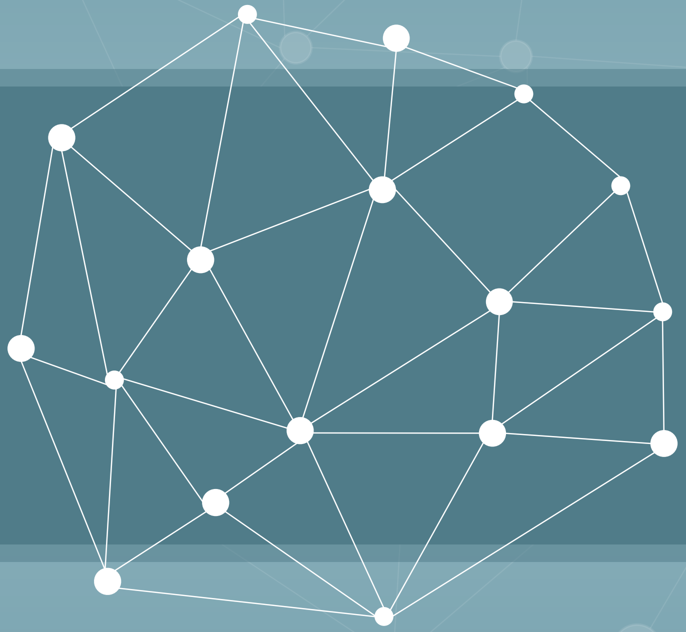
Analysis of correlation between FourSquare Check-ins and weather in New York



Introduction

Why we are doing this?

01

Predict Activities

Based on the weather and the day activities could be recommended

02

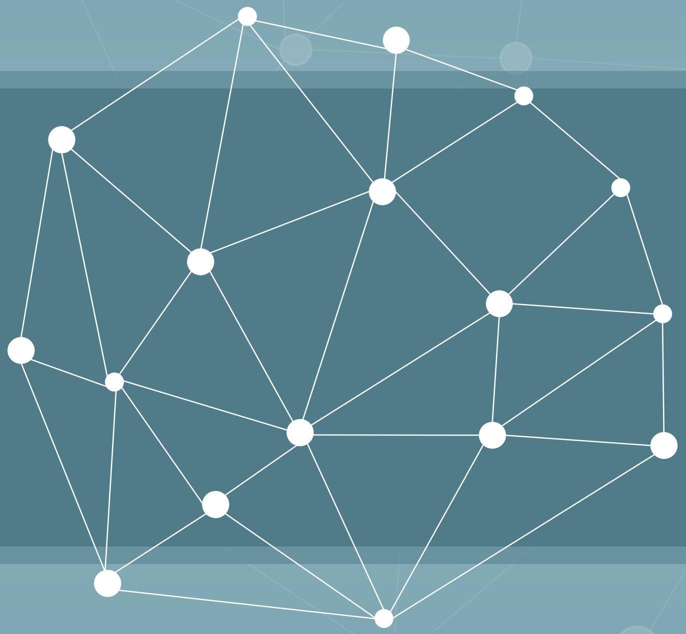
Support Marketing

If you know when people are doing what, this might help travel agencies to better use their marketing budget

03

Get new insights

Unknown insights and correlation could be detected



Data acquisition and cleaning

Data Acquisition and cleaning

01

Get data

First of all, the necessary data needs to be collected. In this case from Kaggle a FourSquare data set and from weatheronline.com the respective weather data.

02

Clean data

The downloaded data, especially the weather data needs to be cleaned and formatted into readable formats. Unnecessary columns need to be removed.

03

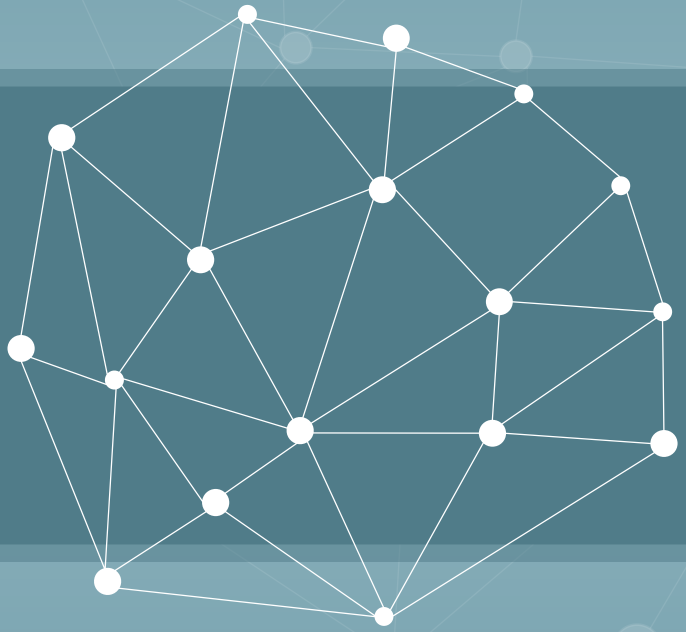
Join data

The FourSquare Check-ins need to be connected to the respective weather data.

