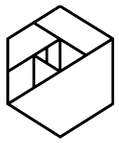
WOMEN IN TECH AND WHERE TO FIND THEM

Exploratory Data Analysis of MTA Turnstiles

Danish James, Flora Xinru Cheng, Luke Newman



METIS

September 27 2019

OBJECTIVES

- Analyze MTA data to:
  - find stations with the highest weekly traffic in May and June 2019
  - Identify stations most likely to attract people who are:
    - women working in or studying technology
    - interested in increasing participation of women in tech

objective, problem
**narrative - what to emphasize to the audience? [put street teams near stns with top weekly counts near targeted groups(uni and tech)]

Background:
WomenTechWomenYes (WTWY) has an annual gala at the beginning of the summer to build awareness and increase participation of women in technology

METHODOLOGY

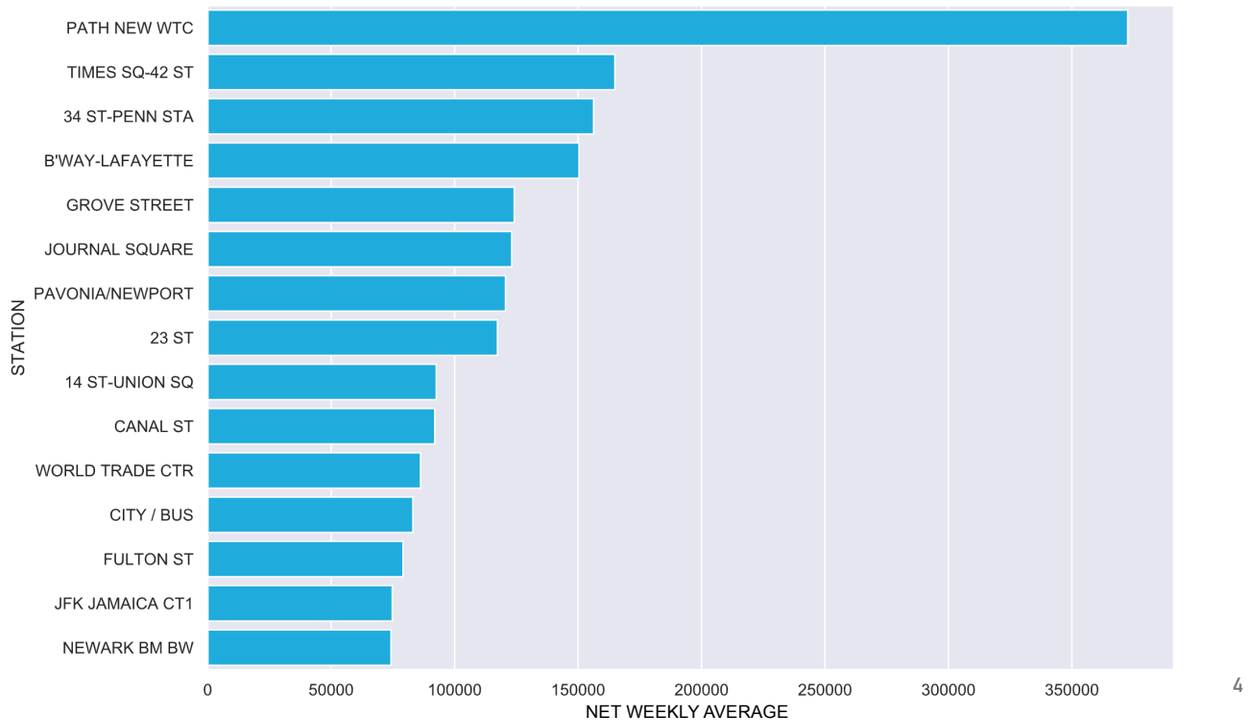
- DATA
  - MTA Turnstiles data (MTA)
  - Additional geographical data:
    - Subway stations (NYC OpenData)
    - Colleges and universities (NYC OpenData)

