Project report (Group 10)

The relation between event headlines, tourism and wealth in Europe

N. Rutz and L. Kury

University of Basel
Databases lecture (cs244)
Autumn Semester 2019

1 Introduction

In the data analysis project, the concepts from the database lecture are applied to an end-to-end scenario. In this hands-on exercise, you are required to perform the most significant steps in integrating multiple data sources to one unified data store which is later on used for analysis purposes. The project is divided in multiple steps: [2]

- Project idea
- Schema integration
- Data integration
- Analysis and visualization

Every step receives its own section in this report, where we will further describe the tasks and challenges faced during this project. In the first two steps ("P0: Project idea" and "P1: Schema integration") we already produced a report. In these section we will only highlight the differences that revealed itself further along the project or cite ourselves.

2 Project idea

Tourism in an important economic branch for many countries. A lot of factors influence the choice a travelers 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. [1]

3 Data sources

This section contains the ER diagrams of the single data sources representing their as-is state in the following order: GDELT, Tourism and Income. Additionally we will highlight the most important attributes and explain their meaning for each data set.

3.1 GDELT 1.0 event database

The GDELT event database is an aggregation of news from worldwide media outlets in English. Events are gathered since March 2013 and are updated every 24 hours. The data of each day is saved in a separate file, which has an average size of about 10 Megabytes (zipped).
The raw data consists of 58 attributes. Among those for example are `sourceurl, dataadded, isrootevent, actorname, ...` and many more. Every event is stored with a `globaleventid` which uniquely identifies an entry in the data set (we discuss the use of it in chapter 4). Some of the most important attributes for our analysis are the month the event occurred (`monthyear`), the type of event (`eventbasecode`) and where the event took place (`adm1code`).

- Source : GDELT, `https://data.gdeltproject.org/events/index.html`
- Size : ~3.6GB/yr (zipped), 63.0GB (3yrs, after integration)
- Format : tab separated values (tsv)

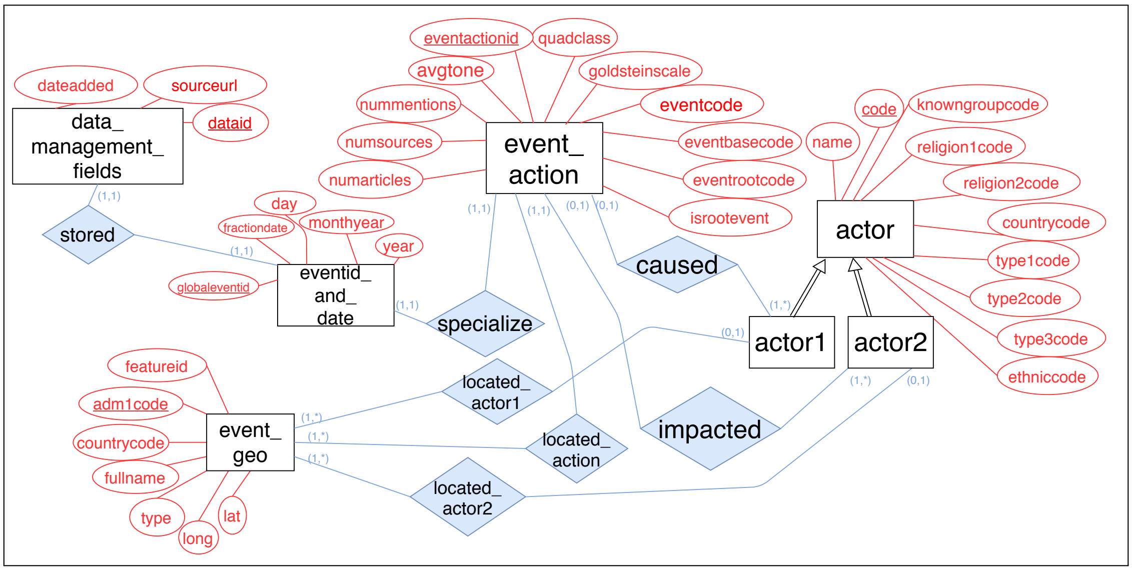


Fig. 1: ERM schema of GDELT as single source data source.

The GDELT data set has a few functions and specifications which cannot be represented in the ER diagram. These are:

- One of the actors can be empty. For some events the `actor1` or `actor2` attribute can be a blank entry for complex or single-actor situations. `actor1` usually represents the acting party and `actor2` represents the affected party.

- The `event_geo` for actors can can be blank, if a headline did not specify the actor country affiliations.
- The `dateadded` attribute of the `data_management_fields` did not exist prior to 2013. This does not affect us, since we do not rely on this attribute to exist.

3.2 Europe tourism dataset

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 tourist data is provided as two different values, **percentage change compared to same period in previous year** and **total number of visitors**. Also the accommodation are split into five different categories (hotels, camping, etc.) for the same timeframe, the timeframe is also available from the data set. Lastly the values are reported as three different entities: **reporting country**, **foreign country** and **total**.

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

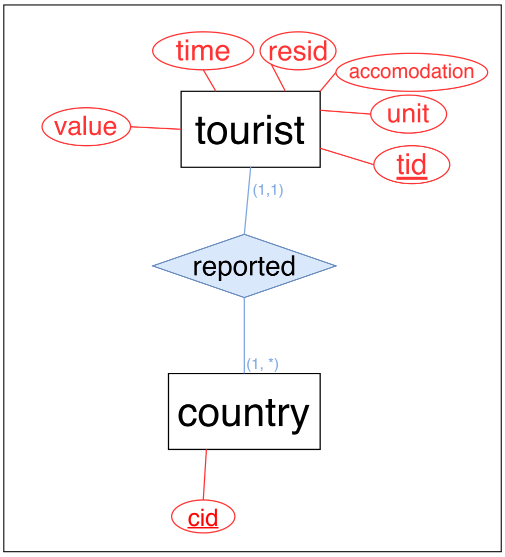


Fig. 2: ERM schema of EuroStat tourism as single data source.

3.3 Europe income dataset

This dataset contains average income values for European citizens, collected over the last ten years. The income is only available as a single value for a whole year. The data observes only people who are 18 years or over and does not differentiate between gender (`sex`), age class (`agegroup`) and household type. The mean and median were calculated for employed and unemployed people alike. The country (`cid`) is the same text string for both the tourism and income data sets.

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

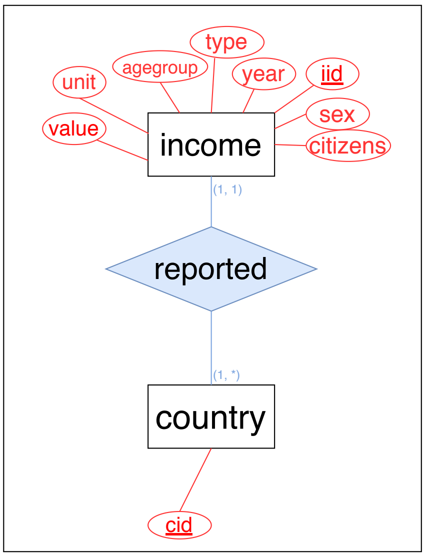


Fig. 3: ERM schema of EuroStat income as single source data.