04

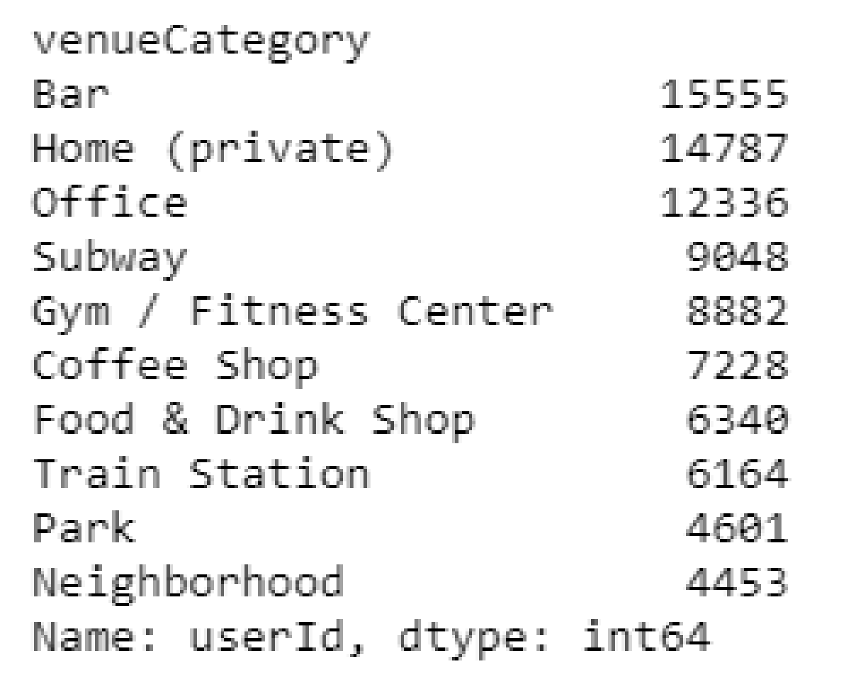
Repeat Process

This process is highly iterative. Every time, I tried a new algorithm I first tried to sort out and improve the data structure.

