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

1 Introduction

Every year, several million people are visiting New York. Some of these tourists use foursquare check-ins (https://support.foursquare.com/hc/en-us/articles/201065340-Check-ins) to keep a record of all places they visited. This may give their friends advises for secret places. Furthermore, everyone who has ever planned a larger trip knows that especially outdoor activities highly rely on the weather condition. This makes it often necessary to plan simultaneously for different weather situations. The check of a possible correlation between this check-in data of New York and the respective weather data at this time could give insights on, which locations are best for which weather condition and predict possible destinations for tourists.

1.1 Problem

This project aims to find a correlation between the FourSquare check-in data of New York and the weather condition on these days and based on that predict possible destinations based on the weather condition.

1.2 Interest

This analysis might be interesting for people planning their next holiday trips to New York to make easy recommendations for possible activities based on day, daytime, season and weather. Furthermore, any travel agencies could use the data for more specific marketing activities for travels to/around New York.

2 Data acquisition and cleaning

2.1 Data sources

The data that is used for this analysis can be gathered from two different sources. The first source is kaggle (https://www.kaggle.com/chetanism/foursquare-nyc-and-tokyo-checkin-dataset) where the check-in data from New York and Tokyo from 12 April 2012 to 16 February 2013 can be downloaded.
For New York, the data set contains 227,428 entries in the following 8 columns:

- User ID (anonymized)
- Venue ID (Foursquare)
- Venue category ID (Foursquare)
- Venue category name (Fousquare)
- Latitude
- Longitude
- Timezone offset in minutes (The offset in minutes between when this check-in occurred and the same time in UTC)

- UTC time

The weather data is scrapped from the API of Worldweatheronline (https://www.worldweatheronline.com/). The service provider is offering realtime, future and historical weather data. In this case especially the historical data is relevant.
Based on several parameters (start date, end date, time interval and extras as e.g. utcDateTime) the API can give back historical weather data of New York. To get a most accurate experience I scrapped the weather data on an hourly time interval.

2.2 Data cleaning

The data cleaning for the check-in data from kaggle was quite comfortable as there were cleaning activities necessary whereas the weather data was more challenging to prepare. As the data was gathered from the API interface it was provided as a json which needed to be converted to a data frame. Especially the json containing the hourly weather data for each hour which was embedded in the original request was difficult to extract. The next step was to remove all unnecessary columns including some additional information which are not necessary for this project (e.g. uvIndex, windspeedMiles and tempF as the temperature was given in degree Fahrenheit and degree Celsius and I decided to go with the Celsius scale. The next step was the cleaning of the time. As we are trying to find a correlation between weather and activities it is crucial to proper match the time of the activity with the respective weather condition at this time. Therefore, both datetime values were converted to UTC datetime. In the last step the data was clustered into bins to improve the accuracy. For this project the exact date of the activity is not relevant but the respective day, the time of the day and the season of the year is.
This results three bins (season, day, day_ow) with the following values:

- Season_bin: ['winter', 'spring', 'summer', 'autumn']
- Day_bin: ['morning', 'noon', 'afternoon', 'evening', 'night']
- Day_ow_bin: ['Monday', 'Tuesday', 'Wednesday', 'Thursday', 'Friday', 'Saturday', 'Sunday']

2.3 Feature selection

After cleaning the data, 219,501 entries were still left within 6 columns (venueCategory, tempC, sunHour, season_bin, day_bin, day_ow_bin). The columns were selected based on the problem described in the chapter before. The venueCategory is the category which should be predicted. The other columns are the independent columns which should influence the venueCategory. E.g. if the weather is sunny and without any rain I would expect the people to do more outdoor activities where as in cold and rainy days the people are less active or at least more focused on indoor activities. Furthermore, I would expect the people to be more active on Friday or Saturday nights especially regarding going out for dinner or going out for party.

