A novel application of big data to measure trends in tourism

Data Manual and Description

Sara Mitchell and Karol J. Borowiecki

Last updated: March 11, 2024

Abstract:

This data manual presents data collected from Tripadvisor as part of the project TOURCO. The collected variables are presented and summarised, and the data is validated through a simple comparison with official statistics.

Recommended citations:

* Sara Mitchell and Karol J. Borowiecki (2024). *A novel application of big data to measure trends in tourism.* Data Manual and Description. Mobile Lives Forum (MLF).

This data manual is part of a research project whose findings are summarised in a report and a paper:

* Borowiecki, Karol J. and Pedersen, Maja U. (2024) *A novel application of big data to measure trends in tourism: France, Spain and Denmark, January 2016 - March 2022.* Report. Mobile Lives Forum (MLF).

* Borowiecki, Karol J., Pedersen, Maja U. and Mitchell, Sara B. (2023). *Using big data to measure cultural tourism in Europe with unprecedented precision.* Discussion Papers on Economics, Working paper No 5/2023, University of Southern Denmark.

Acknowledgements: The project is funded and scientifically supervised by the Mobile Lives Forum, as part of its research program on the mobility transition. The Mobile Lives Forum is a research institute created by SNCF.

Contents

1 Introduction 2

2 The data 4

3 Descriptive statistics 8

4 Validating Tripadvisor data 9

5 References 12

6 Tables 13

7 Figures 18

1 Introduction

Official tourism statistics is notorious for being 1) over-aggregated and available usually at the country level (at the best, it is available at the region level), 2) lacking information about the tourist (at the best, the data records whether the tourist is domestic or foreign), 3) available with a lag of at least several months, and often 4) available only at the annual level.

In response to this, we suggest a unique and unorthodox approach that is computer-science driven and relies on big data collected from a leading travel portal (Tripadvisor). The novel approach enables us to obtain a systematic, reliable and consistent approximation for tourism flows, and this with unprecedented precision, frequency, and depth of information. In comparison with mainstream tourism statistics, our approach delivers 1) information on tourism flows at the attraction-level (not country-level), 2) detailed information about the tourist, including from what location (city) she comes from, how she travelled (solo, business, family, etc.), and the travel history for several previous years, 3) information as good as in real-time, and 4) at a daily frequency.

Ultimately, the aim of the project is to provide a comprehensive overview of the patterns and changes in travel flows in selected European destinations. With the data at hand, we will be able to construct unique and real-time databases for a period before the onset of Covid-19, for the months during the pandemic and lockdowns, as well as after the gradual re-opening of the society in a post-Covid-19 Europe. Particularly, we are seeking to look into questions whether tourism activity decreased during the pandemic, whether travelling distances changed due to the pandemic, and whether the pandemic pushed tourism flows to the periphery. In addition to this, the data helps identifying the factors of tourist attractiveness at specific, disagregated locations using state-of-the-art econometric methods.

This document provides a detailed description of the data collection process, an overview of the data collected from Tripadvisor, and validation tests comparing the trends in

Tripadvisor data to those in official statistics.

The data presented in this manual is also presented in Borowiecki et al. (2023); Borowiecki and Pedersen (2024) where the data it is described, validated and analyzed. The paper both presents a visual and a formal validation of the data, before it is used to make inferences.

Section 2 provides a non-technical overview of the data collection process, how the data is organised, the variables contained in each data entity, and how new variables were generated. Section 3 provides summary statistics, charts, and maps for the Tripadvisor data for France, Spain, Denmark, and the total. Section 4 provides validation tests that compare various trends in tourist activity in the Tripadvisor data to those in official statistics (e.g., Eurostat). Please contact the authors for a technical manual for the data scraping process, including detailed description of the scraping process, the structure of the code, the approach used to scrape, challenges faced in the scraping process, and examples of the code.

2 The data

We used a purpose-built Python web scraping program to collect data from Tripadvisor.com on tourist attractions and reviews for Denmark, France, and Spain. We divided the data into four data entities: list of attractions, attraction reviews, user profiles, and user travel history.

