MTA TURNSTILE DATA ANALYSIS

MOROOJ ALDEEB

TURNSTILE DATA ANALYSIS

Introduction

BACK STORY

WE HAVE DOWNLOADED WEATHER DATA SET FROM KAGGLE WEBSITE AND WE JOINED THE WEATHER DATA FOR THE SAME PERIOD TAKEN FROM THE TURNSTILE DATA BY DATE COLUMN. WE COMBINED TWO ADDITIONAL COLUMNS FROM THE WEATHER DATA SET (WEATHER TYPE AND SEVERITY) INTO THE TURNSTILE COLUMNS.

RESEARCH QUESTION

- WHAT ARE THE BUSIEST STATIONS BASED ON ENTRIES (IN 4 MONTHS IN 2018)FOR AWARENESS LEAFLETS ABOUT WEATHER CONDITIONS IN CROWDED ?

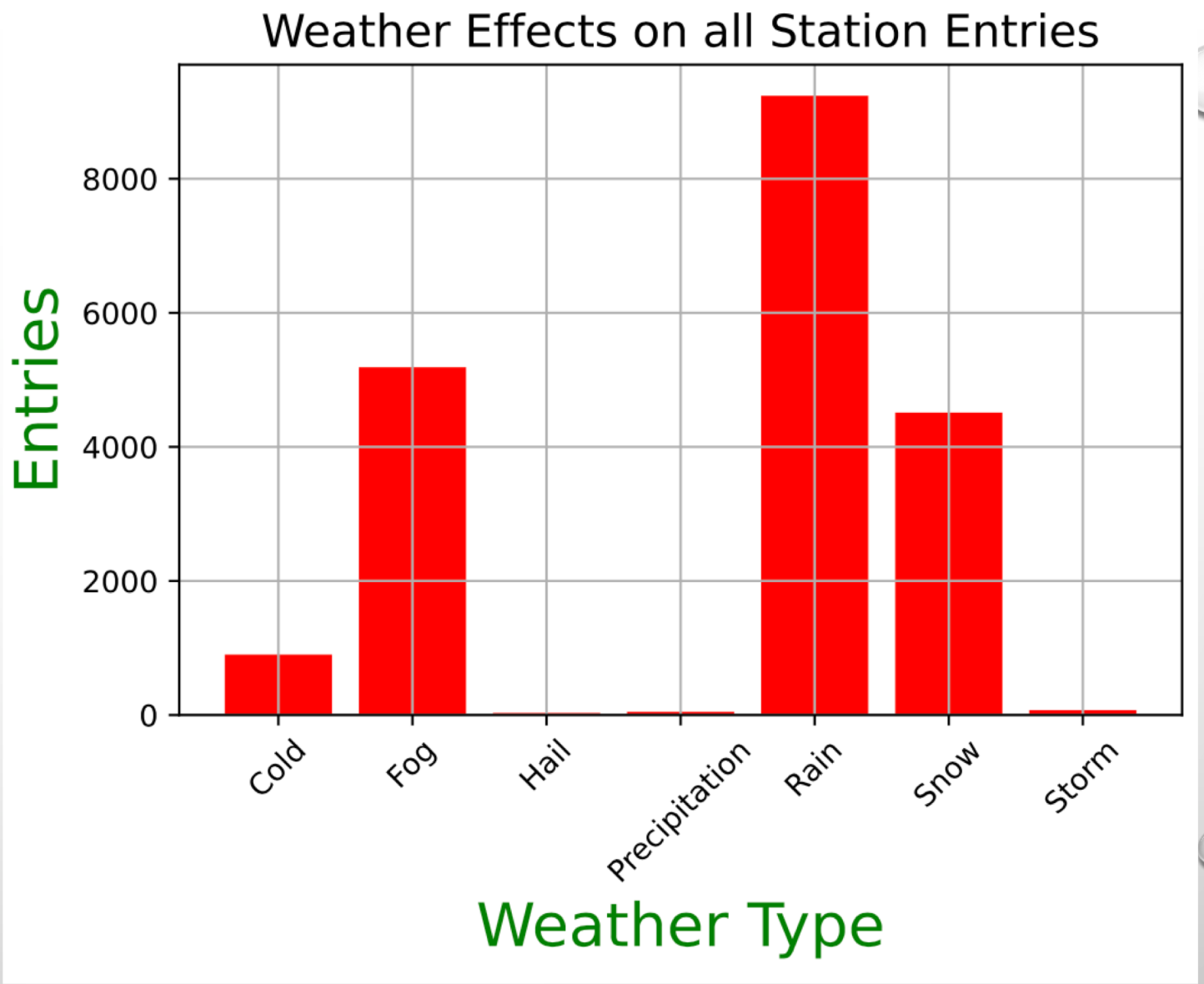
- IS THERE ANY RELATIONSHIP BETWEEN SIZE OF THE TRAIN STATION AND THE TRAFFIC ?

