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.