MTA Turnstile Data Analysis

—

Team: Nick Sherwin, Shuo Jia, and Fahimeh Khaleghi

Objective

To maximize the number of signatures obtained at subway station entrances/exits via street marketing teams, focusing on those individuals who will attend the gala and contribute to WTWY’s cause.

Methodology

Data sources + references:

- MTA Data
- New York Home Price $ Values Data (Zillow)
- Mapping NYC’s Top 10 Most Funded Zip Codes

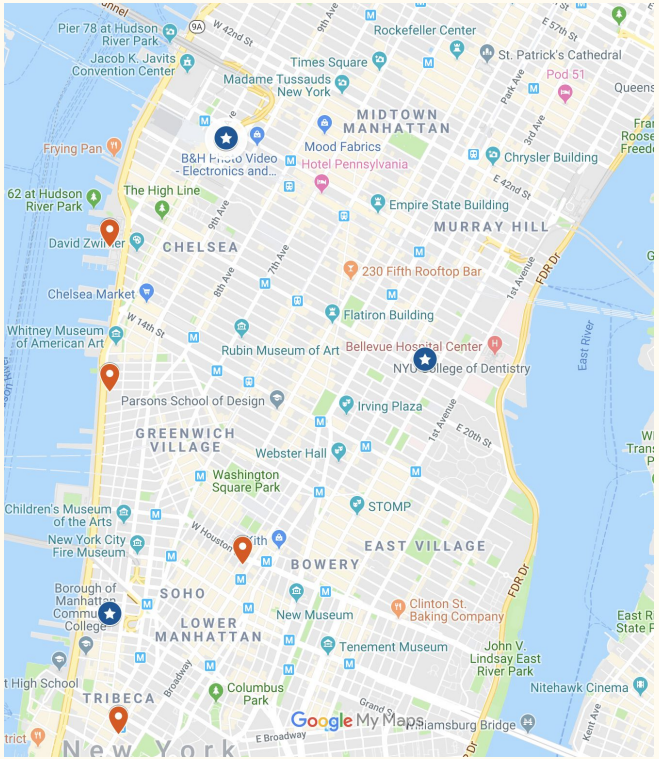
Tools:

- Pandas, Numpy, Matplotlib, Seaborn, Datetime, and Dateutil

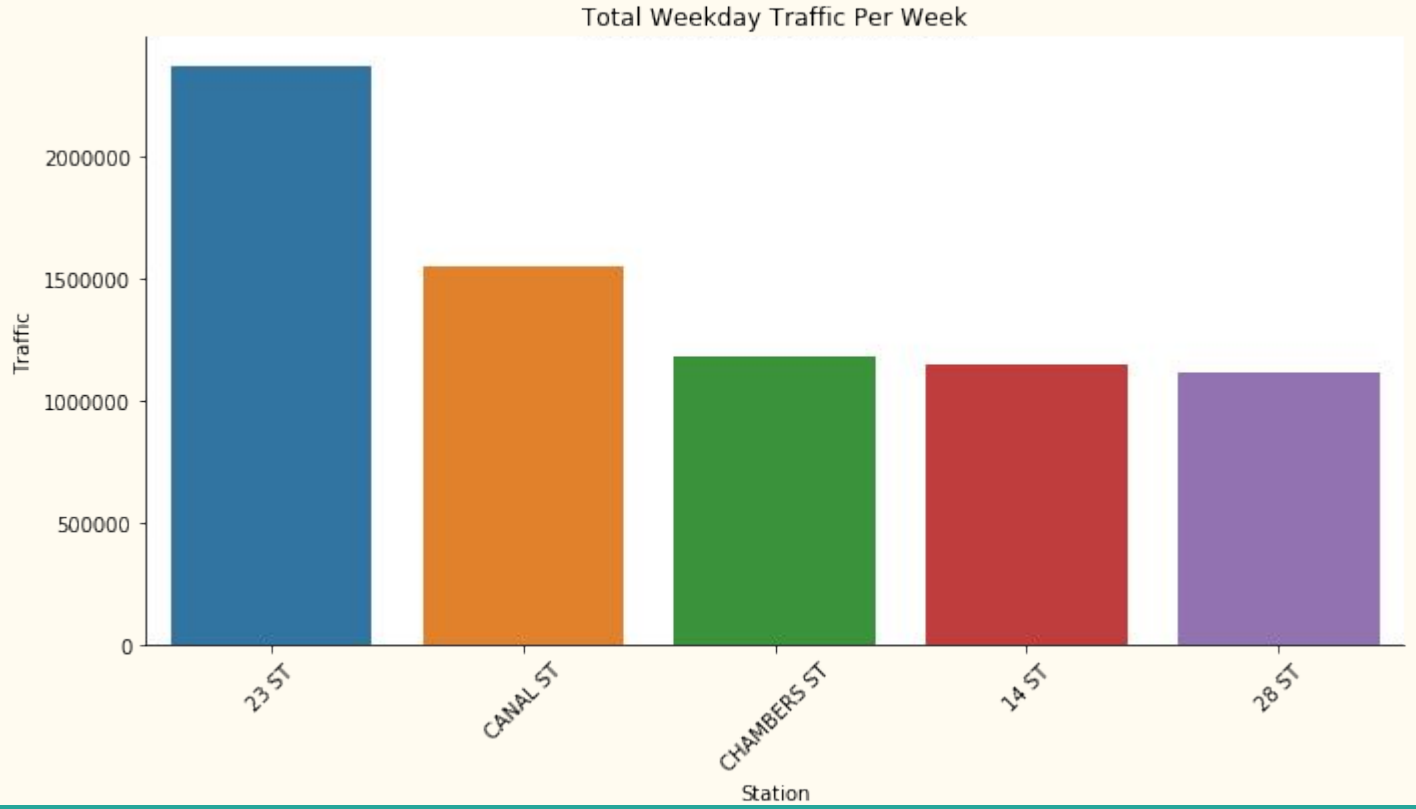
Methodology & Workflow

- Quality over quantity
- Focus on March to May 2019

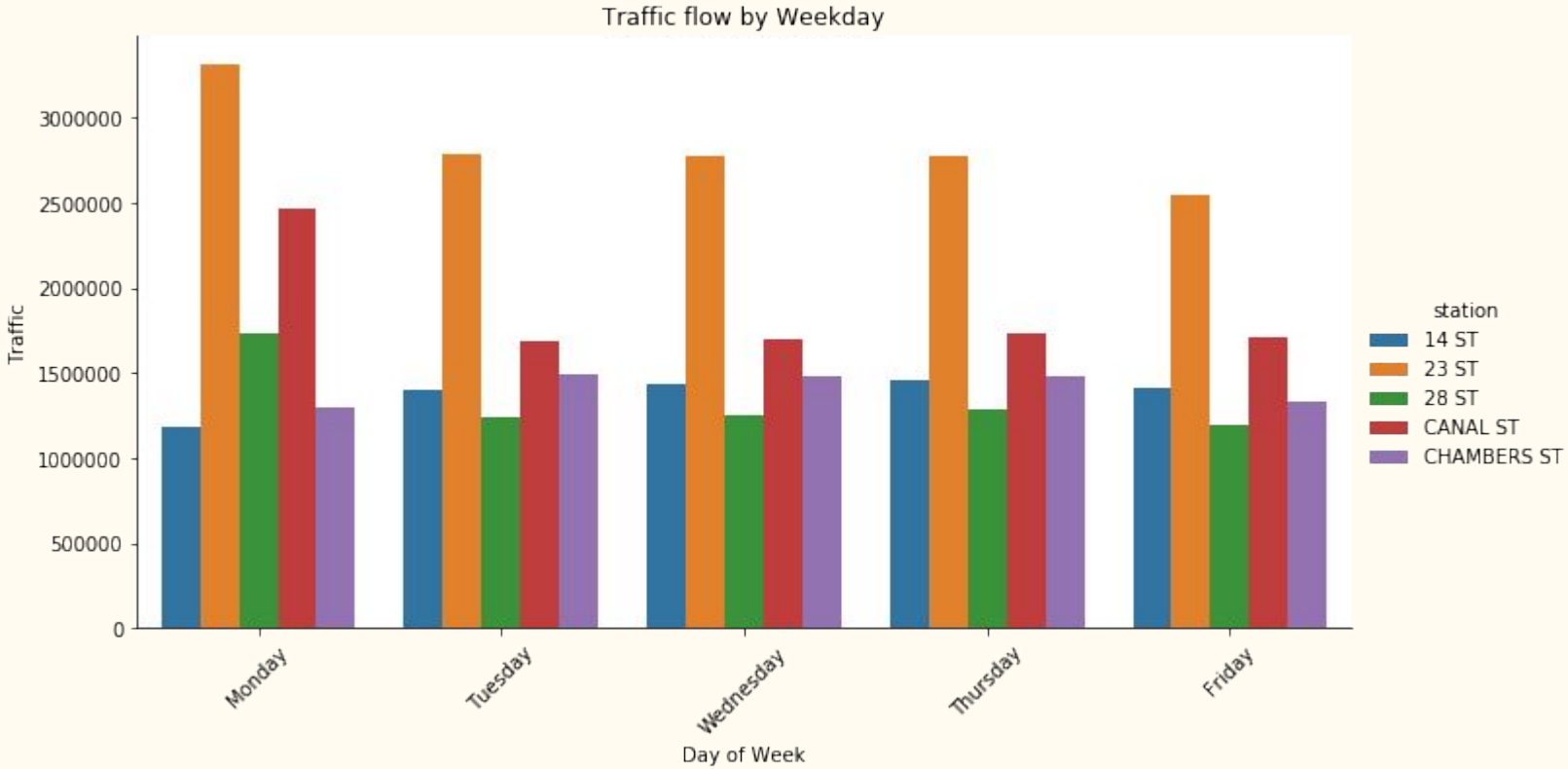
- Google Maps + Zillow + NYC Investment Data
- MTA Traffic Patterns
- Correlation between sources



