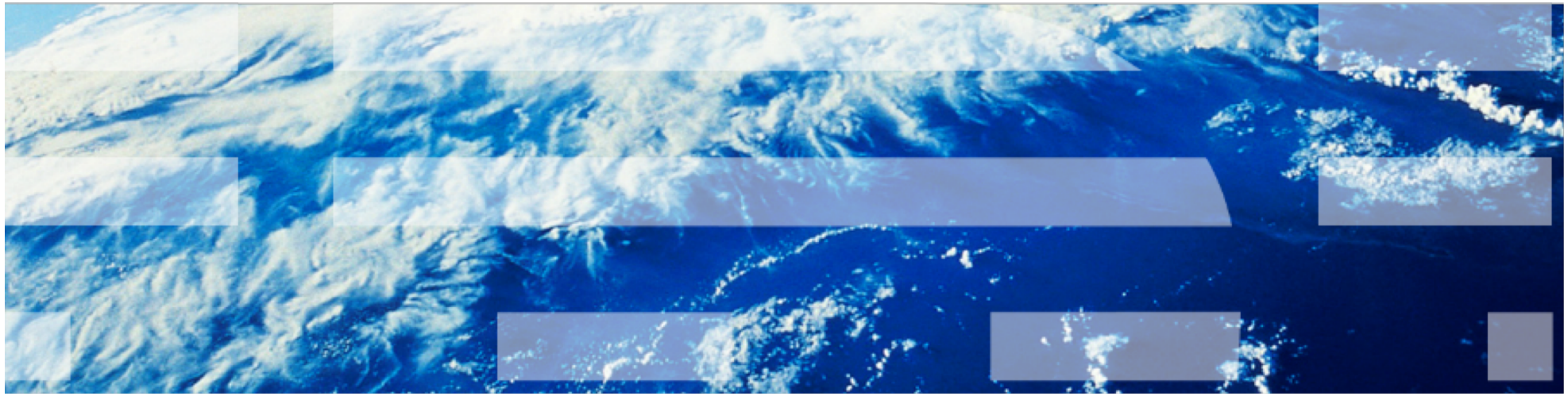
E6893 Big Data Analytics:

Relationship among Crimes, Public Commuters and Sentiment in Manhattan

Project ID: 201812-9
Team Members (with UNI): Yunan Lu (yl4021)
Zhicheng Ding (zd2212)
Lin Bai (lb3161)

