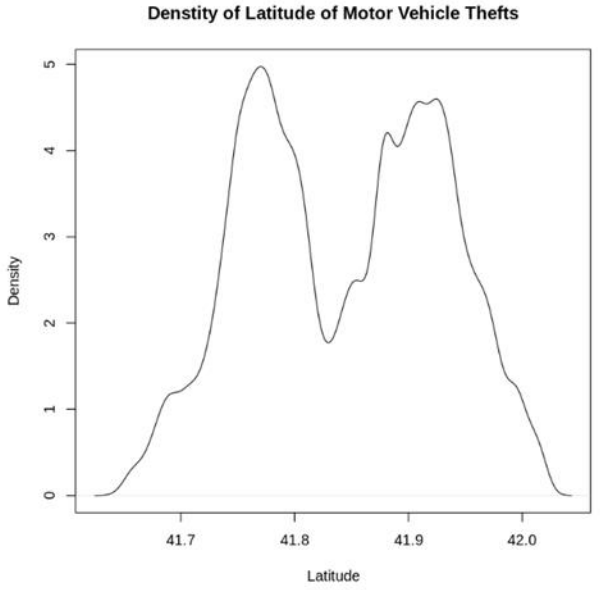
+ Code + Text

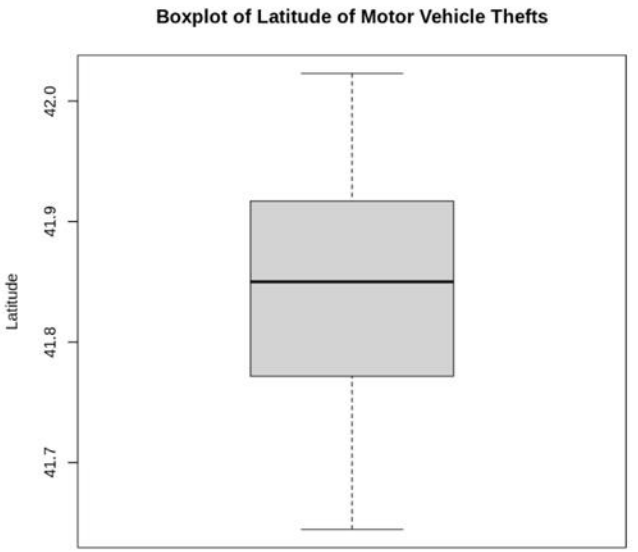
RAM Disk

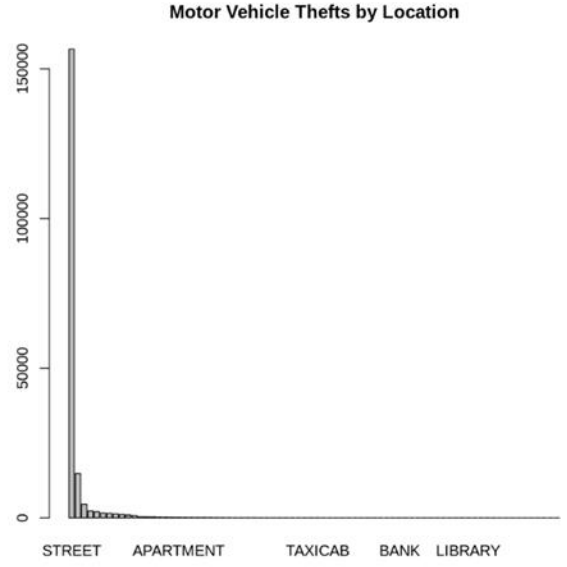
Basic Plots

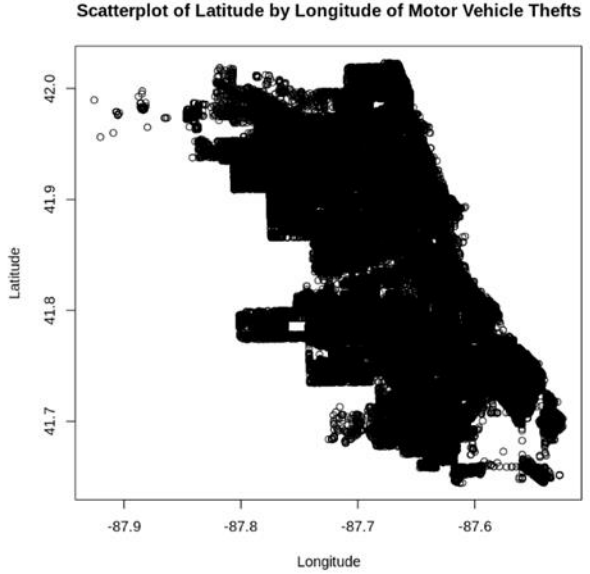
```
# Histogram
hist(mvt$Latitude, main='Histogram of Latitude of Motor Vehicle Thefts', xlab = 'Latitude')

# Kernel Density plot
plot(density(mvt$Latitude[!is.na(mvt$Latitude)]), main = 'Denstity of Latitude of Motor Vehicle Thefts', xlab = 'Latitude')

# Box Plot
boxplot(mvt$Latitude, main='Boxplot of Latitude of Motor Vehicle Thefts', ylab='Latitude')

# Bar Plot
LocDesc <- table(mvt$LocationDescription)
barplot(LocDesc[order(LocDesc, decreasing=TRUE)], main='Motor Vehicle Thefts by Location')

# Scatter Plot
plot(x=mvt$Longitude, y=mvt$Latitude, main='Scatterplot of Latitude by Longitude of Motor Vehicle Thefts', xlab='Longitude', ylab='Latitude')

# SmoothScatter Plot
smoothScatter(x=mvt$Longitude, y=mvt$Latitude, main='Smooth Scatterplot of Latitude by Longitude of Motor Vehicle Thefts',