4 Data integration

For our database backend we opted to use PostgreSQL. We like this solution because it is free and open source software. Instead of installing the database software directly on our machines, we decided to use Docker[1] to make our database

[1] Docker is a set of platform as a service products that use OS-level virtualization to deliver software in packages called containers.

solution platform independent. Furthermore we used Python as our programming language of choice, to handle the raw data files. This includes downloading and extracting the zip and csv files from the GDELT website, reading the files into Pandas dataframes and splitting our data into the entities defined by our ER model (compare figure 4). As a preprocessing step before inserting we used Pandas to drop null values from columns which we later used as primary key and deleted duplicate entries in the `event_geo`, since there were a lot of them (99.9%). The filtered and processed dataframe was then exported to a filebuffer, which is then used to bulk insert the data into the database. We use the `psycopg2` package to interact with the PostgreSQL server. The package provides a wrapper around the SQL `COPY FROM` function. This enables us to insert the 60GB of GDELT data in 5hours on a single thread. The tourist and income data set consisted of 440'000 and 7'680 rows respectively. Because of the small row count, the insertion process was done row by row in a matter of 10 minutes. The only issue with the EuroStat datasets we encountered was that the `value` columns which should have only consisted of doubles (datatype) but sometimes contained a marker text string. We decided that the simplest solution was to manually delete the text strings in these positions with a regular expression before inserting them into the database with Python.

Because we used the bulk insertion method, many of the required constraints for our database weren't established yet. To do so, our data required further cleaning. From this point on we used SQL statements to change our data into the desired state. The last step was to add all the primary and foreign key constraints to the specified attributes. The code used to alter the database tables can be found in the *git* repository[2].

```sql
-- OPERATE ON EVENT_GEO

ALTER TABLE event_geo ADD column id serial;

DELETE FROM event_geo
WHERE id IN
    (SELECT id
    FROM
        (SELECT id,
         ROW_NUMBER() OVER( PARTITION BY adm1code
        ORDER BY id DESC ) AS row_num
        FROM event_geo ) t
        WHERE t.row_num > 1 );

ALTER TABLE event_geo DROP column id;
ALTER TABLE event_geo add primary key (adm1code);
.
```

[2] https://git.scicore.unibas.ch/databases-hs19/gruppe-10/blob/master/schema/alter_database.psql

.

.

After cleaning, normalization, unification and transformation of the data with the steps described as above, the structure of our database is now equal to the integrated ERM diagram (see figure 4).

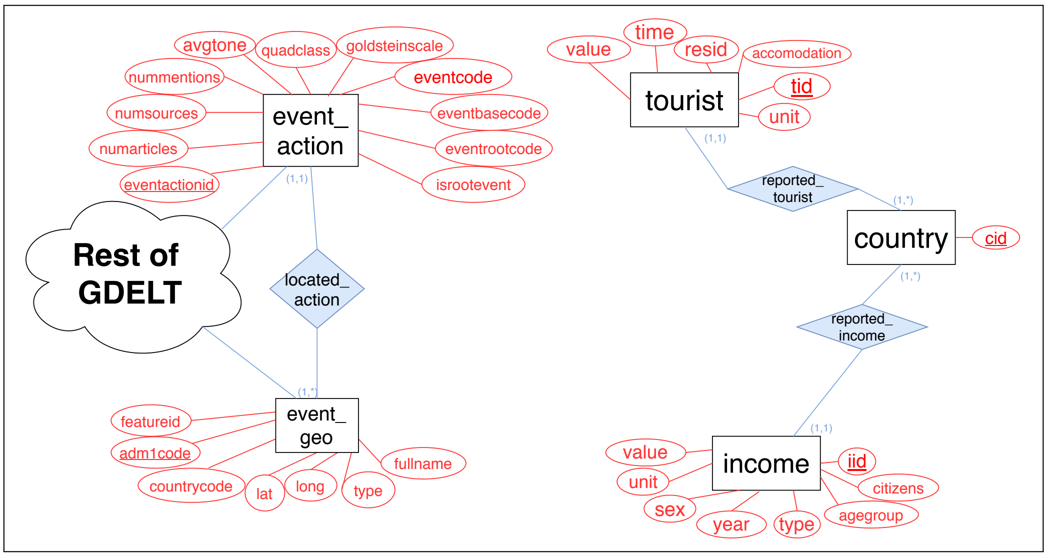


Fig. 4: ERM schema of integrated data sets.

Note that the placeholder "Rest of GDELT" in the figure above refers to the remaining entities and relations shown in figure 1.

Table 1: Overview of database statistics

Table	Data Length	Index Length	Rows
actor1	1,466,368	557,056	19,725
actor2	1,400,832	532,480	18,903
country	8,192	24,576	44
data_management_fields	28,890,382,336	4,630,896,640	206,138,210
event_action	22,558,842,880	18,536,636,416	206,168,080
event_geo	466,944	131,072	4,451
eventid_and_date	12,418,596,864	13,899,202,560	206,168,110
income	1,196,032	196,608	7,680
tourist	79,953,920	34,390,016	442,800
10 in total	63,952,314,368	37,102,575,616	

In this state, the database now holds 63'952'314'368 bytes of data. After running some simple test queries against our database, we noticed that they take a long time to complete. Therefore we decided to add indexes for attributes often used in queries, besides just the primary keys. The additional indexes (including primary keys) let the size requirement for the database grow for another 37'102'575'616 bytes. The database including all indexes now consumes 94.1GiB on disk. Dumping the database in this state with standard compression enabled, yields a 8.74GiB file.

4.1 Relational schema

With the integrated ER model described above, we deduct the relational schema for our database. We merged all the relations where possible and defined the necessary primary and foreign keys.

- `eventid_and_date`(globaleventid, fractiondate, day, monthyear, year)
- `data_management_fields`(dataid, globaleventid, sourceurl, dateadded)
- `actor`(code,name, knowngroupcode, religion1code, religion2code, countrycode, type1code, type2code, type3code, ethniccode)
- `actor1`(code, adm1code, name, knowngroupcode, religion1code, religion2code, countrycode, type1code, type2code, type3code, ethniccode)
- `actor2`(code, adm1code, name, knowngroupcode, religion1code, religion2code, countrycode, type1code, type2code, type3code, ethniccode)
- `event_action`(eventactionid,actor1code, actor2code, adm1ode, globaleventid, eventcode, eventbasecode, eventrootcode, isrootevent, goldsteinscale, quadclass, avgtone, nummentions, numsources, numarticles)
- `event_geo`(adm1ode, featureid, countrycode, lat, long, type, fullname)
- `income`(iid, cid, value, citizens, unit, agegroup, type, year, sex)
- `tourist`(tid, cid, value, time, resid, accomodation, unit)
- `country`(cid)

5 Analysis and visualization

The GDELT database serves as the starting point for the analysis. The entries are filtered for negative headlines associated with terrorism, violence and so on. Afterwards 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 multiple month (tourism) or a whole year (income). The overlapping time frame of our data sets are the years 2013 to 2019, but we specifically look at the events between 2015 and 2017. 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. To visualize the query results we used Tableau because of its powerful feature sets and ease of use.

