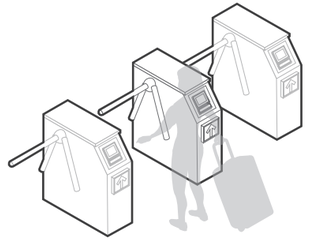


WOMEN TECH WOMEN YES

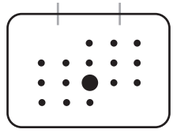
2020 Summer Gala EDA

Da Guo, Darien Mitchell-Tontar & Al Yoo

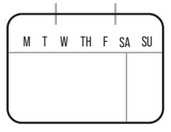
PROBLEM RECAP

Best subway stations to place your teams in order to collect email addresses from those who are most likely to attend.

INVESTIGATION OUTCOME



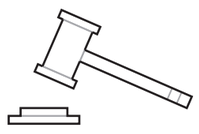
The busiest stations by daily average



The busiest stations by weekday

The optimal time to visit busiest stations

Highest female income across NYC boroughs



Recommendation based on data analyzed

PROCESS

	C/A	UNIT	SCP	STATION	LINENAME	DIVISION	DATE	TIME	DESC	ENTRIES	EXITS
0	A002	R051	02-00-00	59 ST	NQR456W	BMT	2019-04-27	00:00:00	REGULAR	7035249	2384833
1	A002	R051	02-00-00	59 ST	NQR456W	BMT	2019-04-27	04:00:00	REGULAR	7035269	2384840
2	A002	R051	02-00-00	59 ST	NQR456W	BMT	2019-04-27	08:00:00	REGULAR	7035292	2384875
3	A002	R051	02-00-00	59 ST	NQR456W	BMT	2019-04-27	12:00:00	REGULAR	7035392	2384951
4	A002	R051	02-00-00	59 ST	NQR456W	BMT	2019-04-27	16:00:00	REGULAR	7035651	2385020
5	A002	R051	02-00-00	59 ST	NQR456W	BMT	2019-04-27	20:00:00	REGULAR	7035930	2385070
6	A002	R051	02-00-00	59 ST	NQR456W	BMT	2019-04-28	00:00:00	REGULAR	7036100	2385087
7	A002	R051	02-00-00	59 ST	NQR456W	BMT	2019-04-28	04:00:00	REGULAR	7036119	2385088
8	A002	R051	02-00-00	59 ST	NQR456W	BMT	2019-04-28	08:00:00	REGULAR	7036125	2385103
9	A002	R051	02-00-00	59 ST	NQR456W	BMT	2019-04-28	12:00:00	REGULAR	7036197	2385155
10	A002	R051	02-00-00	59 ST	NQR456W	BMT	2019-04-28	16:00:00	REGULAR	7036372	2385198
11	A002	R051	02-00-00	59 ST	NQR456W	BMT	2019-04-28	20:00:00	REGULAR	7036621	2385240
12	A002	R051	02-00-00	59 ST	NQR456W	BMT	2019-04-29	00:00:00	REGULAR	7036746	2385256
13	A002	R051	02-00-00	59 ST	NQR456W	BMT	2019-04-29	04:00:00	REGULAR	7036754	2385260
14	A002	R051	02-00-00	59 ST	NQR456W	BMT	2019-04-29	08:00:00	REGULAR	7036789	2385385
15	A002	R051	02-00-00	59 ST	NQR456W	BMT	2019-04-29	12:00:00	REGULAR	7036956	2385602

Raw Data

MTA Turnstile - Feb 23 to June 29, 2019

NYC Census - American Community Survey 2015 5-year estimates

Source: http://web.mta.info/developers/turnstile.html
https://www1.nyc.gov/site/census/index.page

+

In [ ]:
```
# Cleaning whitespace in column titles
df.rename(columns=lambda x: x.strip(), inplace=True)
```

In [ ]:
```
# Adding a "WEEKDAY" column for future reference
df.DATE = pd.to_datetime(df.DATE)
df['WEEKDAY'] = df.DATE.dt.dayofweek
days = {0:'Mon',1:'Tues',2:'Weds',3:'Thurs',4:'Fri',5:'Sat',6:'Sun'}
df.WEEKDAY = df.WEEKDAY.apply(lambda x: days[x])
```

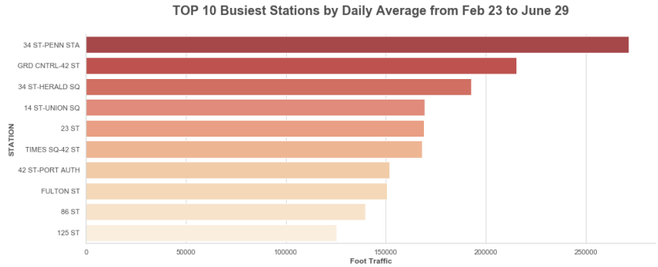
In [ ]:
```
# Adding a "DELTA" column to calculate total foot traffic of all turnstiles
df['DELTA'] = (df['ENTRIES'] - df['ENTRIES'].shift(-1)).abs() + (df['EXITS'] - df['EXITS'].shift(-1)).abs()
```