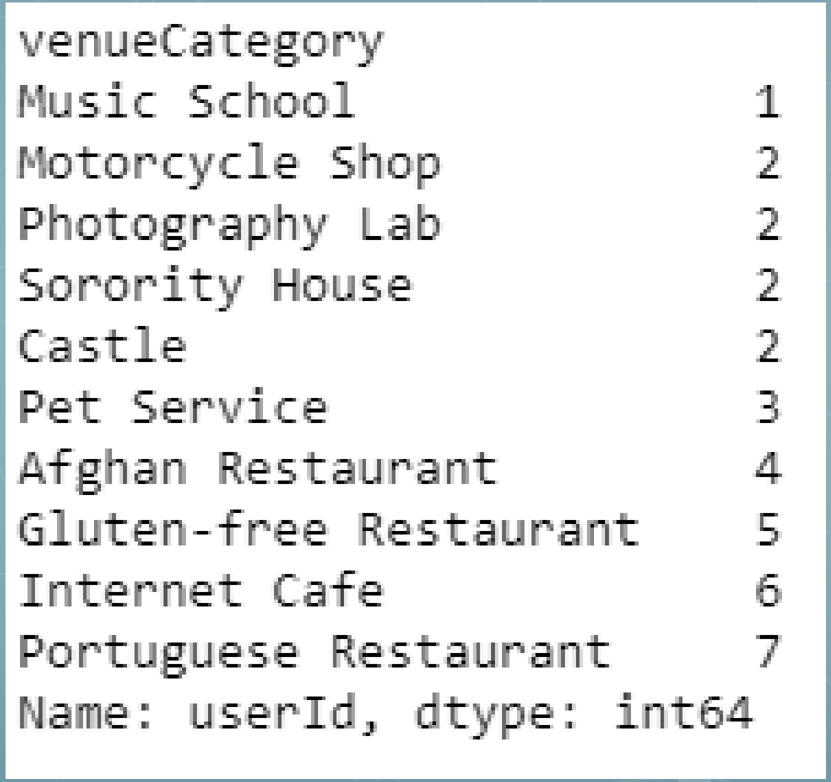


Exploration

Top / least 10 categories

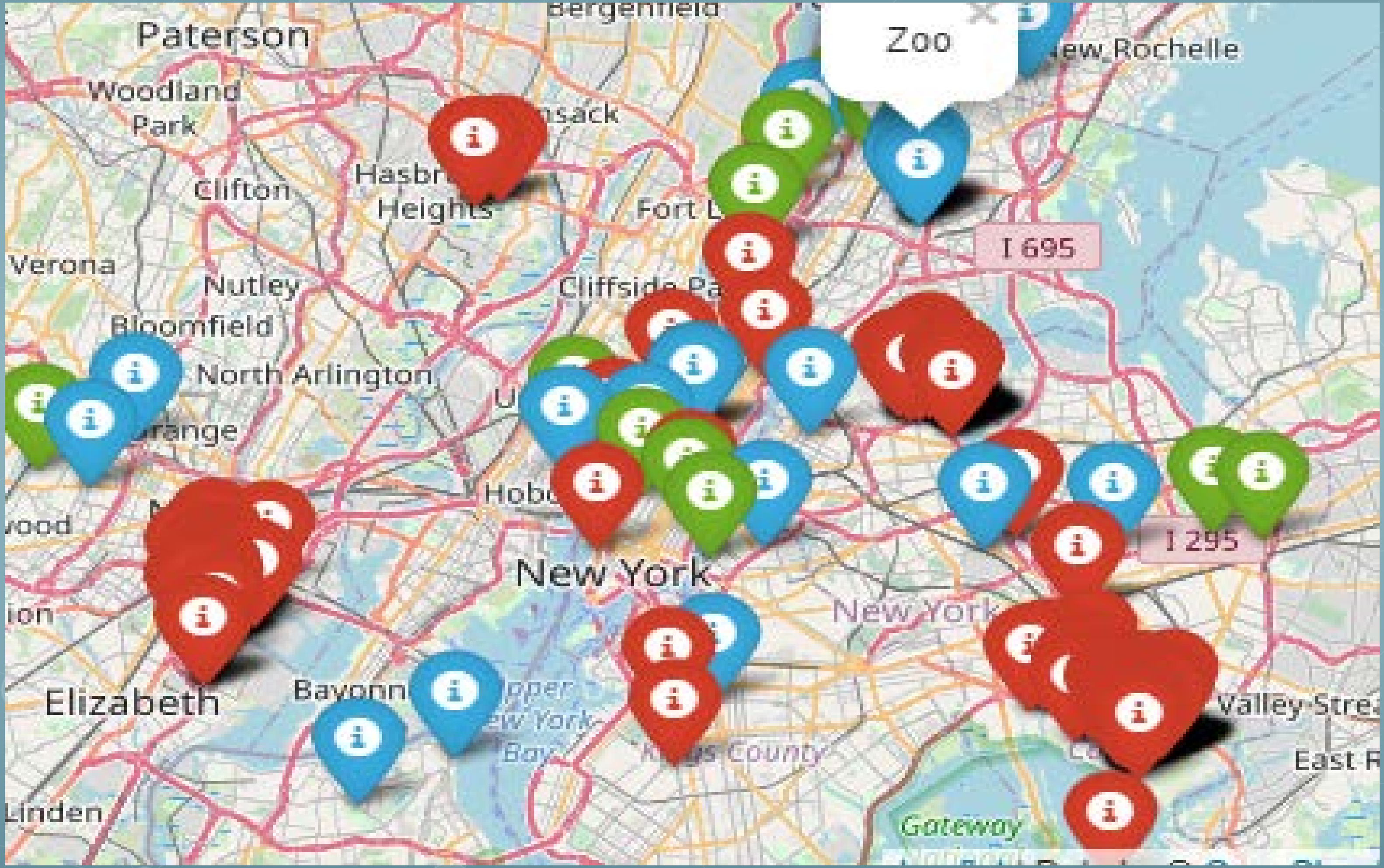


```
venueCategory
Bar                      15555
Home (private)           14787
Office                   12336
Subway                    9048
Gym / Fitness Center      8882
Coffee Shop               7228
Food & Drink Shop         6340
Train Station             6164
Park                      4601