- DOES THE DAILY WEATHER AFFECT THE TRAIN TRANSPORTATION?

APPROACH

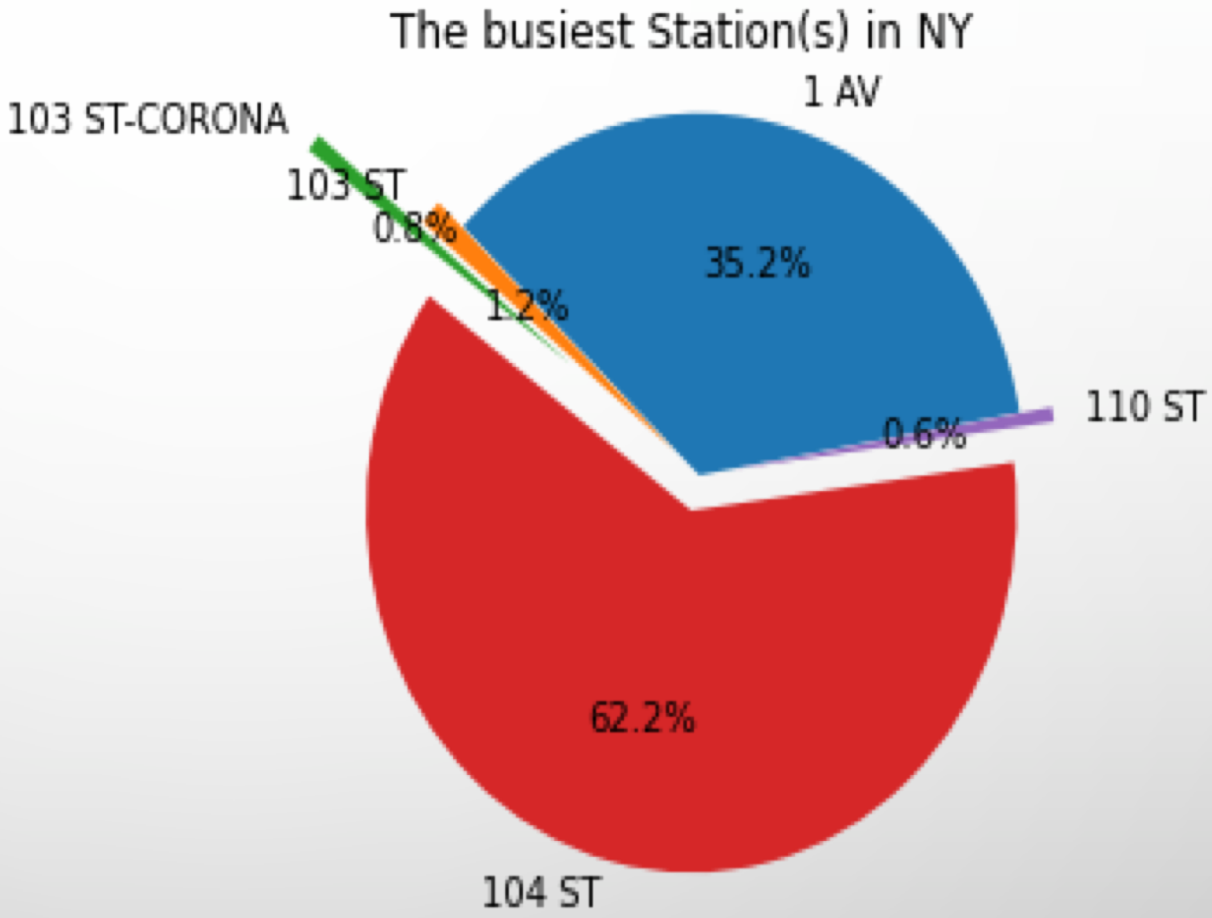
WE INVESTIGATED WHETHER THE WEATHER AFFECTS THE NUMBER OF PEOPLE UTILIZING RAILWAYS FOR DAILY TRANSPORTATION. FOR THIS, WE GROUPED DATA BY THE “WEATHER TYPE” AGAINST THE COUNTS OF ENTRIES WITH RESPECT TO THE WEATHER TYPE.

