Databases Autumn 2019

Data Analysis Project
P0: Project Idea

The relation between event headlines, tourism and wealth in Europe.

October 6, 2019

Nikolai Rutz, Luc Kury

1 Introduction

Tourism in an important part in today's economics of many countries. A lot of factors influence the choice a travel destination, one of them can be recent events at the chosen location, such as terrorist attacks, environmental disasters, political tension and so on. Do these events have an impact on the overall popularity of the vacation place? Further we want to analyse the influence of economical wealth of a country, on the amount of visiting tourists.

2 Datasets

This section introduces the data sets we would like to use and analyse in the course of this project.

2.1 GDELT 1.0 Event Database

The GDELT event database is an aggregation of news from worldwide media outlets in English. The data is amongst others, categorised by country, region and topic. Event headlines are gathered since March 2013 and are updated every 24 hours. The headlines of each day are saved in a separate

file, which has an average size of about 10 Megabytes (zipped). The challenge with this data set is, that is is split in daily chunks. The first task is going to be to join all the data together in an automatic fashion. Secondly there seem to be a lot of gaps in the the rows for certain attributes. Thirdly there is the challenge of identifying the negatives headlines.

- Source : GDELT, `https://data.gdeltproject.org/events/index.html`
- Size : ∼3.6GB/yr (zipped)
- Format : CSV

2.2 Nights spent at tourist accommodation establishments

The data shows the number of foreign visitors who spent a night at tourist accommodation establishments in the corresponding European country. The first entry in the data set dates back to 1990, the latest entries are from August 2018. The data is structured in monthly visitor count per country. This data will hopefully give us a good indication of how tourism fluctuated over a certain time period.

- Source : Eurostat, `https://ec.europa.eu/eurostat/web/products-datasets/product?code=tour_occ_nim`
- Size : 634.0kB (zipped)
- Format : XLS, CSV, TSV

2.3 Mean and median income by broad group of citizenship

This dataset contains average income values for European citizens, collected over the last ten years. The data observes only people who are 18 years or over and does not differentiate between gender, age class and household type. The mean and median were calculated for employed and unemployed people alike. We well use this data set as an indicator of how wealthy the average citizen in an European country is, and by extent how expensive it is to go on vacation there.

- Source : Eurostat, `https://ec.europa.eu/eurostat/web/products-datasets/product?code=ilc_di15`
- Size : 17.4kB (zipped)
- Format : XLS, CSV, TSV

3 Analysis Goals

The GDELT database servers as the starting point for the analysis. The entries are filtered for negative headlines associated with terrorism, violence, environmental disasters and so on. After that the filtered results are categorised by their geographical location. Due the the nature of the other two data sets, the categorisation will only consider European countries over the interval of a month. The overlapping time frame of our data sets are the years 2013 to 2019, therefore the evaluation can only be done in this time frame. The next step is to correlate the change in the amount of tourists to the overall amount of negative headlines around that time. Finally we want to correlate the average wealth of a country to the amount of touristic travel a country receives.
We expect that locations with many negative headlines will suffer from a drop in tourism after a certain amount of time. Further we think that wealthier countries generally receive less tourists, since it will be more expensive for the majority of the other countries to travel.

4 Tools

Since our data sets are only available as text files, We will use the Python programming language and the Pandas library to reconstruct the data sets into a programmatic form. To store the actual data, we will use a relational database, which implements SQL, like PostgreSQL. The results of our analysis will be visualised with matplotlib in a series of different plots. We would also like to use Jupyter Notebooks to explore the data and at the same time document the process.