Neighborhood              4453
Name: userId, dtype: int64
```

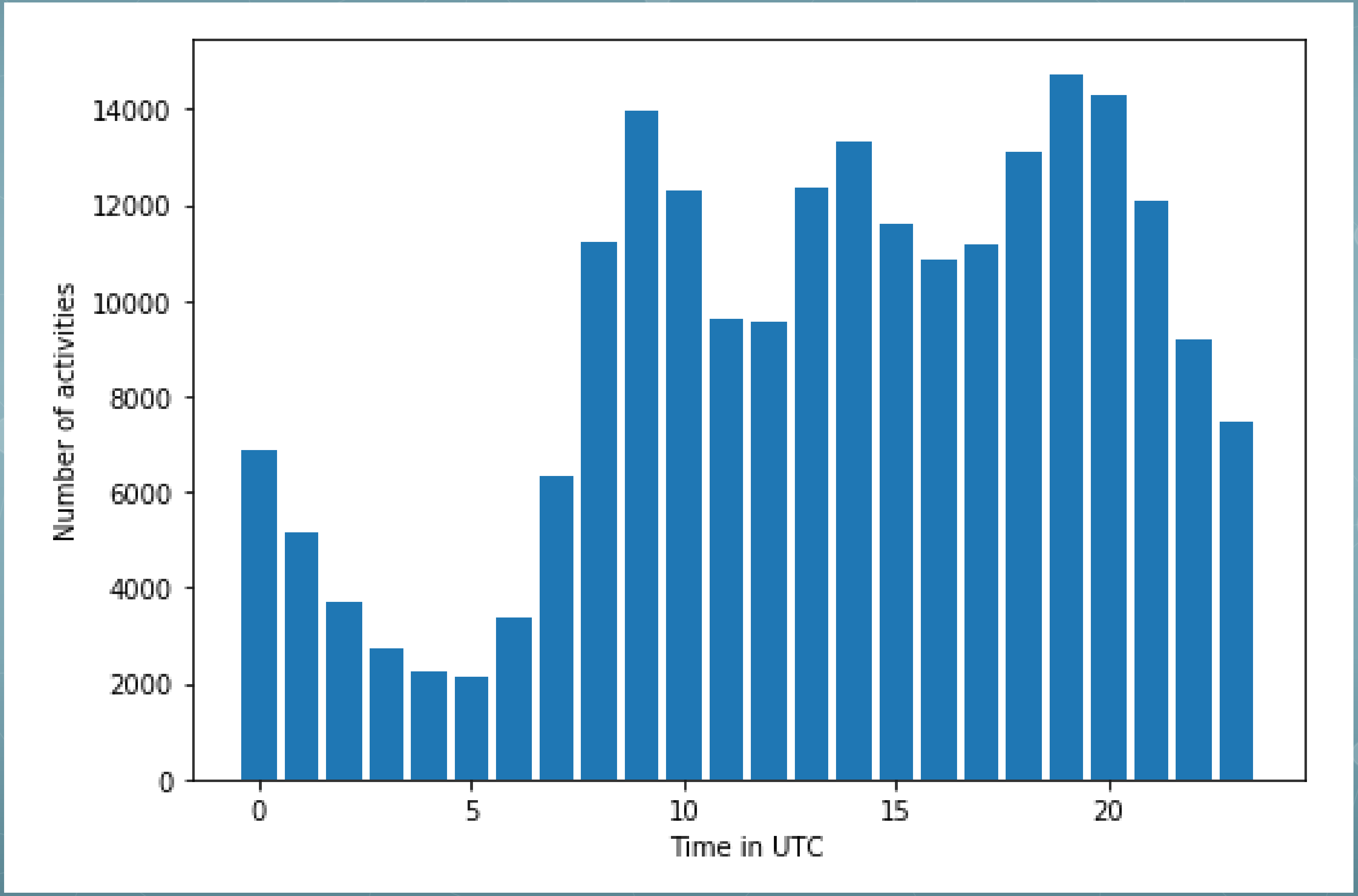


```
venueCategory
Music School              1
Motorcycle Shop           2
Photography Lab           2
Sorority House            2
Castle                    2
Pet Service               3
Afghan Restaurant         4
Gluten-free Restaurant    5
Internet Cafe             6
Portuguese Restaurant     7
Name: userId, dtype: int64
```