In [ ]:
```
# Removing all "RECOVR AUD" for potential duplicate entries
df = df[df.DESC != 'RECOVR AUD']
```

In [ ]:
```
# sets the boundaries for the outliers according to the 1.5 * IQR rule

q1 = df.DELTA.describe()['25%']
q3 = df.DELTA.describe()['75%']

upper = q3 + 5*(q3 - q1)
lower = q1 - 5*(q3 - q1)

df = df[(df.DELTA < upper) & (df.DELTA > lower)]
```

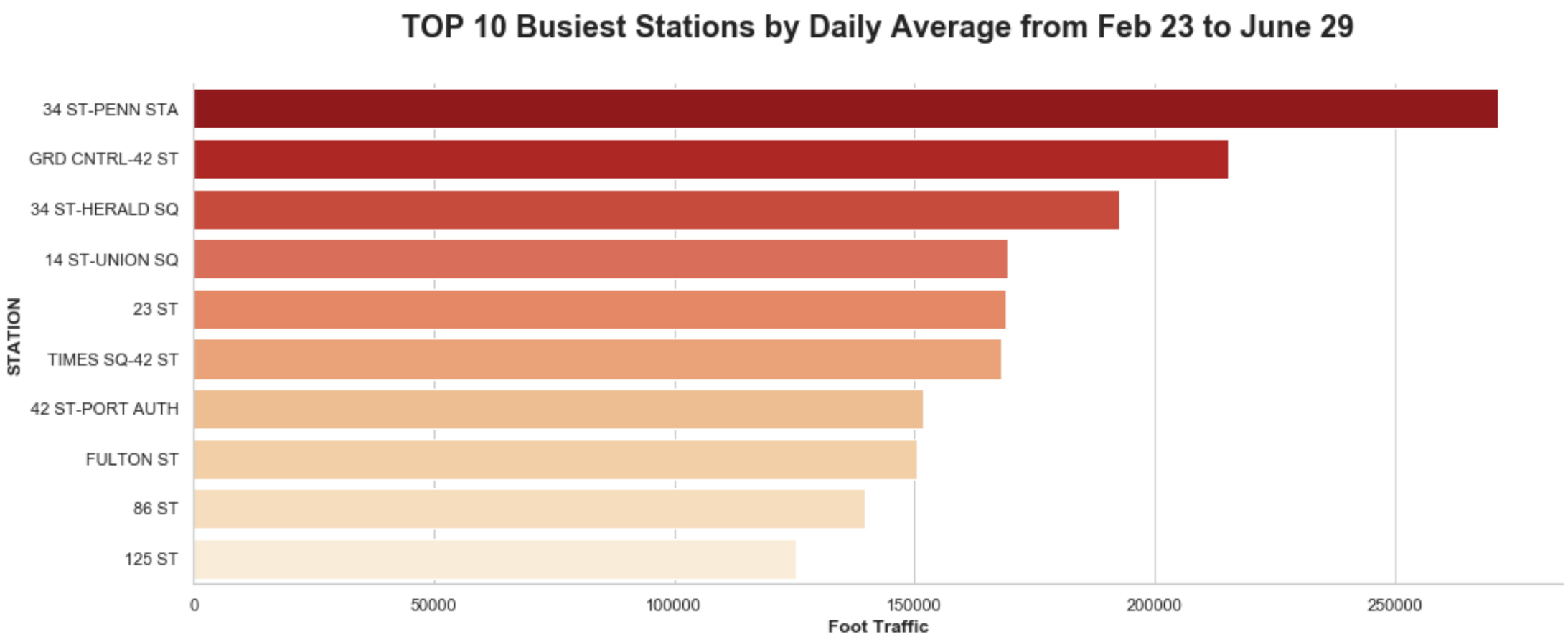
Cleaning Data

=

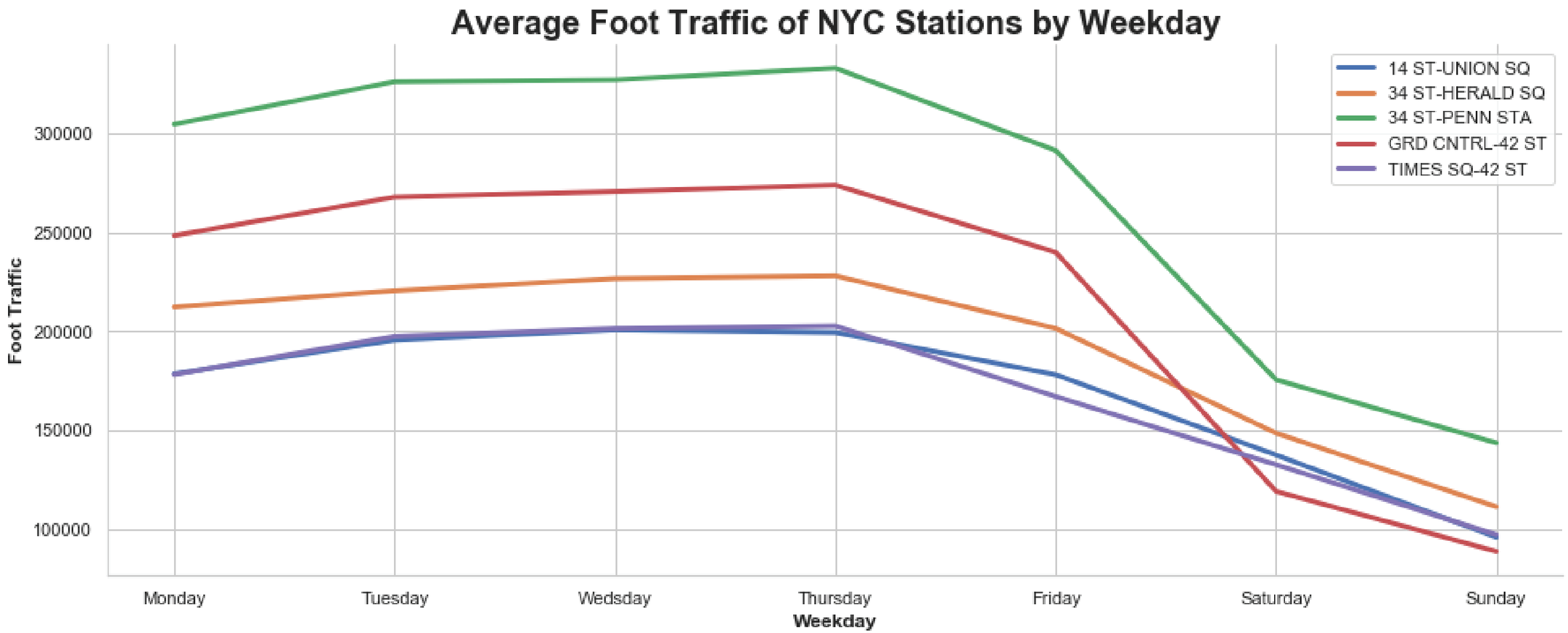


Data Analysis and Conclusions

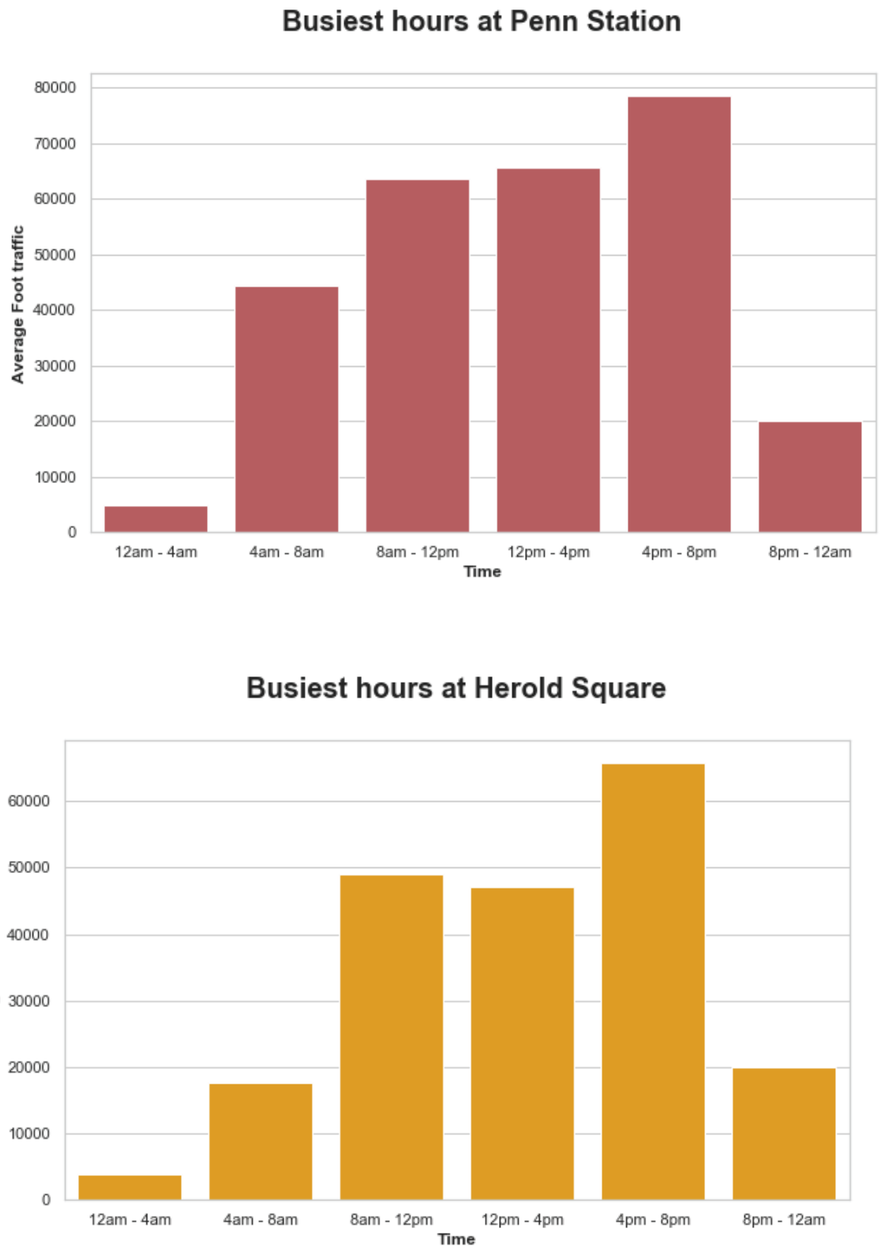
