

Team-2

Günel Alizada

Salim Kılınç

Süleyman Emre Öksüz

Eren Özdił

Onur Can CIKLA



Exploratory Data Analysis: Maximizing Audience Reach for Women in Tech





Project's Purpose

A fictitious organization called Women Tech Women Yes (WTWY) is attempting to organize a summer gala in New York City in order to "increase the participation of women in technology, and to concurrently build awareness and reach."

Our Mission

We are requested to provide advice as a data science team on where to put the survey teams in order to enhance event efficacy.





python

pandas

matplotlib

seaborn

TOOLS AND
DATA SETS
USED

DATA

Steps

The 3 steps that we followed in order to have a conclusion



Step# 1
Data Cleaning



Step # 2
Data Analysis



Step # 3
Data Visualization

Step# 1

Data Cleaning

In general, it is recommended to perform certain data cleaning steps on a dataset, such as formatting columns, removing duplicate entries, and empty spaces

We applied those rules to the MTA_Subway_Hourly_Ridership__Beginning_February_2022 dataset and MTA_Daily_Ridership_Data__Beginning_2020 for further usage.



Step# 2

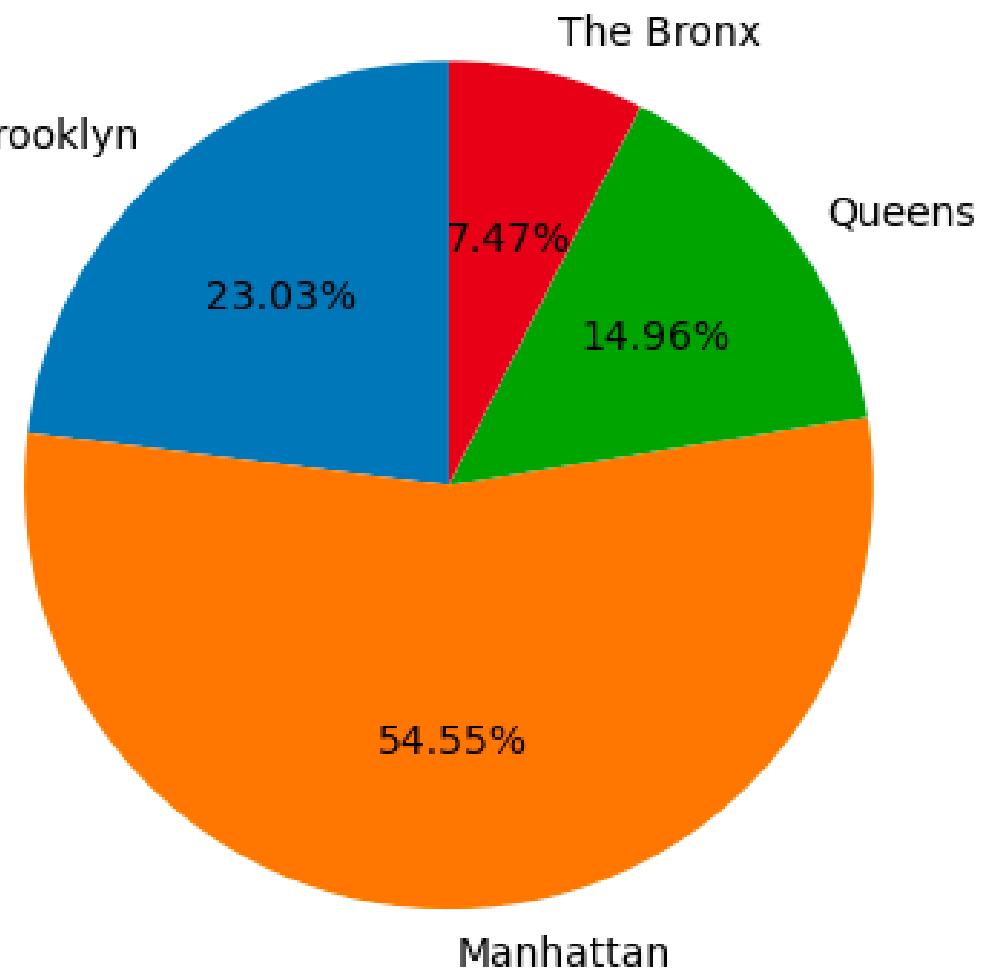
Data Analysis

The procedure of data analysis is the same regardless of the method used; we aggregate passengers by station and identify the busiest boroughs.



Ridership distribution by borough in 2023

Ridership Distribution by Borough



borough	ridership
Brooklyn	123405408
Manhattan	292307203
Queens	80165108
The Bronx	40017820