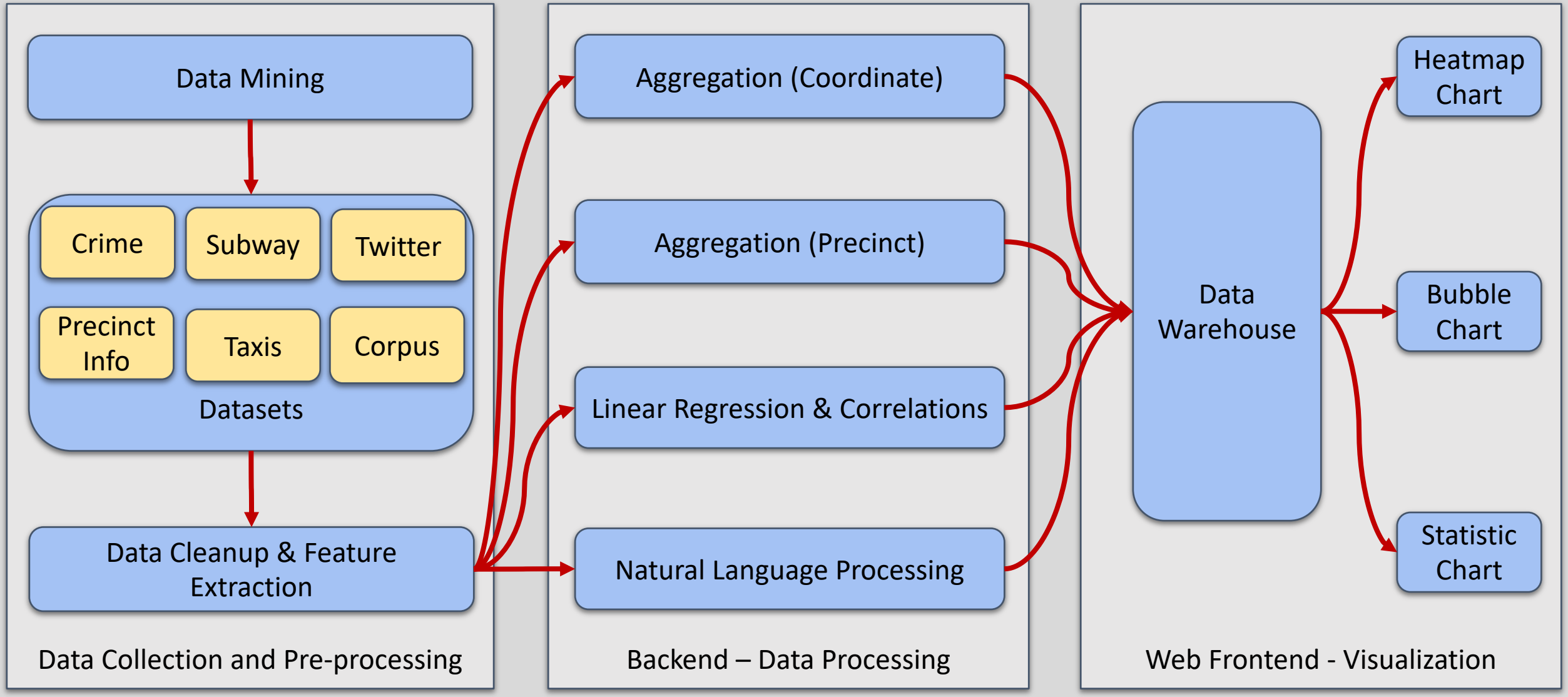


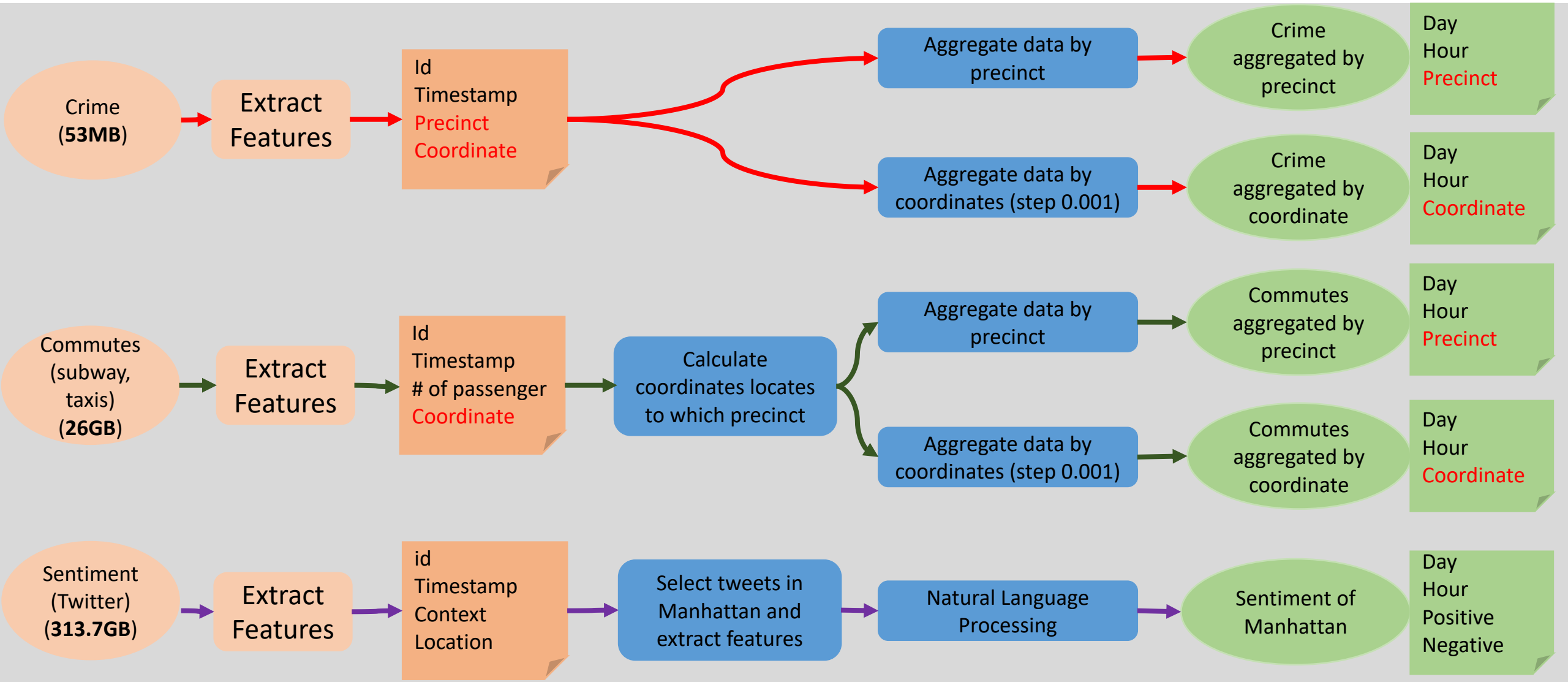
Outline

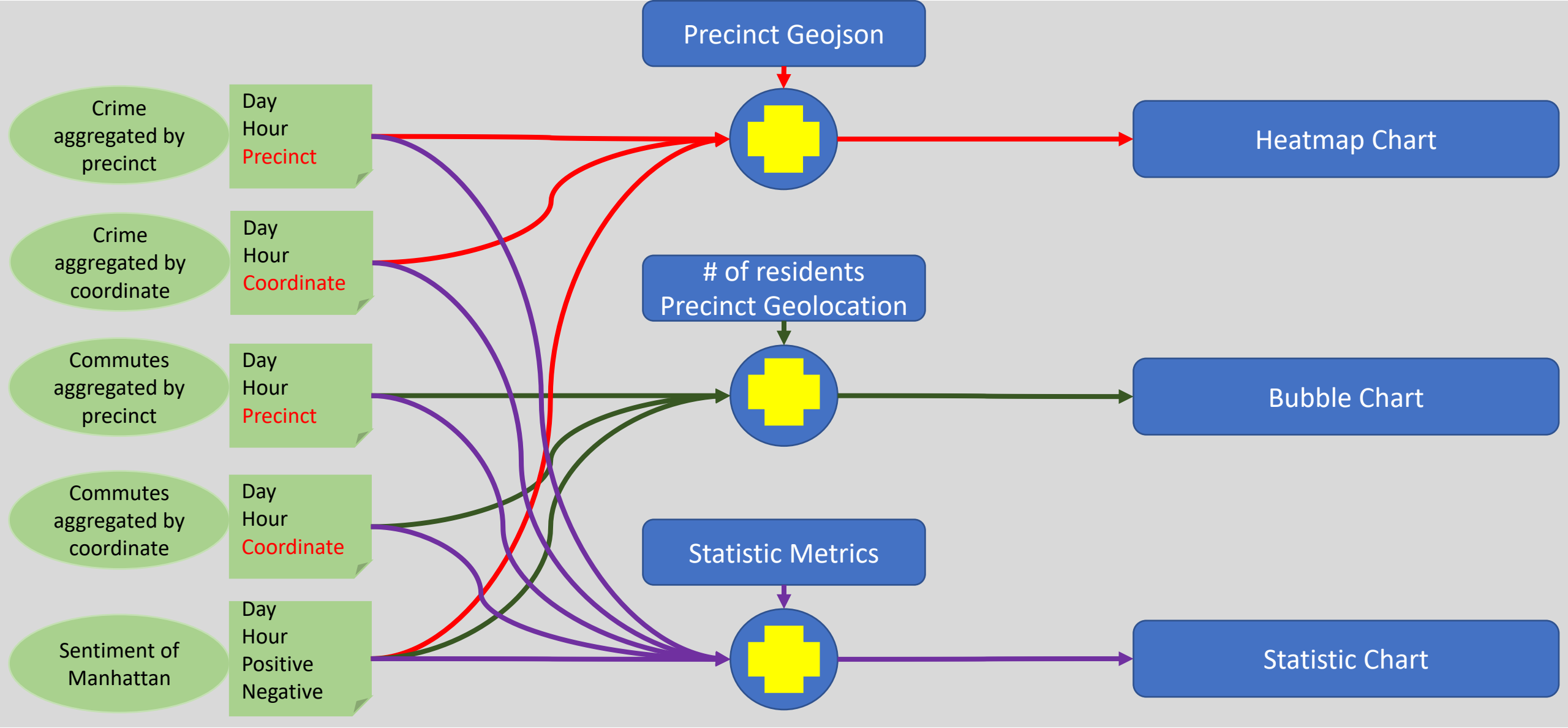
- Motivation
- Data Analysis – Dataset, Algorithm, Tools
  - Commuters
    - Taxi
    - Subway
  - Crimes
  - Sentiment
    - Twitter
  - Pearson Correlation
- Visualization
  - Web UI (Heatmap plot, Bubble plot, Statistic plot)

Motivation

- Every day, large amount of people travel via public transportation (e.g., taxi, subway). At the same time, crimes happens everywhere.
- Does it exist any pattern between the number of commuters and the frequency of crimes?
- Whether people's sentiment reflect the trend of crimes during the day among the week?







Commuters - MTA

Dataset (900MB)

- Turnstile Data (Subway) – http://web.mta.info/developers/turnstile.html
- Data range: Jan 04, 2014 – Jan 03, 2015
- Dataset Description
  - Main info: it contains the inflow and outflow of each subway station.
  - Attributes:
    - Id
    - Timestamp
    - Population
    - Entry or Exit

- Station location – http://web.mta.info/developers/data/nyct/subway/Stations.csv

- Connection between Turnstile Data and Station location – http://web.mta.info/developers/turnstile.html

Commuters - MTA

Data clean-up

- Normalized the turnstile table before and after Oct. 11 with the same format
- Match the station location table with the turnstile table based on the station name
- Standardized the “TIME” column’s value
- Calculated the hourly enter and hourly exit number
- Aggregate the data into a week.
