Project Benson

MTA Turnstile Data Analysis

Team: Connor Stefan, Prashant Tatineni, and Rosie Hoyem

The Challenge

WomenTechWomenYes (WTWY) wants to leverage data to optimize recruitment efforts at subway stations for an upcoming event. They would like to target specifically women in tech industries.

Assumptions

> Traffic = entries + exits

> Gala is being held early June, thus subway data from May 2016 (4 weeks)

> Target participant/contributor is a woman working in the tech industry

> Further assume target demographic will use subway stations near their place of residence; not considering major commute stations as commuters would be less likely to attend the Gala

Solution Steps

1. Locate high density residential areas for women working in tech industries (US census data) and rank the top stations
2. Merge the subways stops (City of New York dataset) on the Census data in ArcGIS
3. Use MTA data (May 2016) to determine which stops near these locations are the most popular and rank the top stations
4. Create a combined rank of total volume and female tech workers

The Data

Leveraging Census Data

Versions of this table are available for the following years:

2015
2014
2013
2012
2011
2010
2009

1 - 11 of 11

	Bronx County, New York	Kings County, New York	New York County, New York	Queens County, New York	Richmond County, New York
	Estimate	Estimate	Estimate	Estimate	Estimate
Total:	563,903	1,167,448	884,457	1,104,930	209,113
Female:	284,891	584,677	439,087	514,432	99,062
Management, business, science, and arts occupations:	82,000	243,800	263,314	191,012	45,815
Management, business, and financial occupations:	23,580	72,178	105,841	64,218	11,781
Computer, engineering, and science occupations:	3,387	14,845	19,499	11,601	2,344
Education, legal, community service, arts, and media occupations:	38,357	115,486	109,629	74,427	20,683
Healthcare practitioners and technical occupations:	16,676	41,291	28,345	40,766	11,007
Service occupations:	107,015	166,236	64,777	143,846	19,715
Sales and office occupations:	83,148	150,542	100,987	154,050	30,819
Natural resources, construction, and maintenance occupations:	1,455	2,658	1,254	3,191	477
Production, transportation, and material moving occupations:	11,273	21,441	8,755	22,333	2,236

Source: U.S. Census Bureau, 2011-2015 American Community Survey 5-Year Estimates

Heat Map of Census Data

Subway Stations and Census Tracts

Source: City of New York DCP, US Census Bureau

Getting Technical:

Regex used on station names to make a merge between the subway geographic data and the turnstile data possible

```
import re

capitalizer = lambda x: x.upper()
census['name'] = census['name'].apply(capitalizer)

clean_ave = lambda x: re.sub('AVE','AV', x)
census['name'] = census['name'].apply(clean_ave)

clean_numbers1 = lambda x: re.sub(r'(\b\d+)(RD\b)',r'\1', x)
census['name'] = census['name'].apply(clean_numbers1)

clean_numbers2 = lambda x: re.sub(r'(\b\d+)(TH\b)',r'\1', x)
census['name'] = census['name'].apply(clean_numbers2)

clean_numbers3 = lambda x: re.sub(r'(\b\d+)(ST\b)',r'\1', x)
census['name'] = census['name'].apply(clean_numbers3)

clean_numbers4 = lambda x: re.sub(r'(\b\d+)(ND\b)',r'\1', x)
census['name'] = census['name'].apply(clean_numbers4)

clean_dashes = lambda x: re.sub(r'(\w*)(\s-\s)(\w*)',r'\1-\3', x)
census['name'] = census['name'].apply(clean_dashes)

remove_parans = lambda x: re.sub(r'\(.*\)',r'', x)
census['name'] = census['name'].apply(remove_parans)

census.replace('CONCOURSE', 'CONC', inplace=True)
census.replace('AVENUE', 'AV', inplace=True)
census.replace('WASHINGTON', 'WASH', inplace=True)
census.replace('JUNCTION', 'JCT', inplace=True)
census.replace('CONCOURSE', 'CONC', inplace=True)
census.replace('WOODHAVN', 'WOODHAVEN', inplace=True)
census.replace('CENTER', 'CTR', inplace=True)
census.replace('QUEENSBRIDGE', 'QNSBRIDGE', inplace=True)
census.replace('W 4 ST-WASHINGTON SQ ', 'W 4 ST-WASH SQ', inplace=True)
```

MTA Turnstile Data

> 4 weeks worth of data (May 2016)

> over 779,000 rows

>Very messy!