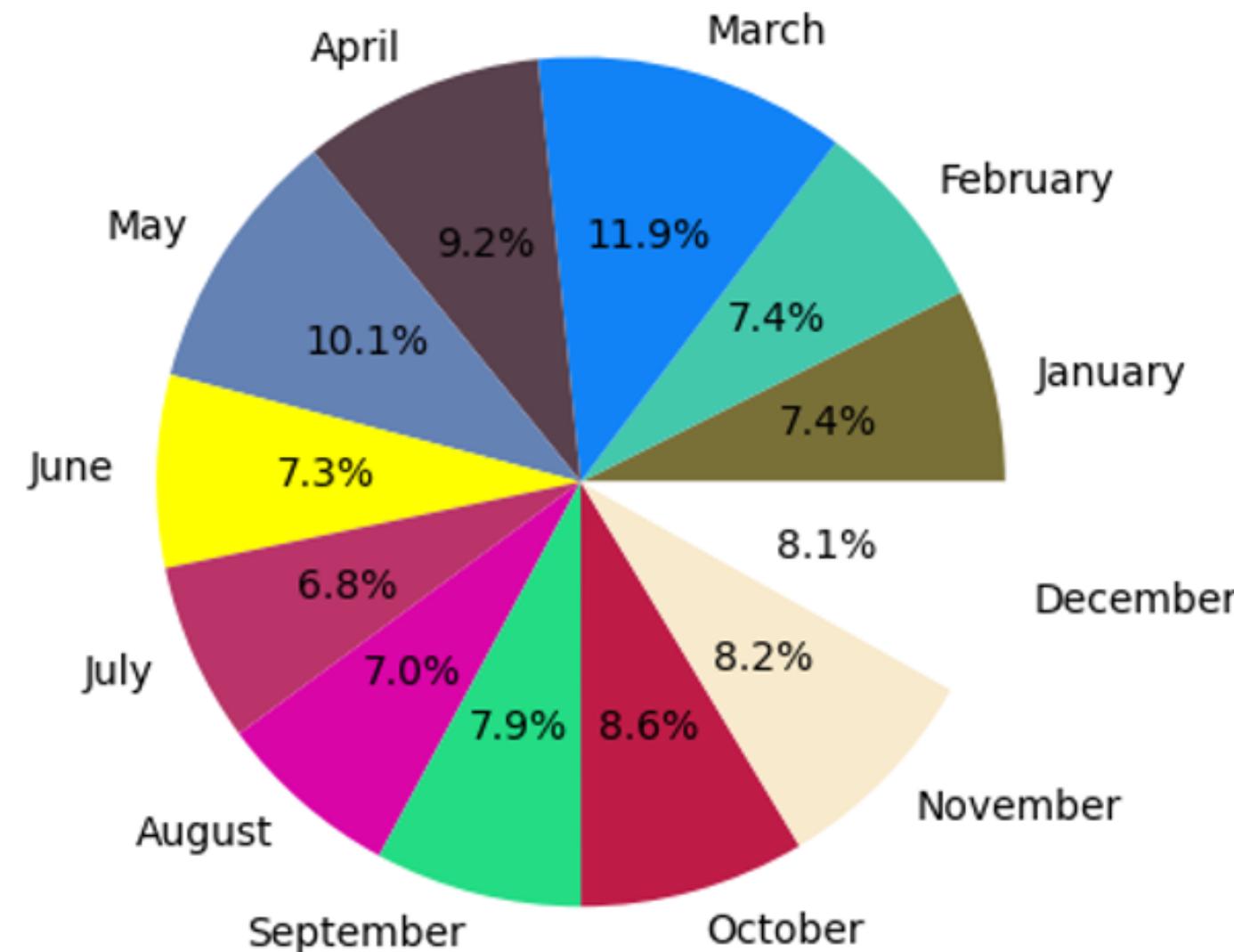
Step# 2

Data Analysis

To determine the most suitable month for organization, we calculated the total riderships by boroughs on a monthly basis.

Ridership distribution by month in 2022

Total Riderships per month data



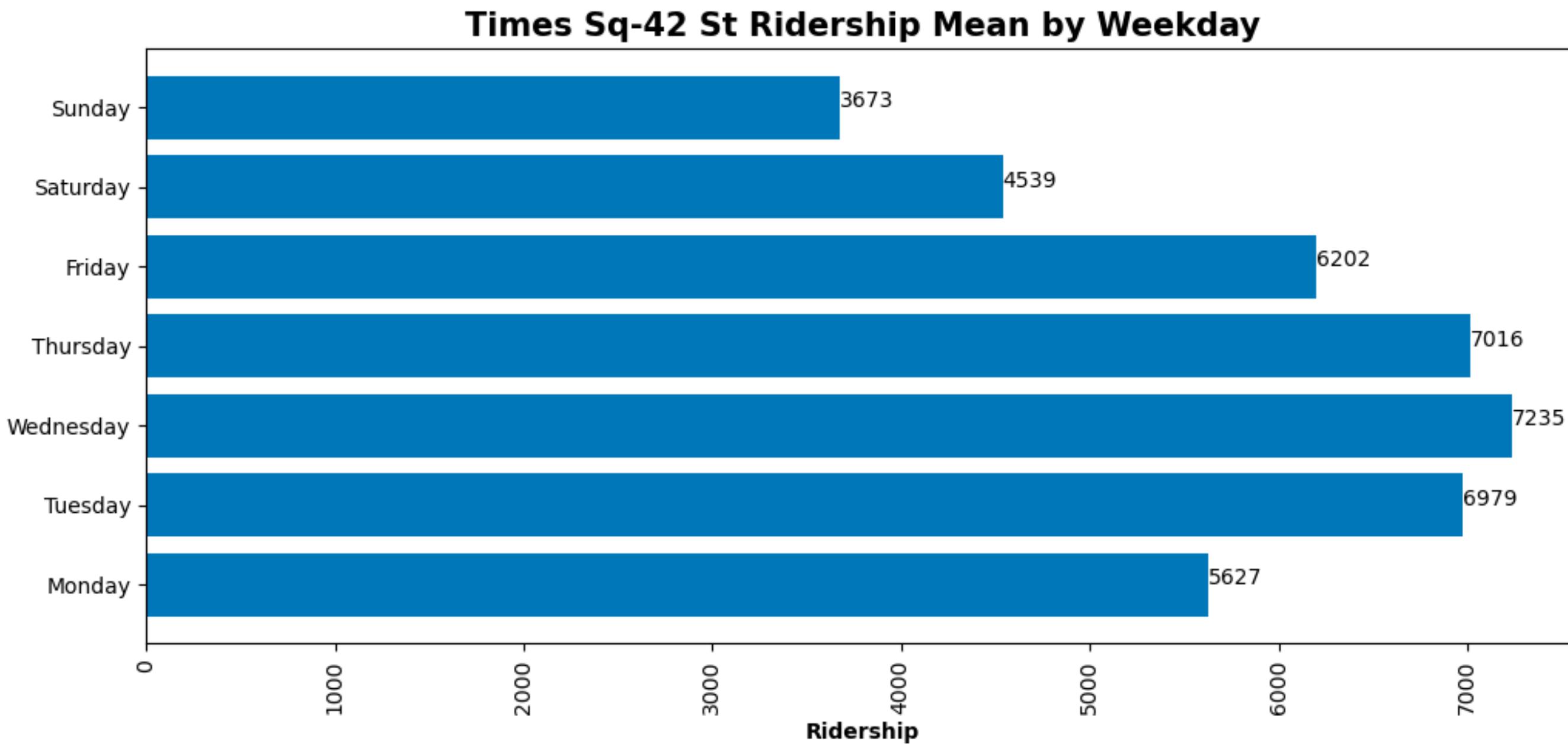
Month	Total Ridership
January	63,734,924
February	70,181,466
March	87,293,930
April	83,273,427
May	87,695,742
June	87,977,903
July	80,514,845
August	82,948,894
September	91,019,574
October	95,120,931
November	90,905,300
December	91,838,943