BUSIEST STATIONS

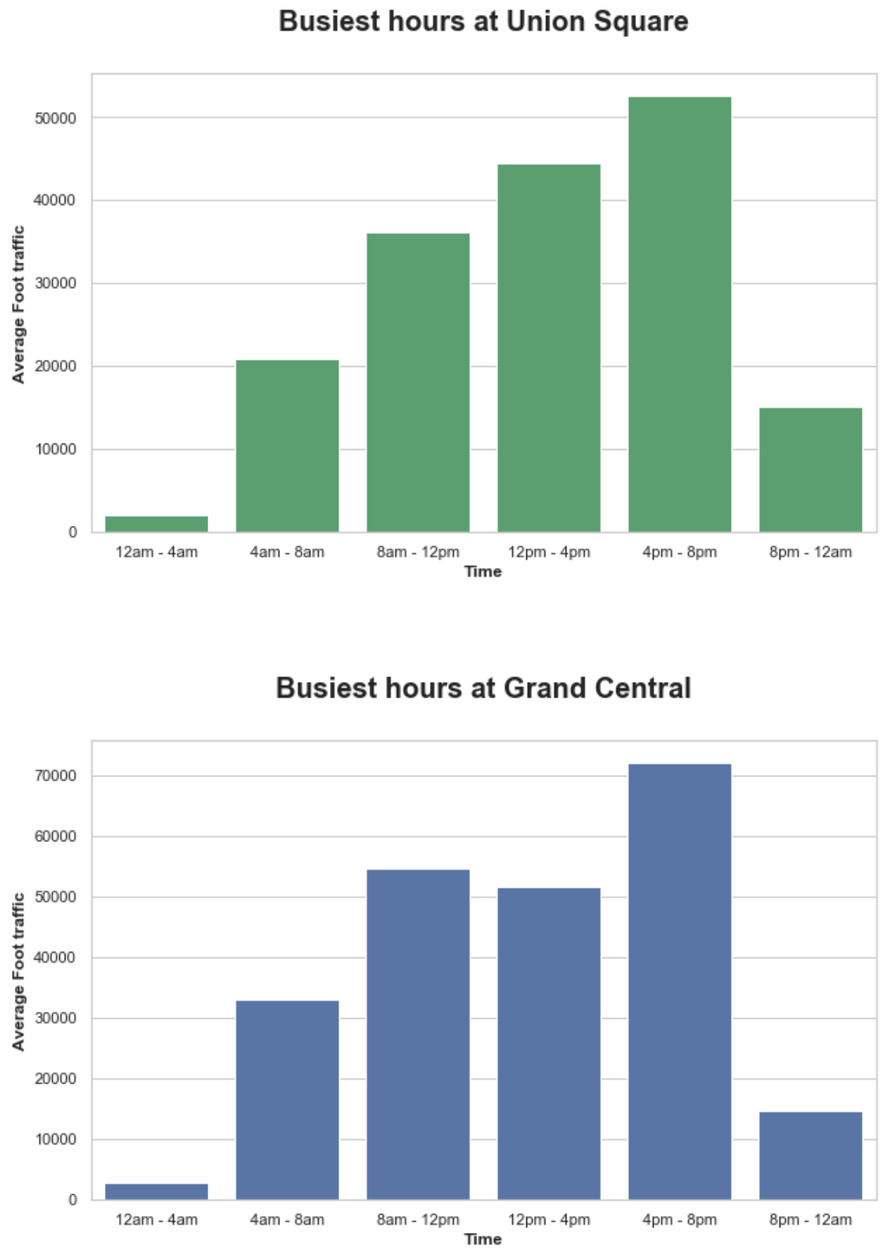


BUSIEST WEEKDAYS

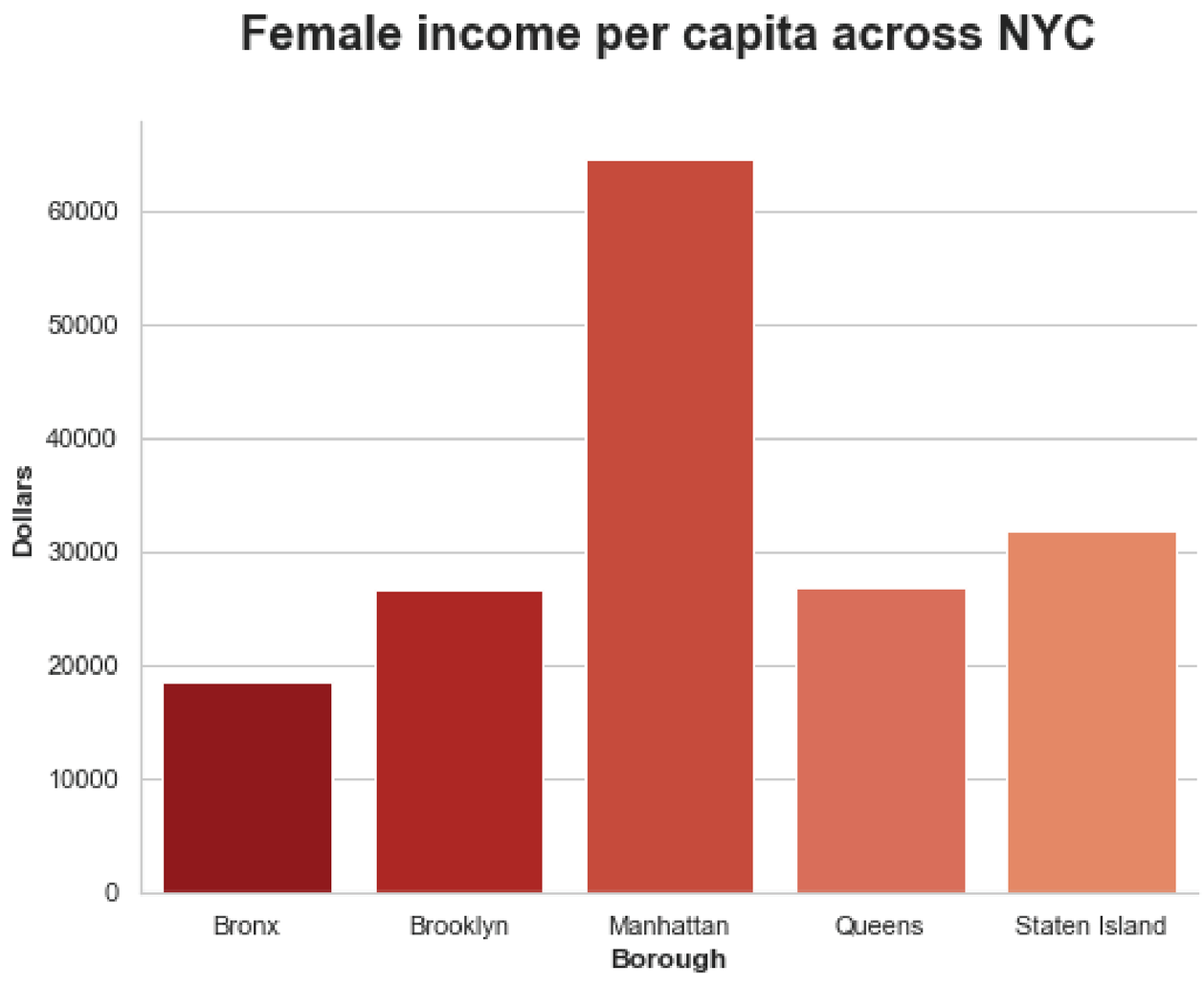


BUSIEST HOURS

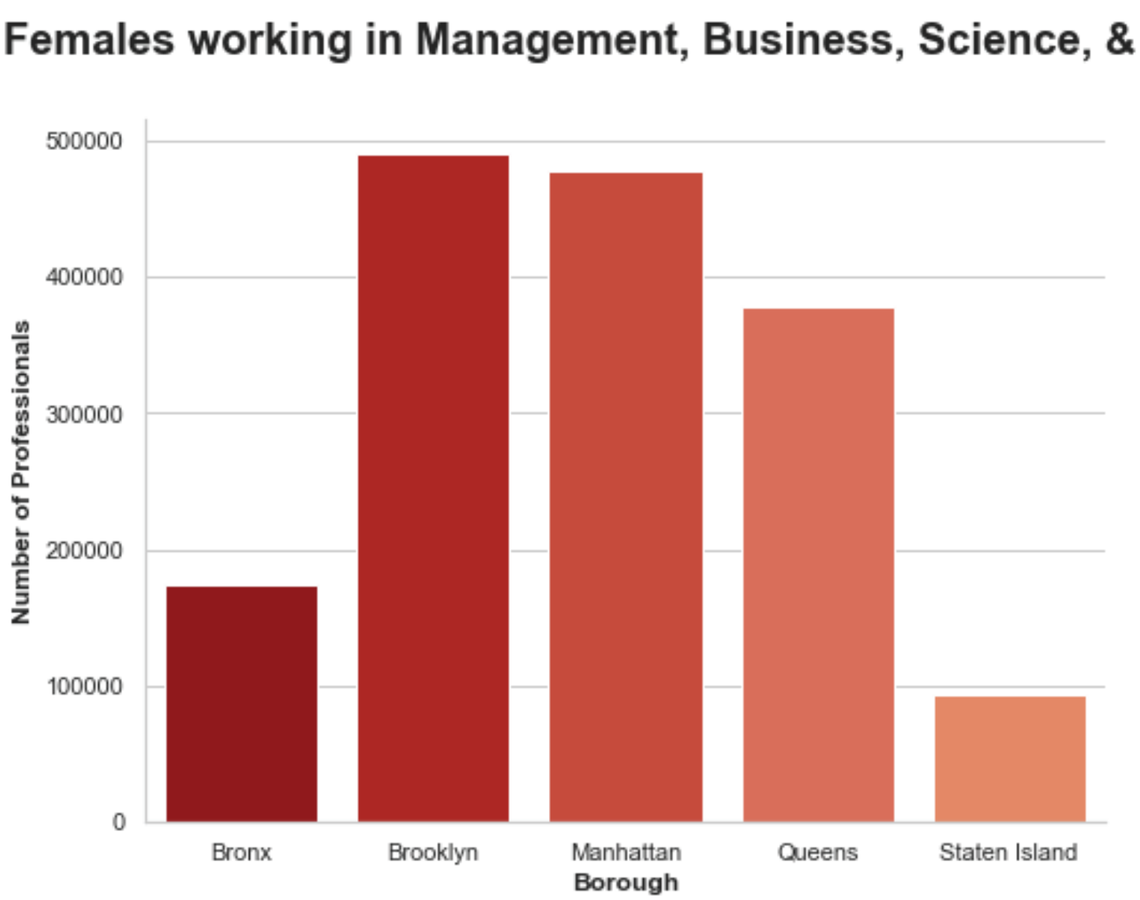