              xlab='Longitude', ylab='Latitude')
```

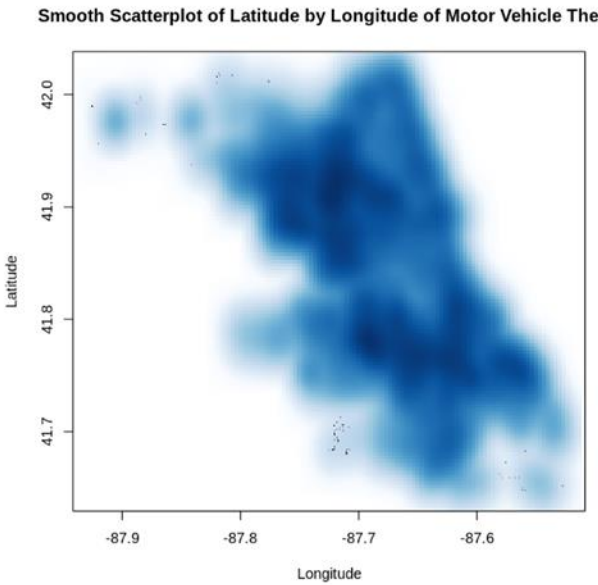

Histogram of Latitude of Motor Vehicle Thefts
Frequency
0
5000
10000
15000
20000
41.7
41.8
41.9
42.0
Latitude


Denstity of Latitude of Motor Vehicle Thefts
Density
0
1
2
3
4
5
41.7
41.8
41.9
42.0
Latitude


Boxplot of Latitude of Motor Vehicle Thefts
Latitude
41.7
41.8
41.9
42.0


Motor Vehicle Thefts by Location
0
50000
100000
150000
STREET
APARTMENT
TAXICAB
BANK
LIBRARY


Scatterplot of Latitude by Longitude of Motor Vehicle Thefts
Latitude
41.7
41.8
41.9
42.0
-87.9
-87.8
-87.7
-87.6
Longitude


Smooth Scatterplot of Latitude by Longitude of Motor Vehicle Thefts
Latitude
41.7
41.8
41.9
42.0
-87.9
-87.8
-87.7
-87.6
Longitude

Basic Data Analysis & Data Visual

colab.research.google.com/drive/1JusFPRPcZ2F-v5QeapGjGjSEb4hzLKBm#scrollTo=FnR4iwsbLV0X

Incognito (2)