3 Explorative Analysis

In the following section I will provide a little overview of the performed explorative analysis to get familiar with the data:

3.1 Top 10 check-ins

First I checked the dataset on the different type of categories of activities that are included in the dataset. The top 10 check-in categories can be found below:

```
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
```

3.2 Least 10 check-ins

In the next step I was also curious about the activities that were used very seldom with the least 10 which can be found in the next graphic:

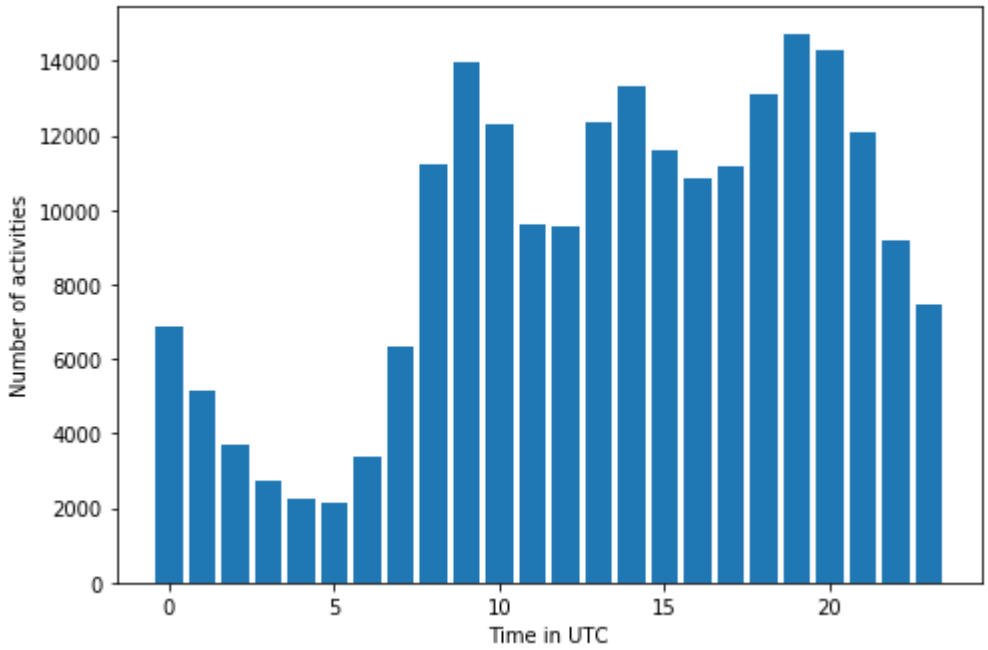
```
venueCategory
Music School              1
Motorcycle Shop           2
Photography Lab           2
Sorority House            2
Castle                    2
Pet Service               3
Afghan Restaurant         4
Gluten-free Restaurant    5
Internet Cafe             6
Portuguese Restaurant     7
Name: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

3.3 Map of check-ins of different activities

In the next step I plotted a map with several activity locations to get a rough overview of the parts of the city which are heavily used for specific activities:

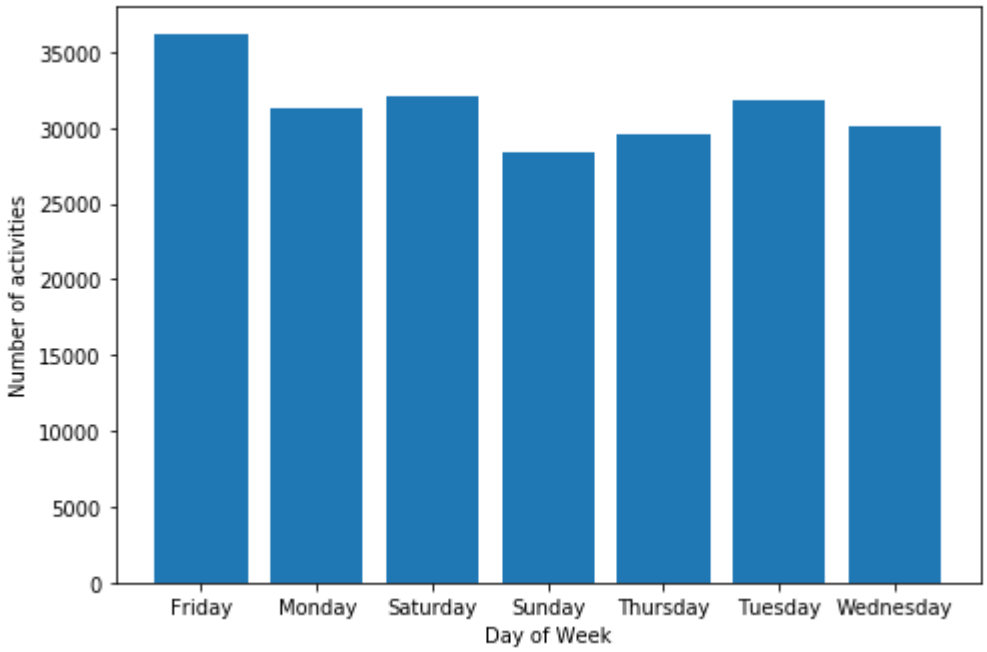
3.4 Check-ins per time