Step# 2

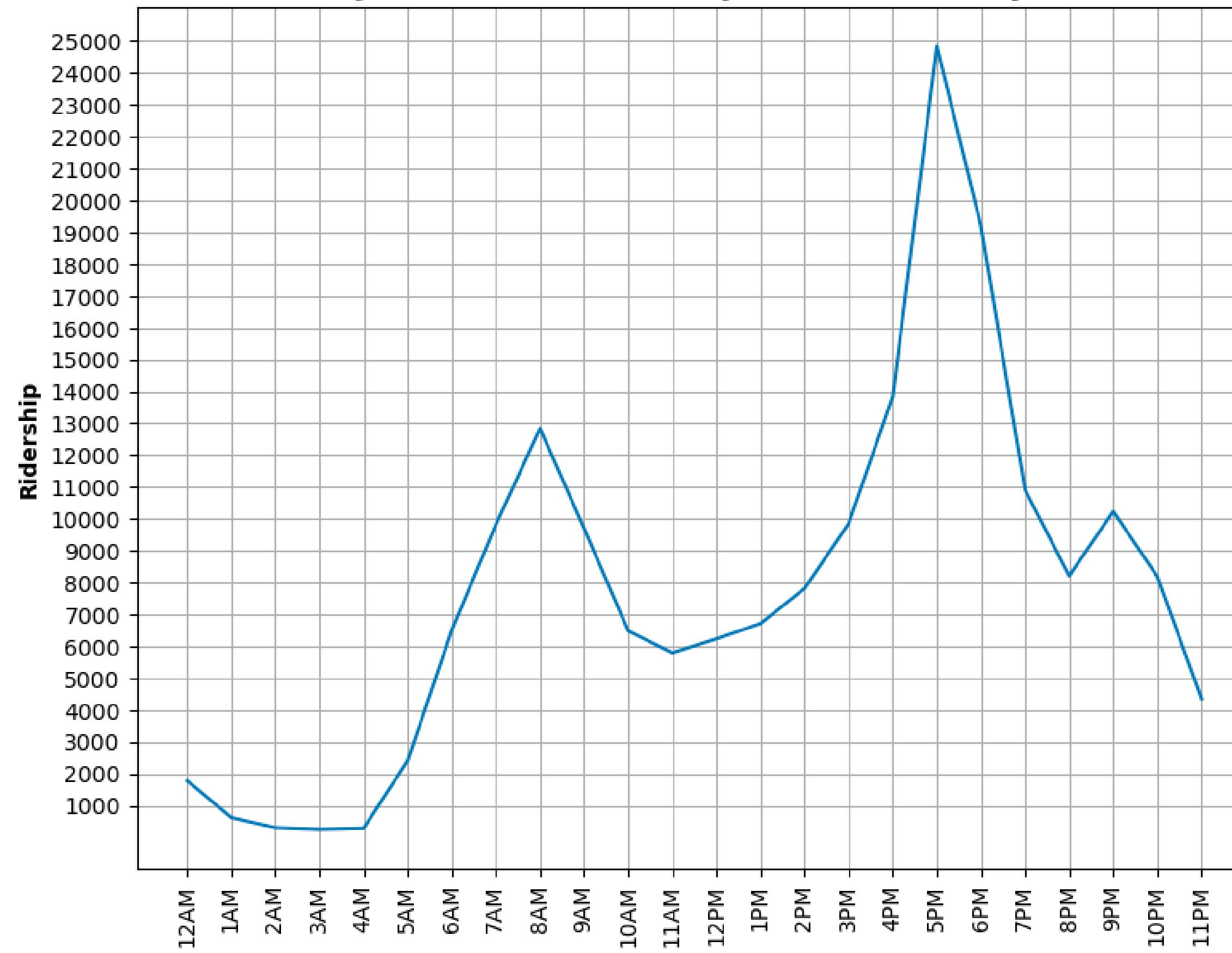
Data Analysis

We calculated the mean of passenger counts at stations by days and total riderships by hours grouped by stations, in order to determine the most suitable hour.

