EDA Project

T5004

Abdulaziz Alshehri

Exploratory Data Analysis (EDA) Project

- What **patterns** people follows in the NYC subway?
- **Relationship** between the NYC **subway** and **taxies**?

DATASETS

Station Name (common column)

MTA

	C/A	UNIT	SCP	STATION	DATE	TIME	ENTRIES	EXITS
0	A002	R051	02-00-00	LEXINGTON AVE	12/19/2015	03:00:00	5460344	1843674
1	A002	R051	02-00-00	LEXINGTON AVE	12/19/2015	07:00:00	5460357	1843686
2	A002	R051	02-00-00	LEXINGTON AVE	12/19/2015	11:00:00	5460448	1843797
3	A002	R051	02-00-00	LEXINGTON AVE	12/19/2015	15:00:00	5460702	1843864
4	A002	R051	02-00-00	LEXINGTON AVE	12/19/2015	19:00:00	5461157	1843945

STATIONS

	Name	Latitude	Longitude
0	Astoria-Ditmars Blvd	40.775036	-73.912034
1	Astoria Blvd	40.770258	-73.917843
2	30 Av	40.766779	-73.921479
3	Broadway	40.761820	-73.925508
4	36 Av	40.756804	-73.929575

TAXIES

	tpep_pickup_datetime	tpep_dropoff_datetime	passenger_count	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude
0	2016-01-01 00:00:00	2016-01-01 00:00:00	2	-73.990372	40.734695	-73.981842	40.732407
1	2016-01-01 00:00:00	2016-01-01 00:00:00	5	-73.980782	40.729912	-73.944473	40.716679
2	2016-01-01 00:00:00	2016-01-01 00:00:00	1	-73.984550	40.679565	-73.950272	40.788925
3	2016-01-01 00:00:00	2016-01-01 00:00:00	1	-73.993469	40.718990	-73.962242	40.657333
4	2016-01-01 00:00:00	2016-01-01 00:00:00	3	-73.960625	40.781330	-73.977264	40.758514

CLEANING MTA – Wrong Data

```
df_mta['DAILY_ENTRIES'] = df_mta[['ENTRIES']].diff()
```

Wrong Subtraction between two different turnstiles

	TURN_ID	DATE	TIME	STATION	ENTRIES	EXITS	DAILY_ENTRIES	DAILY_EXITS	DAILY_ENTRIES_EXITS
250	A002R05102-00-0059 st	2016-02-05	19:00:00	59 ST	5530435	1867071	879.0	83.0	962.0
251	A002R05102-00-0059 st	2016-02-05	23:00:00	59 ST	5530780	1867104	345.0	33.0	378.0
252	A002R05102-00-00lexington ave	2015-12-19	03:00:00	LEXINGTON AVE	5460344	1843674	-70436.0	-23430.0	-93866.0
253	A002R05102-00-00lexington ave	2015-12-19	07:00:00	LEXINGTON AVE	5460357	1843686	13.0	12.0	25.0
254	A002R05102-00-00lexington ave	2015-12-19	11:00:00	LEXINGTON AVE	5460448	1843797	91.0	111.0	202.0