FEMALE INCOME

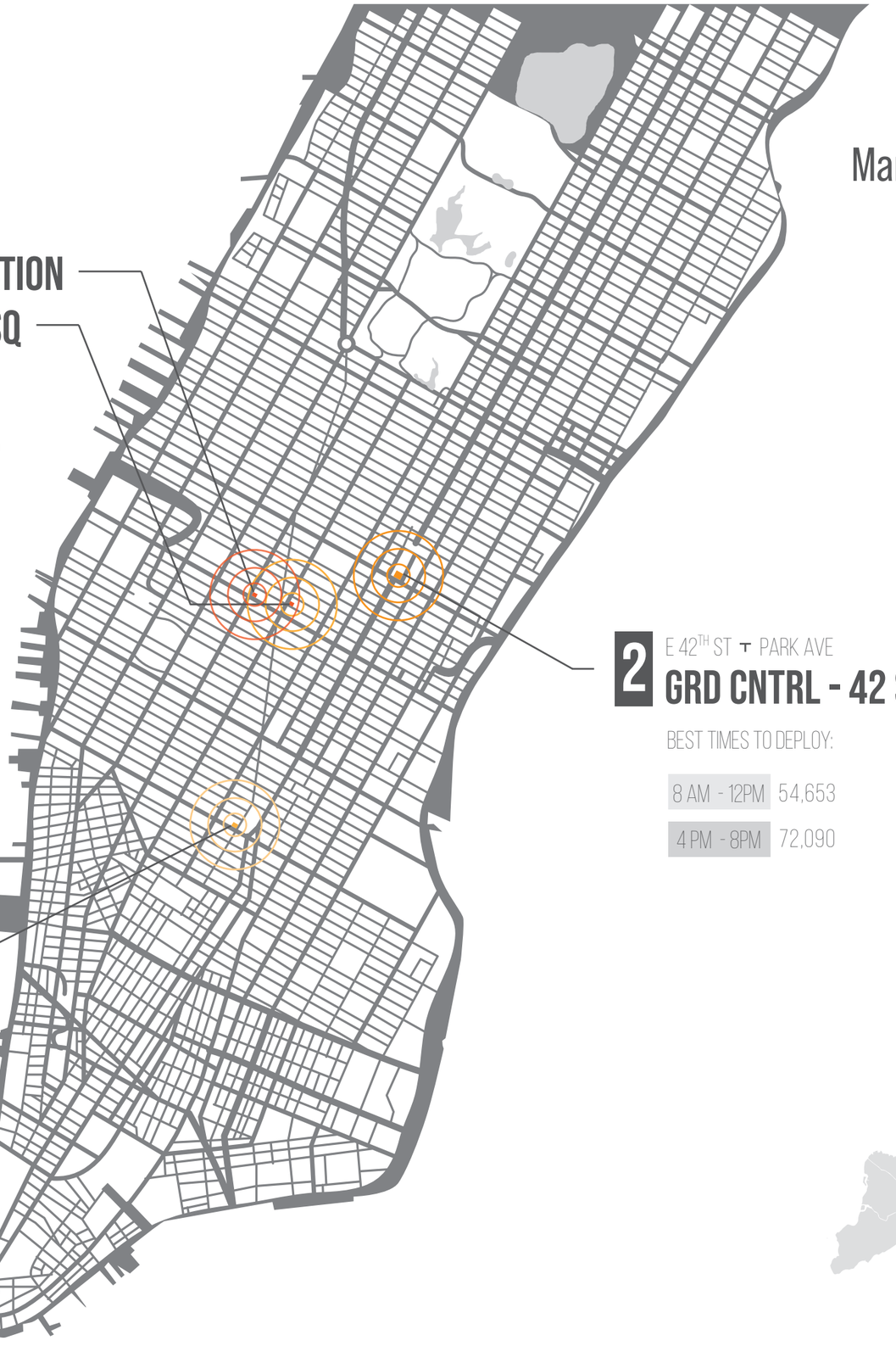


FEMALE OCCUPATION



Source: https://www.kaggle.com/muonneutrino/mapping-new-york-city-census-data/data

