Exploratory data analysis MTA turnstile in NYC

Advertising campaign for new resort in nyc

Back Story - Client Email

Hello

We are Emily and Henry
Marketing Team of Palma Resort Company

A new resort that provides the family with entertainment, relaxation and great times.
We will open the resort on April 1 and will conduct an advertising campaign pre and during first week opening
We have a budget of 5 billboards that we want to distribute in the most crowd subway stations in different divisions where it will cover a wide geographical area
As we said, we will start opening the resort's doors on April 1
The AD shows that the resort is a beautiful destination to spend wonderful time with the family during the spring break

In your opinion, as a data analyst, can you help us with:
determine most crowded stations in different divisions where the billboards will be distributed

The Target :

determine most crowded stations in different divisions where the billboards will be distributed

Data

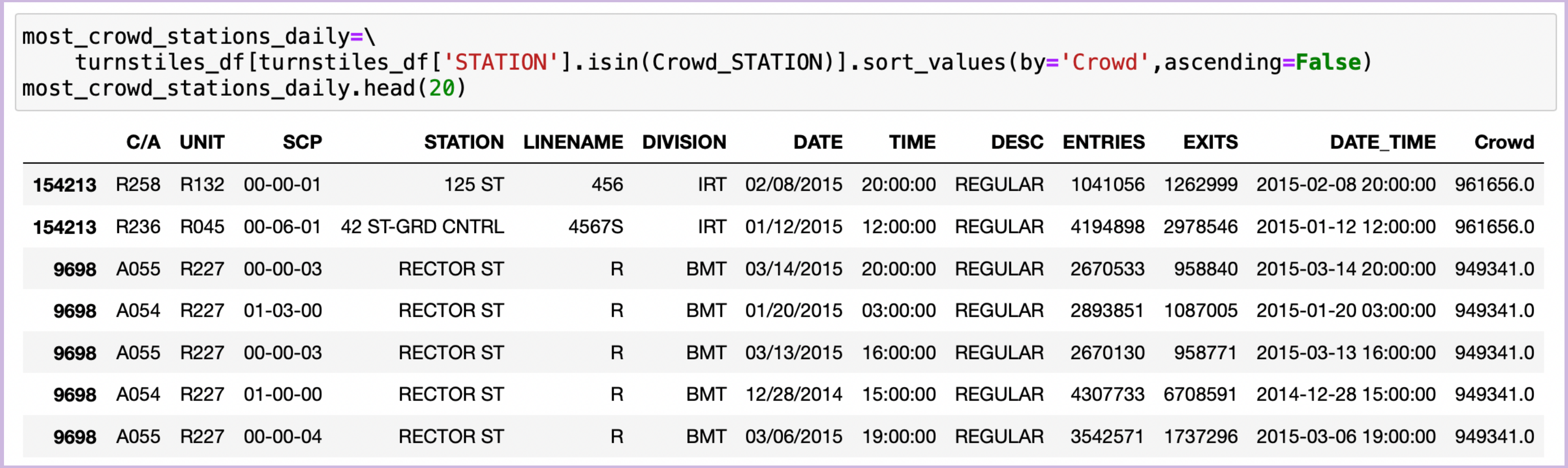
The dataset contain MTA turnstile data with 3 months worth of data for january ,february and march .

Algorithms

Perform a thorough Exploratory Data Analysis of the MTA turnstile data; clean, explore, aggregate, and visualize the data as appropriate to address the client's needs.

Tools

Numpy and Pandas for data manipulation Matplotlib and Seaborn for plotting , SQLalchemy