- TOOLS
  - Python, Pandas, NumPy
  - Matplotlib, Seaborn, GeoPandas
  - GitHub

*minimum: MTA data, which stations have highest traffic flow (entries + exits)
— date time —> each day of the week
— outliers due to tourists
— justification for manual top 3 rec.
—1 week;—> March to June
next step: location; — before plotting, top 3 to do rainbow plot
then tech; then universities (might go to further work section)

RESULTS

➤ Plot 1: Top 15 stations with highest weekly ridership values



RESULTS

- Plot 2: Geographical map of top stations and local universities

identify top 3, label on map (manually?)

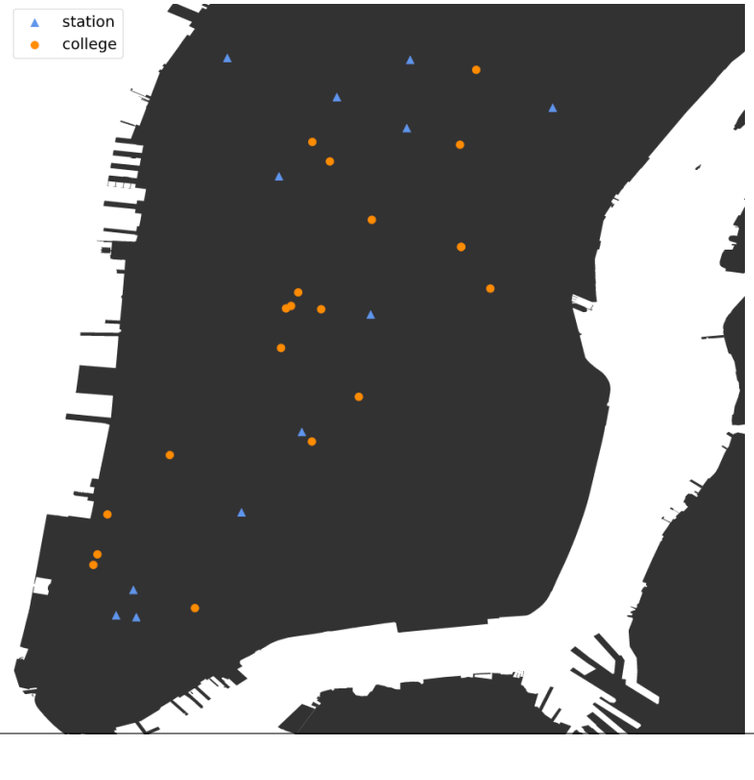
RESULTS

- Plot 2: Geographical map of top stations and local universities

identify top 3, label on map (manually?)

RESULTS

- Plot 2: Geographical map of top stations and local universities



identify top 3, label on map (manually?)

RESULTS

- Plot 2: Geographical map of top stations and local universities

identify top 3, label on map (manually?)

CONCLUSION

- Top stations close to universities
- Put street teams near these stations:
  - 14 ST-UNION SQ
  - 23 ST (8TH AVE Line)
  - FULTON ST

**narrative - what to emphasize to the audience? [put street teams near stns with top weekly counts near targeted groups(uni and tech)]
*recommendations, interesting insights - keep audience in mind
* 3 bullet points ideal

