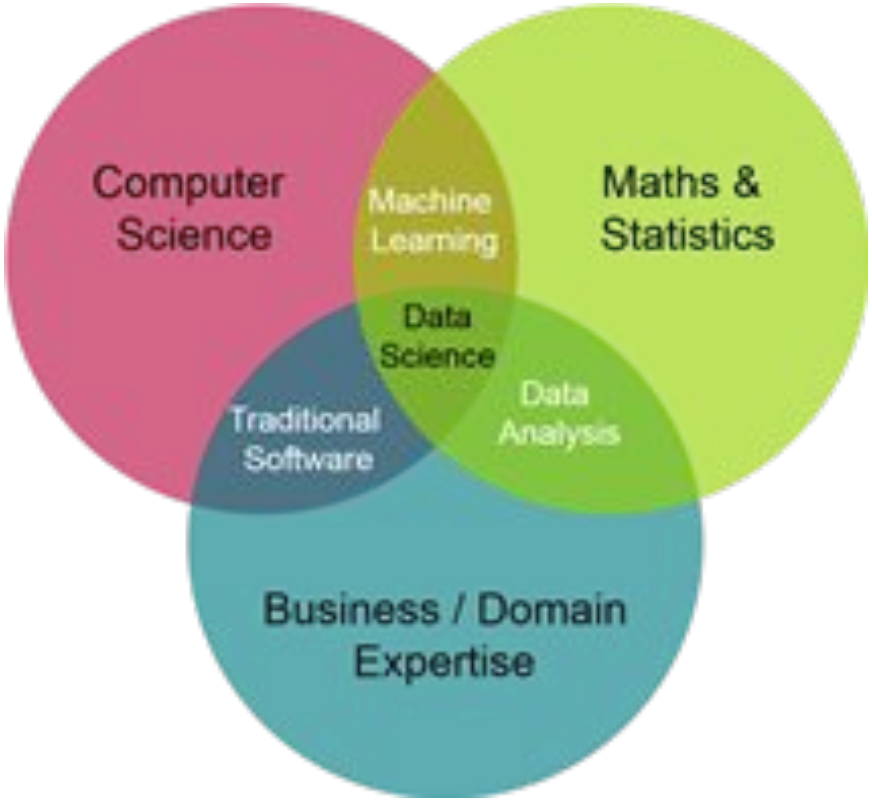
Big Data and Math Modeling:

Using Python to Analyze The NYC Subway System

Lauren Shareshian
Oregon Episcopal School
Portland, OR

What is Data Science?



Tons of Data Science education is popping up

From the free online materials: UC Berkeley Data 8

To the undergraduate majors: Columbia University

To the expensive bootcamps: Metis, UCSD

What is Big Data?

- Volume - Lots of it
- Velocity - New data continuously coming in
- Variety - Data comes in all types of formats

"Big data" refers to the **use of analytics to extract value** from data, and seldom to a particular size of data set.

MTA NYC Subway Data Set

- Publicly available
- Published weekly
- Info on entries/exits through every turnstile in 4 hour intervals

http://web.mta.info/developers/turnstile.html

How big is the data set?

```
In [1]: import pandas as pd

        data = pd.read_csv('subway.csv')
        data.shape

Out[1]: (197209, 11)
```

If you don't know any programming

Fear not! There's a cleaned up excel spreadsheet:

	Weekday Entries 0-4 am	Weekday Entries 4-8 am	Weekday Entries 8-12 pm	Weekday Entries 12-4 pm	Weekday Entries 4-8 pm	Weekday Entries 8-12 am	Weekend Entries 0-4 am
59 ST	1334	7154	15292	16655	32353	9690	1729
5 AV/59 ST	586	418	1601	5140	7053	3698	35
57 ST-7 AV	844	1143	5044	6823	16815	6667	1443
49 ST	828	644	3515	4610	12580	5889	1407
TIMES SQ-42 ST	3678	5648	14784	17593	46960	21404	6971

How big is this data set?

- 376 stations
- 4,695 turnstiles
- Each turnstile reports rider data every four hours (6 times per day) for one week (7 days) 6*7*4,695 = 197,190 rows of data

What does the data set look like?

```
data[(data['STATION'] == '34 ST-PENN STA') & (data['DATE'] == '06/12/2017')]
```

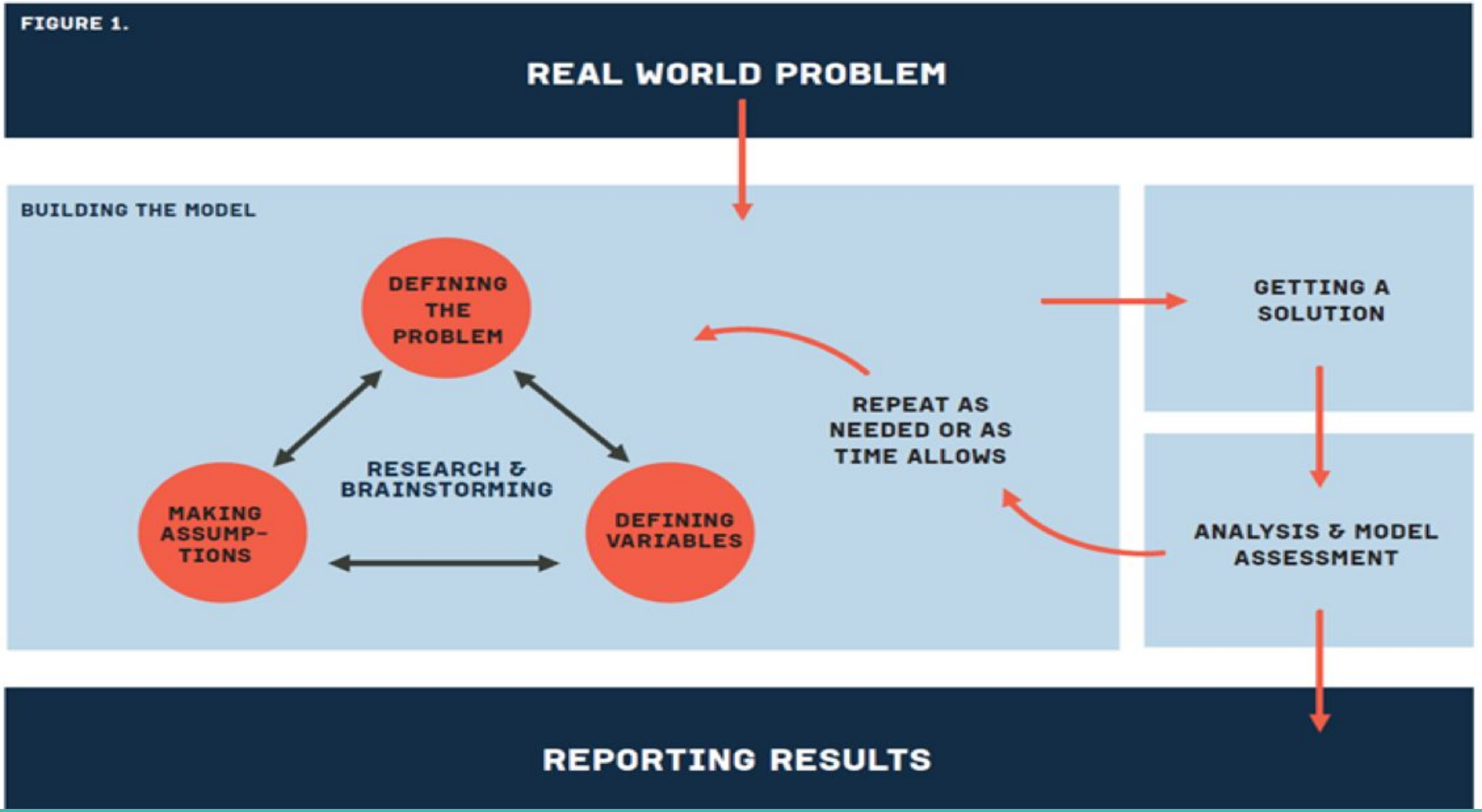
	C/A	UNIT	SCP	STATION	LINENAME	DIVISION	DATE	TIME	DESC	ENTRIES	EXITS
49612	N067	R012	00-00-00	34 ST-PENN STA	ACE	IND	06/12/2017	00:00:00	REGULAR	1829493	1553798
49613	N067	R012	00-00-00	34 ST-PENN STA	ACE	IND	06/12/2017	04:00:00	REGULAR	1829495	1553801
49614	N067	R012	00-00-00	34 ST-PENN STA	ACE	IND	06/12/2017	08:00:00	REGULAR	1829676	1553947
49615	N067	R012	00-00-00	34 ST-PENN STA	ACE	IND	06/12/2017	12:00:00	REGULAR	1829944	1554414
49616	N067	R012	00-00-00	34 ST-PENN STA	ACE	IND	06/12/2017	16:00:00	REGULAR	1829981	1554571