5.1 Changes in tourists for a specific country due to negative headlines

The first step in this correlation is to get the amount of negative headlines for a specific country. To identify negative headlines we use the `eventbasecode` attribute from the `event_action` table, which assigns a numerical code to the recorded event. An extension of the `eventbasecode` attribute is the `eventcode` attrribute, which further specialized the type of event. For example an event marked with the `eventbasecode` 13, would be categorize with the keyword *THREATEN*. The `eventcode` 137 specializes the *THREATEN* event as a *Threaten with violent repression* event. For our purposes the attribute `eventbasecode` suffices to extract the relevant headlines. In detail we consider events marked with the following `eventbasecode` to be negative headlines:

Table 2: Explanation of evenbasecodes

eventbasecode	Description
13	THREATEN
14	PROTEST
15	EXHIBIT FORCE POSTURE
17	COERCE
18	ASSAULT
19	FIGHT

The time frame for this analysis was limited with the `monthyear` attribute to an interval of three years starting January 2015 and ending January 2018. This attribute is located in the `eventid_and_date` table. `actor2` references all attributes for the affected party of the events, including the location described in the `adm1code`.

Together with all that information we can pull out the amount of negative headlines of a specific region to a defined time frame. The SQL query for this extraction looks as follows:

Listing 1 SQL query for negative headline count for a specific country

```sql
SELECT monthyear, COUNT (*) as occurances, eventbasecode
FROM event_action, eventid_and_date, actor2
WHERE eventid_and_date.monthyear >= 201501
AND eventid_and_date.monthyear <= 201801
AND event_action.globaleventid = eventid_and_date.globaleventid
AND event_action.eventbasecode IN ('13', '14', '15', '17',  '18', '19')
AND actor2.adm1code = <Parameter.adm1code>
AND actor2.code = event_action.actor2code
GROUP BY eventid_and_date.monthyear, event_action.eventbasecode
ORDER BY eventid_and_date.monthyear ASC
```

For the `tourist` table we extracted the sum of tourists for a specific country, which is indicated through the `cid` attribute.The chosen time frame is extracted through the `time` attribute. We only considered values that were indicated as numbers through the `unit` attribute. Furthermore the values are only referring to tourists in the reporting country stated by the `resid` attribute. We implemented following SQL query for this extraction:

Listing 2 SQL query for tourist count for a specific country

```sql
SELECT SUM(value) as tourists, time
FROM tourist
WHERE tourist.unit = 'Number'
AND tourist.resid = 'Reporting country'
AND tourist.cid = <Parameter.cid>
GROUP BY tourist.time
```

We did a `LEFT JOIN`-operation with Tableau joining the SQL queries above by the attributes `time` and `monthyear`. Note that the `<Parameter>` values in the queries can be dynamically adjusted in the Tableau file provided in the *git* repository. The joined operation was visualized with a graph plot in Tableau.

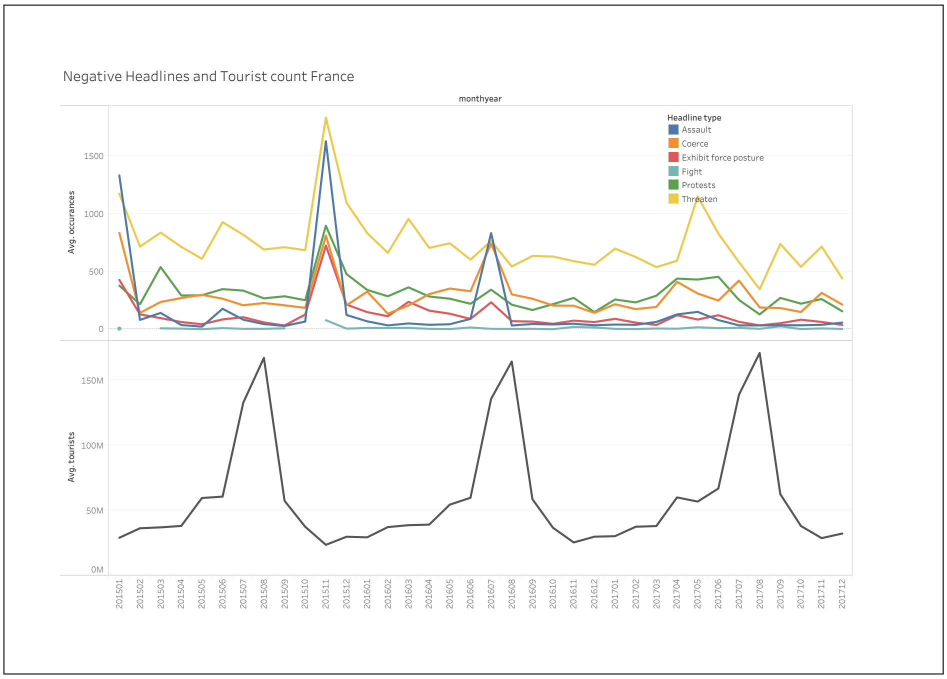


Fig. 5: Evaluation for changes in total amount of tourists in France due to negative headlines.

In this plot we can see the occurrence of negative headlines in the upper half and the count of tourist for the country in the lower half. The x-axis describes the time in months and the y-axis describes the number of tourist and the number of headlines for the corresponding categorization respectively. The different categorizations are indicated by the color in the legend.

We can clearly see that the amount of negative headlines peaked in some points over the chosen time frame. More specific there are three key times that the headline count reached its maximum. The first time is in January 2015. This corresponds to the time where the Charlie Hebdo incident occurred. The second peak in November 2015 during the assassination in Paris. The last rise in July 2016 indicates general negative headlines towards the european soccer championship.

The tourism count has regular surge during the summer months. It can be observed that prior to the summer months there is a steady increase in tourist numbers. After September there is a sharp decrease of foreign visitors.

There is no obvious connection between negative headline count and the desire to travel. Despite the amount of negative headline peaks the tourism seems to be unaffected by it. The maxima and minima over three years are constant. The tourism is much more affected by seasons rather then negative events. This can also be observed for other countries (compare figure 8, 9, 10, 11 in the appendix).