The bar plot clearly shows that stormy days experience the lowest traffic to train stations. On Contrary, train transportation seems unaffected during light rainy days.



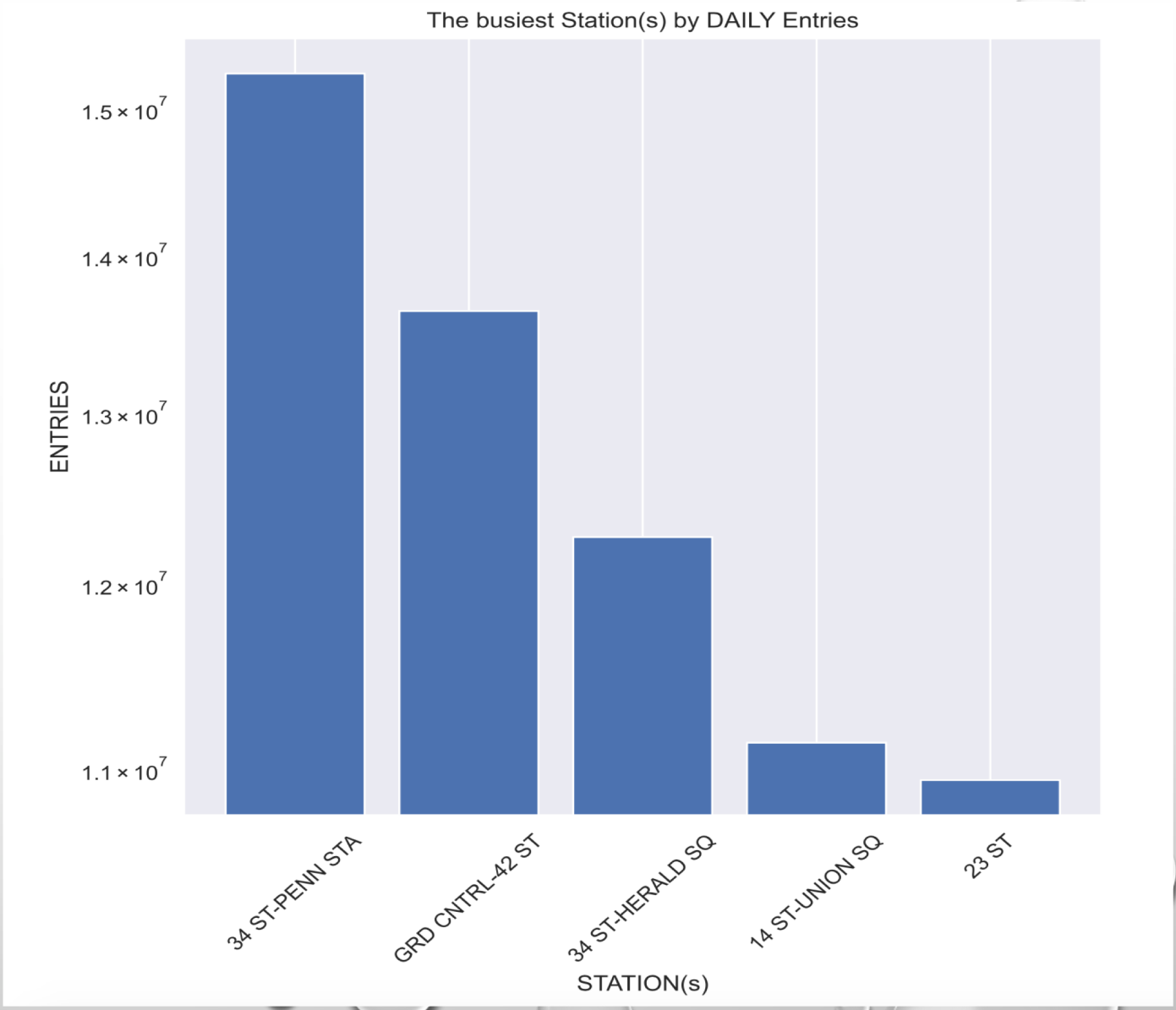
The Top 5 busiest Train Stations by Average of Entries