How does this relate to math modeling?

From this data set we can ask questions that are:

- Interesting
- Purposeful (**extracts value**)
- Collaborative
- Allow for a variety of solutions
- Open ended

FIGURE 1.


REAL WORLD PROBLEM
BUILDING THE MODEL
DEFINING THE PROBLEM
RESEARCH & BRAINSTORMING
MAKING ASSUMP-TIONS
DEFINING VARIABLES
REPEAT AS NEEDED OR AS TIME ALLOWS
GETTING A SOLUTION
ANALYSIS & MODEL ASSESSMENT
REPORTING RESULTS

What are those variables?

Skills developed: Research skills, resourcefulness

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

mtadeveloperresources ›

What are control areas, remote units, and subunits?

4 posts by 4 authors

JS 8/21/15

I'm trying to understand the different fields in the MTA turnstile dataset. I know that the documentation says

```
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)

But what do these things mean? I don't know what a control area is, what a remote unit is, or what a subunit
position is.

Is CONTROL AREA the same thing as a station? UNIT a group of turnstiles? SCP a specific turnstile?

Thanks!
```

What are those variables telling me?

Skills developed: Number Sense

Did 1,829,493 people enter through a Penn Station turnstile at midnight?

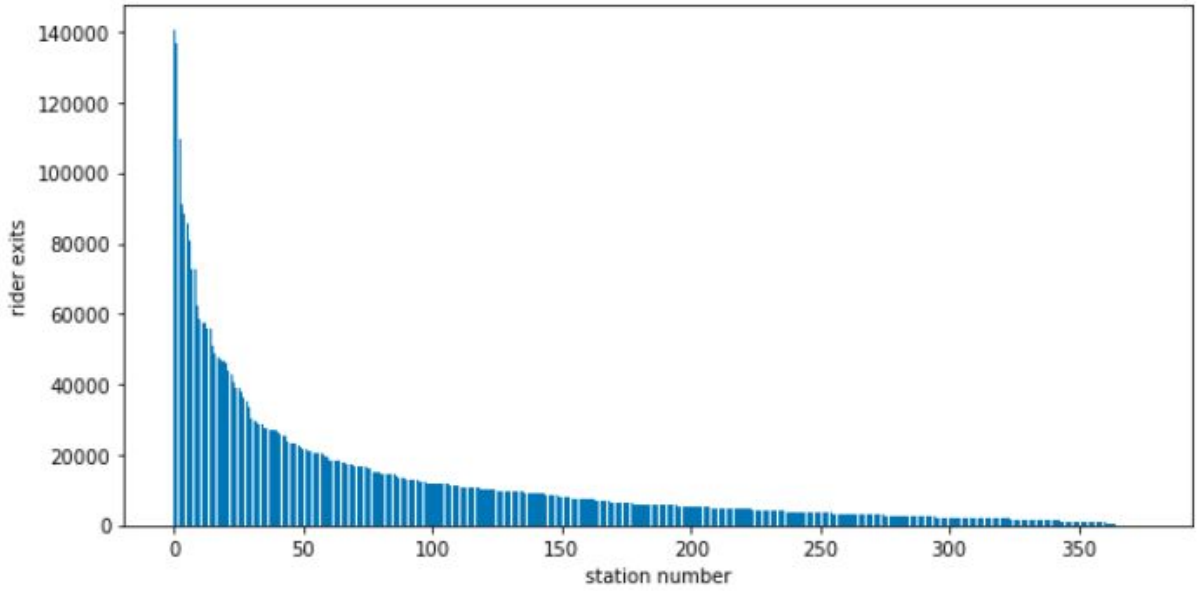
	C/A	UNIT	SCP	STATION	LINENAME	DIVISION	DATE	TIME	DESC	ENTRIES	EXITS
49612	N067	R012	00-00-00	34 ST-PENN STA	ACE	IND	06/12/2017	00:00:00	REGULAR	1829493	1553798
49613	N067	R012	00-00-00	34 ST-PENN STA	ACE	IND	06/12/2017	04:00:00	REGULAR	1829495	1553801

Put the data in a form we can work with

	C/A	UNIT	SCP	STATION	LINENAME	DATE	TIME	ENTRIES	EXITS	ENTRY_DIFF	EXIT_DIFF
49612	N067	R012	00-00-00	34 ST-PENN STA	ACE	06/12/2017	00:00:00	1829493	1553798	0	0
49613	N067	R012	00-00-00	34 ST-PENN STA	ACE	06/12/2017	04:00:00	1829495	1553801	2	3
49614	N067	R012	00-00-00	34 ST-PENN STA	ACE	06/12/2017	08:00:00	1829676	1553947	181	146
49615	N067	R012	00-00-00	34 ST-PENN STA	ACE	06/12/2017	12:00:00	1829944	1554414	268	467
49616	N067	R012	00-00-00	34 ST-PENN STA	ACE	06/12/2017	16:00:00	1829981	1554571	37	157
49617	N067	R012	00-00-00	34 ST-PENN STA	ACE	06/12/2017	20:00:00	1830036	1555166	55	595

Riders on Monday, June 12, 2017

```
import matplotlib.pyplot as plt
%matplotlib inline

plt.bar(station, riders)
plt.xlabel('station number')
plt.ylabel('rider exits')
```



What were the busiest stations?

```
rider_list.sort(reverse = True)
for rider_info in rider_list:
    print(rider_info)
```

```
(136834, '34 ST-PENN STA')
(135077, 'GRD CNTRL-42 ST')
(109563, '34 ST-HERALD SQ')
(91048, 'TIMES SQ-42 ST')
(88400, '14 ST-UNION SQ')
(85480, '23 ST')
(81102, 'FULTON ST')
(72934, '42 ST-PORT AUTH')
(72796, '86 ST')
(62141, '47-50 STS ROCK')
```

Modeling Task

Coding Chicks has an annual gala this summer. Please help us **optimize the placement** of our street teams in the subway. Our goal is to gather the most contact info from women who will **attend the gala and donate**.