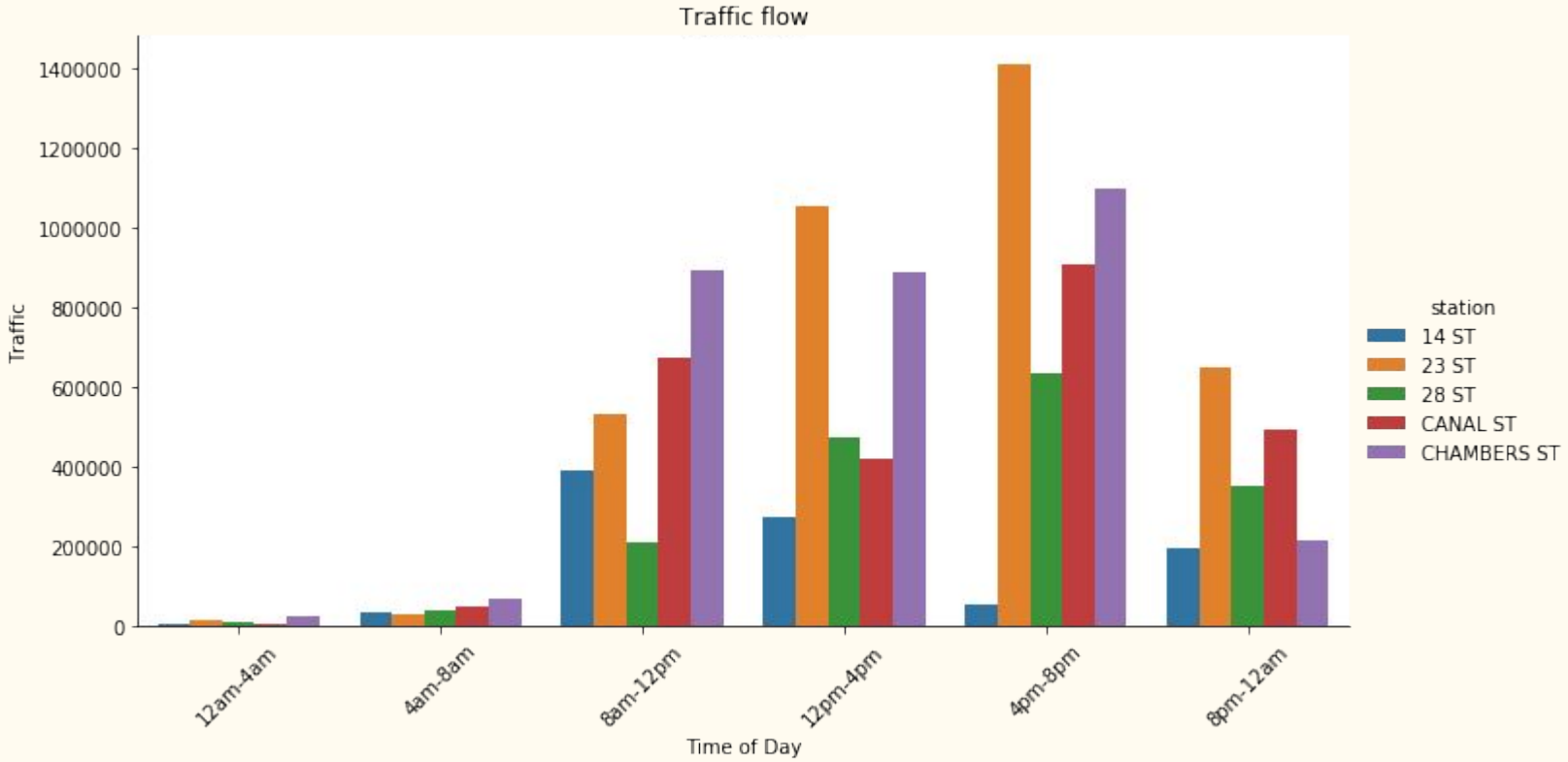
Results & Findings



Results & Findings



Results & Findings



Conclusion

Highest Volume:

- Between 4pm and 8pm at **23rd St**

Most Affluent:

- Between 4pm and 8pm at **Canal St**

Well-Funded:

- Between 4pm and 8pm at **28th St**

Future Work

Combine other data sources:

- US Census Data
- IRS Data
- Weather Data
- Tourism

Questions for *WTWY*:

- Marketing Budget? Past Success with campaigns? Demographic data?