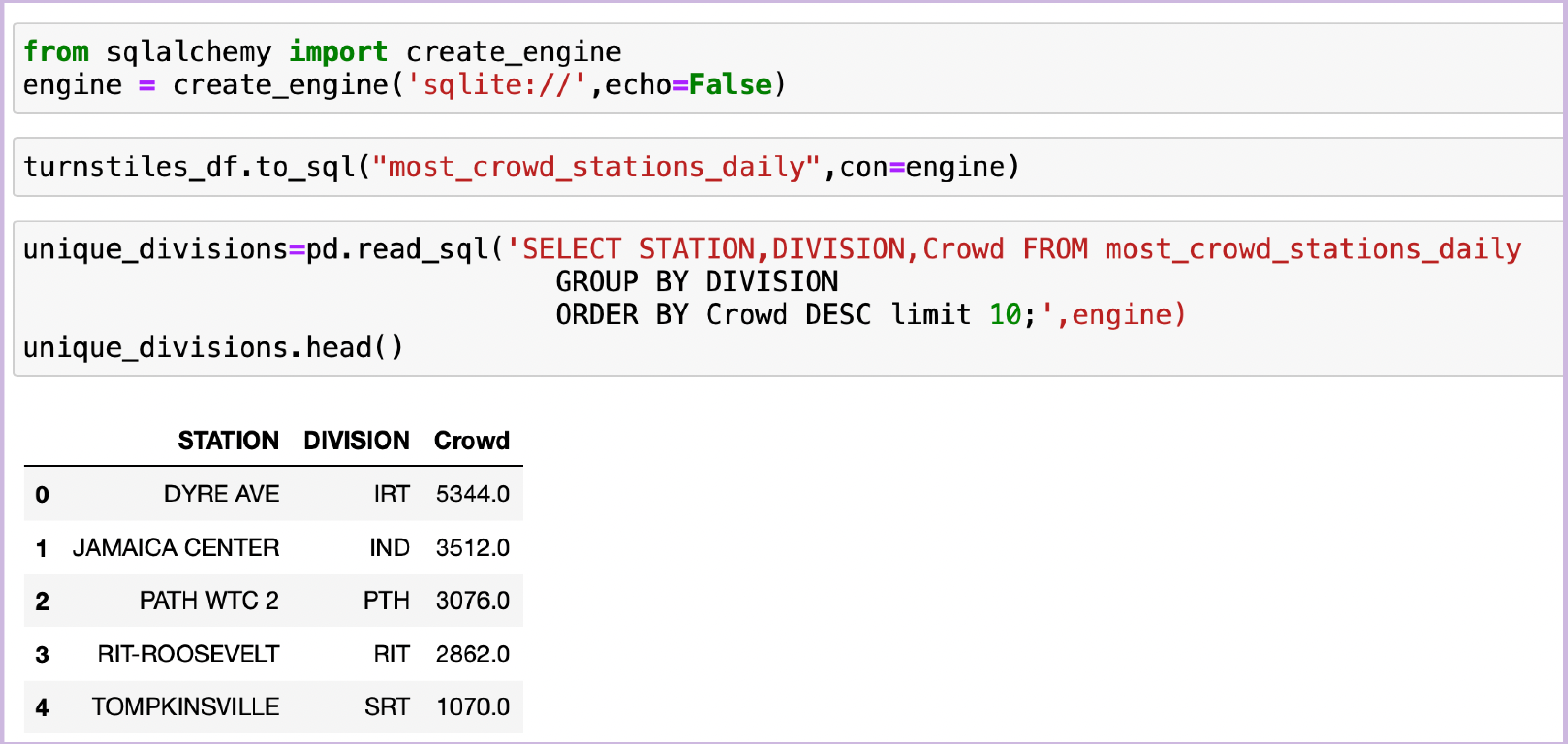
```
most_crowd_stations_daily=\
    turnstiles_df[turnstiles_df['STATION'].isin(Crowd_STATION)].sort_values(by='Crowd',ascending=False)
most_crowd_stations_daily.head(20)
```

	C/A	UNIT	SCP	STATION	LINENAME	DIVISION	DATE	TIME	DESC	ENTRIES	EXITS	DATE_TIME	Crowd
154213	R258	R132	00-00-01	125 ST	456	IRT	02/08/2015	20:00:00	REGULAR	1041056	1262999	2015-02-08 20:00:00	961656.0
154213	R236	R045	00-06-01	42 ST-GRD CNTRL	4567S	IRT	01/12/2015	12:00:00	REGULAR	4194898	2978546	2015-01-12 12:00:00	961656.0
9698	A055	R227	00-00-03	RECTOR ST	R	BMT	03/14/2015	20:00:00	REGULAR	2670533	958840	2015-03-14 20:00:00	949341.0
9698	A054	R227	01-03-00	RECTOR ST	R	BMT	01/20/2015	03:00:00	REGULAR	2893851	1087005	2015-01-20 03:00:00	949341.0
9698	A055	R227	00-00-03	RECTOR ST	R	BMT	03/13/2015	16:00:00	REGULAR	2670130	958771	2015-03-13 16:00:00	949341.0
9698	A054	R227	01-00-00	RECTOR ST	R	BMT	12/28/2014	15:00:00	REGULAR	4307733	6708591	2014-12-28 15:00:00	949341.0
9698	A055	R227	00-00-04	RECTOR ST	R	BMT	03/06/2015	19:00:00	REGULAR	3542571	1737296	2015-03-06 19:00:00	949341.0

This is most crowd stations but some of it share same division !!!
then i did not reache the goal yet