The next figure is about the number of check-ins versus the time of the day. It can easily be identified that during the night (1am – 6am) less activity was recorded:



3.5 Check-ins per weekday

The next figure shows the activity per day. There no major differences can be identified. The number of activities for Friday and Saturday is slightly higher than for the other weeks but all days show more or less the same level of activities:

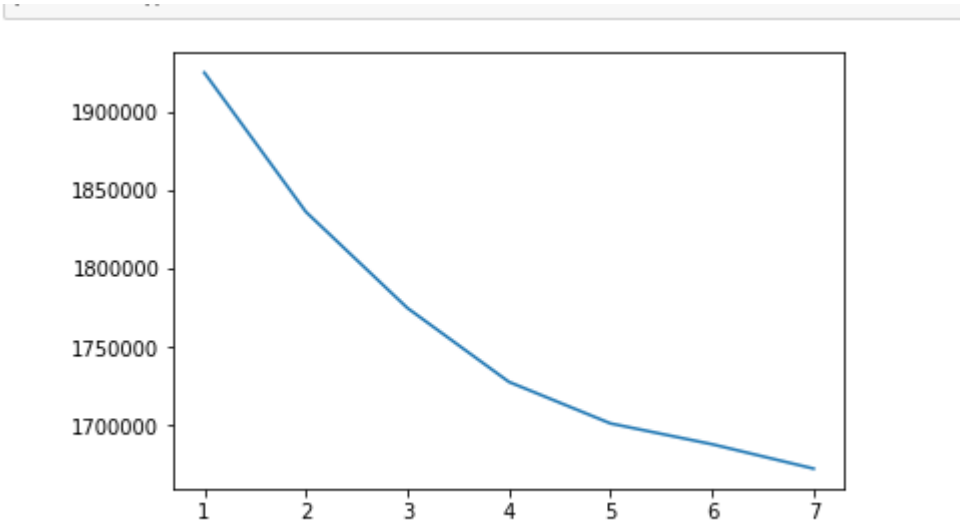


4 Clustering

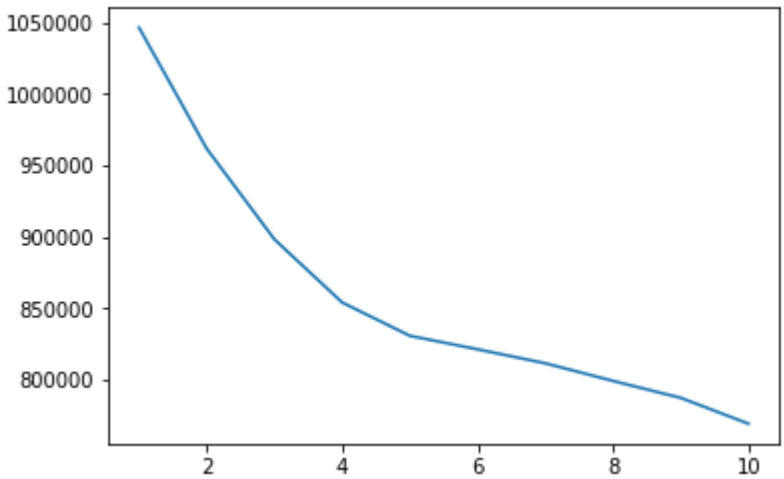
In the next step I tried to cluster the data into different segments to show similarities. This could e.g. mean that there is a group of people going to bars every Friday or Saturday night or maybe another group of people that do some cultural activities as e.g. museum or theatre.

4.1 K-Means / K-Mode

For the clustering I thought about using the K-Means algorithm. Unfortunately, this algorithm is only working with numeric data and as I have several features containing Strings, this algorithm would not work. I found out that there is an algorithm like the K-Means with the ability to also work with non-numeric data which is called K-Mode.
To use the algorithm I deleted all columns that are not necessary, trained the model and ran the algorithm 8 times to select the best parameter.
The result was the following figure:



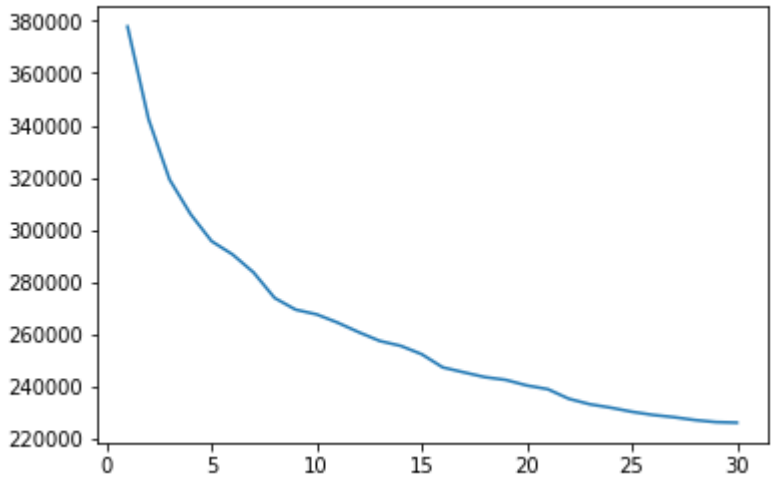
As in this case, the elbow point could not be clearly selected I adapted the data again by rounding the hours of sun per day and the temperature to integers. Running the algorithm again with 11 attempts gave the following output:



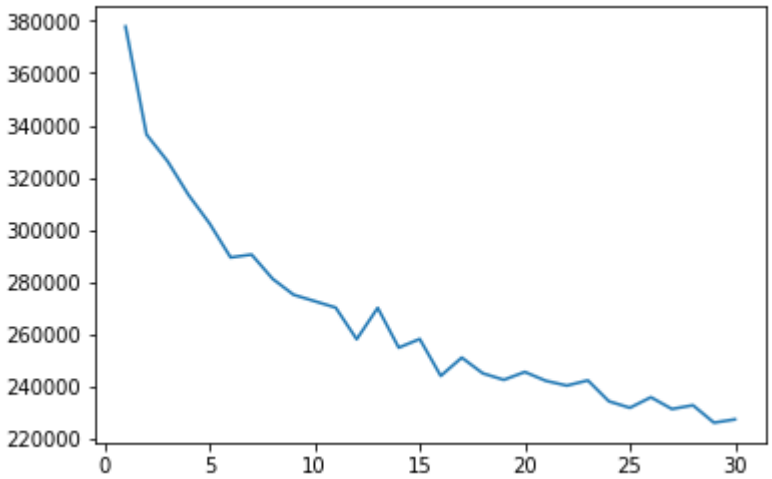
In this case the elbow point could be estimated more clearly.
To further improve the algorithm, I cleaned up the data again by only focusing on the most visited places we identified during the explorative analysis.
Only the check-ins from the following categories are now still used:

['Bar', 'Food & Drink Shop', 'Clothing Store', 'Coffee Shop', 'Pizza Place', 'American Restaurant', 'Deli / Bodega', 'Gym / Fitness Center', 'Italian Restaurant',
'Park', 'Chinese Restaurant', 'Fast Food Restaurant', 'Sandwich Place', 'Mexican Restaurant', 'Bakery','Café','Diner', 'General Entertainment', 'Burger Joint',
'Sushi Restaurant', 'Theater', 'Art Gallery', 'Food Truck', 'Athletic & Sport', 'Ice Cream Shop', 'Spa / Massage', 'Bagel Shop', 'Donut Shop', 'Asian Restaurant',
'French Restaurant', 'Thai Restaurant', 'Japanese Restaurant', 'Seafood Restaurant', 'Indian Restaurant']

This time I ran the algorithm for 30 times to check if a higher number of attempts maybe result in a more precise result:



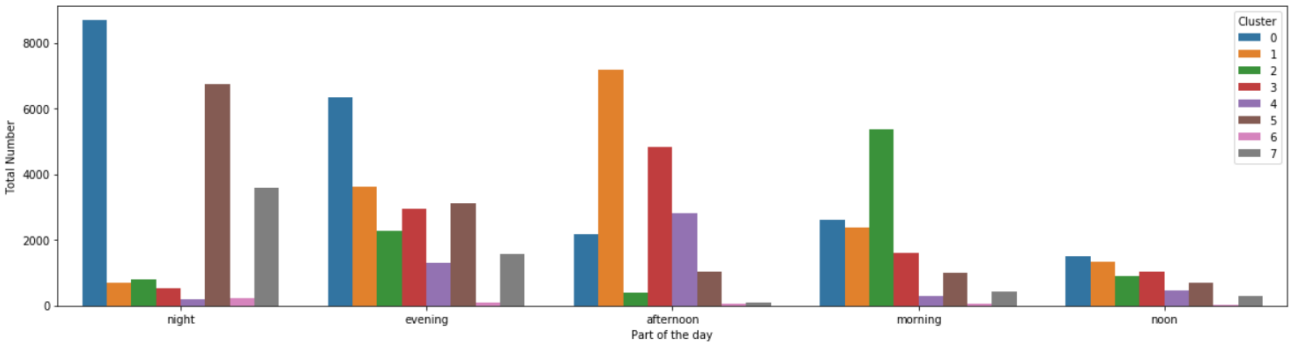
As for the K-Mode algorithm different types are available (Cao, Huang). I switched this time from Cao to Huang. The result for another 30 attempts was the following:



Finally, the best K and the best method was chosen. In this case it was K=8 and method was Huang.

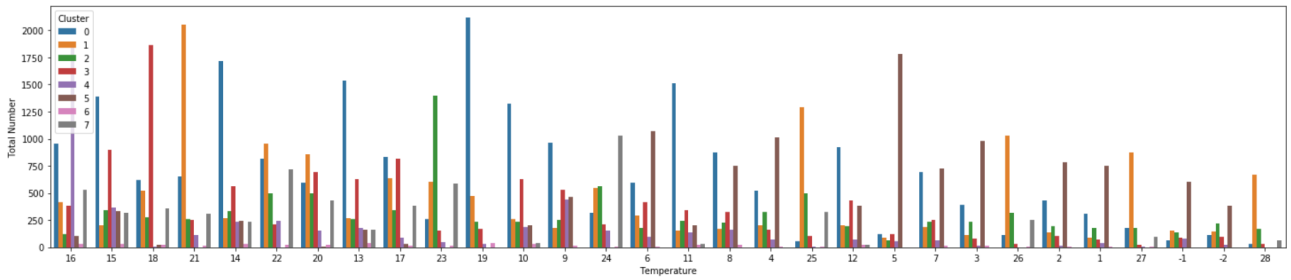
The last step in this section was to visualize the clusters.

The first one was about Total Numbers of activities vs. Part of the day.



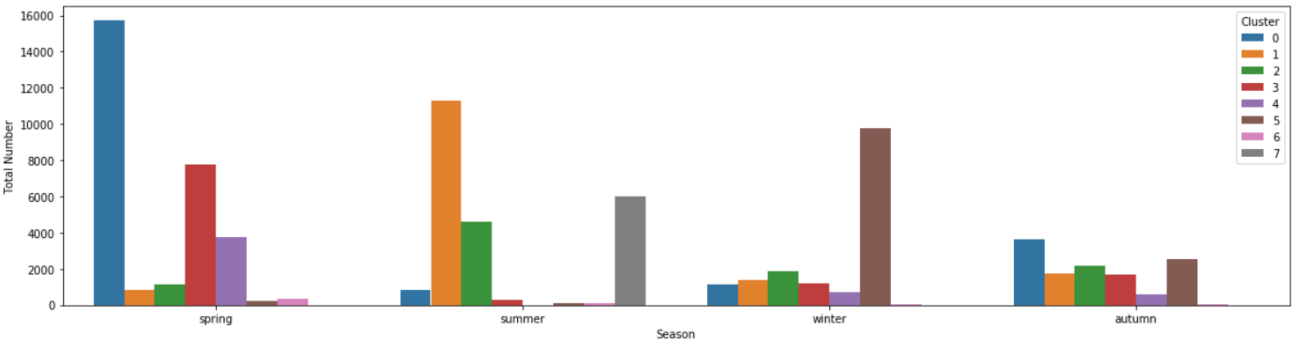
There you can easily see that cluster 0 is most active in the night and the morning, as well as cluster 5 whereas cluster 1 and cluster 3 is more active in the afternoon and cluster 2 early in the morning.

Next one is about Total Numbers of activities vs. Temperature



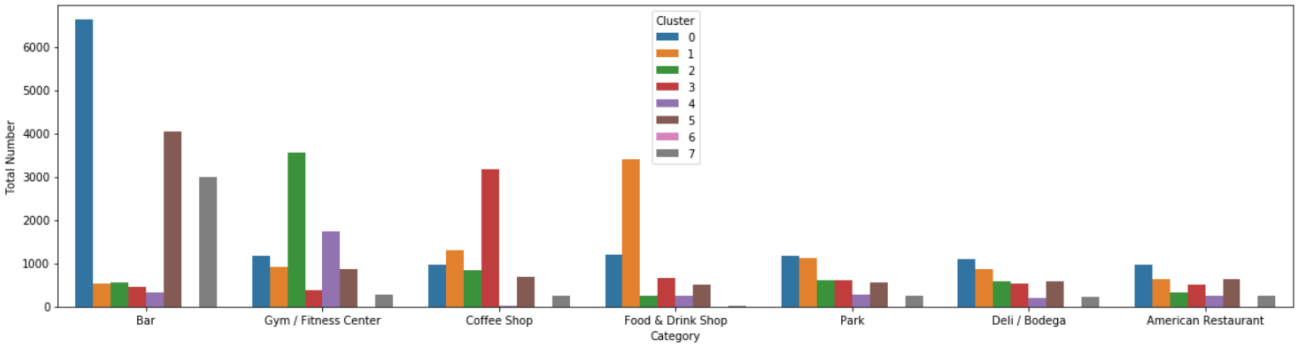
For the temperature you can see that cluster 0 likes middle to high temperature, cluster 1 likes mostly higher tempratures and cluster 5 is mostly active for if the temperature is lower.

Next one is about Total Numbers of activities vs. Season



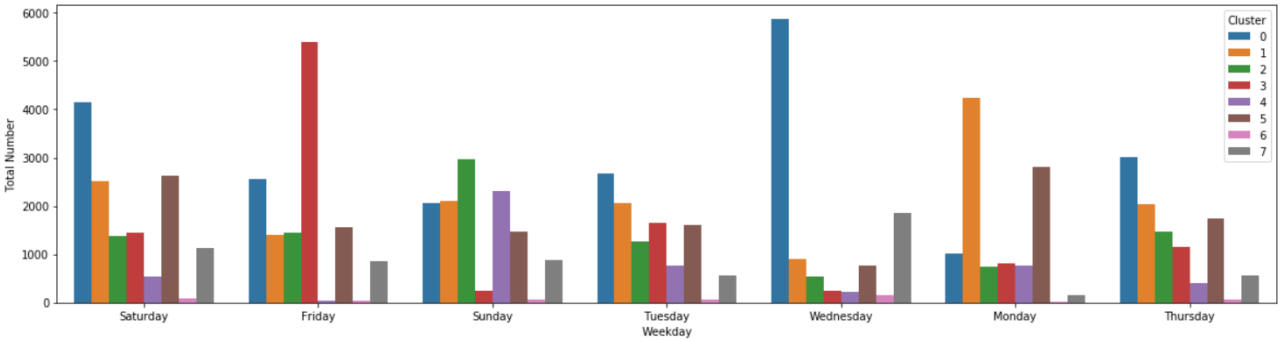
The activity per temperature can also be found in the overview of the activities per season.

Next one is about Total Numbers of activities vs. Category