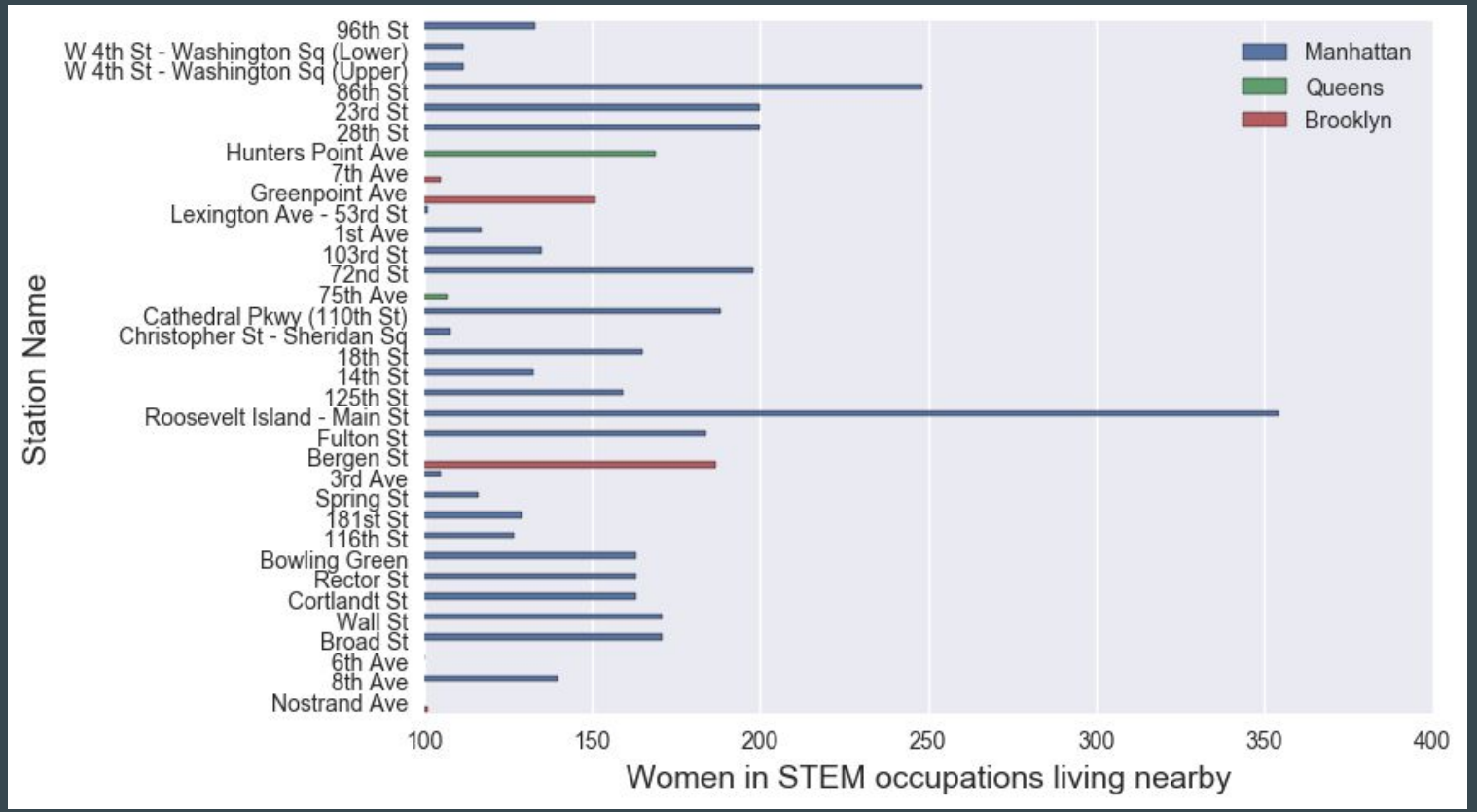
Cleaning the Data

	C/A	UNIT	SCP	STATION	LINENAME	DIVISION	DATE	TIME	DESC	ENTRIES	EXITS	diffEntries	diffExits	machine_change
169	A002	R051	02-00-00	59 ST	NQR456	BMT	05/27/2016	20:00:00	REGULAR	5682728	1924311	780.0	51.0	False
170	A002	R051	02-00-01	59 ST	NQR456	BMT	04/30/2016	00:00:00	REGULAR	5180722	1137502	0.0	0.0	True