- Calculate each point belong to which precinct and then group by precinct.

Tools

- Python: Pandas, Numpy

MTA - Result

Station_ID	HOURLY_ENTRIES	HOURLY_EXITS	Date	HOUR_mo	Station_name	Latitude	Longitude
4	285	517	1/5/2014	0	broadway	40.76182	-73.9255
7	1679	633	1/5/2014	0	lexingtonav/59st	40.76266	-73.9673
9	5628	1814	1/5/2014	0	57st-7av	40.76466	-73.9807
11	12197	7180	1/5/2014	0	timessq-42st	40.75467	-73.9868
12	23284	14310	1/5/2014	0	34st-heraldsq	40.74957	-73.988
13	3233	2590	1/5/2014	0	28st	40.74549	-73.9887
14	5291	3385	1/5/2014	0	23st	40.7413	-73.9893
15	20288	11493	1/5/2014	0	14st-unionsq	40.73574	-73.9906
18	6585	3161	1/5/2014	0	canalst	40.71953	-74.0018
19	6585	3161	1/5/2014	0	canalst	40.71838	-74.0005
21	0	2	1/5/2014	0	cortlandtst	40.71067	-74.011
22	315	452	1/5/2014	0	rectorst	40.70722	-74.0133
23	0	6	1/5/2014	0	whitehallst	40.70309	-74.013

Commuters - Taxi

Dataset (25GB)

- 2014 Yellow Taxi Trip Data – https://data.cityofnewyork.us/Transportation/2014-Yellow-Taxi-Trip-Data/gn7m-em8n
- Dataset Description:
  - Yellow Taxi normally cover most area in Manhattan
  - Attributes:
    - Id
    - Timestamp
    - # of Passenger
    - Coordinate

Commuters - Taxi

Data clean-up

- Simplify the dataset columns to timestamp, # of passenger, coordinate.
- Iteratively manipulated dataset to standardized the datetime format and add corresponding weekday and hour.
- Aggregate the # of passenger based on the location which is achieved by slicing the latitude and longitude with a stepsize [0.001]
- Calculate each point belong to which precinct and then group by precinct.