Basic Data Analysis & Data Visualization in R.ipynb

File Edit View Insert Runtime Tools Help All changes saved

Comment Share

+ Code + Text

RAM Disk

Basic Statistical Test

[9] 0s

```
# Are motor vehicle thefts equally common across police districts?
mvt.district <- table(mvt$District)
mvt.eqprop <- rep(1/length(mvt.district), times=length(mvt.district))
mvt.district
mvt.eqprop
chisq.test(mvt.district, p=mvt.eqprop)
# Motor vehicle thefts are not equally common across police districts.
# Some police districts have more motor vehicle thefts than other police districts.
```

```
    1    10    11    12    13    14    15    16    17    18    19     2    20
 2598  6374  7805  4090  4013  7898  4795  4959  6684  3425  5797  7571  2494
   21    22    23    24    25     3    31     4     5     6     7     8     9
   29  4290     5  3930 12824  6225     1  7073  5568  8400  8831 13058  9848
0.0384615384615385 · 0.0384615384615385 · 0.0384615384615385 · 0.0384615384615385 · 0.0384615384615385 · 0.0384615384615385 · 0.0384615384615385 ·
0.0384615384615385 · 0.0384615384615385 · 0.0384615384615385 · 0.0384615384615385 · 0.0384615384615385 · 0.0384615384615385 · 0.0384615384615385 ·
0.0384615384615385 · 0.0384615384615385 · 0.0384615384615385 · 0.0384615384615385 · 0.0384615384615385 · 0.0384615384615385 · 0.0384615384615385 ·
0.0384615384615385 · 0.0384615384615385 · 0.0384615384615385 · 0.0384615384615385 · 0.0384615384615385

        Chi-squared test for given probabilities

data:  mvt.district
X-squared = 50681, df = 25, p-value < 2.2e-16
```

Basic Data Analysis & Data Visual

colab.research.google.com/drive/1JusFPRPcZ2F-v5QeapGjGjSEb4hzLKBm#scrollTo=ew_nNNGoBApG

Incognito (2)

Basic Data Analysis & Data Visualization in R.ipynb

File Edit View Insert Runtime Tools Help All changes saved

Comment Share

+ Code + Text

RAM Disk

Basic Temporal Analysis

```
[9]  # What is the trend of motor vehicle thefts occurence across the years?
     # Are motor vehicle thefts seasonal?
     options(repr.plot.height=8, repr.plot.width=15)
     mvt$YrMonth <- as.Date(format(as.Date(mvt$Date), "%Y-%m-01"))
     mvt
     mvtByYrMonth <- aggregate(ID ~ YrMonth, data = mvt, FUN=function(x) {length(unique(x))})
     mvtByYrMonth
     startYrMonth <- min(mvtByYrMonth$YrMonth)
     mvtByYrMonth.ts <- ts(mvtByYrMonth$ID, start = startYrMonth, frequency = 12)
     plot(stl(mvtByYrMonth.ts, s.window='periodic'))
```

A data.frame: 191641 × 12

ID	Date	LocationDescription	Arrest	Domestic	Beat	District	CommunityArea	Year	Latitude	Longitude	YrMonth
<chr>	<dttm>	<chr>	<lgl>	<lgl>	<chr>	<chr>	<chr>	<int>	<dbl>	<dbl>	<date>
8951354	2012-12-31 23:15:00	STREET	FALSE	FALSE	623	6	69	2012	41.75628	-87.62164	2012-12-01
8951141	2012-12-31 22:00:00	STREET	FALSE	FALSE	1213	12	24	2012	41.89879	-87.66130	2012-12-01

Basic Data Analysis & Data Visual

colab.research.google.com/drive/1JusFPRPcZ2F-v5QeapGjGjSEb4hzLKBm#scrollTo=m5ZWVh6KafPo

Incognito (2)