5.2 Changes in tourism European Union

In this section we looked at the overall change in tourism in the European Union for a specific time. After analyzing the data in section 5.1 we extracted different time frames were the amount of negative headlines in a specific country peaked the most. We looked up the most significant events:

Table 3: Overview of key events time frame

Start Date	End Date	Event
05/2015	08/2015	Bailout Crisis, Greece
11/2015	02/2016	Assassination Paris, France
06/2016	09/2016	Brexit, UK
09/2017	12/2017	Referendum Kurds, Turkey

The query for the amount of negative headlines is almost the same as in section 5.1. The difference is that the overall combined amount of negative headlines was extracted and grouped by a month. The adm1code of the actor2 table isn't needed anymore:

Listing 3 SQL query for negative headline count

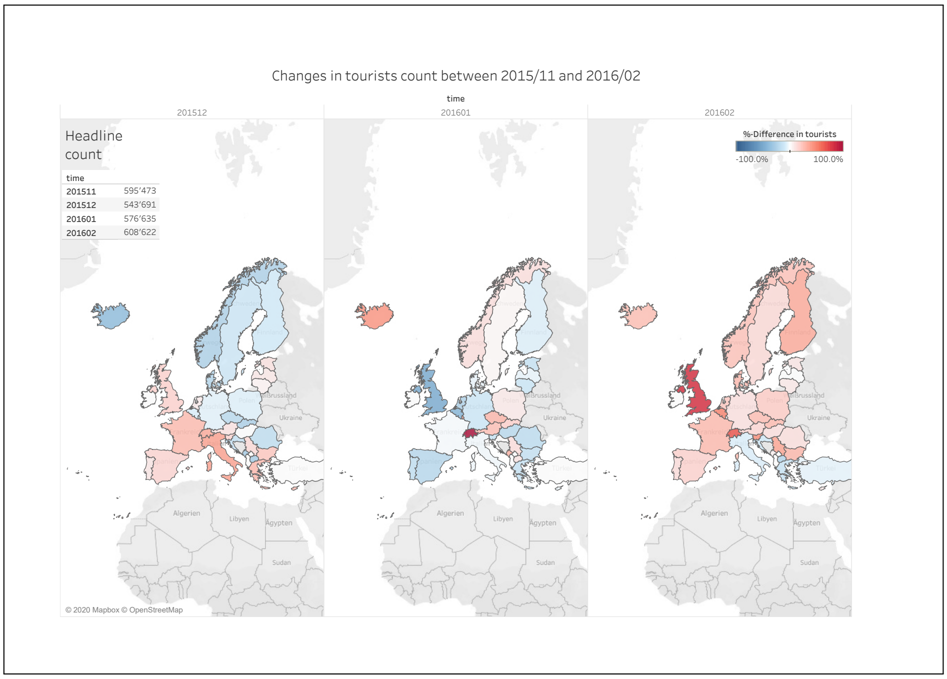
```
SELECT COUNT (*) as occurances, monthyear
FROM eventid_and_date, event_action
WHERE event_action.eventbasecode IN ('13', '14', '15', '17', '18', '19')
AND monthyear <= <Parameter.month_max>
AND monthyear >= <Parameter.month_min>
AND event_action.globaleventid = eventid_and_date.globaleventid
GROUP BY monthyear
```

For the tourist table we extracted the sum of tourists grouped by a specific country in a chosen time frame, that is extracted through the time attribute. Again we only considered values that were indicated as numbers through the unit attribute. Furthermore the values are only referring to tourists in the reporting country stated by the resid attribute:

Listing 4 SQL query for tourist count

```sql
SELECT SUM(value) as tourists, cid, time
FROM tourist
WHERE tourist.time >= <Parameter.month_min>
AND tourist.time <= <Parameter.month_max>
AND tourist.unit = 'Number'
AND tourist.resid = 'Reporting country'
GROUP BY tourist.cid, tourist.time
```

We did a `LEFT JOIN`-operation with Tableau joining the SQL queries above by the attributes `time` and `monthyear`. Note that the `<Parameter>` values in the queries can be dynamically adjusted in the Tableau file provided in the *git* repository. The joined operation was visualized with a graph plot in Tableau.



Changes in tourists count between 2015/11 and 2016/02

Headline count

time	
201511	595'473
201512	543'691
201601	576'635
201602	608'622

Fig. 6: Changes in tourists count between 2015/11 and 2016/02.

We consulted the table 3 to generate four different map plots. In figure 6 we can see three different states of the map of the European Union. The colors represent the severeness of the changes in tourism (percentage difference compared to previous month). The color range is explained in the legend in the top right corner.

The headline count is displayed as a table in the top left corner and indicates the sum of negative headlines for the specific time.

We interpret this map in the context of the assassination in Paris. Event though the headline count is at its highest, the tourism changes is very little in both directions. As we saw in section 5.2 the change in tourism stagnates towards cold seasons. The opposite effect can be observed in figure 13. Despite the negative headline count (in conjunction with the Greece bailout events) a massive increase in touristic activity can be observed in July 2015 all over Europe. Towards the end of summer the countries that are further north, experience the decline in tourism sooner, because of the earlier onset of winter. In conclusion we can state that geopolitical events associated to negative headlines do not influence the tourism in Europe. Seasonal effects are far more dominant.

5.3 Influence of wealth on tourism in the European Union

Last but least we analyzed the correlation between the income of a country over three years, starting January 2015 and ending January 2018 with the tourist count per year in the corresponding country. Again we only considered values that were indicated as numbers through the `unit` attribute. Furthermore the values are only referring to tourists in the reporting country stated by the `resid` attribute. The SQL query for the extraction of the tourist count grouped by the corresponding country per year looks as follows:

Listing 5 SQL query for tourist count per year grouped by country.

```sql
SELECT SUM(value) as tourists, left(tourist.time::text, 4) as year, cid
FROM tourist
WHERE tourist.unit = 'Number'
AND tourist.resid = 'Reporting country'
GROUP BY year, cid
ORDER BY year ASC
```

To receive the per year income of the `income` table, the attribute `citizens` needs to be indicated as a reporting country and the attribute `type` specifies the mean equalised net income. The `value` and the `cid` attribute are selected with following SQL query:

Listing 6 SQL query for income per year grouped by country

```
SELECT year::text, value as income, cid
FROM income
WHERE citizens = 'Reporting country'
AND type = 'Mean equivalised net income'
ORDER BY year ASC
```

We did a `LEFT JOIN`-operation with Tableau joining the SQL queries above by the attributes `cid` of both tables. The joined operation was visualized with a graph plot in Tableau.

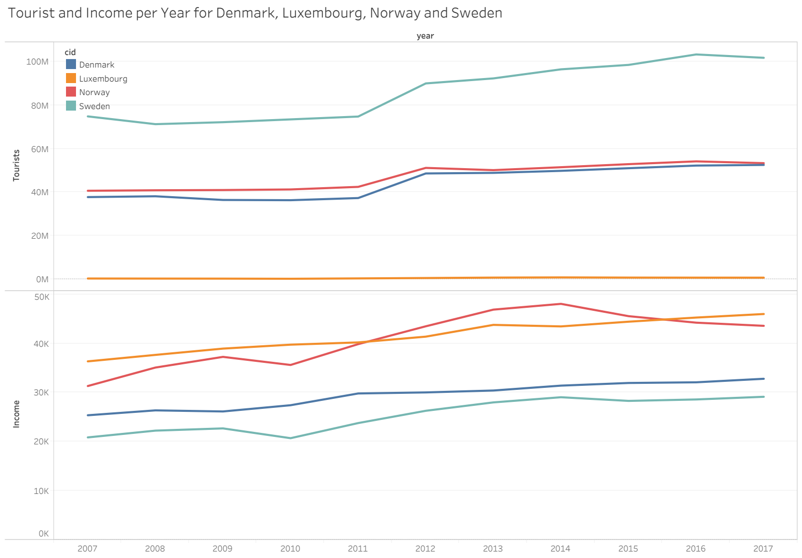


Fig. 7: Correlation for tourist and income per year for Denmark, Luxembourg, Norway and Sweden.