Tools

- Python: Pandas, Numpy
- Google Cloud: By using compute engine with large memory size to avoid "OutofMemory" issue when manipulating large dataset

TAXI - Result

weekday	pick_hour	sum	lat	long
3	2	5	40.49929	-74.2394
4	0	2	40.49929	-74.2394
2	12	1	40.49929	-74.0994
2	12	1	40.49929	-74.0894
2	12	6	40.49929	-74.0394
3	0	1	40.49929	-74.0394
4	13	2	40.49929	-74.0394
6	12	3	40.49929	-74.0394
6	23	1	40.49929	-74.0094
2	15	1	40.49929	-73.9794

After data clean-up steps, the original dataset of 25GB were shrank into 45MB

Crimes

Dataset (53MB)

- 2014 New York City Crimes – https://www.kaggle.com/adamschroeder/crimes-new-york-city#Crime_Column_Description.csv
- Dataset Description:
  - 2014-2015 Crimes reported in all 5 boroughs of New York City
  - Attributes:
    - Id
    - Timestamp
    - Precinct number
    - Coordinate

Crimes

Data clean-up

- Simplify the dataset to weekday, hour, precinct, latitude, longitude
- Aggregated the data based on precinct number.
- Aggregate the number of precinct based on the location which is achieved by slicing the latitude and longitude with a stepsize [0.001]

Tools

- Python: pandas, numpy

Crimes Result

map_plot_criminal_lat_long

day	hour	sum	lat	long
3	3	1	40.4988	-74.2439
4	3	1	40.4988	-74.2439
2	13	1	40.4988	-74.2419
6	11	1	40.4988	-74.2419
1	17	1	40.4988	-74.2409
2	18	1	40.4988	-74.2409
4	23	1	40.4988	-74.2409
0	0	1	40.4998	-74.2449
2	14	1	40.4998	-74.2449
5	9	1	40.4998	-74.2449
6	4	1	40.4998	-74.2449
2	23	1	40.4998	-74.2409
1	20	1	40.4998	-74.2399
2	22	1	40.4998	-74.2399
3	18	2	40.4998	-74.2399
4	15	1	40.4998	-74.2399

Sentiment

Dataset (total: 313.7GB)
2014 Twitter Dataset –
https://archive.org/details/twitterstream?and[]=year%3A%222014%22

- Dataset Description:
  - Twitter post of all year
  - Attributes:
    - Context
    - Timestamp
    - User
    - Description
    - Location

Sentiment

Data clean-up

- Simplify the dataset to weekday, hour, precinct, latitude, longitude
- Aggregated the data based on precinct number.
- Aggregate the number of precinct based on the location which is achieved by slicing the latitude and longitude with a stepsize [0.001]