Questions?

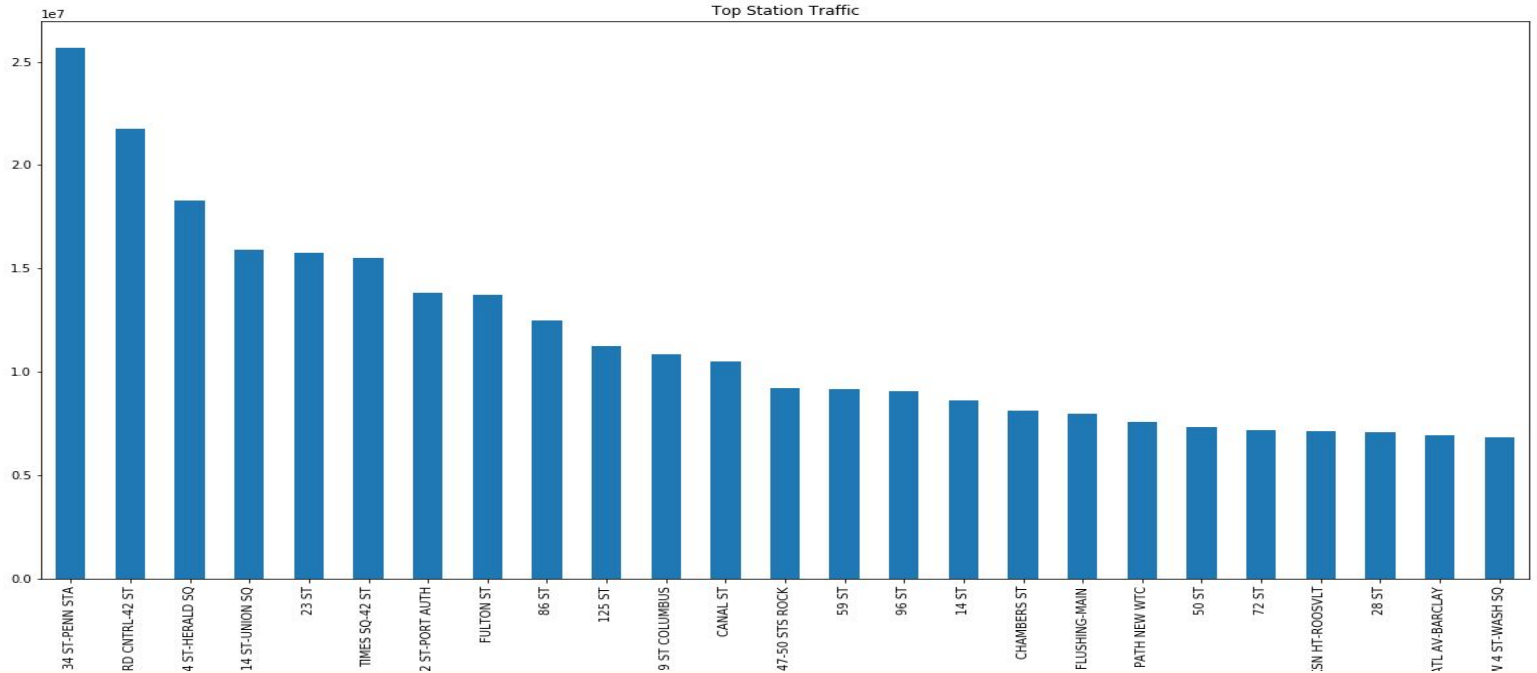
Appendix

Google Maps API Code

In [14]:
```
zips = pd.DataFrame(data=None)
gmaps_base_url = 'https://maps.googleapis.com/maps/api/geocode/json?'
api_key = 'AIzaSyBDYTJ7GVHC7R_-zhmw_48Apkan3mWoOj0'
scontext = None
```

In [18]:
```
for station in mta_stations:
    try:
        search_criteria = {'address': station + ' station, New York, NY',
                           'key' : api_key
                           }
        url = gmaps_base_url + urllib.urlencode(search_criteria)
        uh = urllib.urlopen(url, context=scontext)
        data = uh.read()
        js = json.loads(str(data))
        dicts = js['results'][0]['address_components']
        zip_dict = (item for item in dicts if item["types"] == [ "postal_code" ]).next()
        zip_code = zip_dict['long_name']
        zips = zips.append(pd.Series((station, zip_code)), ignore_index=True)
    except:
        print station
```

Overall Time Series Plots - Total Traffic



Overall Time Series Plots - Distribution