The data for the attraction module was initially collected by computer scientist Anne Møller Madsen from end of January to start of March 2022. The data for the review module was initially collected from end of January to start of the start of April for reviews written in English and in the local language (e.g., Spanish for attractions in Spain). Reviews in other languages (German, French, Italian, Spanish, Dutch, Swedish, Portuguese, Danish, Russian, Polish, and Czech) were collected for attractions in Denmark, France, and Spain during July 2022. The data for the user profile module was collected in June 2022. The data for the travel history module was collected from the end of June 2022 to the start of July 2022. The review, user profile, and travel history module data was updated and extended by computer scientist Sofus Hesseldahl Laubel in September 2022.

For the **attraction module**, we collected a list of all tourist attractions listed on the respective country's *Things to Do* webpage at the time of data scraping. For each attraction, we collected the attraction name, the URL for the attraction Tripadvisor page, the attraction's Tripadvisor within-country ranking, the overall rating, the number of reviews, the attraction type category, the attraction location, the duration (if provided), the price (if provided), and a brief "About" text describing the attraction. As some attractions have the same name (e.g., "Pl. de la Liberté" appears multiple times in France), we generated a universally unique identifier (UUID) so that each attraction could be uniquely identified within the dataset.

Tripadvisor provides a list of more than 200 attraction types.[1] Attraction types are not

[1] Attractions can be added by general users or businesses via an online form. In this form, users are

mutually exclusive; that is, an attraction can be assigned to multiple attraction type categories. Tripadvisor aggregates the attraction type categories into 20 groups. We follow Tripadvisor's system of aggregation, and we generate indicators for each of the 20 groups.

The **review module** is composed of the reviews for each respective attraction. For each review, we collected data on the attraction universally unique identifier (UUID), the username, the user profile URL, the user-assigned rating, the title of the review, the text of review, the type of tourist (if specified), date the review was published, and the date of experience or visit (if specified). Each username was assigned a UUID, which we use to anonymise the user data and link the review module to other user-related modules. We collect all reviews published between 15/11/2001 and 09/04/2022.

The user profile module contains basic information on the users who wrote at least one review for at least one attraction in our sample of countries. This module includes the user UUID, the user profile URL, the user's self-reported location (if provided), the total number of contributions, and the date the user joined Tripadvisor. Users are not required to provide this data on any specific level of territorial aggregation. Users may indicate, for example, their country, their NUTS2 region, their city, or no location at all.

The **user travel history module** contains data on all reviews written by users included in the user profile module. This travel history module only contains reviews written for attractions (i.e., we exclude reviews written for hotels, restaurants, etc.) over the period from 15/11/2001 to 09/04/2022. This module contains the user UUID, the name of the attraction, the attraction UUID (if the respective attraction is included in the attraction module), rating, the date the review was published, and the date of experience or visit. In order to assure anonymity of each respective user in our travel history module we aggregate their visits and summarise each user's number of attractions visited, average rating given, average publishing date of review, and average date of experience. The average

asked how the attraction should be categorised, but the attraction type categories can be edited by users after the attraction is added to Tripadvisor (subject to Tripadvisor approval). User are limited to the list of of attraction types provided by Tripadvior.

publishing date of review and average date of experience are calculated by converting each date into a corresponding numerical value, followed by taking the average of these values.

We use *OpenStreetMap* data to identify the latitude and longitude of locations (user and attraction) in the statistical software *R* using the *geocode* function in the *tidygeocoder* package Cambon et al. (2021). This procedure identifies the latitude and longitude for the centre point of the administrative area. A small fraction of locations (<1%) could not be automatically identified using this method. We manually checked and updated the spelling of the remaining unidentified locations using *OpenStreetMap* and repeated the geocoding procedure. Where possible, we manually identified the latitude and longitude using Google Maps.

Out of 6,981,081 users, 4,777,985 users (68.4%) included their location. We were able to identify the latitude and longitude for about 99.9% of those user locations.