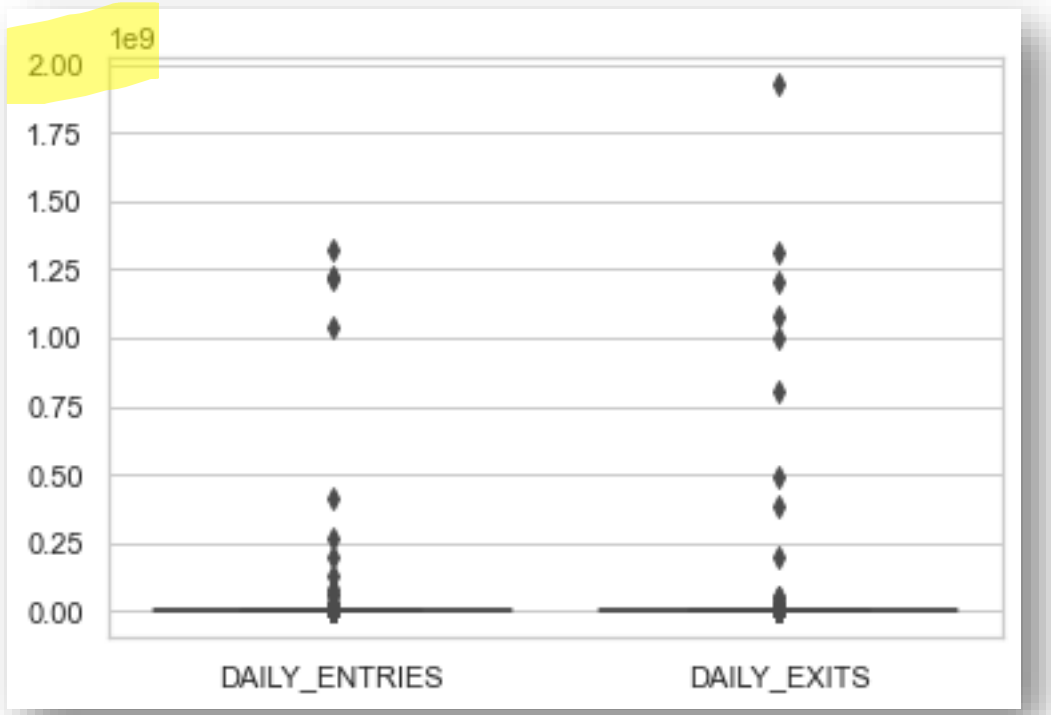
CLEANING MTA – Off-Duty Turnstiles

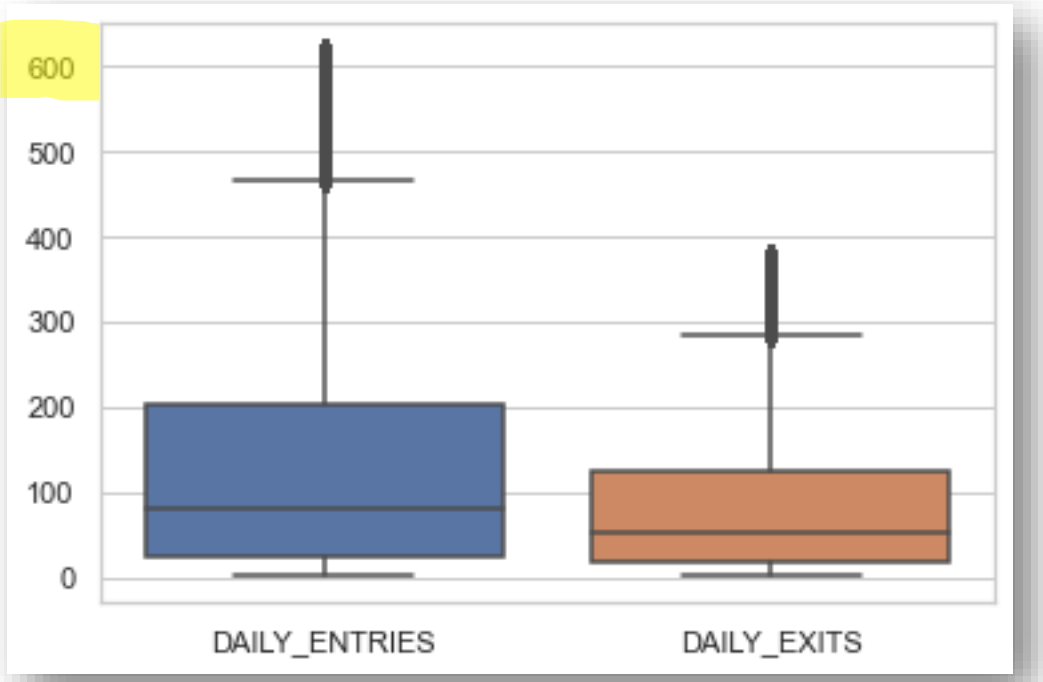
Turnstiles values **not changing**

TURN_ID	DATE	TIME	STATION	ENTRIES	EXITS	DAILY_ENTRIES	DAILY_EXITS	DAILY_ENTRIES_EXITS
A002R05102-05-0059 st	2015-12-26	07:00:00	59 ST	1239	0	0.0	0.0	0.0
A002R05102-05-0059 st	2015-12-26	11:00:00	59 ST	1239	0	0.0	0.0	0.0
A002R05102-05-0059 st	2015-12-26	15:00:00	59 ST	1239	0	0.0	0.0	0.0
A002R05102-05-0059 st	2015-12-26	19:00:00	59 ST	1239	0	0.0	0.0	0.0
A002R05102-05-0059 st	2015-12-26	23:00:00	59 ST	1239	0	0.0	0.0	0.0
...	...	...	...	...	...	...	...	...
TRAM2R46900-05-01rit-roosevelt	2016-02-05	08:00:00	RIT-ROOSEVELT	5554	231	0.0	0.0	0.0
TRAM2R46900-05-01rit-roosevelt	2016-02-05	12:00:00	RIT-ROOSEVELT	5554	231	0.0	0.0	0.0
TRAM2R46900-05-01rit-roosevelt	2016-02-05	16:00:00	RIT-ROOSEVELT	5554	231	0.0	0.0	0.0
TRAM2R46900-05-01rit-roosevelt	2016-02-05	16:00:05	RIT-ROOSEVELT	5554	231	0.0	0.0	0.0
TRAM2R46900-05-01rit-roosevelt	2016-02-05	20:00:00	RIT-ROOSEVELT	5554	231	0.0	0.0	0.0