As can be seen from the PIE Chart, 104 ST is the busiest station by the average ENTRIES in the first fourth months of 2018 followed by 1 AV



The busiest Stations by DAILY Entries

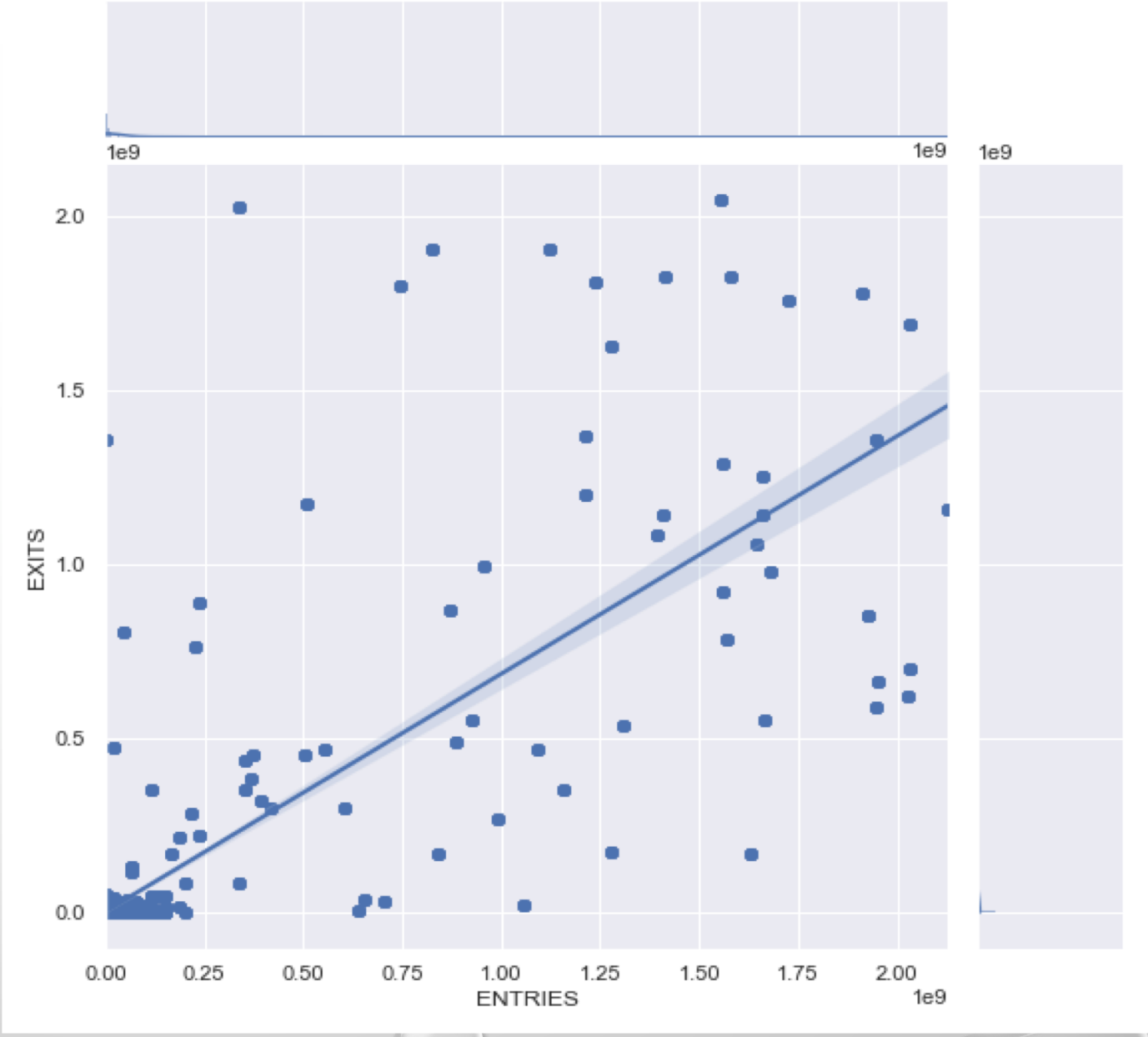
WE HAVE CONSTRUCTED THE DAILY ENTRIES FOR ALL STATIONS, SORTED THEM IN DESCENDING ORDER AND PICKED THE TOP 5 STATIONS BY DAILY ENTRIES, SHOWN BELOW.