We calculated the Vincenty ellipsoidal geodesic distance from attraction location to the user location. These calculations were made in R using the *geodist* package. This method calculates the distance "as the crow flies", taking into account the shape of the earth as an oblate spheroid. The distance is measured in kilometers. This process results in attraction-user location pairs with complete latitude and longitude, as well as the distance from attraction location to user location. We also generate a binary variable equal to one if a user's location is in the same country as the attraction and equal to zero otherwise, which we use to compare trends in domestic vs foreign tourism activity.

Finally, we extract useful information from each review by applying the Linguistic Inquiry and Word Count (LIWC) software. LIWC analyses text and calculates the percentage of words that fall into one or multiple linguistic, psychological and topical categories. In this way, we can analyse objective measures concerning the text reviews (e.g. length of review text, proportion of words with more than 6 characters), but also whether reviews are written in a e.g. joyful, anxious, or angry tone. Additionally, the application of

LIWC helps in anonymizing users while still maintaining important information from their review.

The reviews from Tripadvisor come in multiple languages, and applying LIWC across the languages is not possible since only one dictionary can be used at a time. Therefore, we subset the data according to the languages that LIWC support and analyze each subset using the relevant LIWC directory. Each LIWC directory comes with a different set of psychometric variables and it is therefore not necessarily recommended to compare values across languages.

3 Descriptive statistics

The data entities contain data on 143,387 attractions, 7,915,171 reviews, and 6,981,081 unique users. Figure 1 maps all attractions considered, while Figure 2 maps the reported location of all visitors to the attractions.

We provide some descriptive statistics for attractions that relate to cultural tourism: Museums, Nature & Parks, and Sights & Landmarks, and separately we present descriptive statistics by tourist type.

Although the data includes reviews from as early as 2001, we limit the summary statistics of reviews and the corresponding graphical analysis to 1 Jan 2016 to 31 March 2022.

Our data on reviews of Tripadvisor attractions cover 22 languages, including French, English, Spanish, Italian, Portuguese, German, Dutch, Danish, Russian, Japanese, Mandarin (Chinese Simplified), Taiwanese Mandarin, Swedish, Polish, Norwegian, Korean, Turkish, Greek, Finnish, Czech, Hungarian and Slovakian. Table 1 lists the number of reviews left on Danish, French, and Spanish attractions by language. The languages we have data on covers >96% tourist arrivals to the three countries, according to Eurostat statistics on tourist arrivals by country (Eurostat, 2023).[2]

Table 2 lists the number of non-missing observations for each of the variables available in the attractions data. Table 3 lists the number of observations by attraction type.

Table 4 summarises the numerical variables that are part of the reviews dataset, while 5 summarises the number of non-missing observations of the variables of the reviews dataset.

[2]To be precise, the language data covers 96.2%, 97.4%, and 96.6% of tourists arriving in Denmark, France, and Spain, respectively.

4 Validating Tripadvisor data

In order to convince the reader of the validity of our proxy for tourism activity derived from Tripadvisor data, we compare it with widely recognised mainstream data on tourism activity. Namely, data on tourism activity from the statistical office of the European Commission, Eurostat. For both, our Tripadvisor proxy for tourism activitiy and the Eurostat data, we report absolute values of tourism activity and compare the pre-pandemic (2016-2019) levels of tourism activity to the 2020 level of tourism activity.

Figure 13 depicts tourism activity as proxied by the number of reviews on Tripadvisor. For January and February, the average number of reviews for the 2016-2019 period do not differ substantially compared to the number of reviews in 2020. Simply investigating the patterns for the average number of reviews over 2016-2019, we would expect that the number of reviews in March 2020 would be higher both than January and February 2020. However, number of reviews in March 2020 turned out to be lower than both of these months. Retrospectively, we know that March 2020 was the month where WHO declared the coronavirus spreading for a pandemic, which ultimately led to lockdowns that clearly affected tourism activity negatively.