In this plot we can see the development of the tourist count in the upper and the income change in the lower part of chosen countries. The x-axis indicates time in years between 2007 and 2017 and the y-axis represents the total amount of income and tourism respectively. The different colors mark the chosen countries and are explained in the legend in the top left corner. The chosen countries are located at different spectra of the overall tourism and income data. For example figure 7 depicts countries with the highest income, while figure 16 shows countries with the highest tourism.

For touristic activity in a country we can observe, that a positive trend in income also brings a positive development in tourism. In figure 7 we can observer that prior to a jump in tourist count in the years 2011/2012, the income also steadily increased. The same effect applies to other countries shown in figures 15 and 16. This doesn't hold true for Luxembourg, but its overall tourism count was always very low to begin with. For Greece the overall effect behaves in the same way, but around the year 2009 a sharp increase and then decline happened to the tourism. During the same time period the income only rose slightly and surprisingly decreased over the next years.

5.4 Conclusion

The correlation between negative headlines and tourism count is weak to non existent as we concluded in sections 5.1 and 5.2. The income on the other hand has a direct impact on the tourism. An increase in income results in growth of tourism in the corresponding country.
Indirectly we observed another correlation in regard to tourism, time. The current season dictates the development of the tourism in Europe. Counterintuitively to us the increasing wealth of a country (here denoted by the average income) does not ward off people to travel to this country.

6 Lessons Learned

Overall we were able to realize the the project with the provided guidelines. We felt that the progress was steady, but some of the tasks came with their own challenges for us. Overcoming those problems felt very rewarding and the experience gained in dealing with databases can surely be applied to future projects. If we were to do this project again, there are a few key points that we would do differently:

- Analyze the structure of the data sets carefully. At best, choose data sets that have a good quality to begin with. For example the provided documentation can be incomplete or wrong.
- Carefully decide on the conceptual schema for the integrated database.
- Be aware of bottlenecks when inserting data into the database. Do not insert data row by row after cleaning it. This results in long run-time for the insertion process.
- Slow queries can be sped up by indexing often used attributes.
- Use existing and proven tools, especially when it comes to the visualization.
- Everything takes longer than expected, especially when dealing with big data sets.

A Appendix

A.1 Additional visualizations

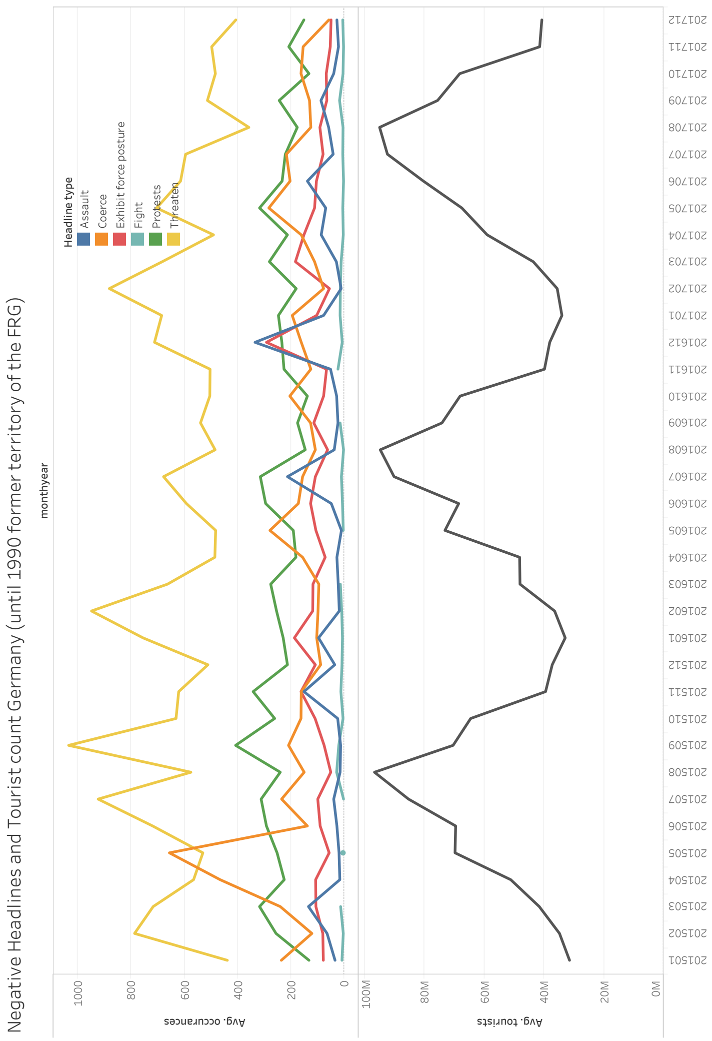


Fig. 8: Total amount of tourists and negative headlines in Germany.

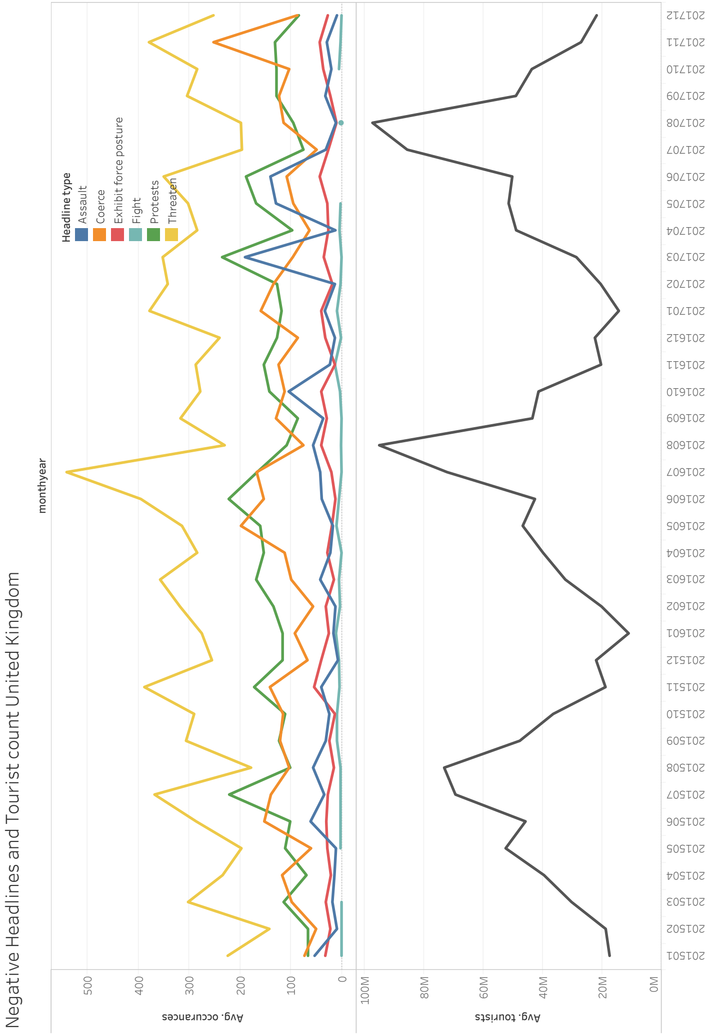


Fig. 9: Total amount of tourists and negative headlines in the United Kingdom.