If we have a look at the different categories used by the clusters you can see that cluster 0 is mostly in bars whereas cluster 2 is mostly going to the gym, cluster 3 can mostly be met in Coffee shops and cluster 1 in Food and Drink stores.

Next one is about Total Numbers of activities vs. Day of the Week.



The comparison of activities and weekdays is a little bit surprising. As cluster 0 is mostly going to bars, Saturday show a lot of activities but the most activities were recorded for Wednesday. Cluster 3 is mostly active on Friday and cluster 1 on Monday.

5 Prediction

The next section is about predicting activities based on the features given to a specific algorithm. The algorithm e.g. could be the Decision Tree, the SVM or the KNN algorithm. In the following section all three were tested and the respective performance was measured.

5.1 Decision Tree Algorithm

Before the decision tree algorithm could be used the used values were reduced to temperature, day of the week and time of the day.
After that, the data was transformed using the LabelEncoder. The data was splitted into a training and a test group and the algorithm was trained. After that a prediction was made and the accuracy was predicted, which was not very high.

```
Accuracy:  0.23320513869568776
```

5.2 SVM

As the Decision Tree Algorithm was not working very well I tried the SVM algorithm. There the same data basis as for the decision tree algorithm was used. The process in general was also the same (training, predicting, measuring). Also this result was not very good.

```
0.06000108955388997
```

5.3 KNN

My last try was the KNN algorithm. There the same data basis was used and transformed with the LabelEncoder. After splitting, the algorithm tried 24 attempts to succeed (1 to 25). The accuracy was as good as for the last two algorithms.

```
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
```

6 Conclusion

As a conclusion you can say that the clustering was working quite well.
From the clustering you could create some personae like the following:

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.

However, the prediction did work but not with a very high accuracy.

7 Outlook

For the future the data quality can be further optimized. Furthermore, the biggest task to be done might her to find out, why the prediction is not working well and try to improve by either change/adapt the data quality or find a better algorithm matching the described problem. Nevertheless, the identified clusters show that there are patterns of different groups having different interests and do different activities at different times.