CORRELATION BETWEEN ENTRIES AND EXITS

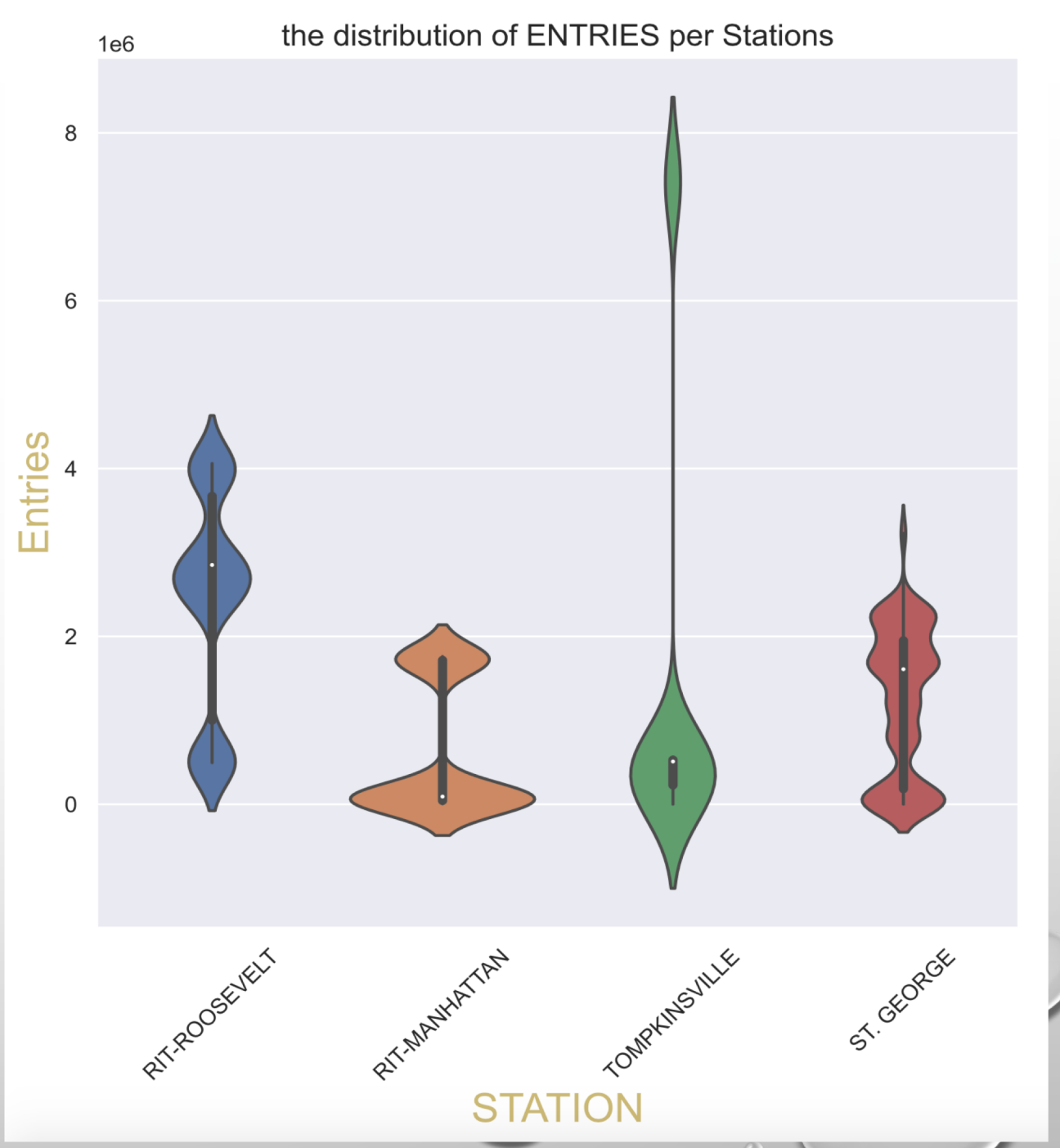
LOGICALLY, THERE SHOULD BE RELATIONSHIP OR CORRELATION BETWEEN ENTRIES AND EXITS IN THE DATA SINCE WHO GOES IN SHOULD GOES OUT FROM THE TRAIN STATION. WE INVESTIGATED THAT CORRELATION STATISTICALLY IN THE DATA BY PLOTTING A SCATTER PLOT AND SUPERIMPOSING THE PEARSON CORRELATION COEFFICIENTS ON THE DATA POINTS. THERE IS APPROXIMATELY 80% POSITIVE RELATIONSHIP BETWEEN ENTRIES AND EXITS WHICH ALIGNS VERY WELL WITH THE LOGIC.