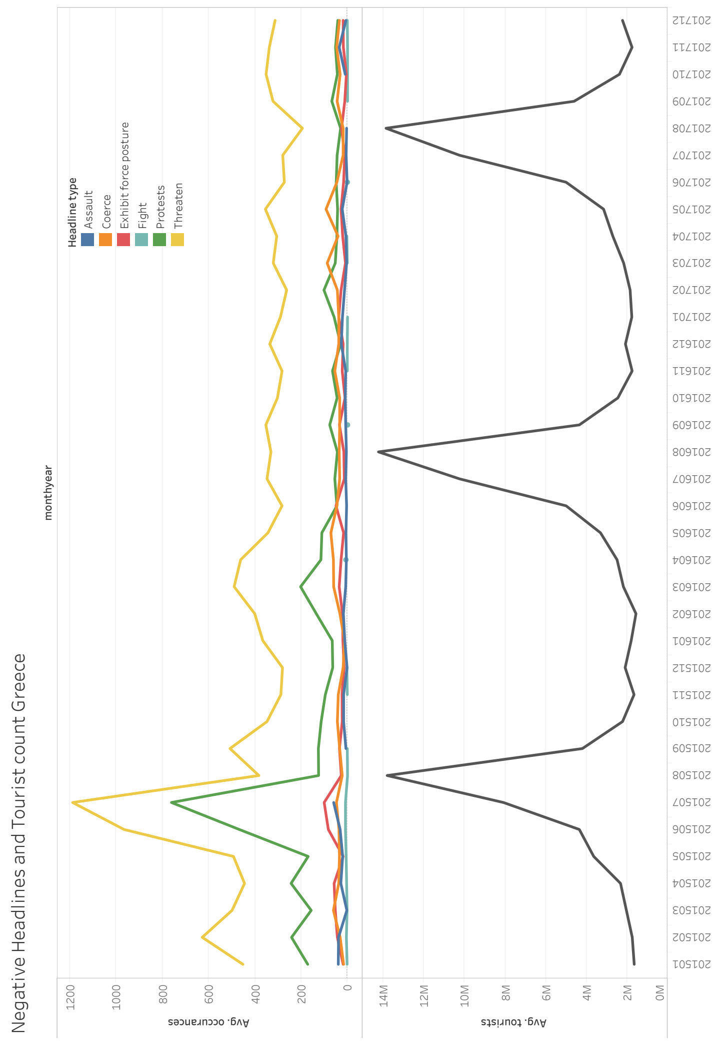


Fig. 10: Total amount of tourists and negative headlines in Greece.

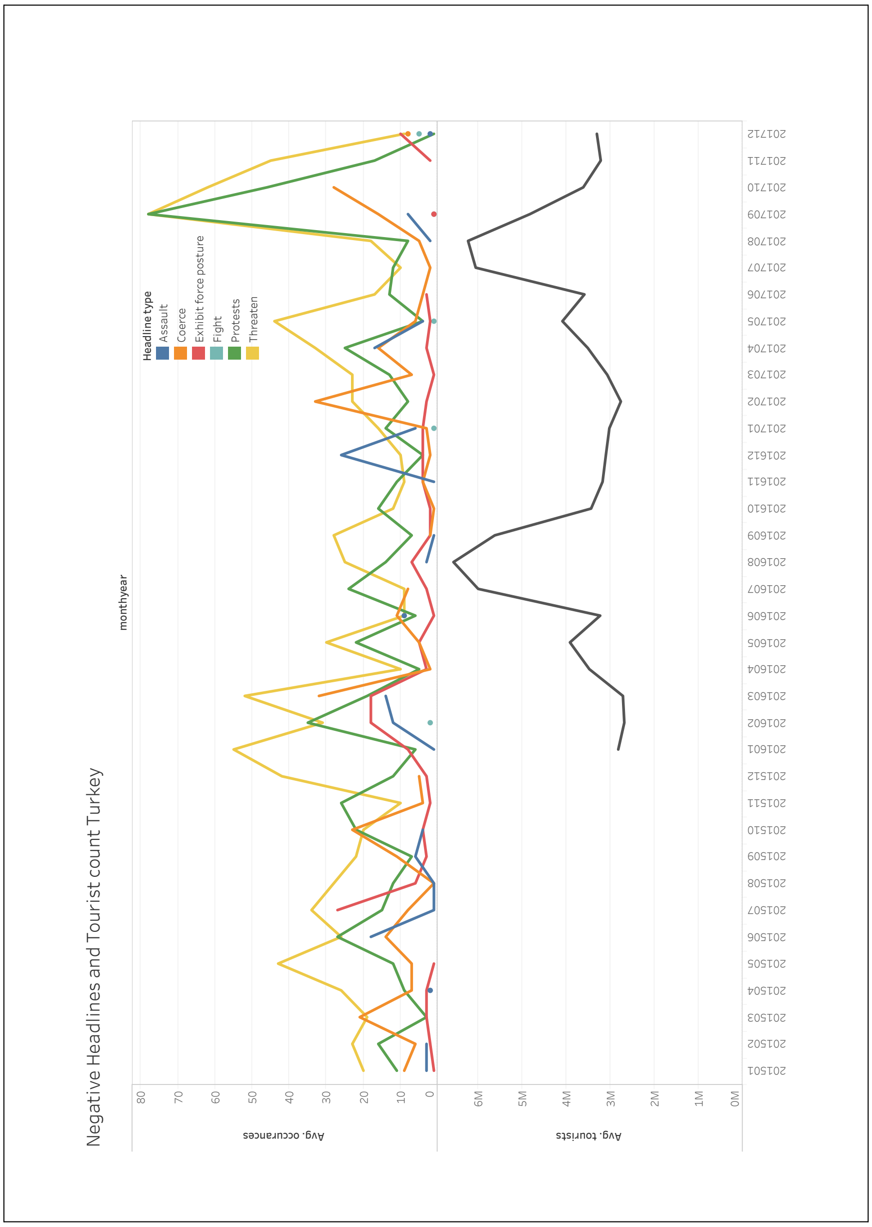


Fig. 11: Total amount of tourists and negative headlines in Turkey.