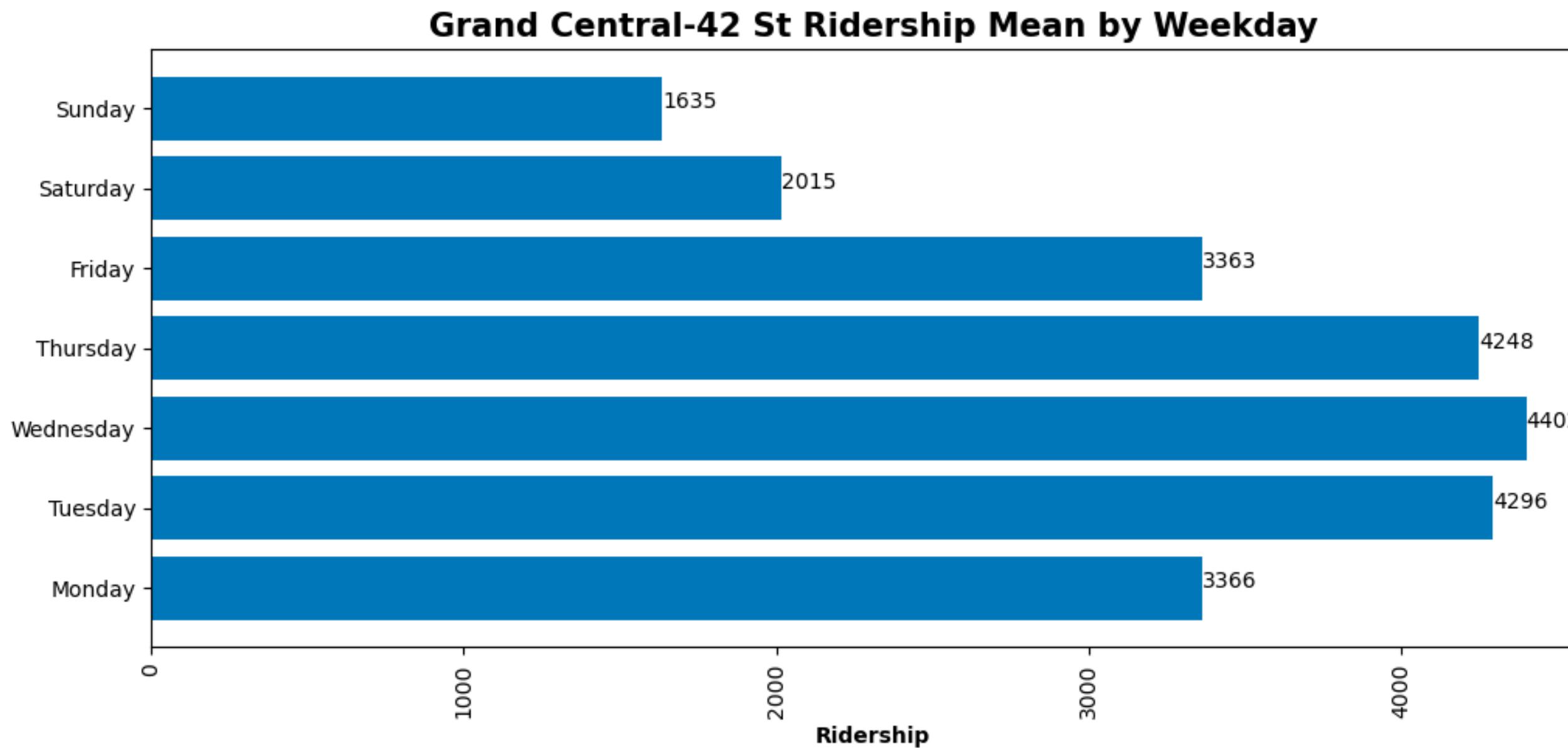
Mean of the passenger counts at stations by days and hours, grouped by stations.