Figure 14 illustrates tourism activity in Europe as reported by official statistics provided by Eurostat. Like the number of reviews left on Tripadvisor, the number of tourist arrivals in January and February 2020 was largely identical to the average number of tourist arrivals in 2016-2019. However, in March 2020 tourist arrivals began to diverge substantially from the average number of tourist arrivals over the 2016-2019 period.

Investigating the percentage-wise change in tourism arrivals from 2016-2019 to 2020 (i.e. panel e in Figure 14) and the percentage-wise change in number of reviews left on Tripadvisor (i.e. Figure 13) reveals a highly comparable pattern. Particularly, the January and February 2020 levels of tourism activity were largely similar with the average 2016-2019 level. Both figures reveal a substantial relative decrease in tourism activity starting in March 2020. In the months of May to August 2020, tourism activity started to recoup, but

slowly started to decrease again from September until December. The similarity between our Tripadvisor proxy for tourism activity and Eurostat's official statistics on tourism activity indicates that our proxy is a reliable proxy for mainstream tourism statistics.

For an overview of the research conducted by Mobile Lives Forum, visit https://forumviesmobiles.org/en/ongoing-research/list.

5 References

Karol J. Borowiecki and Maja U. Pedersen. A novel application of big data to measure trends in tourism: France, Spain and Denmark, January 2016 - March 2022. Report, Mobile Lives Forum (MLF), 2024.

Karol J. Borowiecki, Maja U. Pedersen, and Sara B. Mitchell. Using big data to measure cultural tourism in Europe with unprecedented precision. Discussion Papers on Economics, Working Paper No 5/2023, University of Southern Denmark, 2023.

Jesse Cambon, Diego Hernangómez, Christopher Belanger, and Daniel Possenriede. tidygeocoder: An r package for geocoding. *Journal of Open Source Software*, 6(65):3544, 2021. doi: 10.21105/joss.03544. R package version 1.0.5.

Eurostat. Arrivals at tourist accommodation establishments by country of origin of the tourist. Online data code: TOUR_OCC_ARNRAW. http://data.europa.eu/88u/dataset/ctjvyeqsbf9wsepdmmqrvw, 2023. Last data update: 05/01/2023.

6 Tables

Table 1: Number of reviews by language

(a) Total

Language	Reviews
French	2,655,586
Spanish	1,987,608
English	1,927,717
Italian	399,293
Portuguese	176,503
Dutch	163,006
German	151,241
Russian	104,454
Danish	86,315
Japanese	51,178
Chinese Taiwan	40,168
Chinese Simplified	40,162
Swedish	34,987
Polish	26,093
Norwegian	18,075
Korean	14,499
Turkish	10,268
Greek	10,004
Finnish	6,360
Czech	5,708
Hungarian	4,371
Slovak	1,575

(b) Denmark

Language	Reviews
English	103,266
Danish	66,265
Italian	15,532
Spanish	14,203
French	9,116
German	8,825
Swedish	6,219
Dutch	5,840
Norwegian	4,924
Portuguese	3,939
Chinese Taiwan	3,353
Russian	3,266
Chinese Simplified	3,077
Japanese	2,333
Polish	1,135
Greek	779
Turkish	640
Finnish	538
Korean	496
Czech	319
Slovak	136
Hungarian	119

(c) France

Language	Reviews
French	2,350,322
English	694,201
Spanish	119,055
Italian	112,269
Portuguese	69,758
Dutch	56,511
German	44,455
Russian	31,777
Japanese	26,747
Chinese Simplified	17,502
Chinese Taiwan	17,183
Korean	7,685
Swedish	5,755
Danish	5,611
Turkish	4,744
Greek	4,295
Polish	4,286
Norwegian	2,666
Czech	1,617
Hungarian	967
Finnish	875
Slovak	417