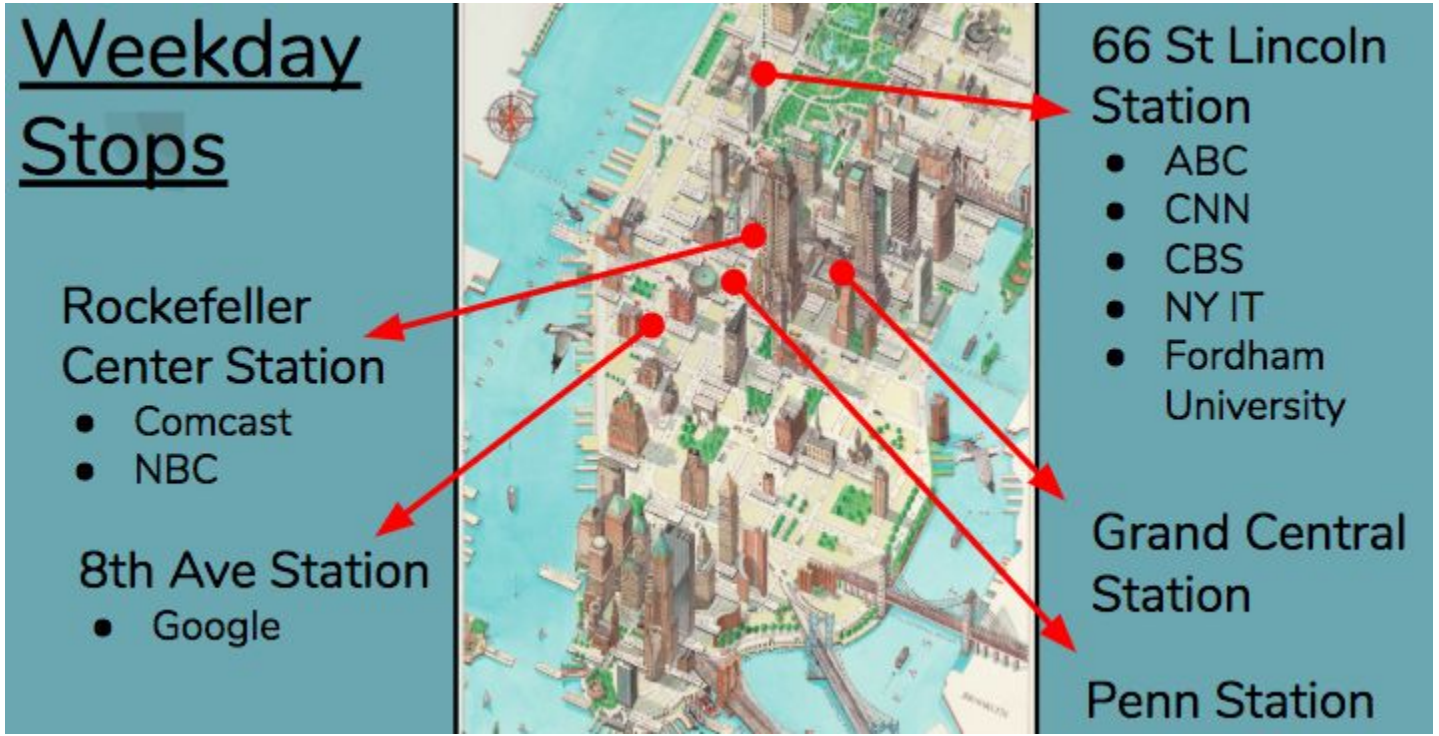
We have **ten volunteers** to advertise in the subway for **four hours each** per day (in one four-hour shift or in 2 two-hour shifts). They can help **seven days** in a row, so we plan on doing all of our advertising during one seven-day blitz.

Please give us a clear, detailed presentation outlining your suggestions. **We will hire the most compelling business solution.**

Lots of complexity to consider

1. Focusing on where women in technology are located.
2. Focusing on where wealthier donors are located.
3. Differentiating between weekday and weekend placement.
4. Differentiating between what subway turnstile entries versus exits tell you.
5. Differentiating between morning and evening placement.
6. Differentiating between tourist and commuter stops.

Student Work: Focusing on tech hubs

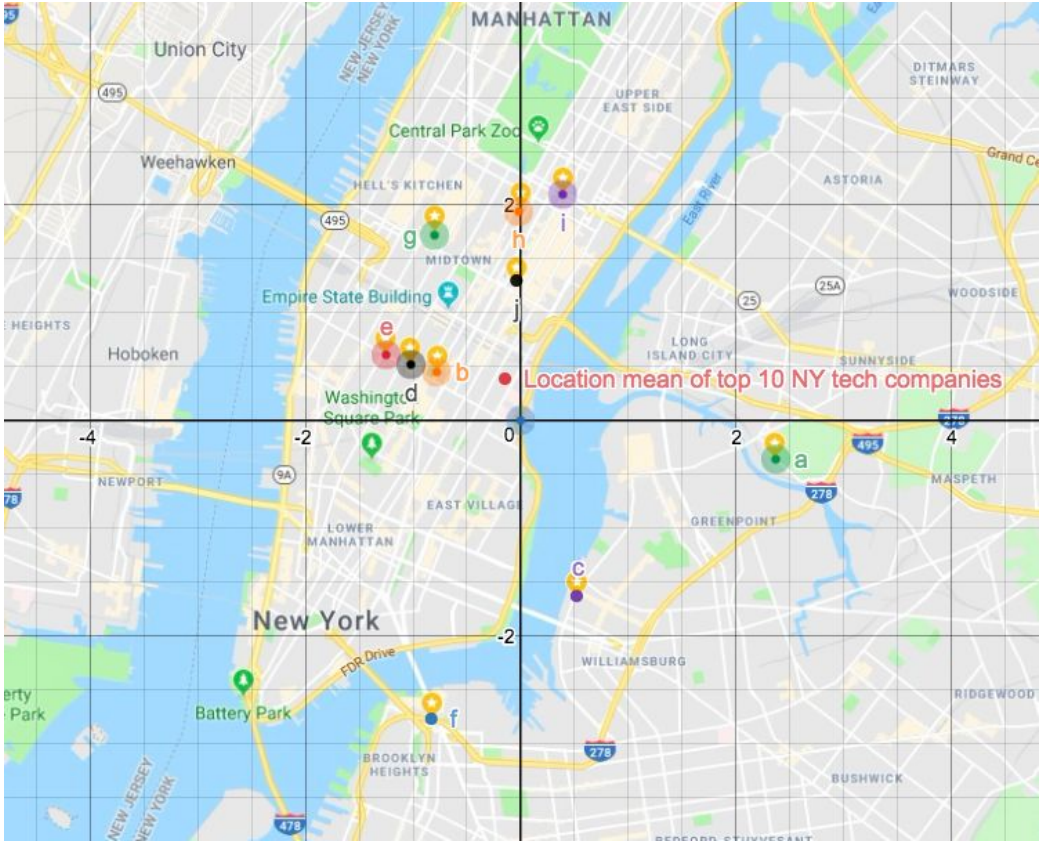


Student Work: Focusing on tech hubs

According to BuiltinNYC, the largest 10 tech companies in New York City are as follows:

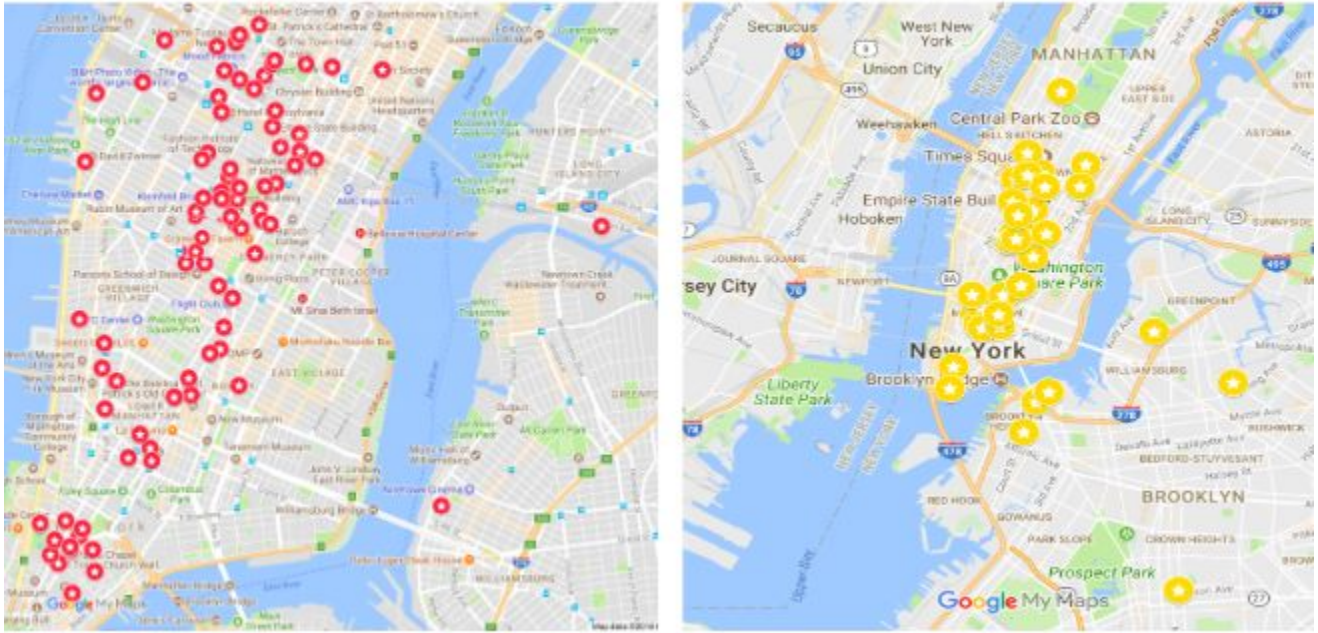
1.	Bloomberg	9,000 employees
2.	Oath	1,400
3.	CA Technologies	1,230
4.	Vice Media	1,217
5.	Blue Apron	890
6.	E*Trade	827
7.	BuzzFeed	730
8.	Yext	675
9.	FreshDirect	657
10.	Etsy	622

Student Work: Focusing on tech hubs

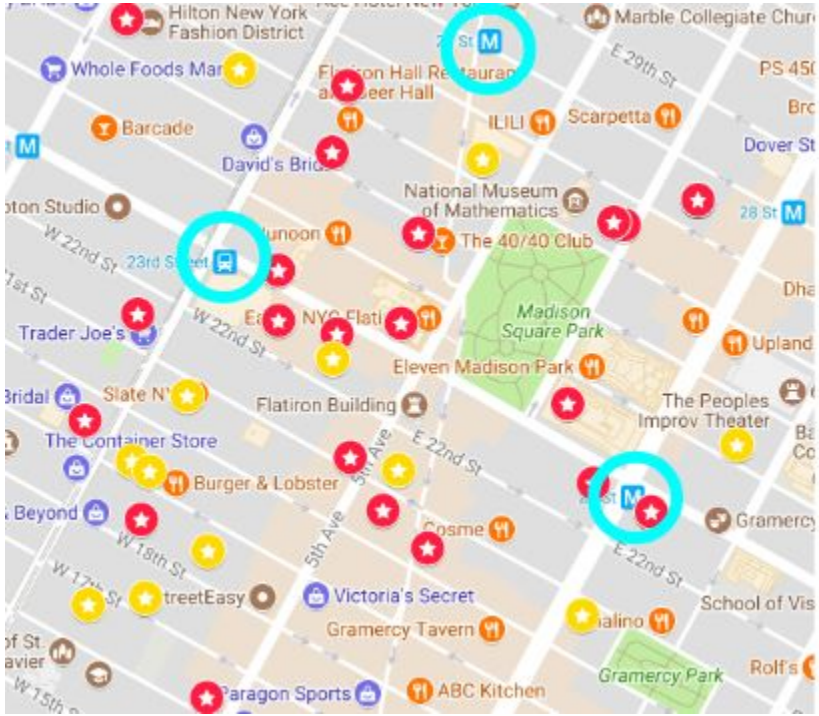


Student Work: Focusing on tech hubs

Established *vs* Startups



Student Work: Focusing on tech hubs



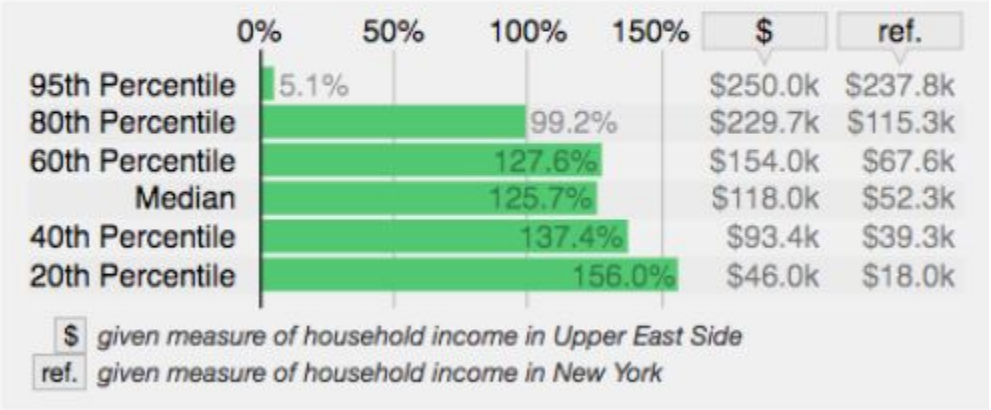
STATIONS

- 23rd St
- 28th St

Student Work: Focusing on Wealthier Donors

Our Strategy: Optimizing Donations + People

This graph shows how much wealthier Upper East Siders in each percentile are than their NY counterparts.