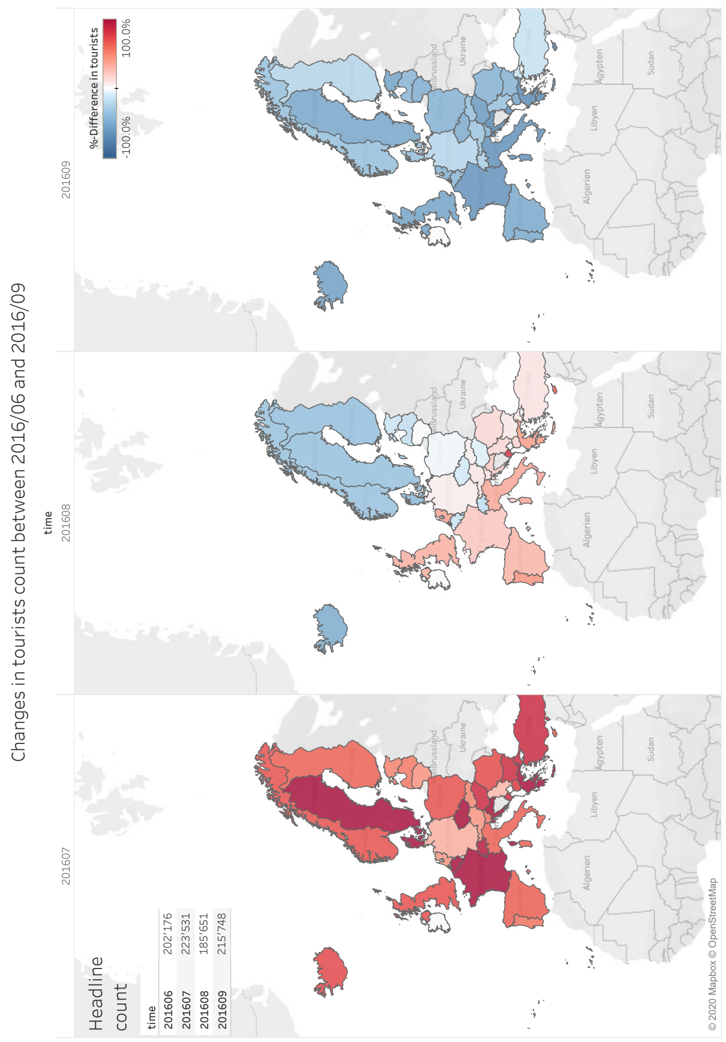
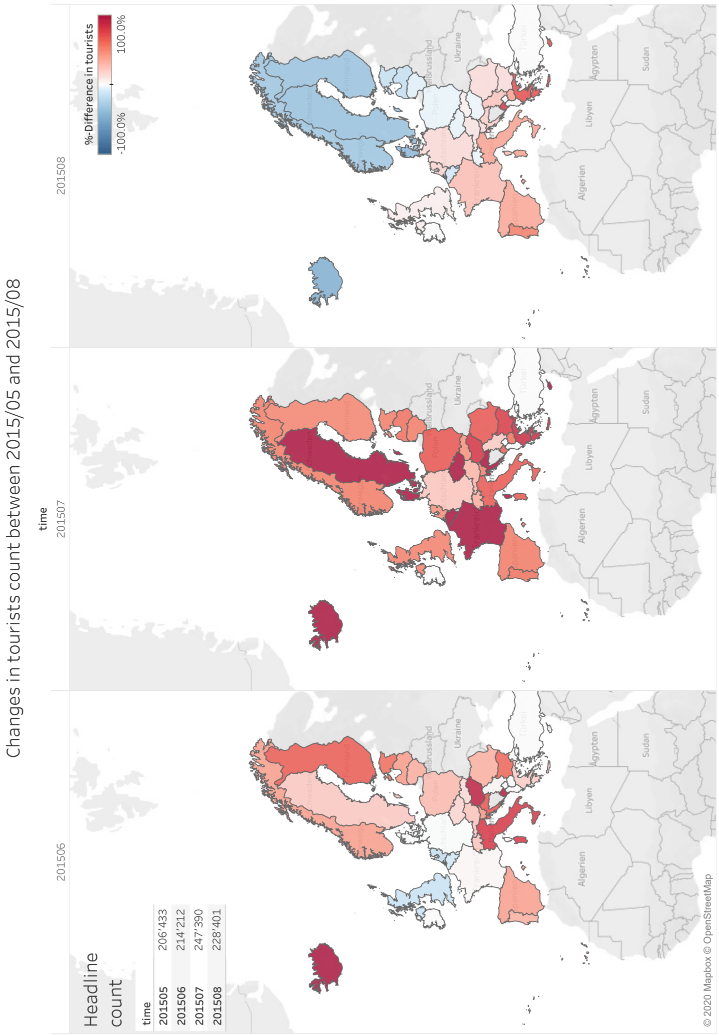


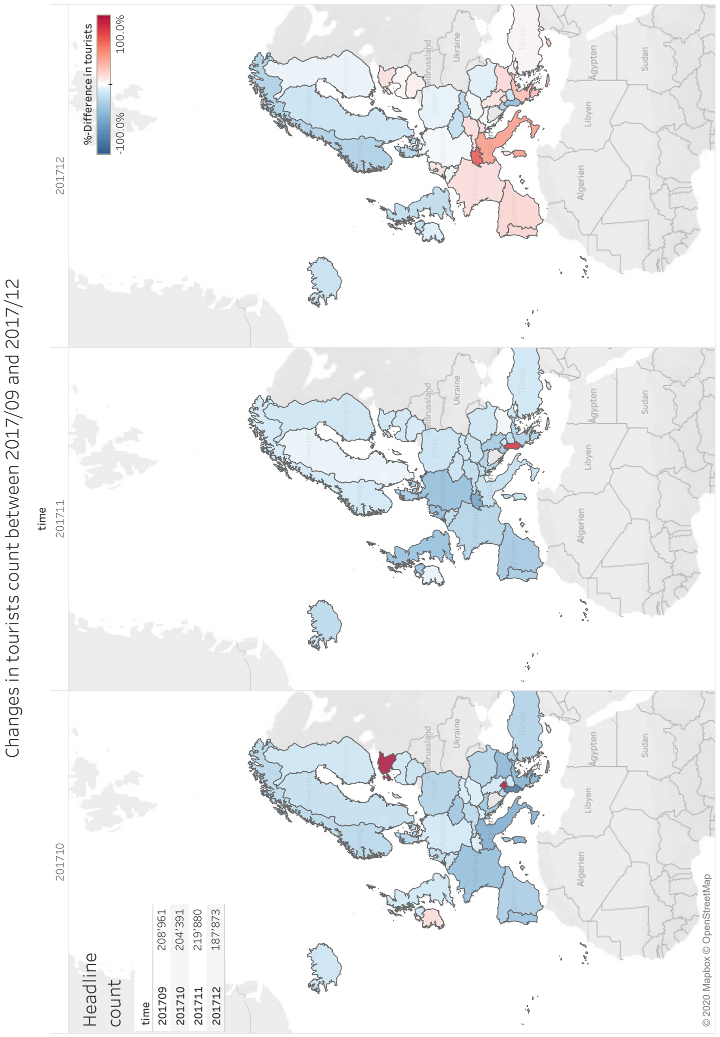
Fig. 12: Changes in tourists count between 2016/06 and 2016/09 (Brexit).