Tools

- Python: pandas, numpy, GCP, NLP

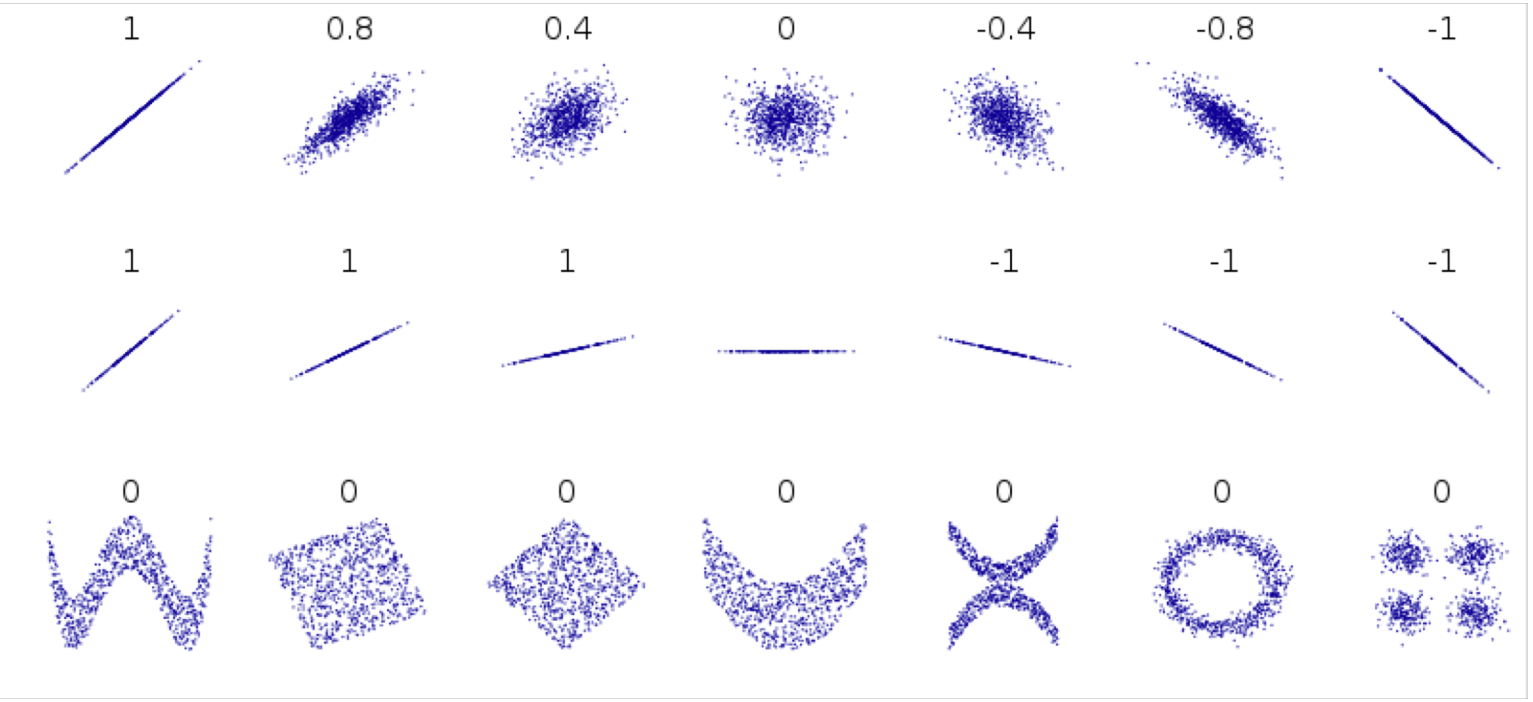
Sentiment Result

sentiment_number

day	hour	negative	neutral	positive
0	0	815	3604	1067
0	1	431	2297	767
0	2	462	2363	725
0	3	680	3140	872
0	4	907	3681	1026
0	5	919	4074	1095
0	6	922	4144	1141
0	7	930	4194	1001
0	8	893	3731	875
0	9	874	3730	850
0	10	718	2762	682
0	11	416	2016	388
0	12	420	1624	330
0	13	386	1509	319
0	14	302	1267	259
0	15	362	1404	295
0	16	388	1404	307
0	17	368	1391	314

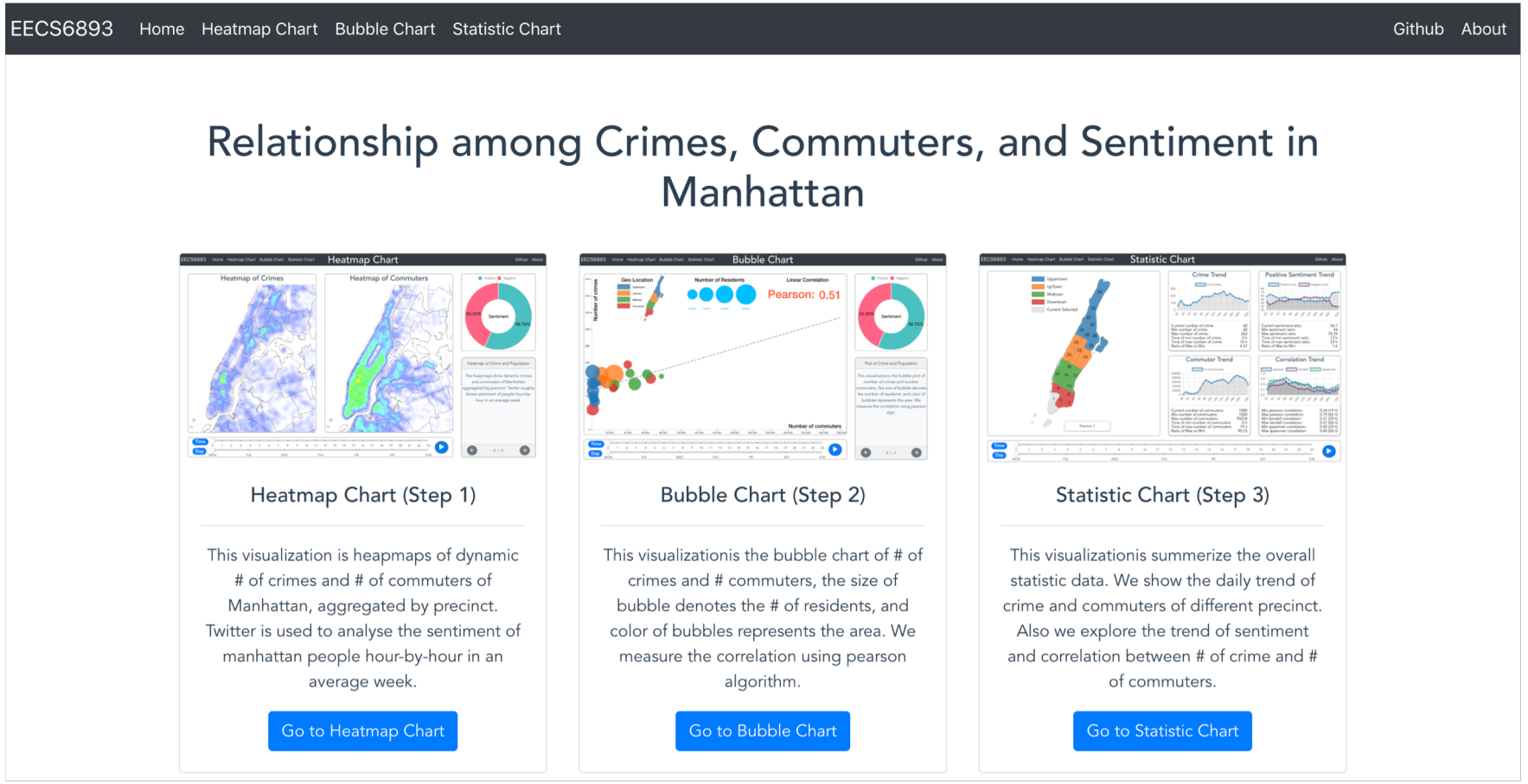
Pearson Correlation

In statistic, the Pearson correlation coefficient (PCC) is a measure of the linear correlation between two variables crimes and commuters.