Student Work: Focusing on Weekends

Huge Debate: Subway entrances vs. exits

- What about the stations that have separate entrances and exits?

- **Targeting entrances:** riders will have time to read pamphlets on train
- But will they be in too much of a rush to make the train?

- **Targeting exits:** riders won't be in a rush to make the train
- But will they discard the pamphlet before reading it?

Student work: Subway entrances vs. exits

Overall Strategy

- Morning: EXITS
- Evening: ENTRIES
- Stops with most people AND near tech firms

Student Work: Finding commuter stops

$$\text{CommuterIndex} = \frac{\text{Weekday Avg}}{\text{Weekday Avg} + \text{Weekend Avg}}$$

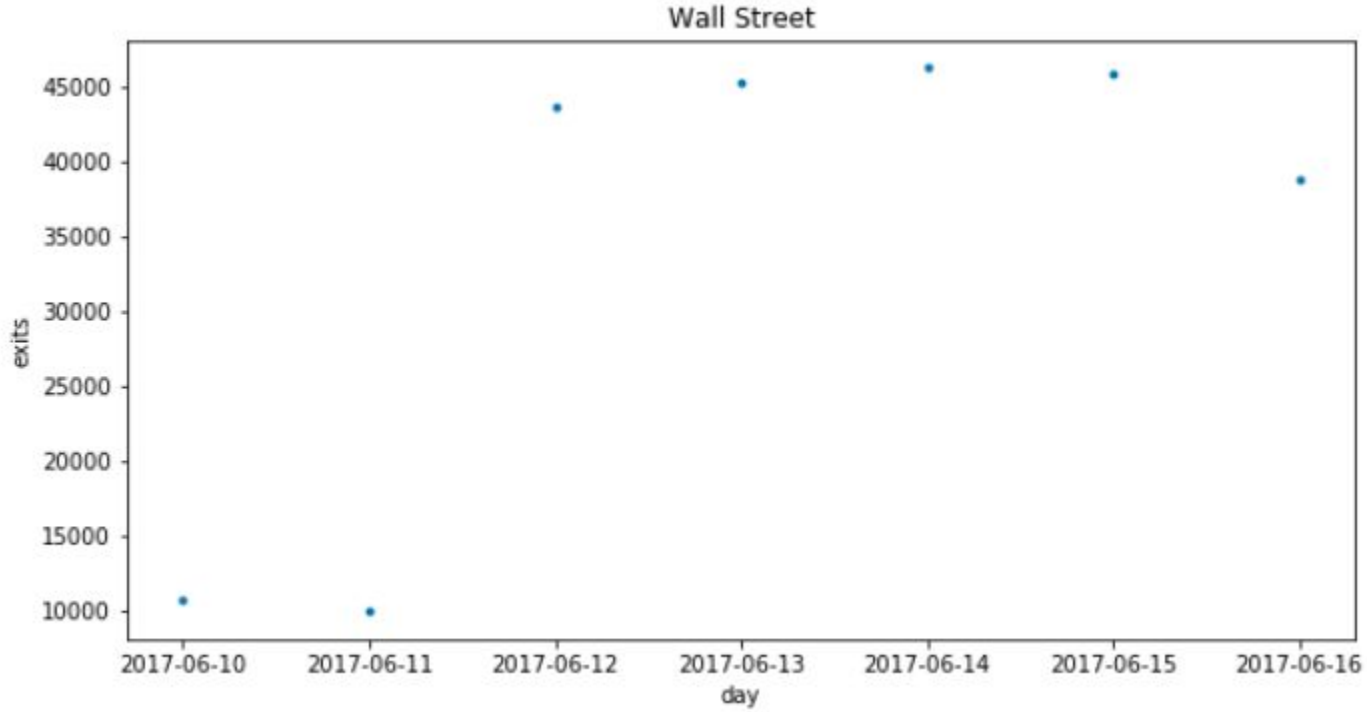
Top commuter exits

```
commuter_list = []
for station, indexes in commuter_dict.items():
    commuter_list.append((np.median(indexes), station))

for info in sorted(commuter_list, reverse = True):
    print(info)
```

```
for info in commuter_list:
    print(info)
```

```
(0.9997511653754915, 'NEW LOTS AV')
(0.9972474538948527, 'PENNSYLVANIA AV')
(0.9970398631758979, 'GREENPOINT AV')
(0.9961351862511307, 'SARATOGA AV')
(0.9921383647798743, 'FLUSHING AV')
(0.9918046924566759, 'MYRTLE-WILLOUGH')
(0.9913849588662121, 'NASSAU AV')
(0.9353417649566577, 'BAY 50 ST')
(0.9308408339103008, 'BOWLING GREEN')
(0.891933474979852, '25 AV')
(0.861607050713121, 'THIRTY ST')
(0.852875091487838, '5 AV/53 ST')
(0.8420157984201581, 'LACKAWANNA')
(0.8378077200096461, 'WALL ST')
(0.8260323427878745, 'NEWARK HW BMEBE')
(0.816057346842834, 'FULTON ST')
```