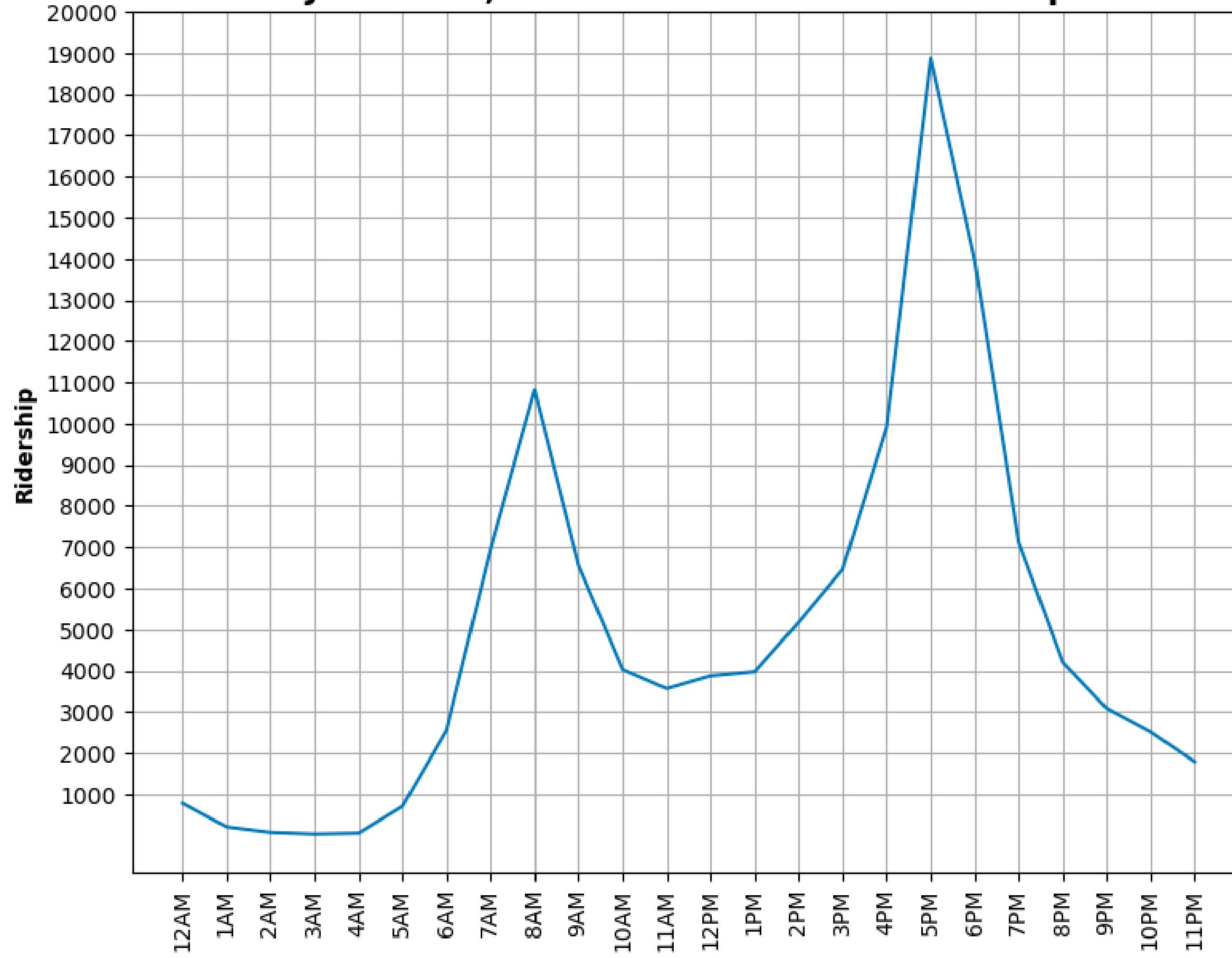
June 13th, Times Sq-42 St Ridership



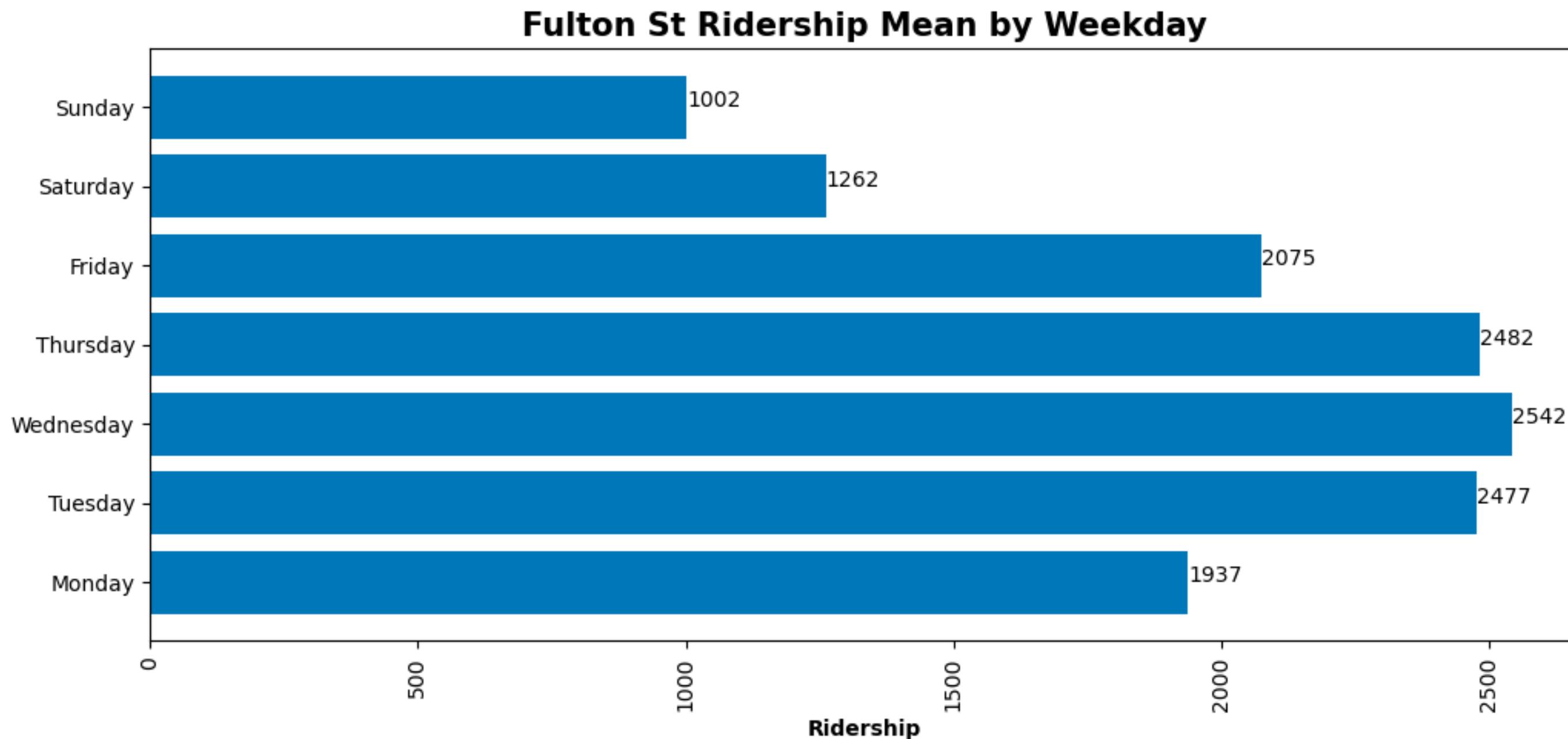
Mean of the passenger counts at stations by days and hours, grouped by stations.



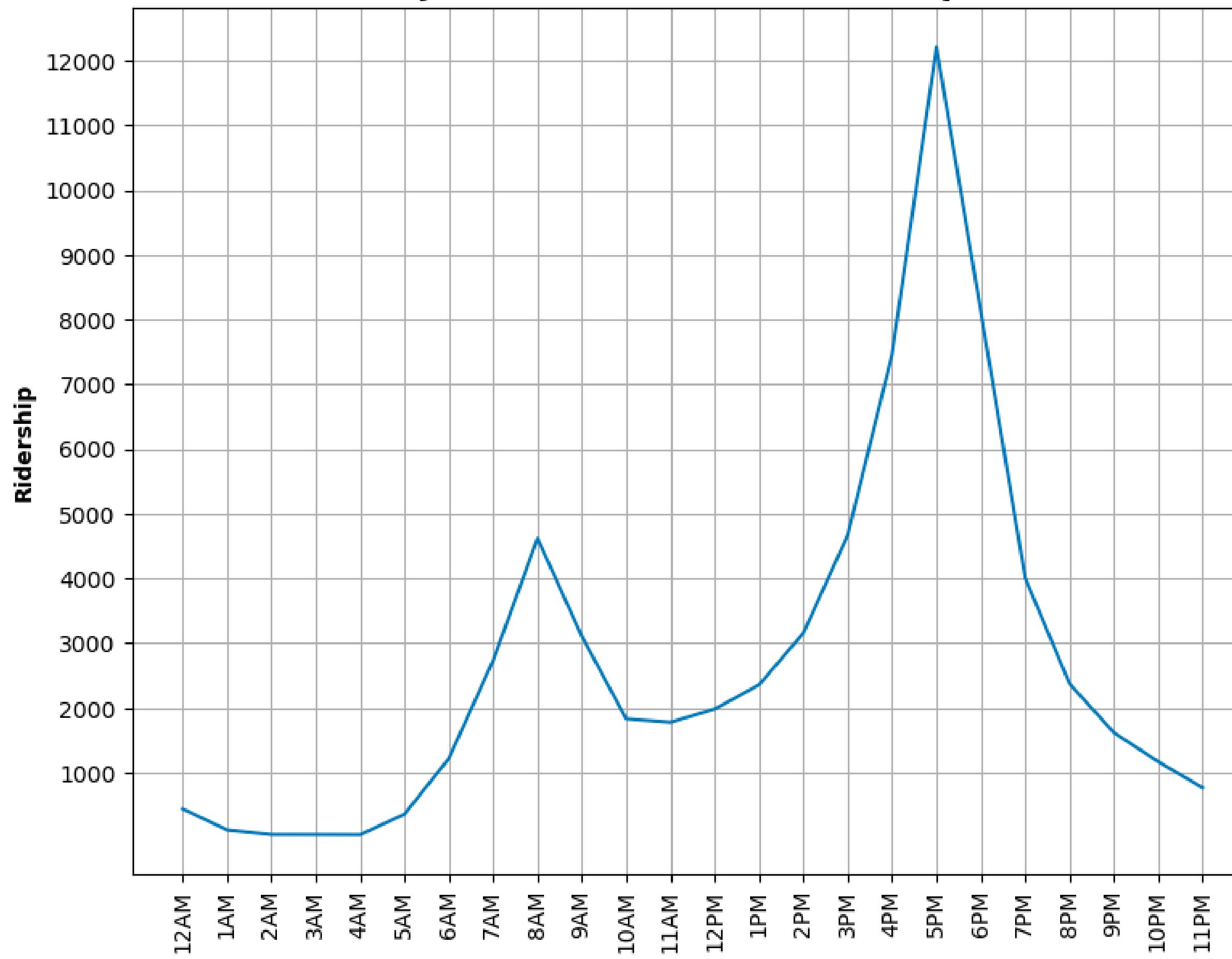
June 13th, Grand Central-42 St Ridership