* Ref: https://en.wikipedia.org/wiki/Pearson_correlation_coefficient

Demo



Link: http://35.231.39.145:8080/

Algorithms & Tools

Algorithms:

- Natural Language Processing
- Calculate point inside/outside polygon
- Calculate geo center of polygon
- Linear Regression
- Pearson Correlation
- Kendall Correlation
- Spearman Correlation

Tools:

- Google Cloud Platform
- Pandas
- Numpy
- Vue.js
- D3.js

Conclusion

- From the map chart, we can see that there are more commuters and crimes in downtown and midtown, but relatively less in uptown and upper town.
- From the bubble chart, we found that downtown, midtown, and uptown have relatively high frequency of criminal and high traffic density during work hour. It gradually reduces to normal at midnight.
- From the Pearson correlation calculated, we can see that the number of crimes and commuters are linearly positive correlated (about 0.5).
- From the sentiment analysis, we can see that for the weekdays, people normally more negative for the work hours and positive for the night. While for the weekend, people would be more positive all the day which may result in safety awareness decreasing and lead number of crimes increase.

Q & A

Youtube Link: https://youtu.be/XA7PlJ9USqc

Appendix – Data Clean-up Process

Field Description

C/A, UNIT, SCP, DATE1, TIME1, DESC1, ENTRIES1, EXITS1, DATE2, TIME2, DESC2, ENTRIES2, EXITS2, DATE3, TIME3, DESC3, ENTRIES3, EXITS3, DATE4, TIME4, DESC4, ENTRIES4, EXITS4, DATE5, TIME5, DESC5, ENTRIES5, EXITS5, DATE6, TIME6, DESC6, ENTRIES6, EXITS6, DATE7, TIME7
, DESC7, ENTRIES7, EXITS7, DATE8, TIME8, DESC8, ENTRIES8, EXITS8

C/A = Control Area (A002)
UNIT = Remote Unit for a station (R051)
SCP = Subunit Channel Position represents an specific address for a device (02-00-00)
DATEn = Represents the date (MM-DD-YY)
TIMEn = Represents the time (hh:mm:ss) for a scheduled audit event
DEScn = Represent the "REGULAR" scheduled audit event (occurs every 4 hours)
ENTRIESn = The comulative entry register value for a device
EXISTn = The cumulative exit register value for a device

Example:
The data below shows the entry/exit register values for one turnstile at control area (A002) from 03/21/10 at 00:00 hours to 03/28/10 at 20:00 hours

A002, R051, 02-00-00, 03-21-10, 00:00:00, REGULAR, 002670738, 000917107, 03-21-10, 04:00:00, REGULAR, 002670738, 000917107, 03-21-10, 08:00:00, REGULAR, 002670746, 000917117, 03-21-10, 12:00:00, REGULAR, 002670790, 000917166, 03-21-
10, 16:00:00, REGULAR, 002670932, 000917204, 03-21-10, 20:00:00, REGULAR, 002671164, 000917230, 03-22-10, 00:00:00, REGULAR, 002671181, 000917231, 03-22-10, 04:00:00, REGULAR, 002671181, 000917231
A002, R051, 02-00-00, 03-22-10, 08:00:00, REGULAR, 002671220, 000917324, 03-22-10, 12:00:00, REGULAR, 002671364, 000917640, 03-22-10, 16:00:00, REGULAR, 002671651, 000917719, 03-22-10, 20:00:00, REGULAR, 002672430, 000917789, 03-23-
10, 00:00:00, REGULAR, 002672473, 000917795, 03-23-10, 04:00:00, REGULAR, 002672474, 000917795, 03-23-10, 08:00:00, REGULAR, 002672516, 000917876, 03-23-10, 12:00:00, REGULAR, 002672652, 000917934
A002, R051, 02-00-00, 03-23-10, 16:00:00, REGULAR, 002672879, 000917996, 03-23-10, 20:00:00, REGULAR, 002673636, 000918073, 03-24-10, 00:00:00, REGULAR, 002673683, 000918079, 03-24-10, 04:00:00, REGULAR, 002673683, 000918079, 03-24-
10, 08:00:00, REGULAR, 002673722, 000918171, 03-24-10, 12:00:00, REGULAR, 002673876, 000918514, 03-24-10, 16:00:00, REGULAR, 002674221, 000918594, 03-24-10, 20:00:00, REGULAR, 002675082, 000918671
A002, R051, 02-00-00, 03-25-10, 00:00:00, REGULAR, 002675153, 000918675, 03-25-10, 04:00:00, REGULAR, 002675153, 000918675, 03-25-10, 08:00:00, REGULAR, 002675190, 000918752, 03-25-10, 12:00:00, REGULAR, 002675345, 000919053, 03-25-
10, 16:00:00, REGULAR, 002675676, 000919118, 03-25-10, 20:00:00, REGULAR, 002676557, 000919179, 03-26-10, 00:00:00, REGULAR, 002676688, 000919207, 03-26-10, 04:00:00, REGULAR, 002676694, 000919208
A002, R051, 02-00-00, 03-26-10, 08:00:00, REGULAR, 002676735, 000919287, 03-26-10, 12:00:00, REGULAR, 002676887, 000919607, 03-26-10, 16:00:00, REGULAR, 002677213, 000919680, 03-26-10, 20:00:00, REGULAR, 002678039, 000919743, 03-27-
10, 00:00:00, REGULAR, 002678144, 000919756, 03-27-10, 04:00:00, REGULAR, 002678145, 000919756, 03-27-10, 08:00:00, REGULAR, 002678155, 000919777, 03-27-10, 12:00:00, REGULAR, 002678247, 000919859
A002, R051, 02-00-00, 03-27-10, 16:00:00, REGULAR, 002678531, 000919908, 03-27-10, 20:00:00, REGULAR, 002678892, 000919964, 03-28-10, 00:00:00, REGULAR, 002678929, 000919966, 03-28-10, 04:00:00, REGULAR, 002678929, 000919966, 03-28-
10, 08:00:00, REGULAR, 002678935, 000919982, 03-28-10, 12:00:00, REGULAR, 002679003, 000920006, 03-28-10, 16:00:00, REGULAR, 002679231, 000920059, 03-28-10, 20:00:00, REGULAR, 002679475, 000920098