(d) Spain

Language	Reviews
Spanish	1,854,350
English	1,130,250
French	296,148
Italian	271,492
Portuguese	102,806
Dutch	100,655
German	97,961
Russian	69,411
Swedish	23,013
Japanese	22,098
Polish	20,672
Chinese Taiwan	19,632
Chinese Simplified	19,583
Danish	14,439
Norwegian	10,485
Korean	6,318
Finnish	4,947
Greek	4,930
Turkish	4,884
Czech	3,772
Hungarian	3,285
Slovak	1,022

Table 2: Number of observations in attraction dataset

Attraction Country	Attractions	Attraction Name	Rating	Reviews	Attraction Type	Location	Duration	Price	About
Denmark	8,977	8,977	6,356	8,977	8,969	8,977	1,986	1	1,578
France	78,070	78,070	58,225	78,070	77,941	78,070	30,620	55	7,354
Spain	56,340	56,257	45,221	53,854	55,714	56,339	19,435	118	18,472
Overall	143,387	143,304	109,802	140,901	142,624	143,386	52,041	174	27,404

Note: se provides the number of non-missing observations for key variables in the attraction dataset.

Table 3: No. Obs. and Country

Attraction type	Overall	Denmark	France	Spain
Boat Tours & Water Sports	8,016	210	3,838	3,968
Casinos & Gambling	321	23	194	104
Classes & Workshops	2,491	45	1,525	921
Concerts & Shows	3,111	201	1,774	1,136
Events	211	23	111	77
Food & Drink	4,610	94	3,325	1,191
Fun & Games	8,613	430	5,409	2,774
Museums	9,245	1,781	4,488	2,976
Nature & Parks	13,348	952	6,727	5,669
Nightlife	11,056	488	3,442	7,126
Outdoor Activities	15,699	637	7,514	7,548
Shopping	13,030	1,547	7,134	4,349
Sights & Landmarks	46,622	3,507	26,980	16,135
Spas & Wellness	6,256	147	3,737	2,372
Tours	23,292	695	12,102	10,495
Transportation	4,081	64	2,377	1,640
Traveler Resources	2,648	59	1,958	631
Water & Amusement Parks	264	15	162	87
Zoos & Aquariums	341	53	191	97
Other	4,558	148	1,414	2,996
NA	763	8	129	626
Total Attractions	143,387	8,977	78,070	56,340

Note: Attractions type categories are not mutually exclusive.

Table 4: Descriptive statistics (reviews)

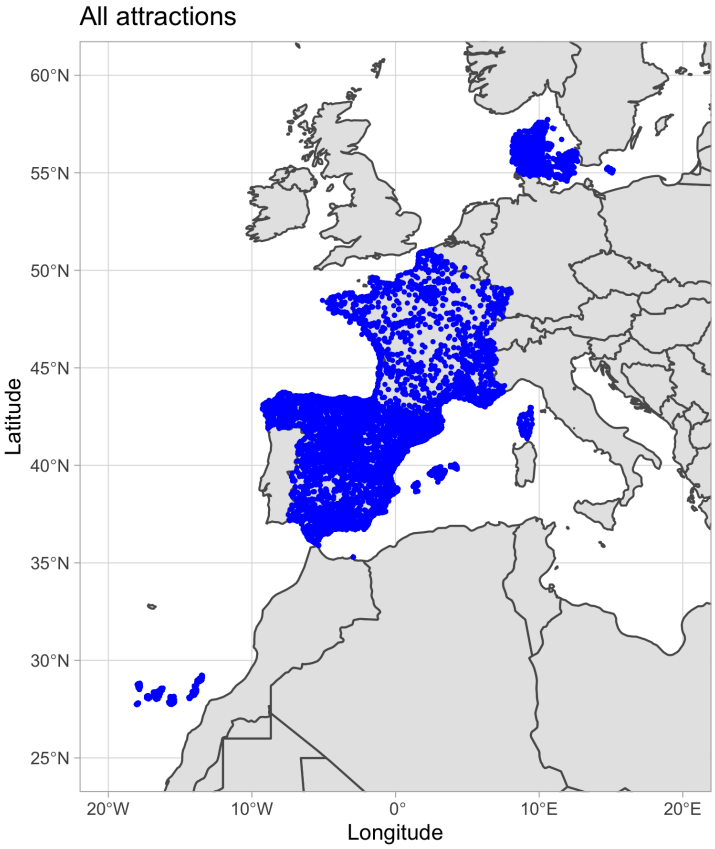
Variable	Obs.	Mean	Std. Dev.	Min.	Max.
Rating	7,913,158	4.41	0.94	1.00	5.00
Distance (km)	3,381,096	2,124.20	3,506.96	0.00	20,025.60
Word count	7,915,171	52.03	48.83	0.00	3,406.00
Words per sentence	7,915,138	16.41	11.42	1.00	750.00
Big words	7,915,138	23.75	11.82	0.00	100.00