CORRELATION BETWEEN ENTRIES AND EXITS

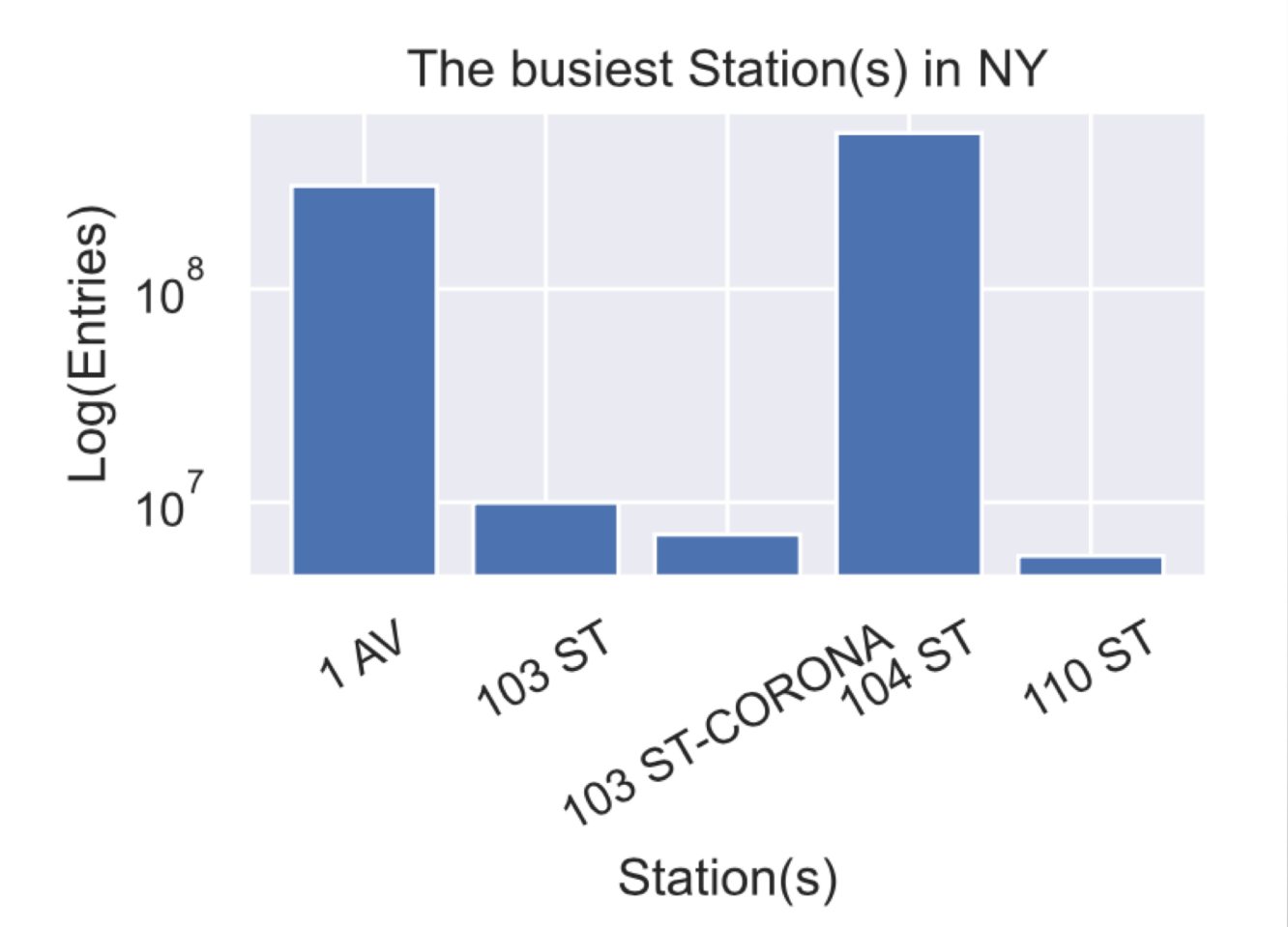


Outliers

To investigate the distribution of ENTRIES pattern(s) per Train station, We constructed the violin plot per station(s) which clearly shows different distribution(s) with variance depending upon the given station. this difference might be attributed to the location of the train station in NY .



THE FIVE BUSIEST STATION(S) IN NY (IN 4 MONTHS)



Datasets and libraries

DATASETS:

1- MTA TURNSTILE DATASET ([06-01-2018] TO [30-04-2018])

2- WEATHER DATASET FROM KAGGLE ([06-01-2018] TO [30-04-2018])

(1926904 ROWS , 25 COLUMNS)

USED TOOLS:

- JUPYTER NOTEBOOK , PYTHON

USED LIBRARIES:

- PANDAS , NUMPY , MATPLOTLIB , SEABORN , SQLITE , SQLALCHEMY

FUTURE WORK

- WE CAN TRY TO PREDICT THE SIZE OF THE TRAIN STATION BASED ON KNOWING THE EXITS AND ENTRIES IN A STATION WHICH MIGHT HELP AUTHORITIES IN DESIGNING BETTER TRAIN STATIONS TO ACCOMMODATE IN ADVANCE THE INFLUX OF PEOPLE AT SPECIFIED GEOGRAPHICAL LOCATION.
- CAN PREDICT THE TRAFFIC FLOW GIVEN A PARTICULAR WEATHER CONDITION.
- CAN PREDICT WEATHER CONDITION TO PUBLISH AWARENESS LEAFLETS LIKE (SMS)

THANK YOU