- Dataset before Oct 11
- This is a flat dataset with multiple (DATE, TIME, ENTRIES, EXITS) columns without station name
- Convert into long dataset

Field Description

C/A, UNIT, SCP, STATION, LINENAME, DIVISION, DATE, TIME, DESC, ENTRIES, EXITS

C/A = Control Area (A002)
UNIT = Remote Unit for a station (R051)
SCP = Subunit Channel Position represents an specific address for a device (02-00-00)
STATION = Represents the station name the device is located at
LINENAME = Represents all train lines that can be boarded at this station
Normally lines are represented by one character. LINENAME 456NQR repersents train server for 4, 5, 6, N, Q, and R trains.
DIVISION = Represents the Line originally the station belonged to BMT, IRT, or IND
DATE = Represents the date (MM-DD-YY)
TIME = Represents the time (hh:mm:ss) for a scheduled audit event
DESc = Represent the "REGULAR" scheduled audit event (Normally occurs every 4 hours)
1. Audits may occur more that 4 hours due to planning, or troubleshooting activities.
2. Additionally, there may be a "RECOVR AUD" entry: This refers to a missed audit that was recovered.
ENTRIES = The comulative entry register value for a device
EXIST = The cumulative exit register value for a device

Example:
The data below shows the entry/exit register values for one turnstile at control area (A002) from 09/27/14 at 00:00 hours to 09/29/14 at 00:00 hours

C/A, UNIT, SCP, STATION, LINENAME, DIVISION, DATE, TIME, DESC, ENTRIES, EXITS
A002, R051, 02-00-00, LEXINGTON AVE, 456NQR, BMT, 09-27-14, 00:00:00, REGULAR, 0004800073, 0001629137,
A002, R051, 02-00-00, LEXINGTON AVE, 456NQR, BMT, 09-27-14, 04:00:00, REGULAR, 0004800125, 0001629149,
A002, R051, 02-00-00, LEXINGTON AVE, 456NQR, BMT, 09-27-14, 08:00:00, REGULAR, 0004800146, 0001629162,
A002, R051, 02-00-00, LEXINGTON AVE, 456NQR, BMT, 09-27-14, 12:00:00, REGULAR, 0004800264, 0001629264,
A002, R051, 02-00-00, LEXINGTON AVE, 456NQR, BMT, 09-27-14, 16:00:00, REGULAR, 0004800523, 0001629328,
A002, R051, 02-00-00, LEXINGTON AVE, 456NQR, BMT, 09-27-14, 20:00:00, REGULAR, 0004800924, 0001629371,
A002, R051, 02-00-00, LEXINGTON AVE, 456NQR, BMT, 09-28-14, 00:00:00, REGULAR, 0004801104, 0001629395,
A002, R051, 02-00-00, LEXINGTON AVE, 456NQR, BMT, 09-28-14, 04:00:00, REGULAR, 0004801149, 0001629402,
A002, R051, 02-00-00, LEXINGTON AVE, 456NQR, BMT, 09-28-14, 08:00:00, REGULAR, 0004801168, 0001629414,
A002, R051, 02-00-00, LEXINGTON AVE, 456NQR, BMT, 09-28-14, 12:00:00, REGULAR, 0004801304, 0001629463,
A002, R051, 02-00-00, LEXINGTON AVE, 456NQR, BMT, 09-28-14, 16:00:00, REGULAR, 0004801463, 0001629521,
A002, R051, 02-00-00, LEXINGTON AVE, 456NQR, BMT, 09-28-14, 20:00:00, REGULAR, 0004801737, 0001629555,
A002, R051, 02-00-00, LEXINGTON AVE, 456NQR, BMT, 09-29-14, 00:00:00, REGULAR, 0004801836, 0001629574,