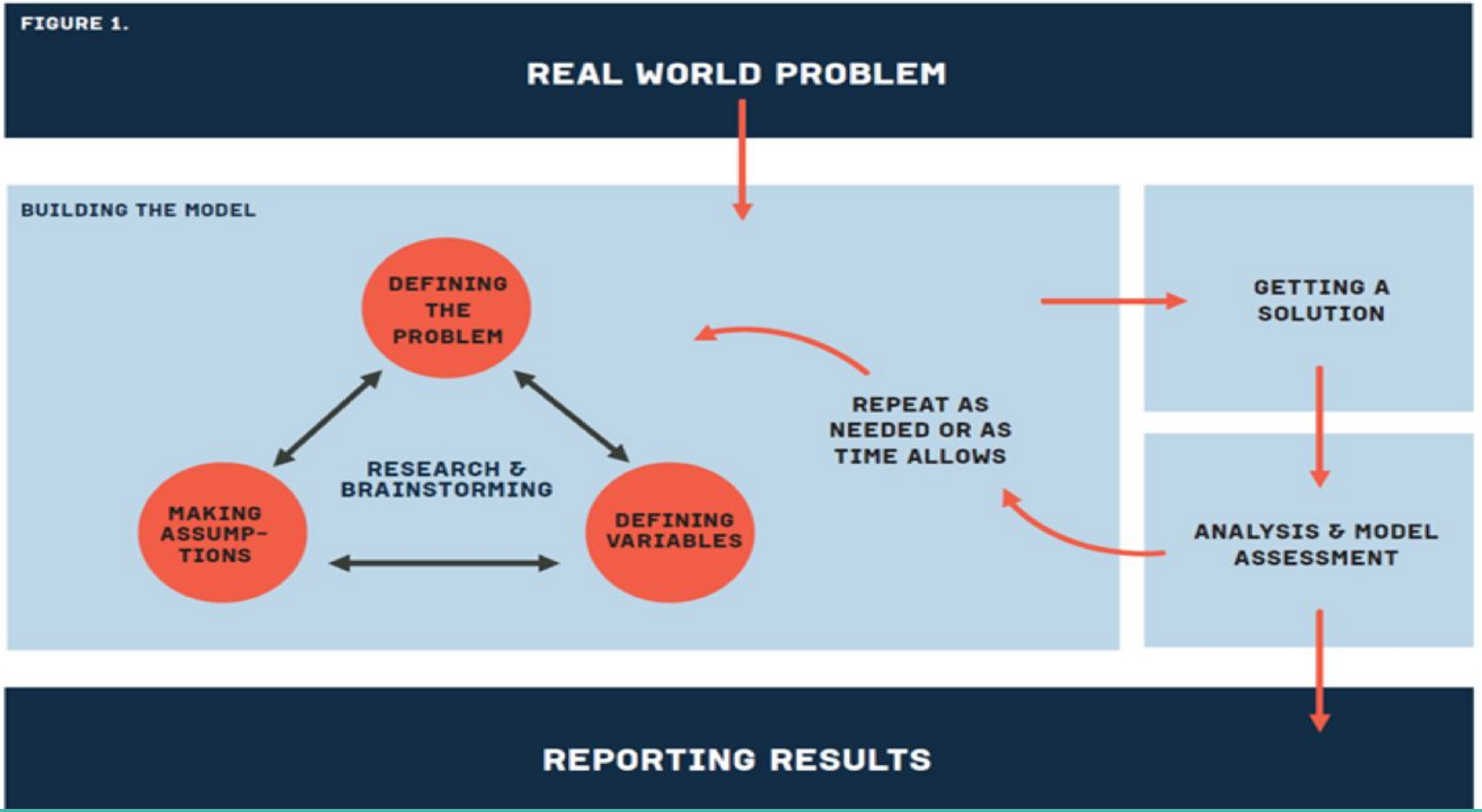
Commuter exit example: Wall Street



Tourist Heavy Exits

```
(0.4747766989931784, 'W 8 ST-AQUARIUM')
(0.4734389561975769, 'AQUEDUCT RACETR')
(0.4710146436977940 6, 'BEACH 67 ST')
(0.4664468561651078 4, 'YORK ST')
(0.4568337591613703 7, 'BEDFORD-NOSTRAN')
(0.4430009743423189 4, '161/YANKEE STAD')
(0.4286186111799190 6, 'ROCKAWAY PARK B')
(0.4166117336849044 5, 'RIT-ROOSEVELT')
(0.4103773584905660 6, 'ORCHARD BEACH')
(0.4077757685352622, 'BROAD CHANNEL')
(0.3824183976261127 4, 'BEACH 105 ST')
(0.374399676457206 4, 'BEACH 98 ST')
(0.3709898349073606 5, 'BEACH 90 ST')
(0.3441427853192559, 'AVENUE N')
```

FIGURE 1.


REAL WORLD PROBLEM
BUILDING THE MODEL
DEFINING THE PROBLEM
RESEARCH & BRAINSTORMING
MAKING ASSUMP-TIONS
DEFINING VARIABLES
REPEAT AS NEEDED OR AS TIME ALLOWS
GETTING A SOLUTION
ANALYSIS & MODEL ASSESSMENT
REPORTING RESULTS

Professional Presentation Requirements

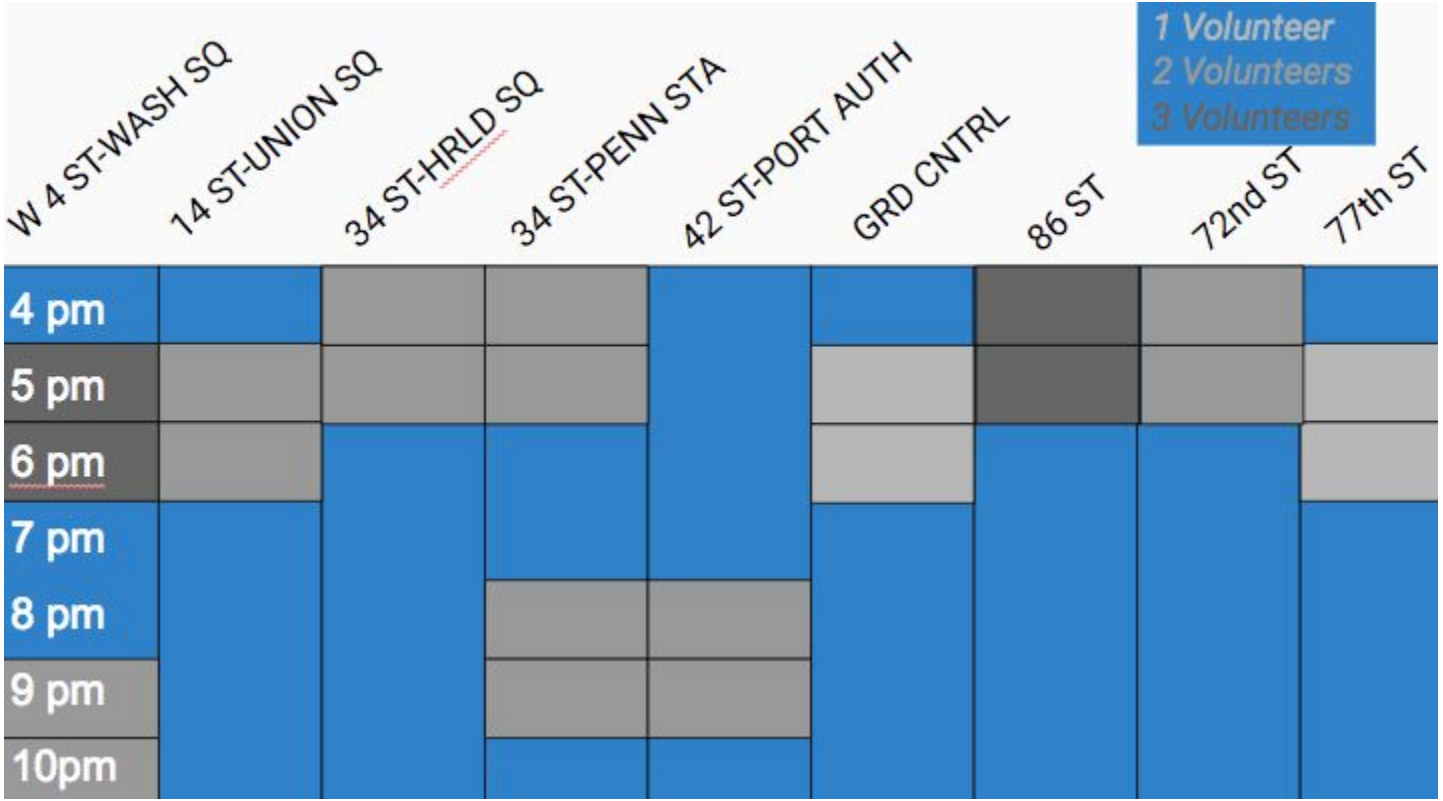
1. You need extremely clear slides and explanations.
2. YOU CANNOT HAVE TOO MUCH TEXT ON YOUR SLIDES.
3. You should create maps and charts to help visualize your suggestions.
4. Your analysis needs to be accurate and thoughtful.
5. Have at least one thing that is unique to your group, or else, why would this organization choose to hire YOUR company?