CLEANING MTA - Outliers

Before removing outliers





After removing outliers

CLEANING TAXI – Zeros and Duplicates

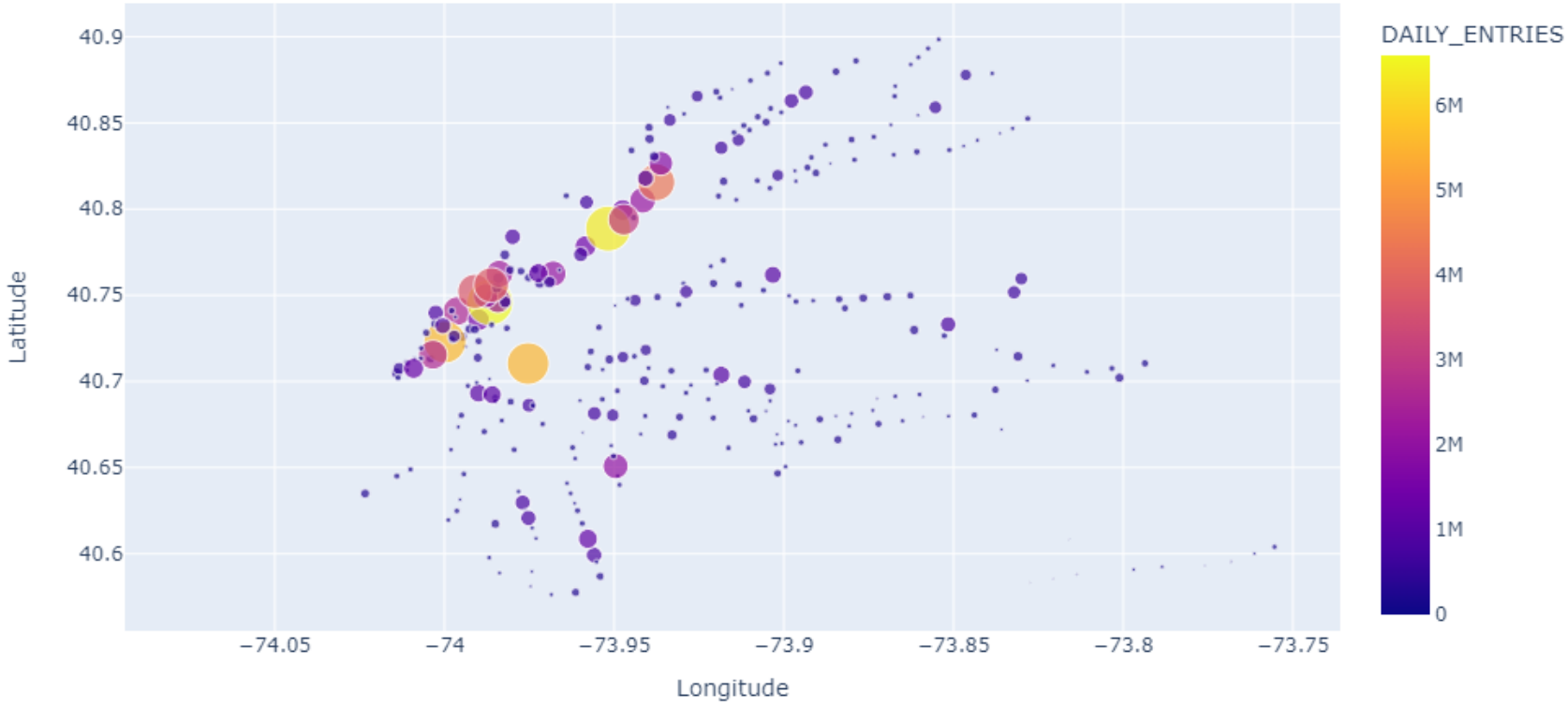
zeros coordinates

	tpep_pickup_datetime	tpep_dropoff_datetime	passenger_count	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude
38	2016-01-01 00:00:19	2016-01-01 00:19:33	1	0.000000	0.000000	0.0	0.0
67	2016-01-01 00:00:41	2016-01-01 00:00:46	5	0.000000	0.000000	0.0	0.0
150	2016-01-01 00:01:34	2016-01-01 00:15:38	1	0.000000	0.000000	0.0	0.0