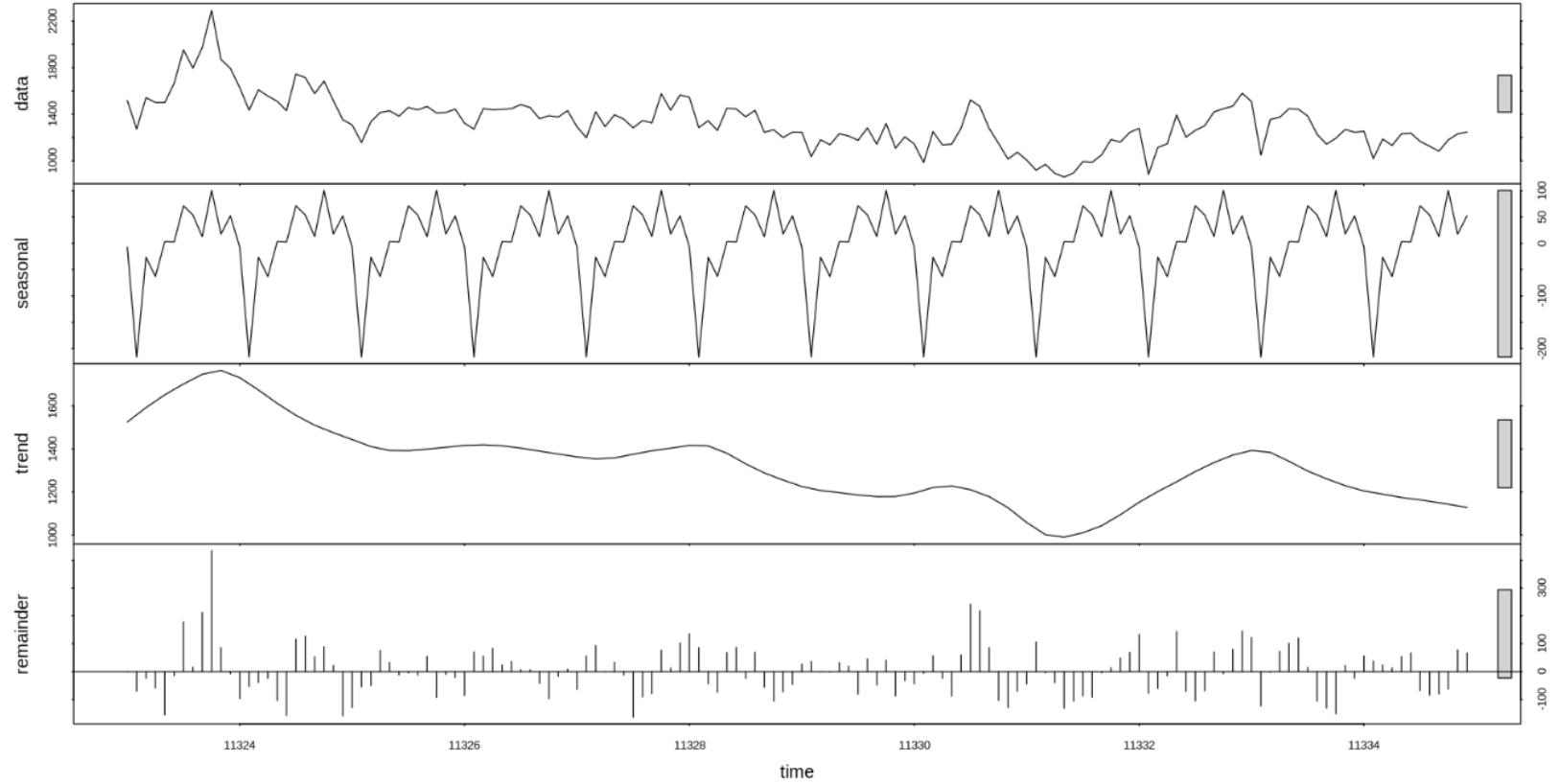
Basic Data Analysis & Data Visualization in R.ipynb

File Edit View Insert Runtime Tools Help All changes saved

Comment Share

+ Code + Text

RAM Disk



Basic Data Analysis & Data Visual

colab.research.google.com/drive/1JusFPRPcZ2F-v5QeapGjGjSEb4hzLKBm#scrollTo=OOoQvT2NMGws

Incognito (2)