- Dataset after Oct 11
- This is a long dataset with station name

Remote	Booth	Station	Line Name	Division
R001	A060	WHITEHALI	R1	BMT
R001	A058	WHITEHALI	R1	BMT
R001	R101S	SOUTH FER	R1	IRT
R002	A077	FULTON ST	ACJZ2345	BMT
R002	A081	FULTON ST	ACJZ2345	BMT
R002	A082	FULTON ST	ACJZ2345	BMT
R003	J025	CYPRESS HI	J	BMT
R004	J028	ELDERTS LA	JZ	BMT
R005	J030	FOREST PA	J	BMT
R006	J031	WOODHAV	JZ	BMT
R006	J032	WOODHAV	JZ	BMT
R007	J034	104 ST	JZ	BMT
R008	J035	111 ST	J	BMT
R009	J037	121 ST	JZ	BMT
R010	N062A	42 ST-PA B	ACENQRS1	IND
R010	N060	42 ST-PA B	ACENQRS1	IND
R011	N063A	42 ST-PA B	ACENQRS1	IND

Remote Unit / Control Area / Station Name Key

	Station ID	Complex ID	GTFS Stop ID	Division	Line	Stop Name	Borough	Daytime Routes	Structure	GTFS Latitude	GTFS Longitude
0	1	1	R01	BMT	Astoria	Astoria - Ditmars Blvd	Q	N W	Elevated	40.775036	-73.912034
1	2	2	R03	BMT	Astoria	Astoria Blvd	Q	N W	Elevated	40.770258	-73.917843
2	3	3	R04	BMT	Astoria	30 Av	Q	N W	Elevated	40.766779	-73.921479
3	4	4	R05	BMT	Astoria	Broadway	Q	N W	Elevated	40.761820	-73.925508
4	5	5	R06	BMT	Astoria	36 Av	Q	N W	Elevated	40.756804	-73.929575
5	6	6	R08	BMT	Astoria	39 Av	Q	N W	Elevated	40.752882	-73.932755
6	7	613	R11	BMT	Astoria	Lexington Av/59 St	M	N W R	Subway	40.762660	-73.967258
7	8	8	R13	BMT	Astoria	5 Av/59 St	M	N W R	Subway	40.764811	-73.973347
8	9	9	R14	BMT	Broadway - Brighton	57 St - 7 Av	M	N Q R W	Subway	40.764664	-73.980658

Station location

Problem:
These two table do not have common columns, the station name are not exactly the same, so we need to find the most common parts from multiple columns to join these two table together.

In [10]:
```
mapping_final
```

Out[10]:

	UNIT	C/A	Station	Line Name	Division	Station_modified	Station ID	GTFS Latitude	GTFS Longitude
0	R001	A060	WHITEHALL ST	R1	BMT	whitehallst	23.0	40.703087	-74.012994
1	R001	A058	WHITEHALL ST	R1	BMT	whitehallst	23.0	40.703087	-74.012994
2	R001	R101S	SOUTH FERRY	R1	IRT	southferry	330.0	40.702068	-74.013664
3	R002	A077	FULTON ST	ACJZ2345	BMT	fultonst	106.0	40.710374	-74.007582
4	R002	A077	FULTON ST	ACJZ2345	BMT	fultonst	172.0	40.710197	-74.007691
5	R002	A077	FULTON ST	ACJZ2345	BMT	fultonst	292.0	40.687119	-73.975375
6	R002	A077	FULTON ST	ACJZ2345	BMT	fultonst	332.0	40.709416	-74.006571
7	R002	A077	FULTON ST	ACJZ2345	BMT	fultonst	412.0	40.710368	-74.009509
8	R002	A081	FULTON ST	ACJZ2345	BMT	fultonst	106.0	40.710374	-74.007582
9	R002	A081	FULTON ST	ACJZ2345	BMT	fultonst	172.0	40.710197	-74.007691
10	R002	A081	FULTON ST	ACJZ2345	BMT	fultonst	292.0	40.687119	-73.975375

In [10]:
```
mapping_final
```

Out[10]:

	UNIT	C/A	Station	Line Name	Division	Station_modified	Station ID	GTFS Latitude	GTFS Longitude
0	R001	A060	WHITEHALL ST	R1	BMT	whitehallst	23.0	40.703087	-74.012994
1	R001	A058	WHITEHALL ST	R1	BMT	whitehallst	23.0	40.703087	-74.012994
2	R001	R101S	SOUTH FERRY	R1	IRT	southferry	330.0	40.702068	-74.013664
3	R002	A077	FULTON ST	ACJZ2345	BMT	fultonst	106.0	40.710374	-74.007582
4	R002	A077	FULTON ST	ACJZ2345	BMT	fultonst	172.0	40.710197	-74.007691
5	R002	A077	FULTON ST	ACJZ2345	BMT	fultonst	292.0	40.687119	-73.975375
6	R002	A077	FULTON ST	ACJZ2345	BMT	fultonst	332.0	40.709416	-74.006571
7	R002	A077	FULTON ST	ACJZ2345	BMT	fultonst	412.0	40.710368	-74.009509
8	R002	A081	FULTON ST	ACJZ2345	BMT	fultonst	106.0	40.710374	-74.007582
9	R002	A081	FULTON ST	ACJZ2345	BMT	fultonst	172.0	40.710197	-74.007691
10	R002	A081	FULTON ST	ACJZ2345	BMT	fultonst	292.0	40.687119	-73.975375

Appendix - Website