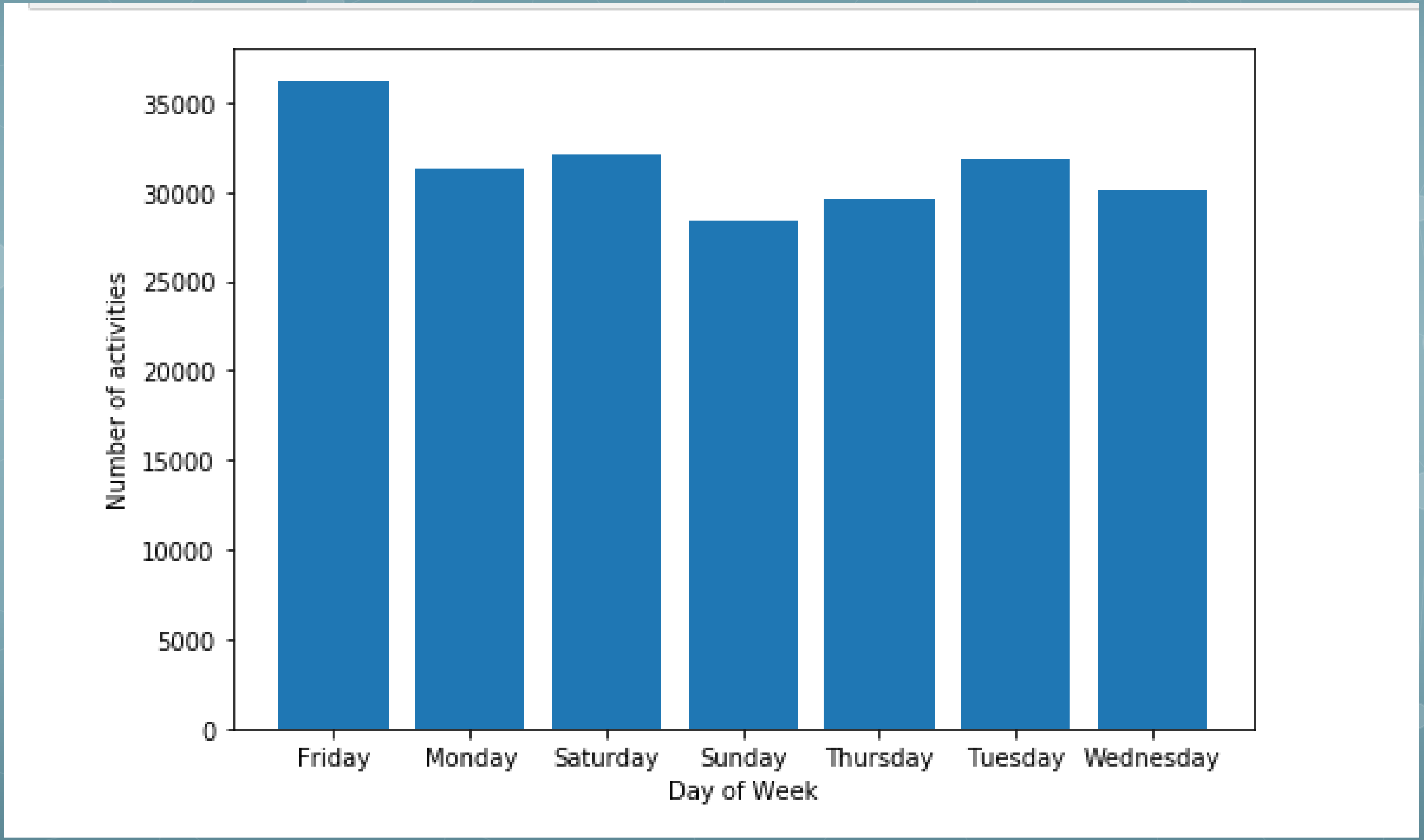
Map of different activities

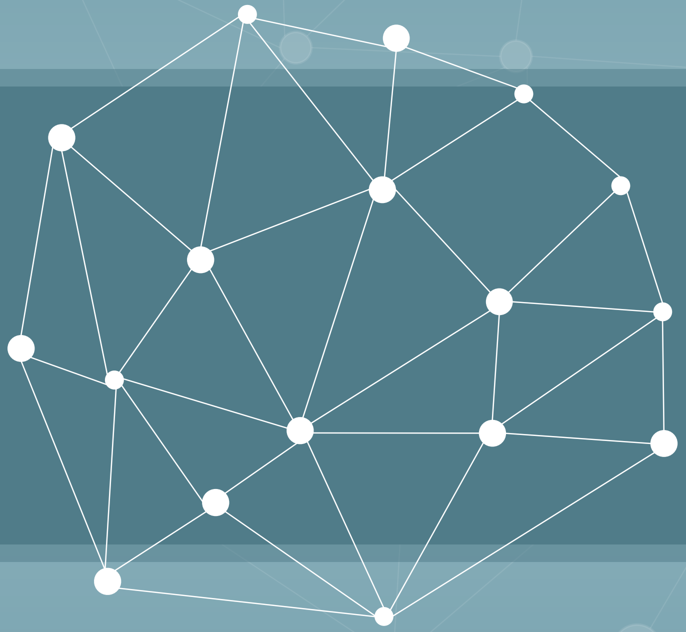


Check-ins per time



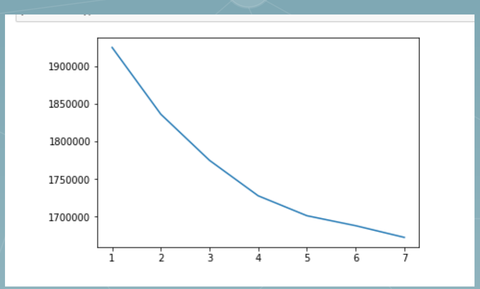
Check-ins per weekday



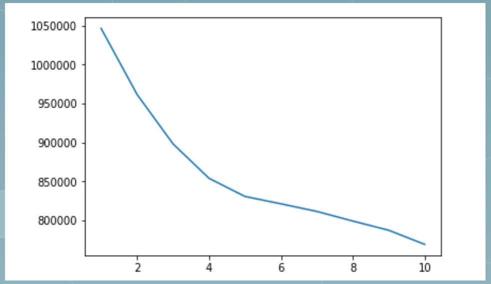


Clustering

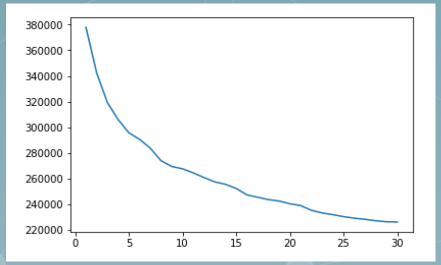
Multiple Attempts



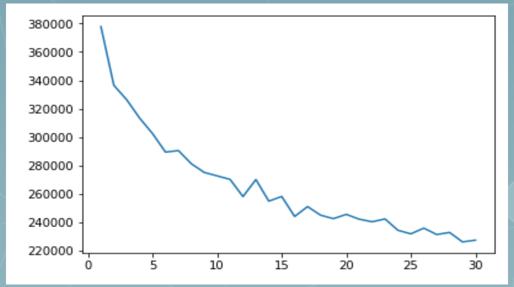
1



2



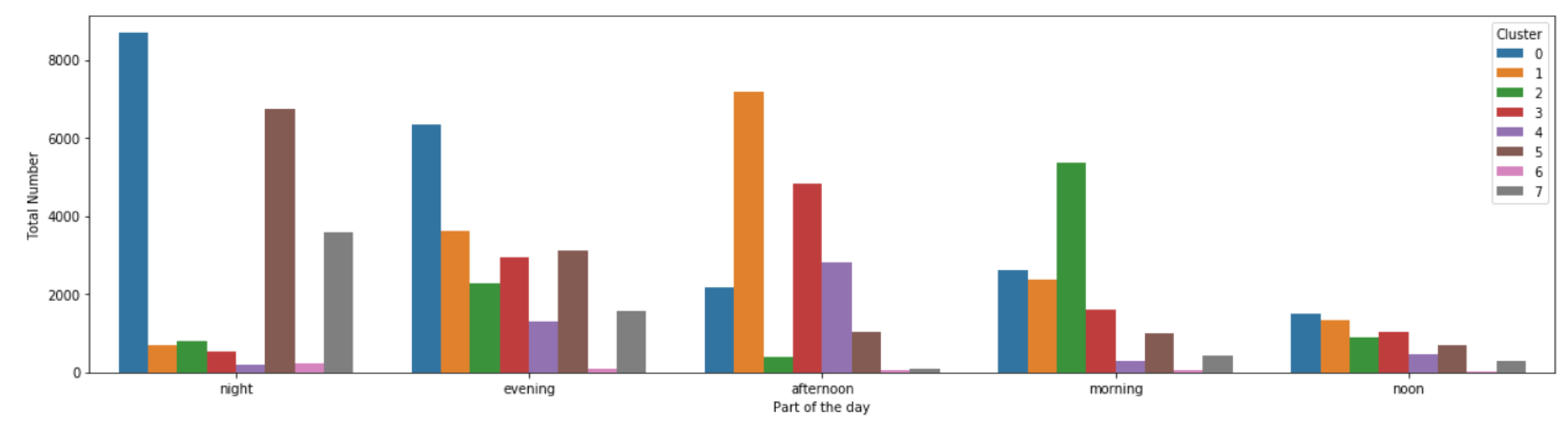
3



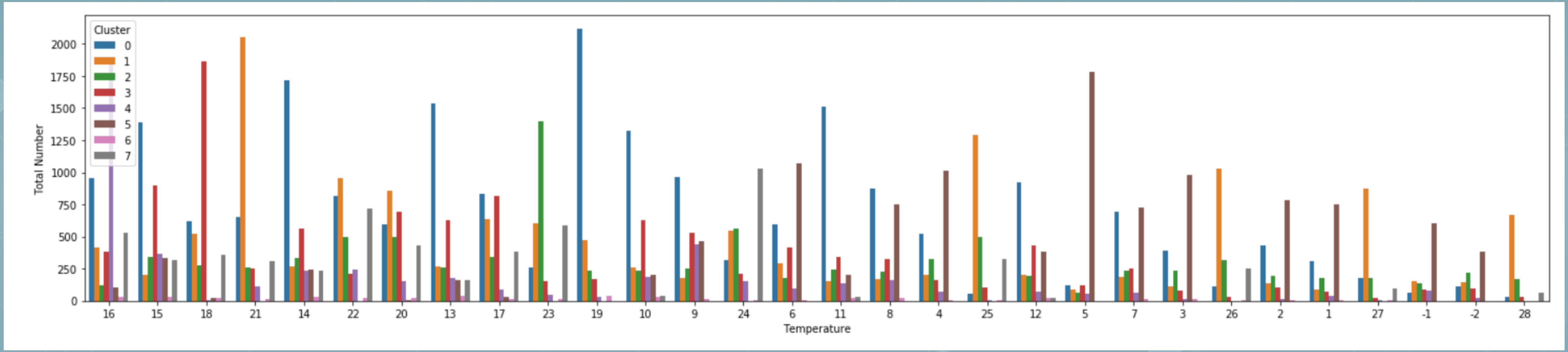
4

Best attempt was with k = 8

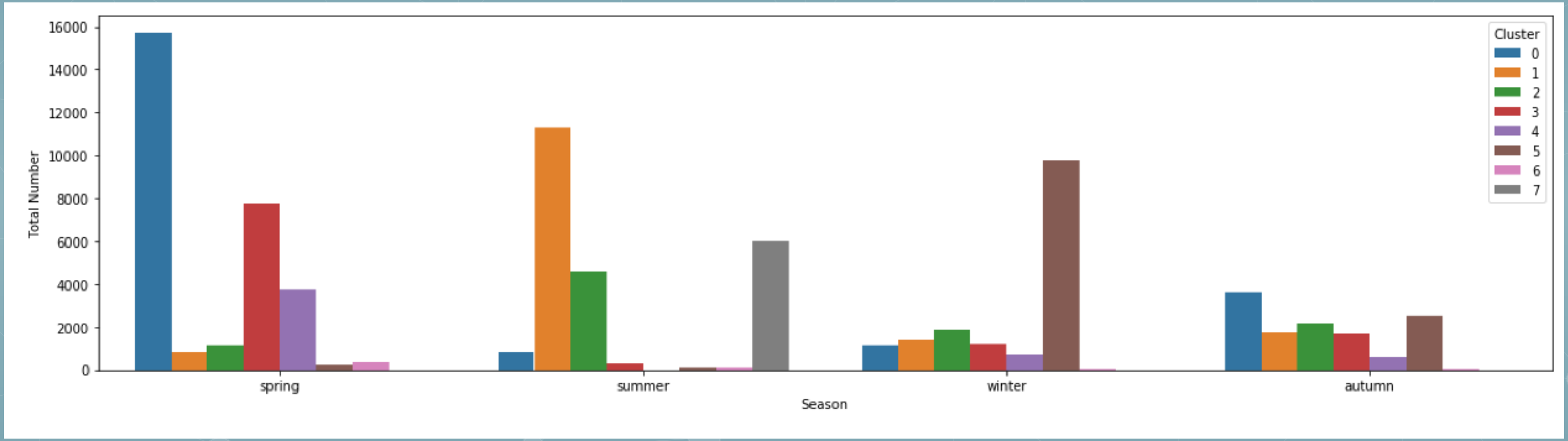
Cluster exploration



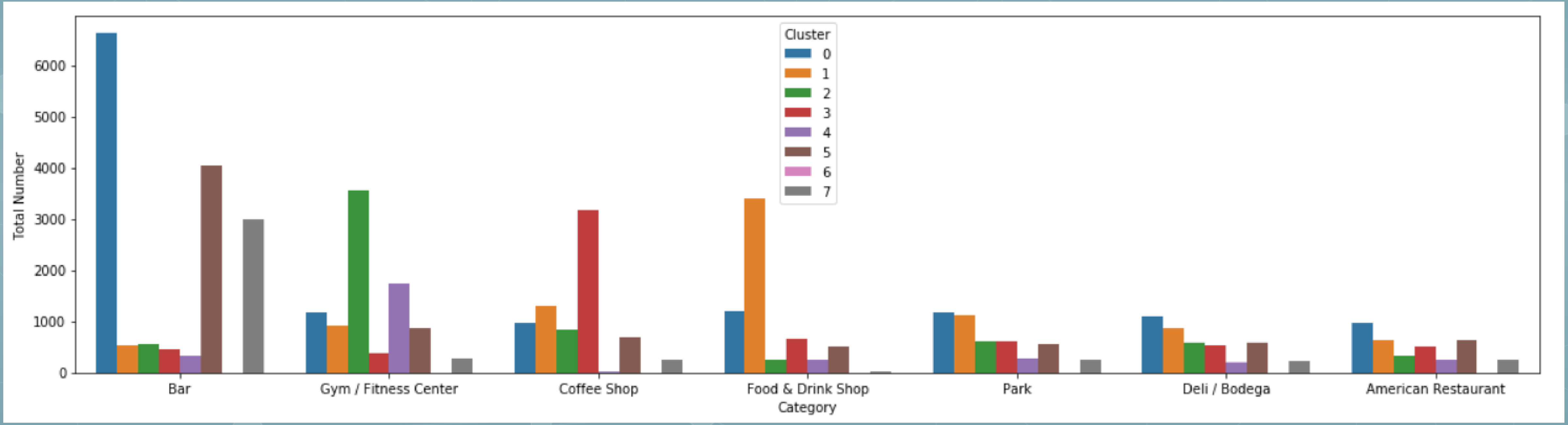
Cluster exploration