Basic Data Analysis & Data Visualization in R.ipynb

File Edit View Insert Runtime Tools Help All changes saved

Comment Share

+ Code + Text

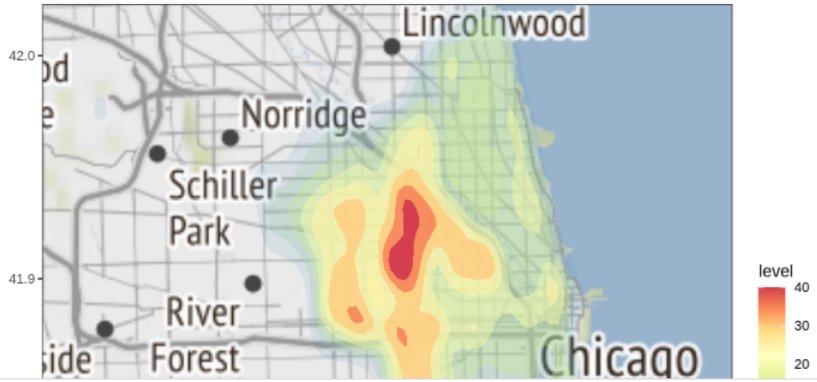
RAM Disk

Basic Spatial Analysis

6s

```
chicago <- get_map(getbb('chicago'), zoom = 10, source = 'stamen')
ggmap(chicago) +
  stat_density2d(data=mvt,  aes(x=Longitude, y=Latitude, fill=after_stat(level), alpha=after_stat(level)), geom="polygon") +
    scale_fill_gradientn(colours=rev(brewer.pal(7, "Spectral"))) +
      theme_bw()
```

```
ℹ Map tiles by Stamen Design, under CC BY 3.0. Data by OpenStreetMap, under ODbL.

Warning message:
“Removed 2276 rows containing non-finite values (`stat_density2d()`).”
```



Basic Data Analysis & Data Visual

colab.research.google.com/drive/1JusFPRPcZ2F-v5QeapGjGjSEb4hzLKBm#scrollTo=OOoQvT2NMGws

Incognito (2)

Basic Data Analysis & Data Visualization in R.ipynb

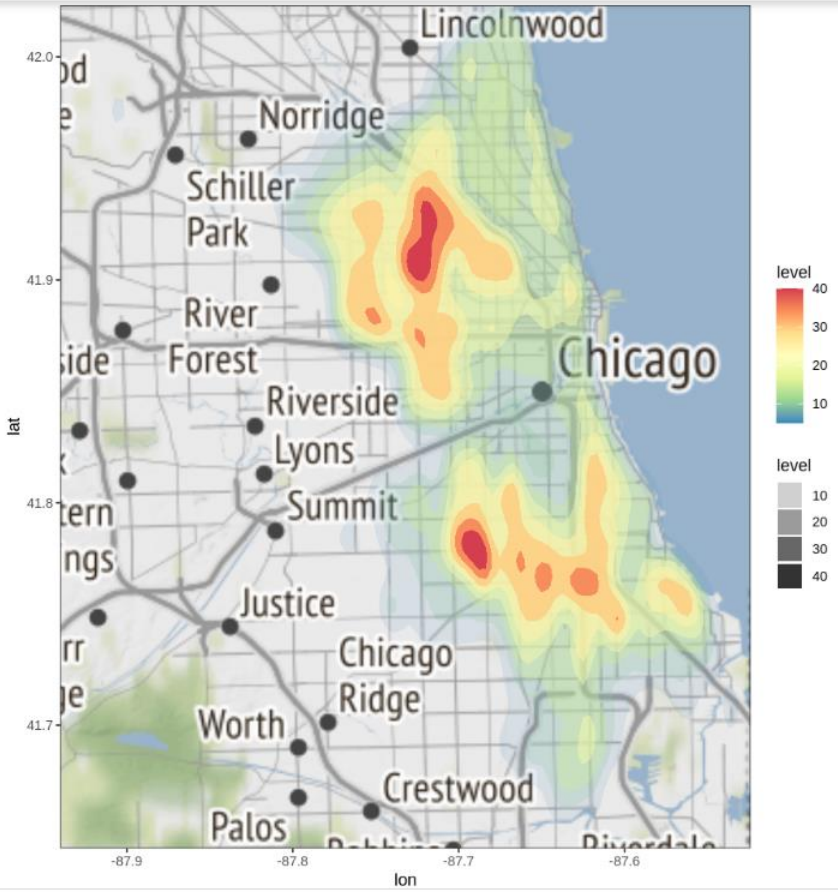
File Edit View Insert Runtime Tools Help All changes saved

Comment Share

+ Code + Text

RAM Disk

6s



Thank You

https://colab.research.google.com/drive/1JusFPRPcZ2F-v5QeapGjGjSEb4hzLKBm?usp=sharing

https://www.linkedin.com/in/joebrillantes/

https://www.facebook.com/rugph/

https://www.meetup.com/r-user-group-philippines/