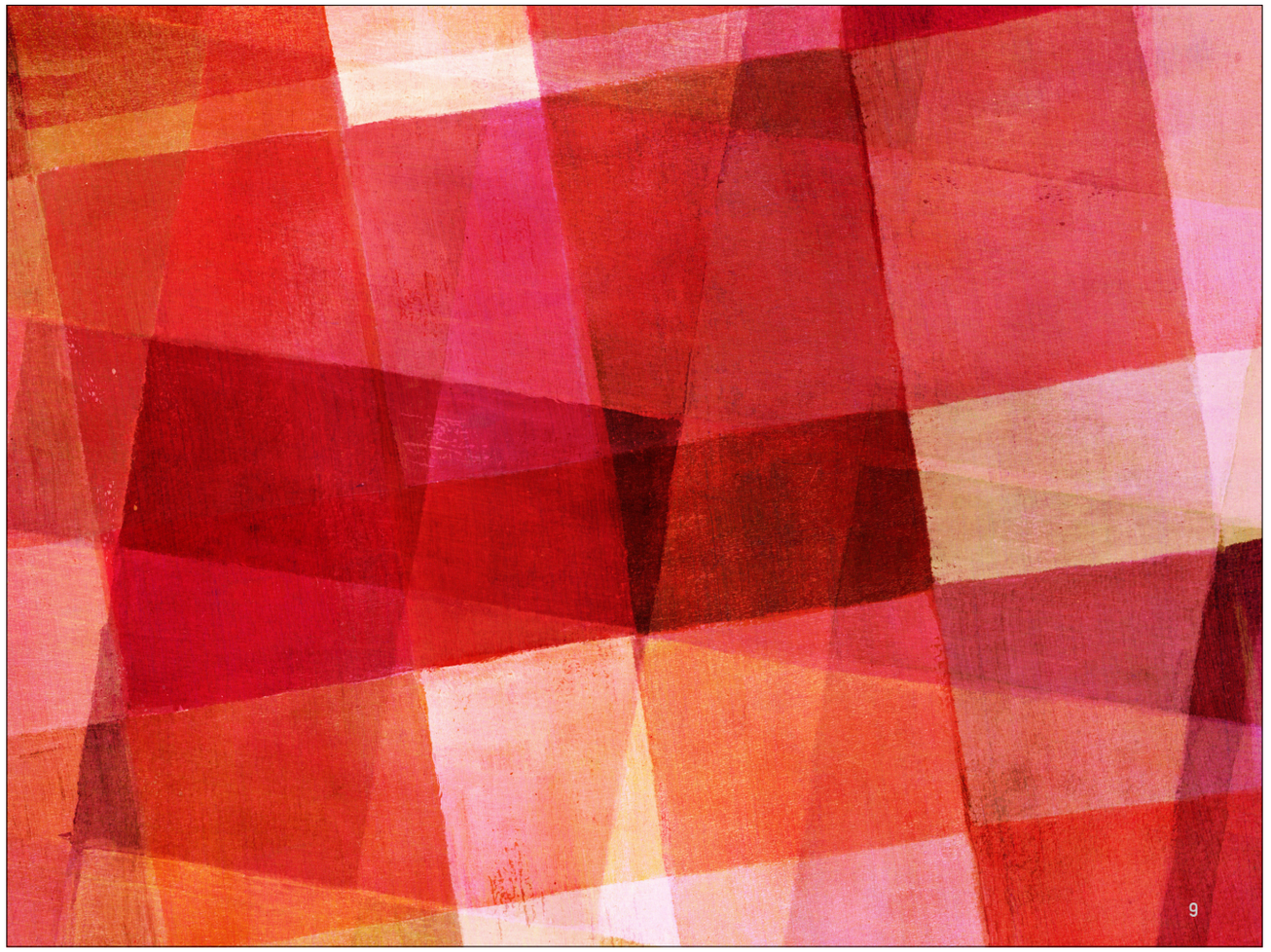
FURTHER WORK

- Plotting location of tech companies in addition to universities
  - Scraping
- Algorithmically compare top stations based on how many tech companies and universities are within a set radius to each

*have dataset of companies but not tech companies specifically
*If we had more time/resources
(pitch, know what we are able to do)

APPENDIX

- Time period: April to June, exclude major holidays and events
- Data cleaning problem: turnstiles resetting
- Reasoning for selecting top recommendations
  - domain knowledge of NYC
  - neighbourhoods
  - outliers - tourism, major transit hubs
- Other factors: Multiple turnstiles per station



This is fine.