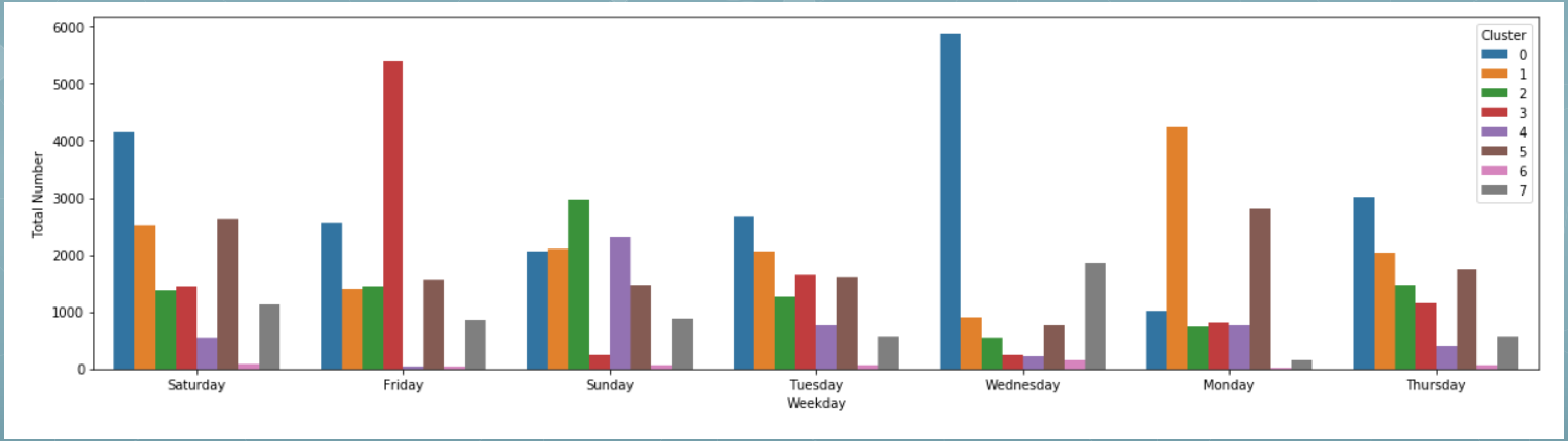
Cluster exploration

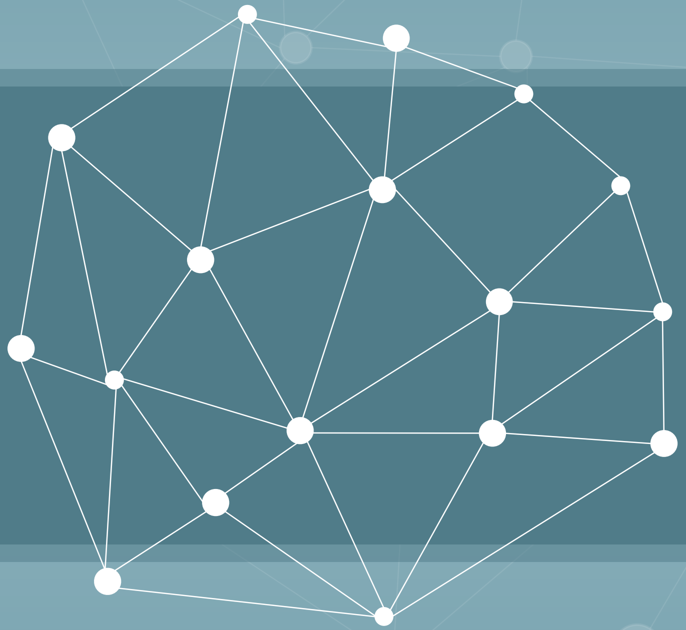


Cluster exploration



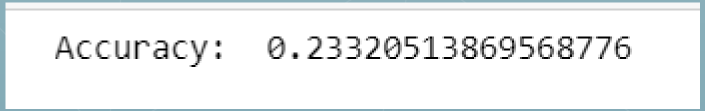
Cluster exploration





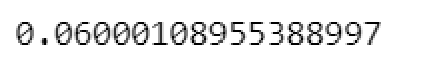
Prediction

Comparison
Decision Tree, SVM, KNN



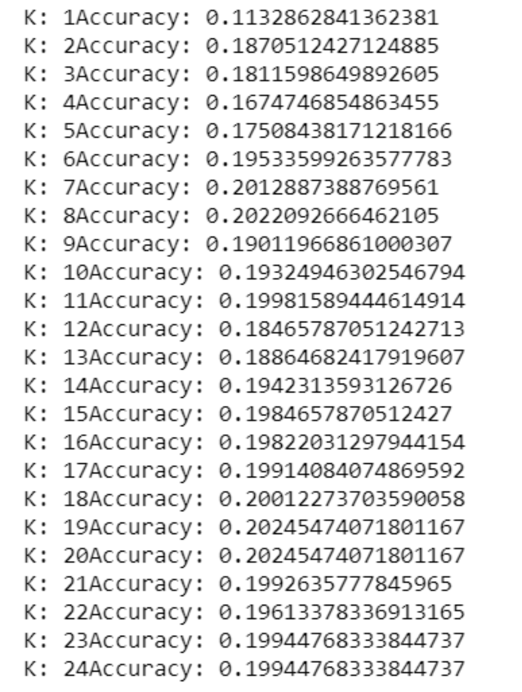
```
Accuracy:  0.23320513869568776
```

Decision Tree



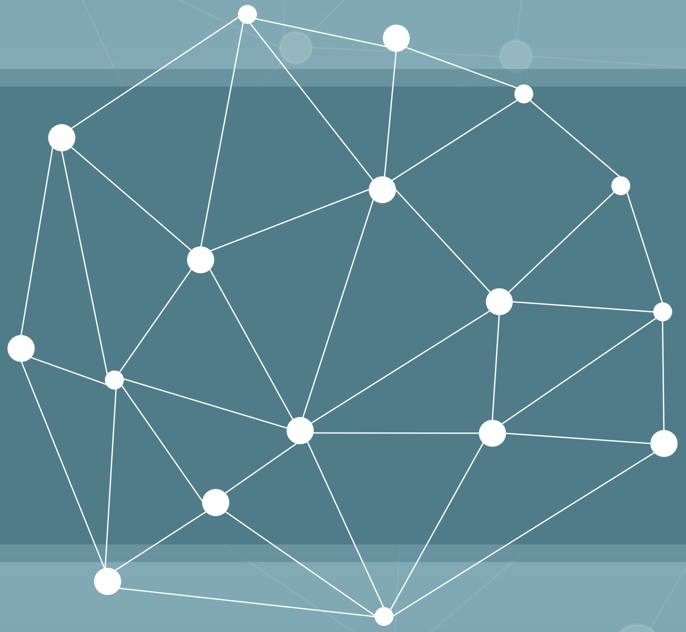
```
0.06000108955388997
```

SVM



```
K: 1Accuracy: 0.1132862841362381
K: 2Accuracy: 0.1870512427124885
K: 3Accuracy: 0.1811598649892605
K: 4Accuracy: 0.1674746854863455
K: 5Accuracy: 0.17508438171218166
K: 6Accuracy: 0.19533599263577783