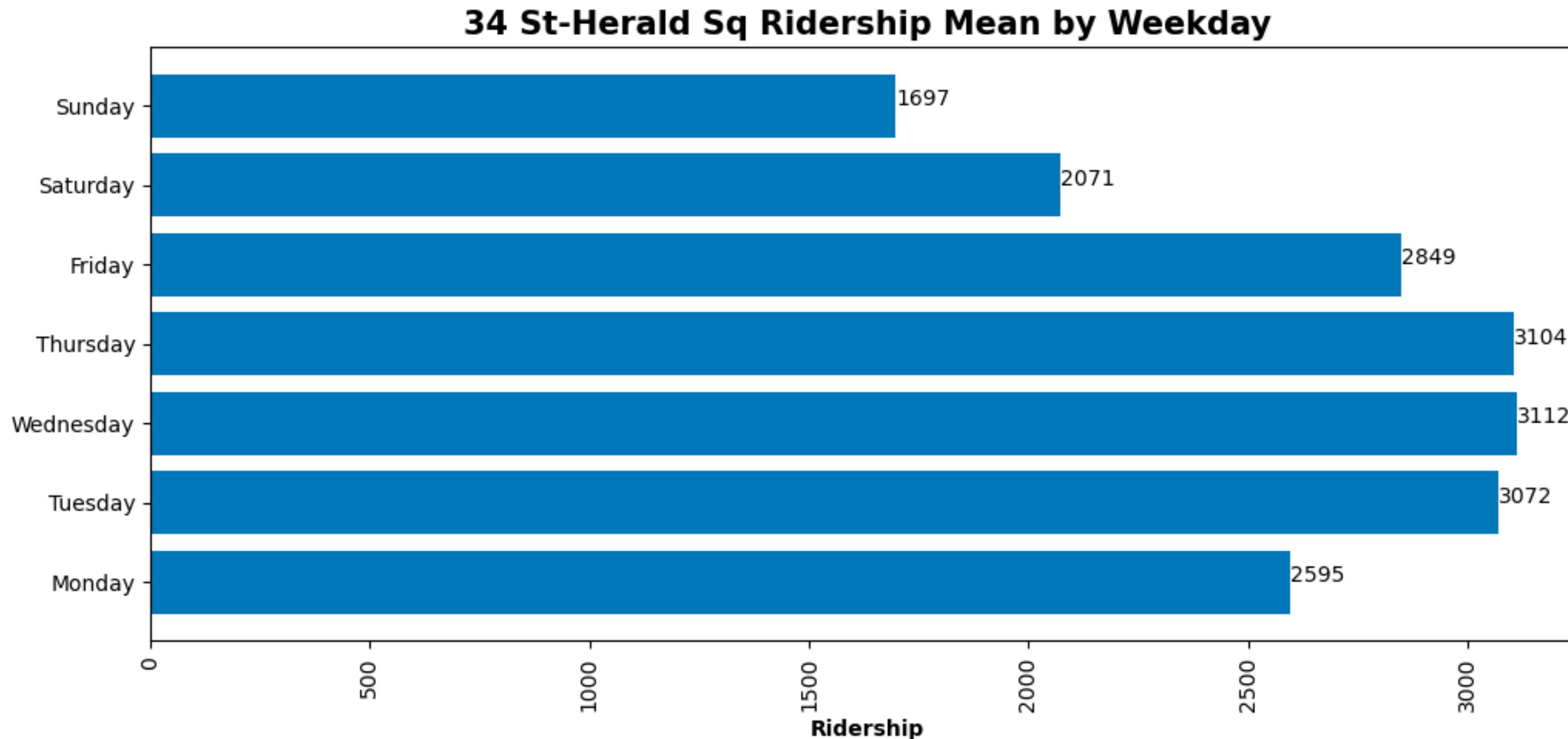
Mean of the passenger counts at stations by days and hours, grouped by stations.