Student Work #1: Concluding summary

Takeaways

- We recommend putting 2 people at each station working shifts of:
  - 7am-9am, 4pm-6pm on weekdays
  - 10am-2pm on weekends
- What sets us apart:
  - We focus on subway exits instead of entrances
  - Weekend stops target residences
  - Weekday stops target both major tech companies AND large commuter stops

Student Work #2: Concluding summary



Many extensions

- Graphical packages
- Google API
- Data cleaning

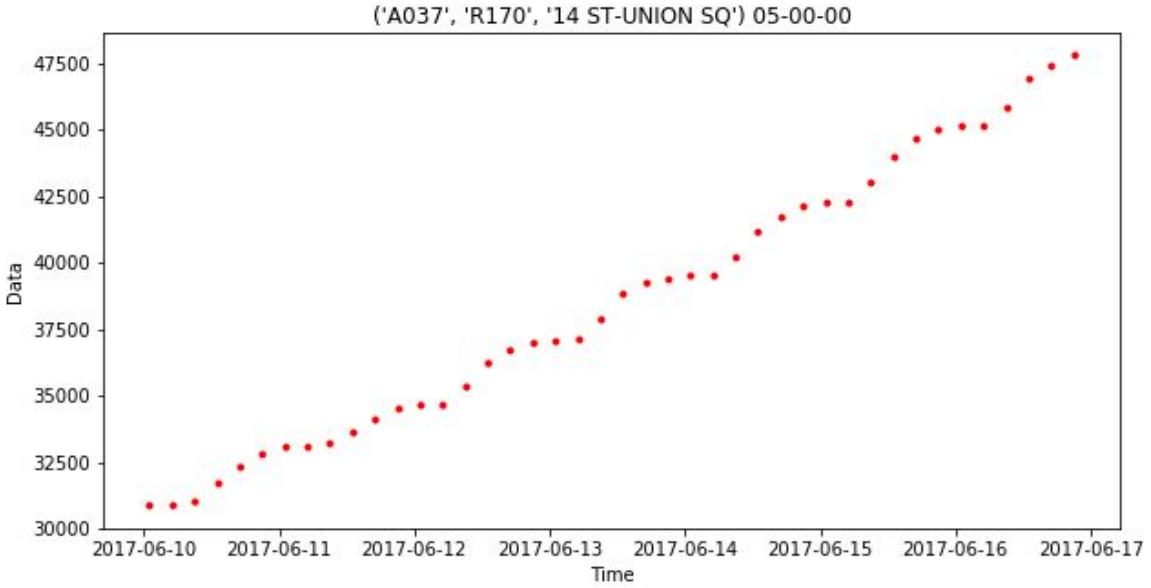
Extension #1: Animations

Extension #2: Working with Google Maps API

```
import requests
import json

url = 'http://maps.googleapis.com/maps/api/geocode/json?'

address = 'Penn Station New York City'

params = {'address': address}
data = requests.get(url, params=params)

js = json.loads(data.text)

print(js['results'])
```

Extension #3: Data Cleaning

If the data is correct, the count at each turnstile should be monotonically increasing like this:

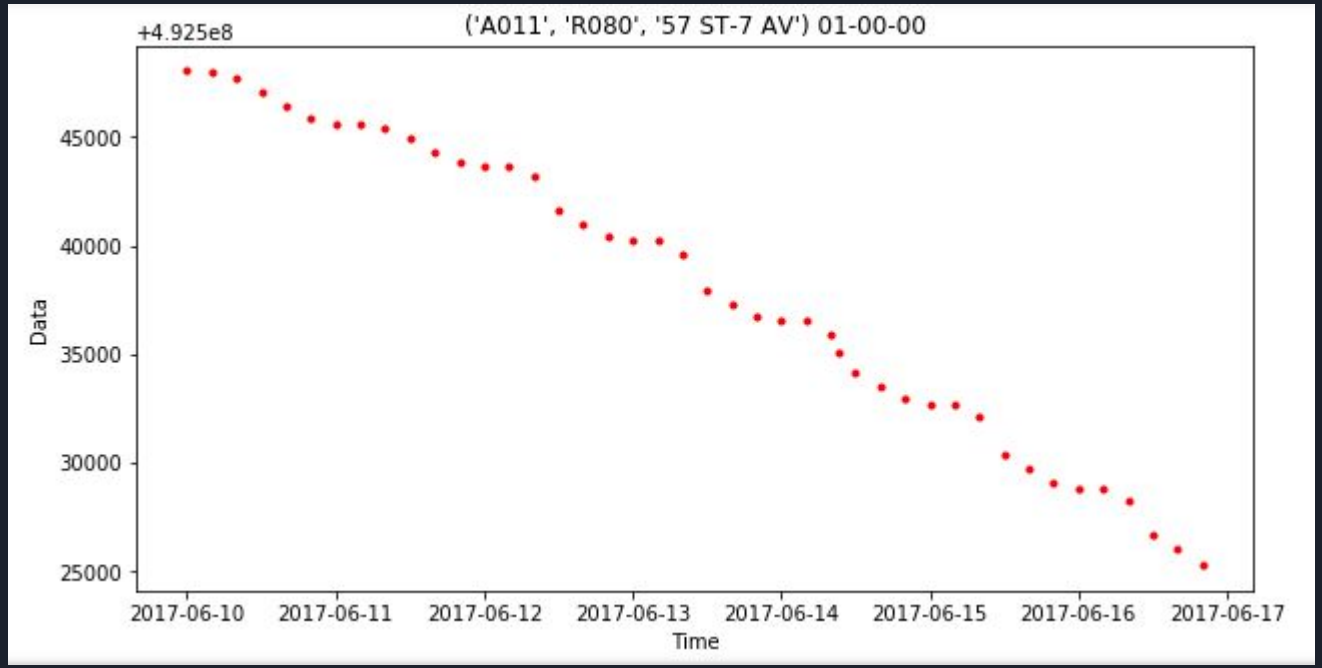


Type of Errors

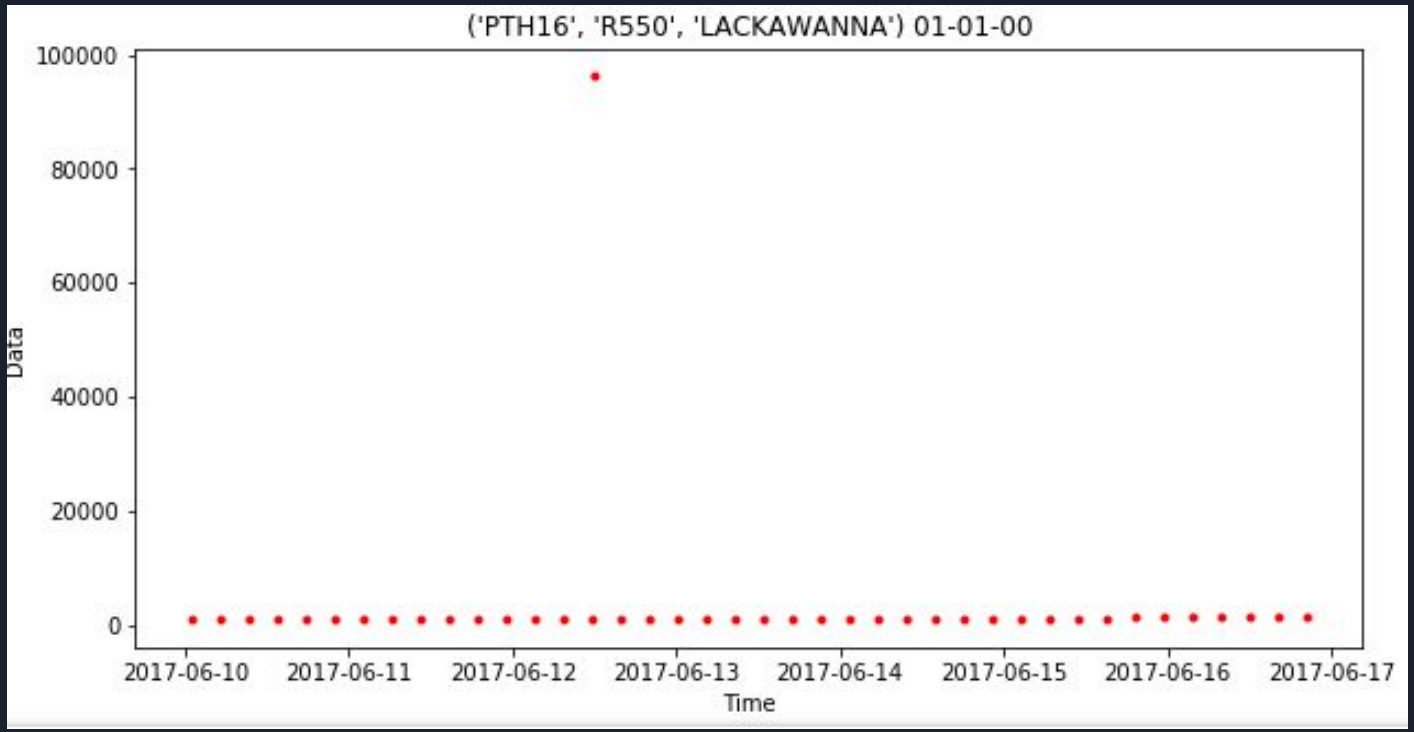
59 turnstiles have incorrect data, in 3 types

1. Monotonic Decreasing
2. Garbage Values
3. Turnstile Resets

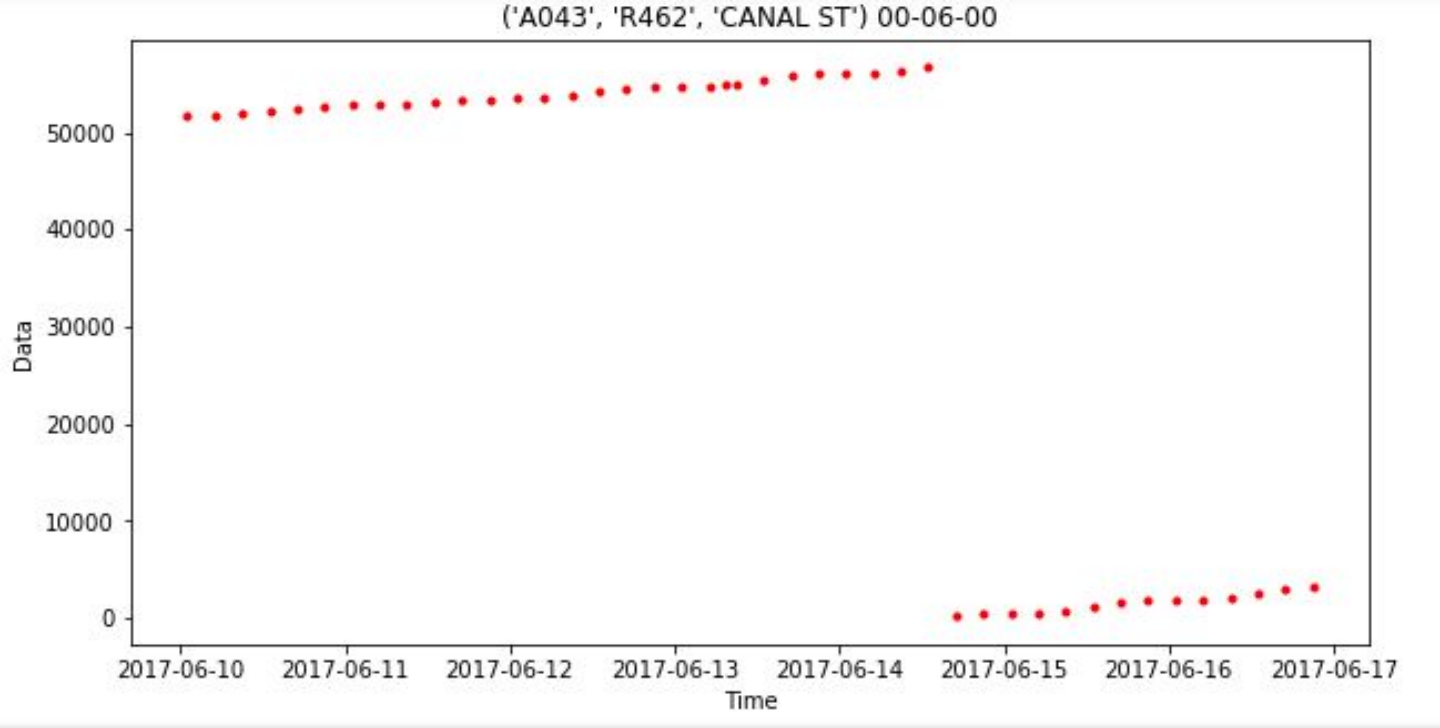
Monotonic Decreasing



Garbage Value



Turnstile Reset



Tech required for data analysis

- Python 3.6 in a Jupyter Notebook (www.anaconda.com/download/)

- matplotlib (plotting)
- **pandas (spreadsheets)**
- NumPy (arrays, math functions, linear algebra)

Free Resources

- UC Berkeley Foundations of Data Science: http://data8.org/
- Think Stats 2e & Think Python 2e, Allen B. Downey http://greenteapress.com/wp/think-stats-2e/
- Python For Everyone, Charles Severance: https://www.py4e.com/html3/
- My course materials:
- https://github.com/laurenshareshian/Python_For_Math_Teachers
- https://github.com/laurenshareshian/Python_Course_Lessons

Thanks!

Lauren Shareshian

Oregon Episcopal School

shareshianl@oes.edu