Headline count

time	
201505	206'433
201506	214'212
201507	247'390
201508	228'401

Fig. 13: Changes in tourists count between 2015/05 and 2015/08 (Bailout Greece).



Headline count

time	
201709	208'961
201710	204'391
201711	219'880
201712	187'873

Fig. 14: Changes in tourists count between 2017/09 and 2017/12 (Referendum Kurds).

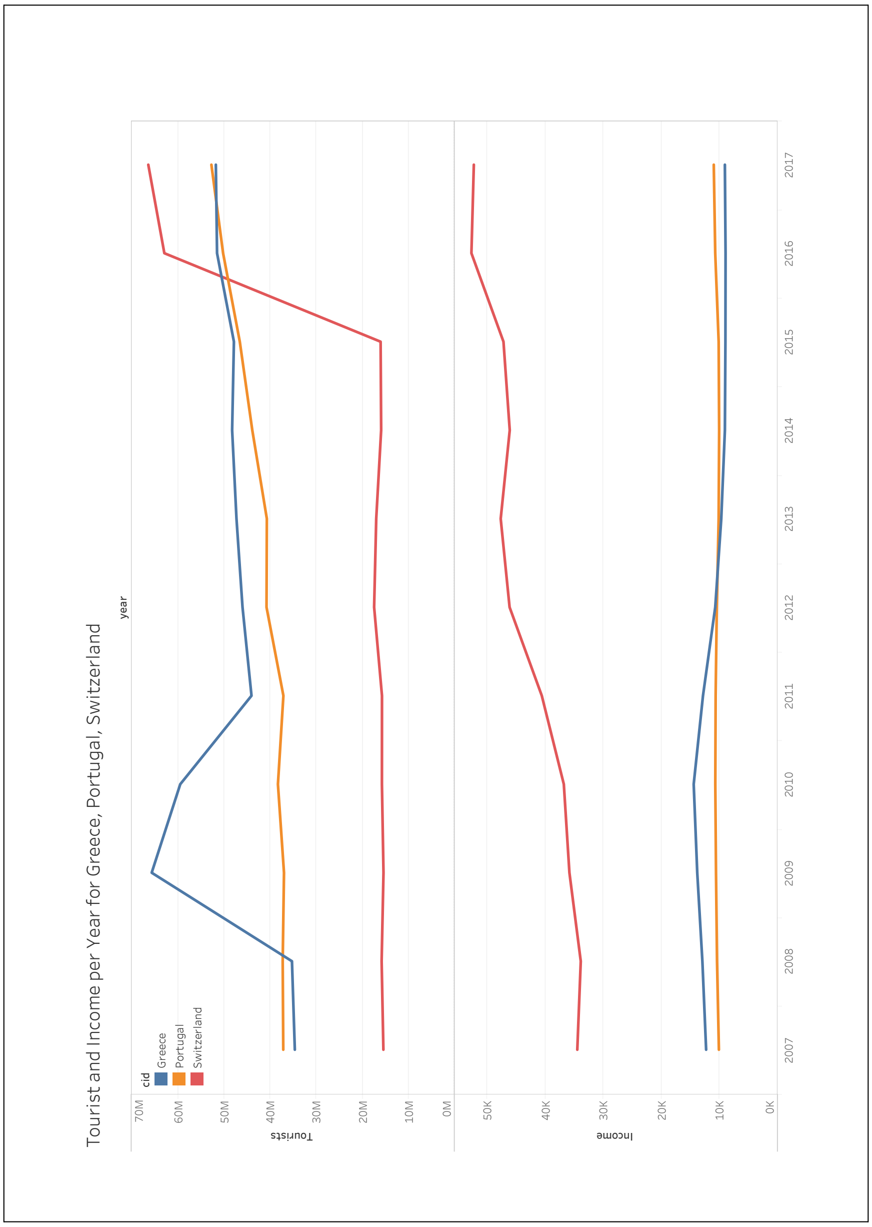


Fig. 15: Correlation for tourist and income per year for Switzerland, Portugal and Greece.

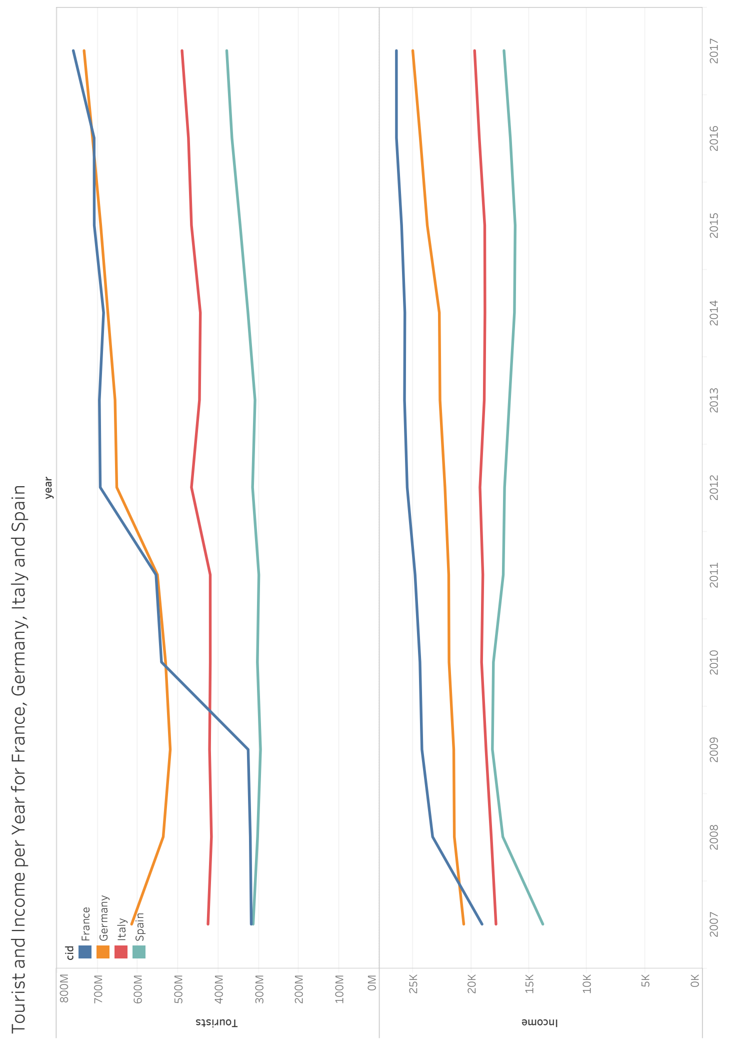


Fig. 16: Correlation for tourist and income per year for France, Italy, Germany and Spain.