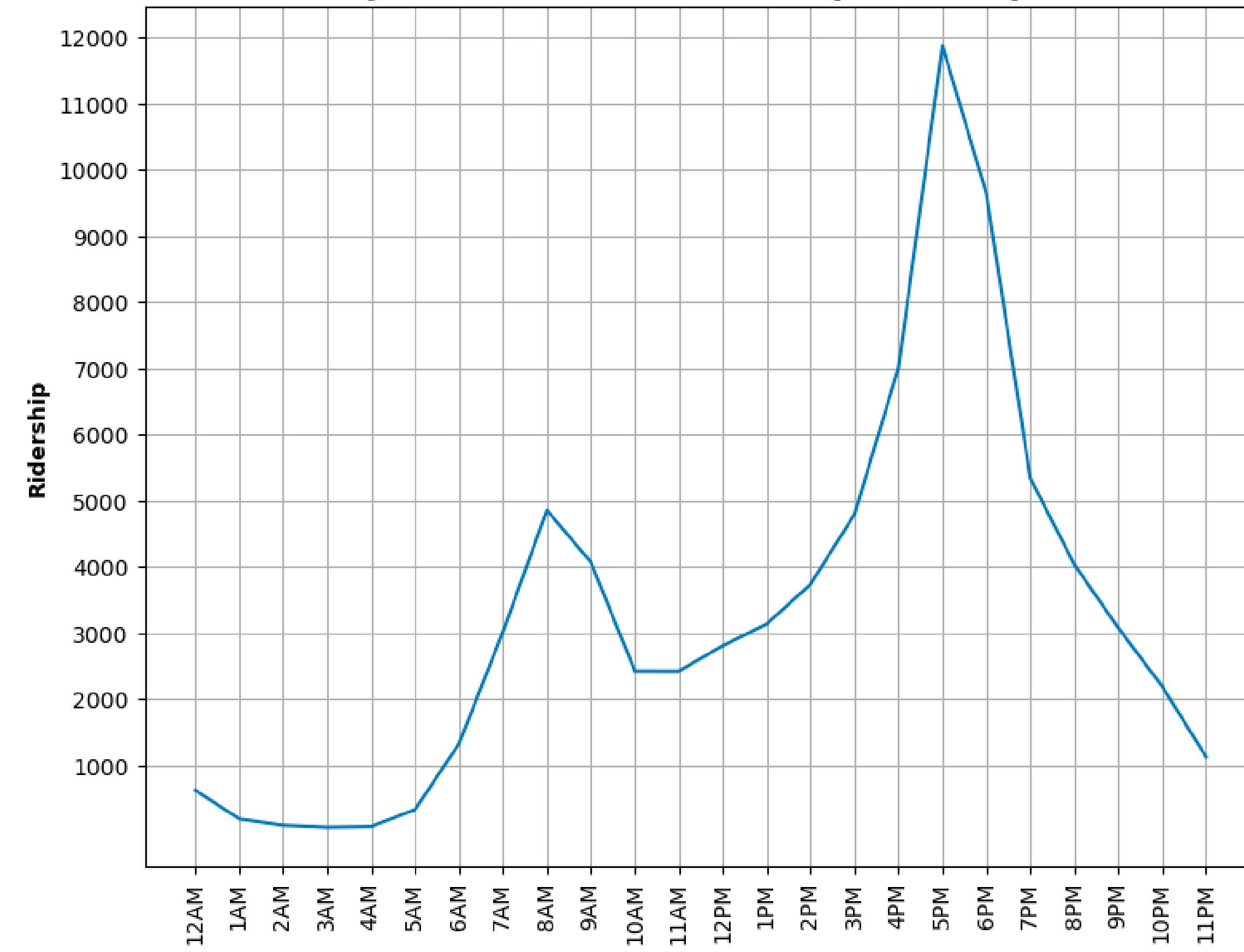
June 6th, Fulton St Ridership



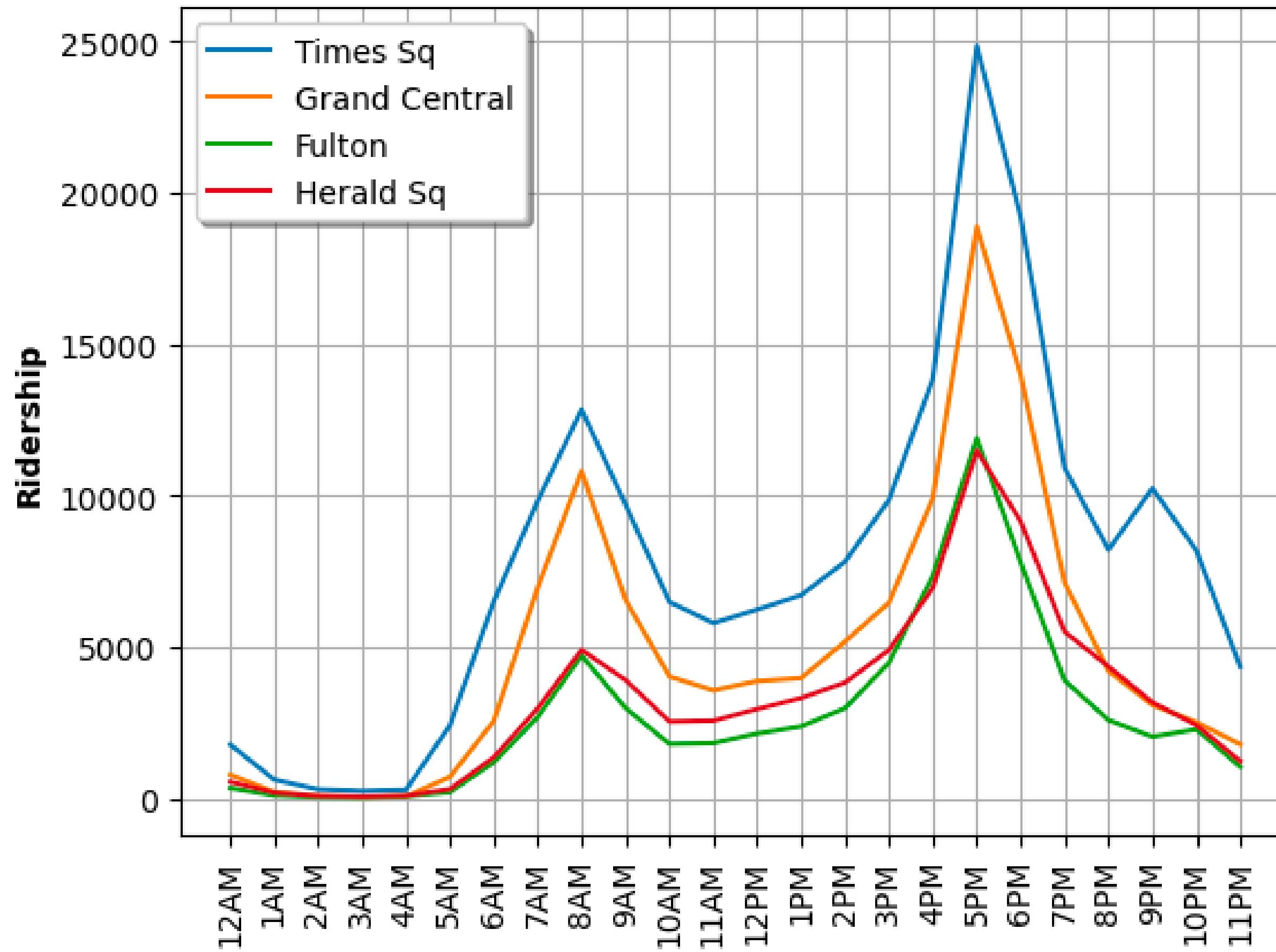
Mean of the passenger counts at stations by days and hours, grouped by stations.