156	2016-01-01 00:01:36	2016-01-01 00:20:36	1	0.000000	0.000000	0.0	0.0
158	2016-01-01 00:01:37	2016-01-01 00:25:25	2	0.000000	0.000000	0.0	0.0

duplicate rows

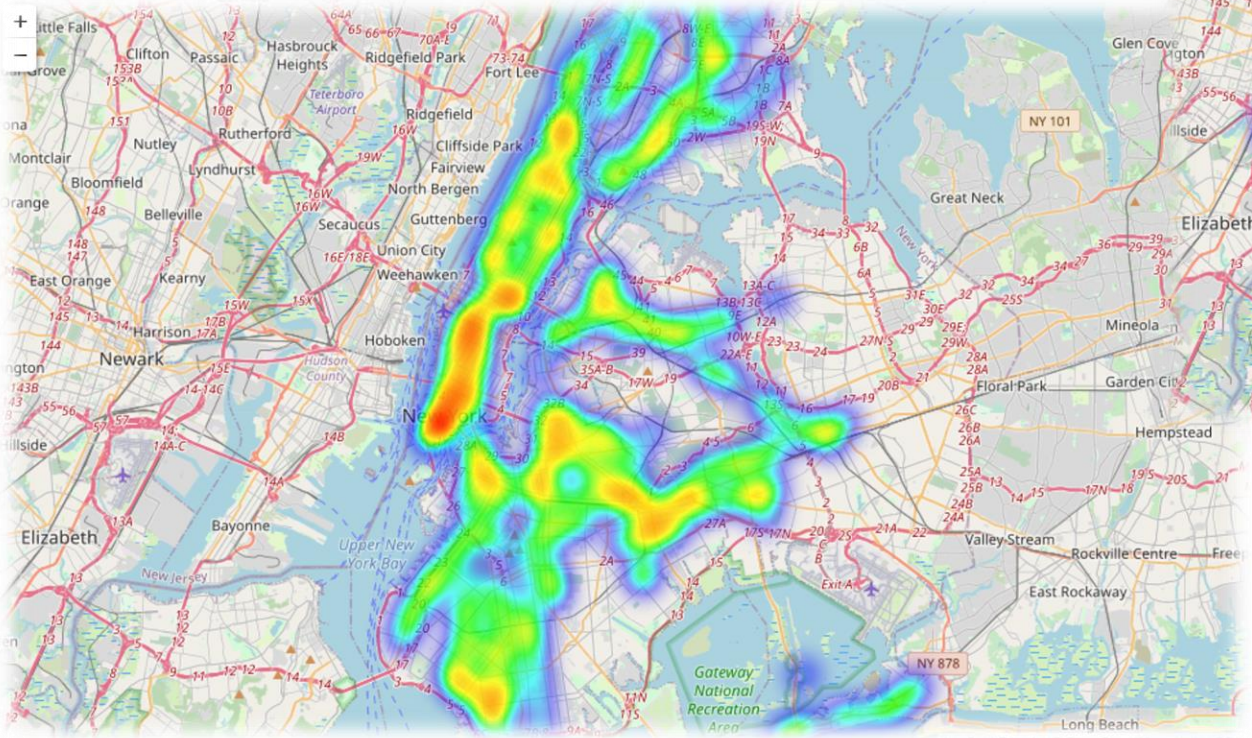
	tpep_pickup_datetime	tpep_dropoff_datetime	passenger_count	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude
1773	2016-01-02 00:50:32	2016-01-02 00:51:16	1	-73.825645	40.712231	-73.83033	40.714161
1774	2016-01-02 00:50:32	2016-01-02 00:51:16	1	-73.825645	40.712231	-73.83033	40.714161