CONCLUSION



Manhattan, NY

1

7TH AVE ⊤ W 33RD ST

34 ST - PENN STATION

34 ST - HERALD SQ

BEST TIMES TO DEPLOY:

8 AM - 12PM	49,099 ~ 63,545
4 PM - 8PM	65,817 ~ 78,584

2

E 42TH ST ⊤ PARK AVE

GRD CNTRL - 42 ST

BEST TIMES TO DEPLOY:

8 AM - 12PM	54,653
4 PM - 8PM	72,090

3

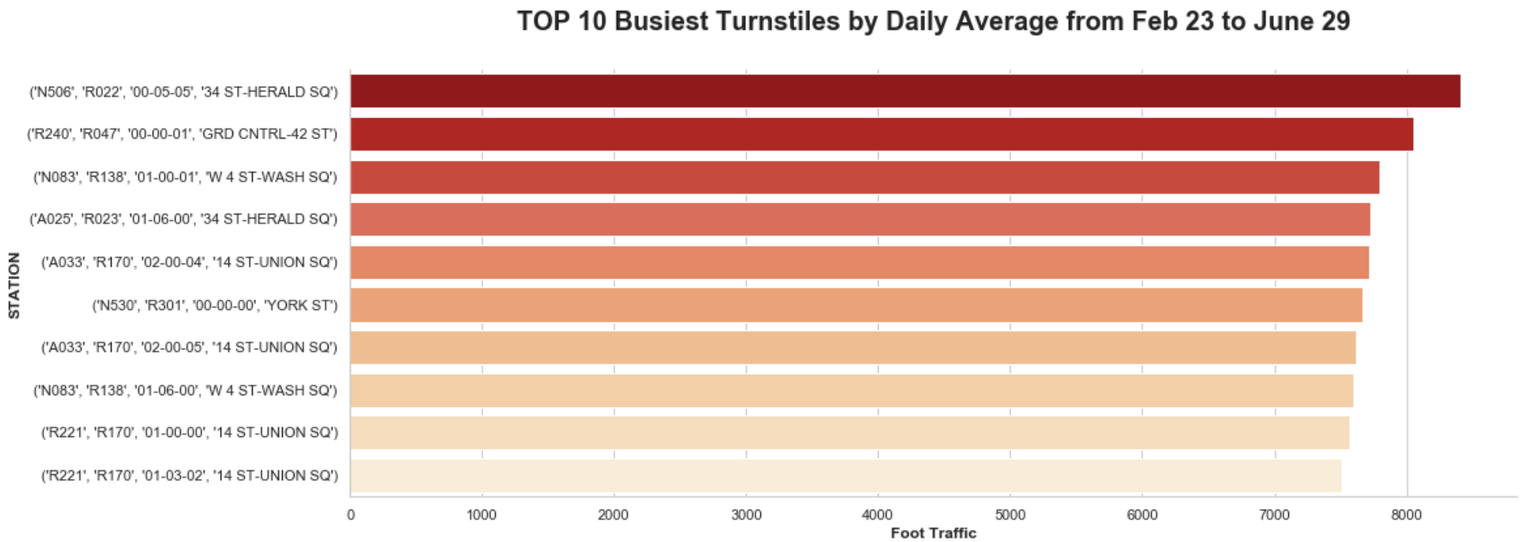
E 14TH ST ⊤ UNION SQ W

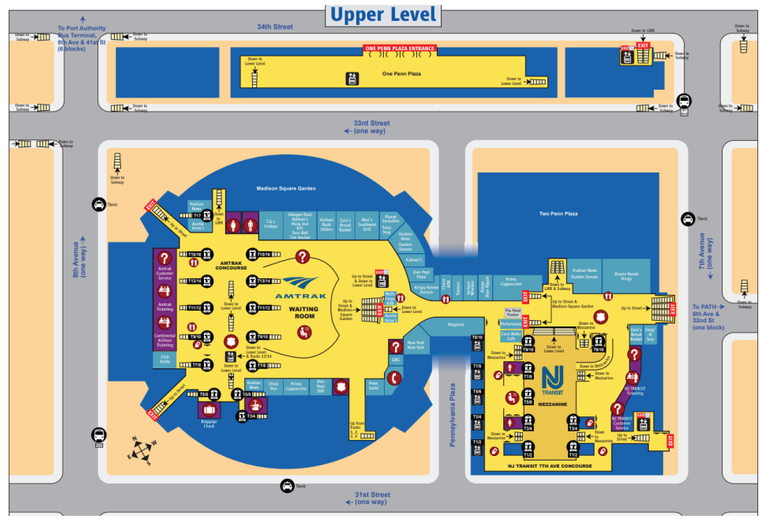
14 ST - UNION SQ

BEST TIMES TO DEPLOY:

12 PM - 4PM	44,563
4 PM - 8PM	52,680

FUTURE





Screenshot of Penn Station Tourist Map

Detailed information on which turnstile to send teams to in order to further optimize the operation

Source: https://www.njtransit.com/pdf/nypenn_map.pdf

THANK YOU

APPENDIX I

Weekday vs. Weekend

	STATION	DELTA
0	34 ST-PENN STA	185672.79
1	GRD CNTRL-42 ST	169192.34
2	34 ST-HERALD SQ	129011.16
3	23 ST	119926.02
4	FULTON ST	114378.89
5	14 ST-UNION SQ	104714.13
6	TIMES SQ-42 ST	104207.20
7	42 ST-PORT AUTH	91954.72
8	86 ST	90364.62
9	PATH NEW WTC	80396.69

APPENDIX II

	C/A	UNIT	SCP	STATION	LINENAME	DIVISION	DATE	TIME	DESC	ENTRIES	EXITS
0	A002	R051	02-00-00	59 ST	NQR456W	BMT	2019-02-23	03:00:00	REGULAR	A 6955483	C 2359112
1	A002	R051	02-00-00	59 ST	NQR456W	BMT	2019-02-23	07:00:00	REGULAR	B 6955494	D 2359125
2	A002	R051	02-00-00	59 ST	NQR456W	BMT	2019-02-23	11:00:00	REGULAR	6955554	2359199
3	A002	R051	02-00-00	59 ST	NQR456W	BMT	2019-02-23	15:00:00	REGULAR	6955714	2359248
4	A002	R051	02-00-00	59 ST	NQR456W	BMT	2019-02-23	19:00:00	REGULAR	6956004	2359292

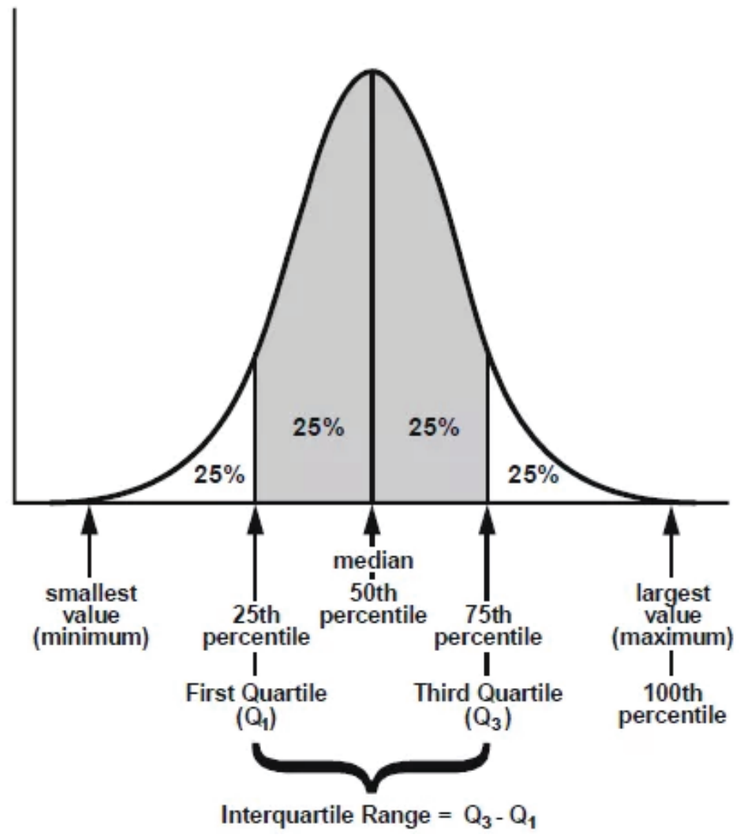
DELTA (Total FootTraffic)

```
# Adding a "DELTA" column to calculate total foot traffic of all turnstiles
df['DELTA'] = (df['ENTRIES'] - df['ENTRIES'].shift(-1)).abs() + (df['EXITS'] - df['EXITS'].shift(-1)).abs()
```

Total Foot Traffic = [A − B] + [C − D]

APPENDIX III

```python
# sets the boundaries for the outliers according to the 1.5 * IQR rule

q1 = df.DELTA.describe()['25%']
q3 = df.DELTA.describe()['75%']

upper = q3 + 5*(q3 - q1)
lower = q1 - 5*(q3 - q1)

df = df[(df.DELTA < upper) & (df.DELTA > lower)]
```



Source: https://i2.wp.com/makemeanalyst.com/wp-content/uploads/2017/05/IQR-1.png?resize=431%2C460