K: 7Accuracy: 0.2012887388769561
K: 8Accuracy: 0.2022092666462105
K: 9Accuracy: 0.19011966861000307
K: 10Accuracy: 0.19324946302546794
K: 11Accuracy: 0.19981589444614914
K: 12Accuracy: 0.18465787051242713
K: 13Accuracy: 0.18864682417919607
K: 14Accuracy: 0.1942313593126726
K: 15Accuracy: 0.1984657870512427
K: 16Accuracy: 0.19822031297944154
K: 17Accuracy: 0.19914084074869592
K: 18Accuracy: 0.20012273703590058
K: 19Accuracy: 0.20245474071801167
K: 20Accuracy: 0.20245474071801167
K: 21Accuracy: 0.1992635777845965
K: 22Accuracy: 0.19613378336913165
K: 23Accuracy: 0.19944768333844737
K: 24Accuracy: 0.19944768333844737
```

KNN

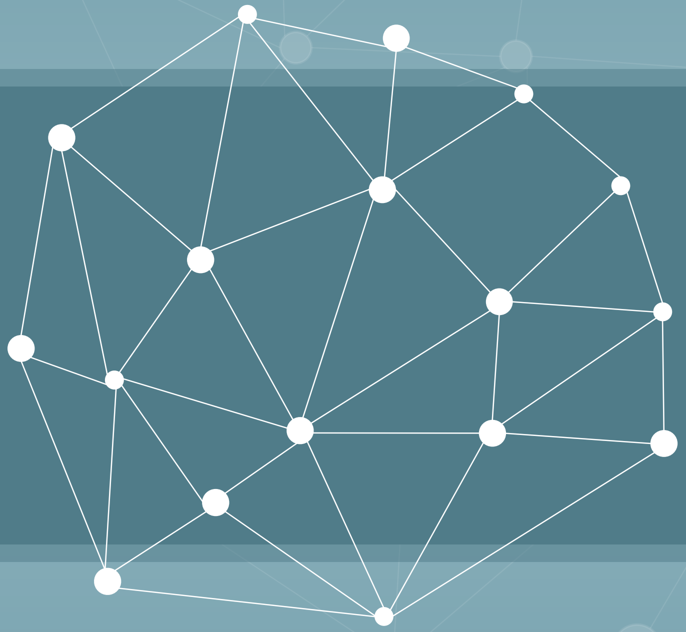


Conclusion

Conclusion

Clustering works quite well, prediction needs to be improved. Identified clusters:

- Cluster 0: The people in this class like to go to bars on warm spring nights, especially on Wednesday.
- Cluster 1: The people in this class like to go Food and Drink Shops on hot summer days, especially on Monday.
- Cluster 5: The people in this class like to bars on cold winter nights. A special day with higher activities as on others could not clearly be examined.



Outlook

Outlook

Improvement of prediction algorithm and cleaning of data that is used for algorithms.