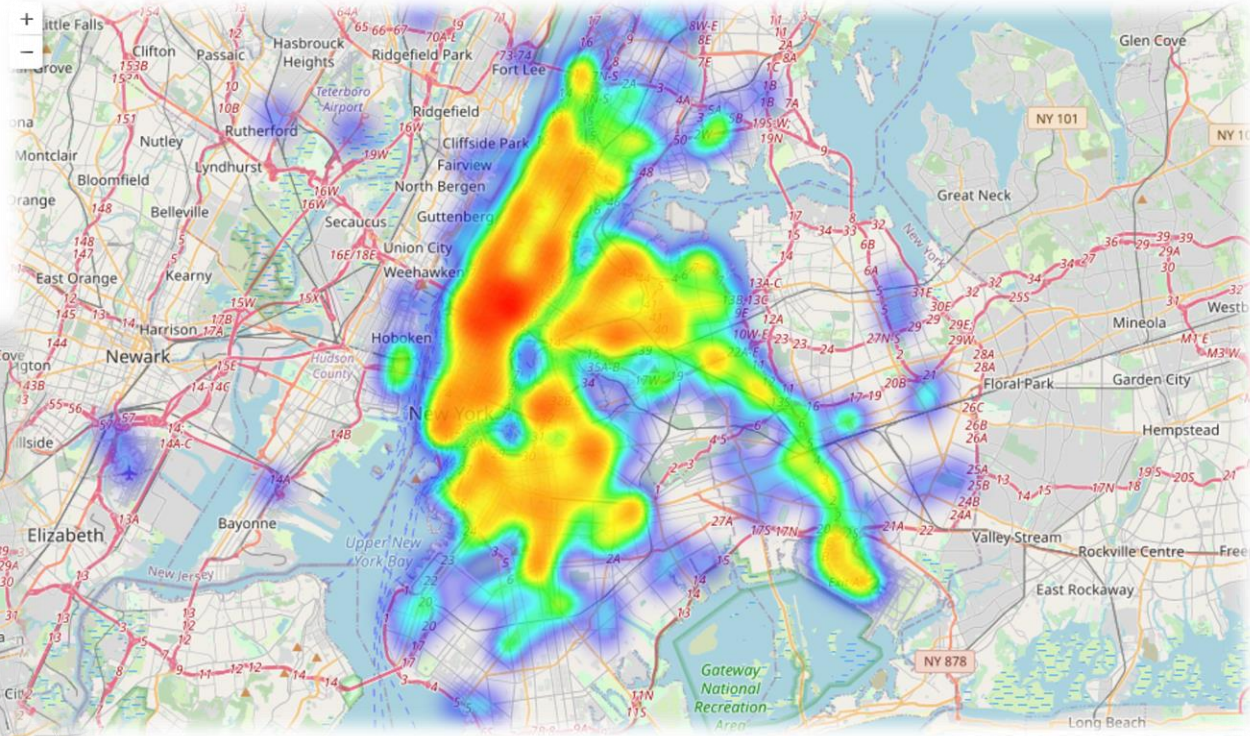
RESULTS - daily_entries



RESULTS – Taxi vs Subway Heatmap

subway entries heatmap





taxi pickups heatmap

CONCLUSIONS

- MTA and Stations datasets joined to get stations locations.
- Stations dataset didn’t require any cleaning.
- Plotting stations their entries show that Manhattan is the busiest.

RECOMENDATIONS

- Based on our analysis we recommend for Taxi & Limousine Commission to direct taxis' drivers to more crowded stations (e.g., times sq 42 ST, etc.).

FURTHER WORK

- This analysis can be enhanced in many ways e.g. extending the exploratory of relationship between taxies and subway demand.