STATION	LINENAME	DIVISION	DATE	TIME	DESC	ENTRIES	EXITS	diffEntries	diffExits	machine_change	counterReset
57 ST-7 AV	NQR	BMT	04/30/2016	00:00:00	REGULAR	888004887	493739559	0.0	0.0	True	False
57 ST-7 AV	NQR	BMT	04/30/2016	04:00:00	REGULAR	888004721	493739477	0.0	0.0	False	True
57 ST-7 AV	NQR	BMT	04/30/2016	08:00:00	REGULAR	888004670	493739166	0.0	0.0	False	True

Results

High Density Areas for Females in Tech



Source: US Census Bureau

PROOF (Rosie's House on 7th Ave)

Top Ranked Stops By Female Stem Residents

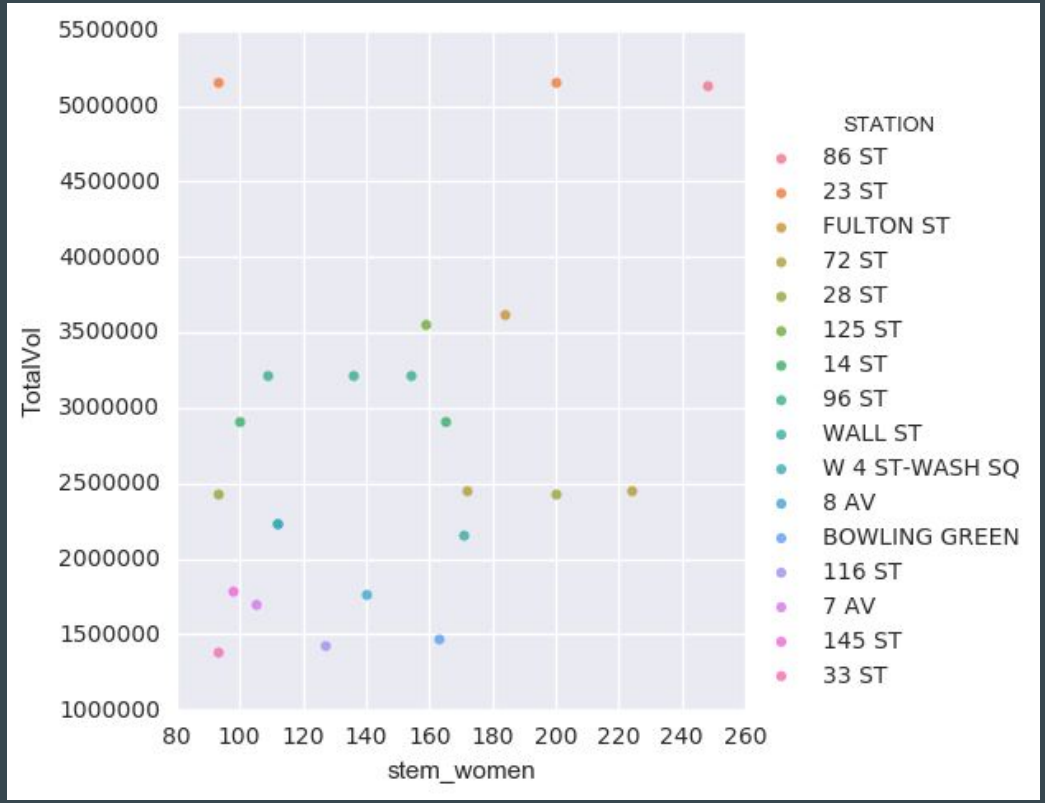
Rank	Stop	Borough
1	ROOSEVELT ISLAND	Manhattan
2	86th ST	Manhattan
3	72nd ST	Manhattan
4	28th ST	Manhattan
5	23rd ST	Manhattan

Top Ranked Stops by Total Volume

Rank	Stop	Borough
1	34th ST-PENN STATION	Manhattan
2	GRD CENTRAL-42th ST	Manhattan
3	34 ST-HERALD SQ	Manhattan
4	23th ST	Manhattan
5	86th ST	Manhattan