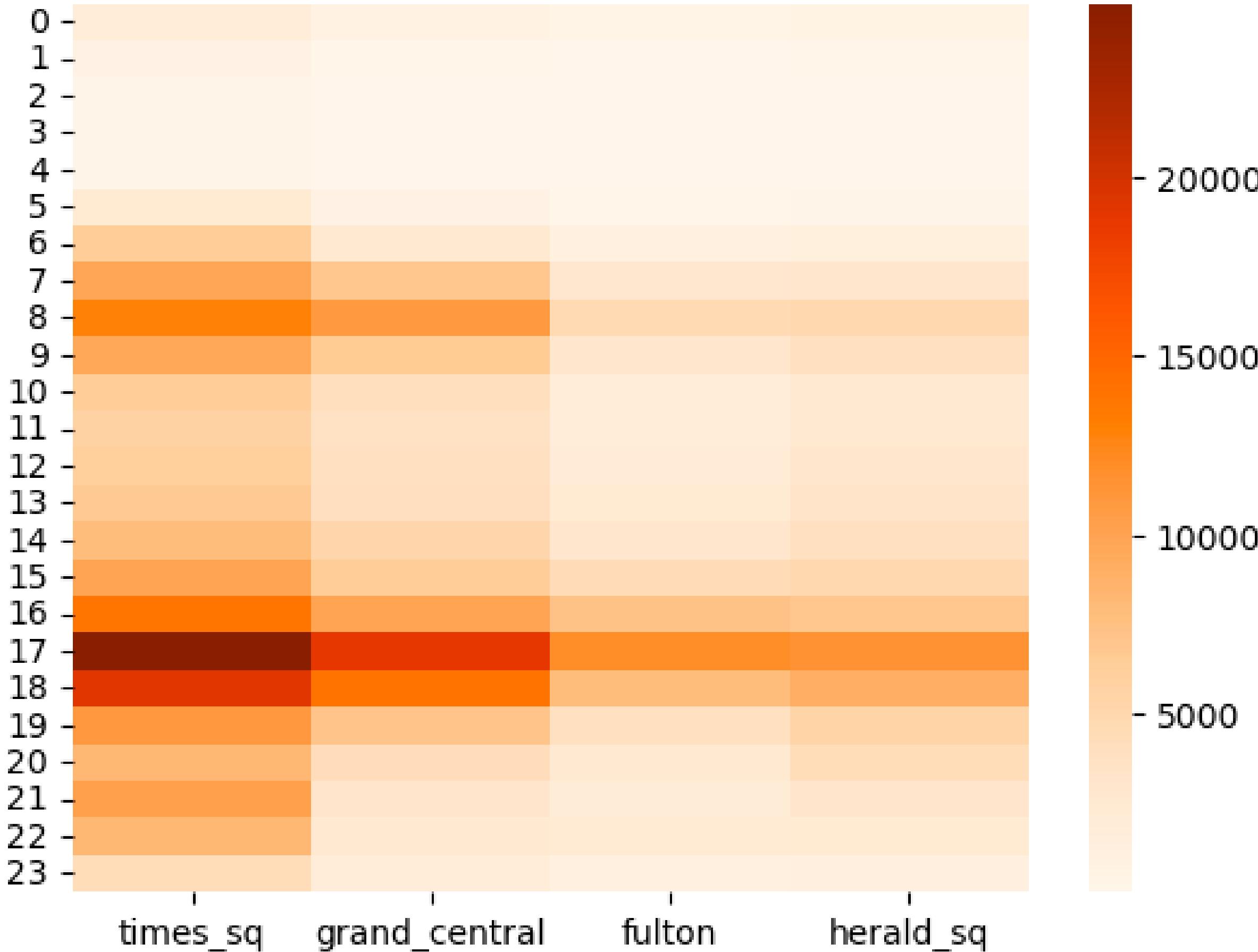
June 6th, 34 St-Herald Sq Ridership



June 13th, Ridership Distribution at Most Crowded Stations



June 13th, Ridership Heatmap at Most Crowded Stations





Step# 3

Conclusion

Based on our analysis, we have determined that focusing primarily on the Manhattan borough would yield the best results.

Considering the two busiest time frames and three different days, we have identified the top three locations to send survey teams. The first choice is a location that includes the busiest station, 42nd Street Times Square, during the morning rush hour (7 a.m. - 10 a.m.) and evening peak hours (4 p.m. - 8 p.m.) on Wednesdays.

Step# 3

Conclusion

The second choice is Grand Central Station during the same time frames on Wednesdays. The third choice is Fulton Station during the specified time frames on Wednesdays.

For optimal efficiency, it is advisable to conduct the survey in the month of June, as it provides the highest ridership.