By using sqlalchem

I wrote this query



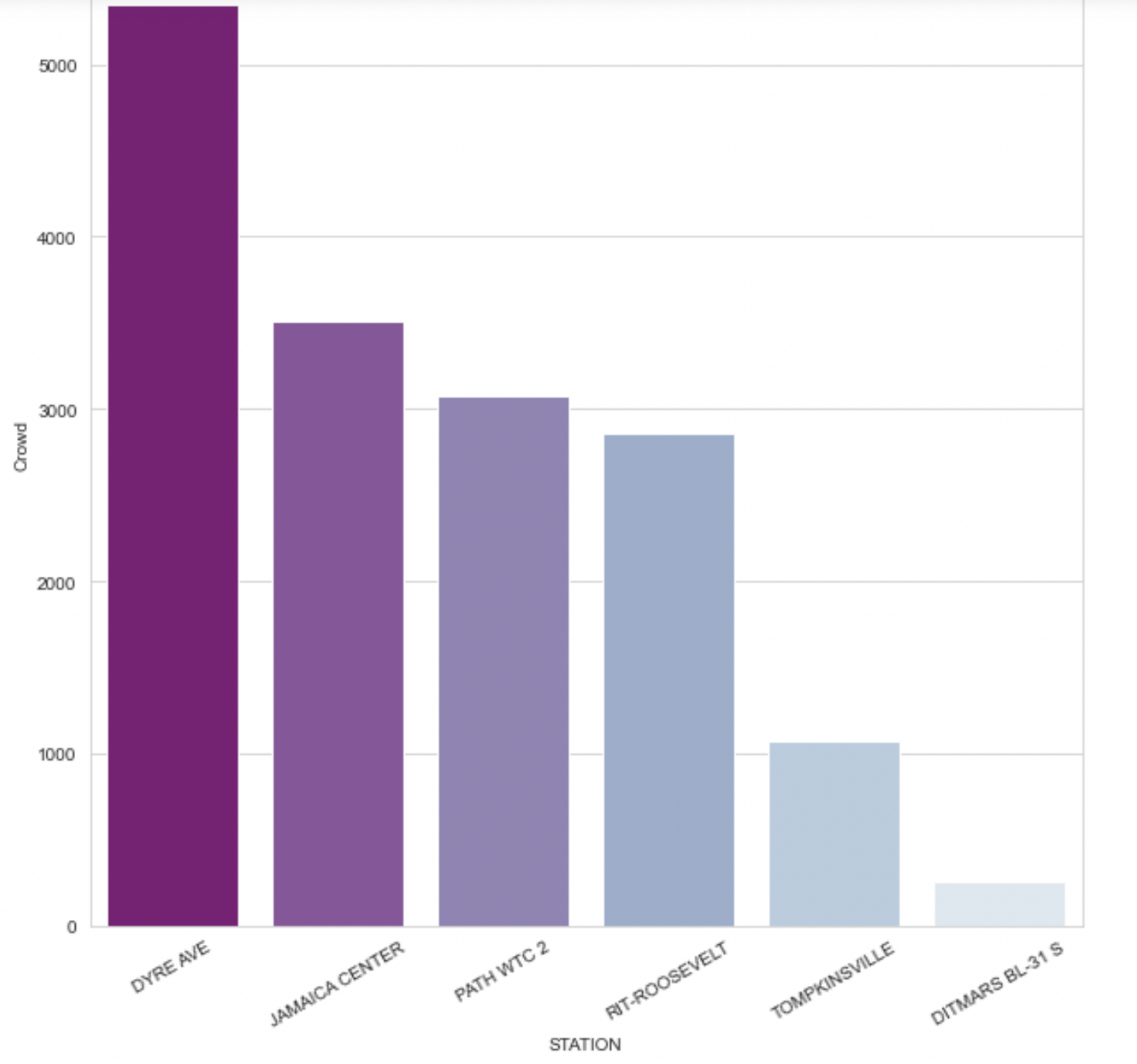
```
from sqlalchemy import create_engine
engine = create_engine('sqlite://',echo=False)
```

```
turnstiles_df.to_sql("most_crowd_stations_daily",con=engine)
```

```
unique_divisions=pd.read_sql('SELECT STATION,DIVISION,Crowd FROM most_crowd_stations_daily
                                GROUP BY DIVISION
                                ORDER BY Crowd DESC limit 10;',engine)
unique_divisions.head()
```

	STATION	DIVISION	Crowd
0	DYRE AVE	IRT	5344.0
1	JAMAICA CENTER	IND	3512.0
2	PATH WTC 2	PTH	3076.0
3	RIT-ROOSEVELT	RIT	2862.0
4	TOMPKINSVILLE	SRT	1070.0

Most Crowded Stations in Unique Divisions



The main point in this exploratory data analysis I looking for most crowded stations in different divisions:
the resulte is :

- **DYRE AVE** from IRT
- **JAMAICA CENTER** from IND
- **PATH WTC 2** from PTH
- **RIT-ROOSEVELT** from RIT
- **TOMPKINSVILLE** from SRT

By Exploratory data analysis i reach my main target and and deliver it to Palma Resort Company

THANK YOU for listening