Source: City of New York Metro Transit Authority

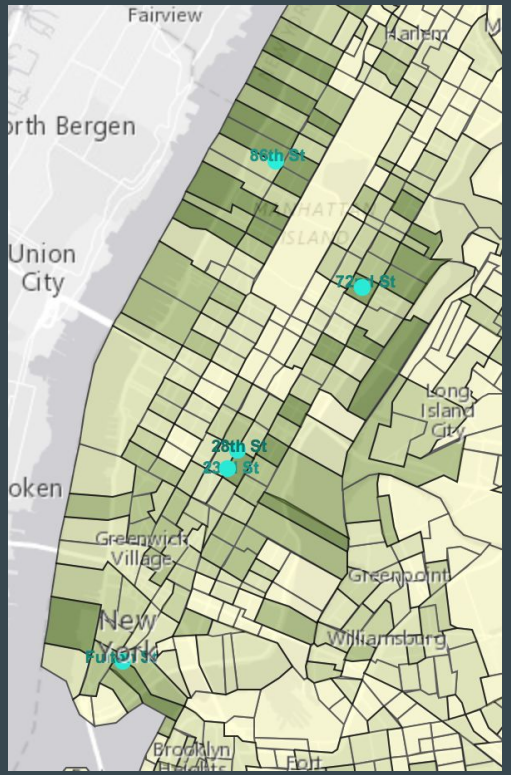
Top Stations by Total Volume and Female STEM Residents



Source: City of New York Metro Transit Authority, US Census Bureau

Combined Rank Stops

Rank	Station Name	Borough
7	86th ST	Manhattan
9	23rd ST	Manhattan
17	FULTON ST	Manhattan
22	72nd ST	Manhattan
24	28th ST	Manhattan



If time permitted...

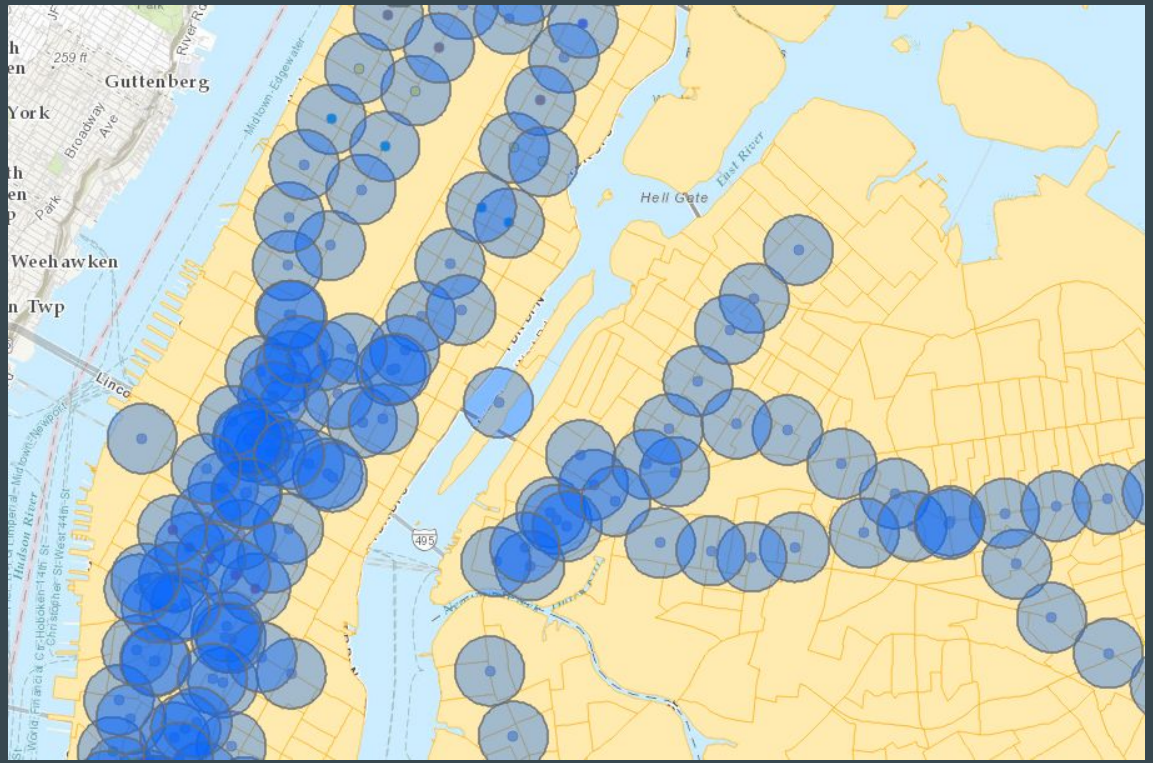
> Detailed time analysis

> More in depth data cleaning!

> Target Demographic information

> Revising census tract approach

Ex: Buffer Analysis



Source: City of New York DCP, US Census Bureau