Table 5: Number of observations in review dataset

Attraction country	Users	Attractions	Rating	Type of user	Date published	Date of experience	Country of attraction	Country of user	Distance
Denmark	254,320	254,320	254,229	254,320	254,320	40860	254,320	171,464	166,942
France	3,578,698	3,578,698	3,577,559	3,578,698	3,578,698	563,879	3,578,698	2,494,748	2,453,016
Spain	4,076,978	4,082,153	4,081,370	4,082,153	4,082,153	1,562,608	4,082,153	2760768	761,138
Overall	7,909,996	7,915,171	7,913,158	7,915,171	7,915,171	2,167,347	7,915,171	5,426,980	3,381,096

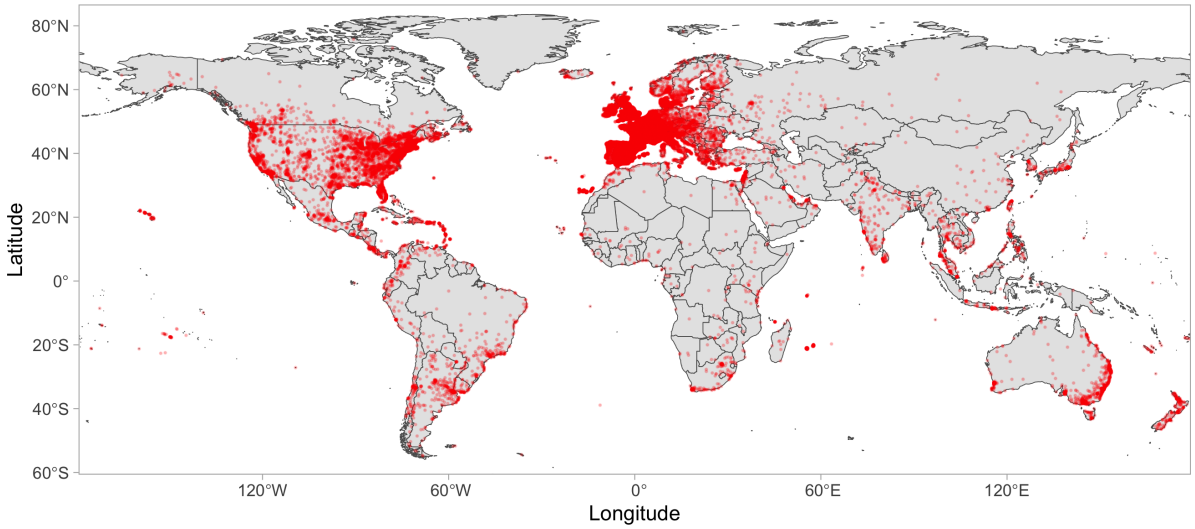
7 Figures

Figure 1: Map of Tripadvisor locations



Source: Tripadvisor (2022) data on locations that were reviewed at least once between 01.01.2016 and 01.04.2022, excluding the Azores, Madeira, and overseas territories.

Figure 2: Map of unique user locations



Source: Tripadvisor (2022) data on unique user locations of reviewers who posted at least one review on a Danish, French, or Spanish attraction between 01.01.2016 and 01.04.2022

Figure 3: Overall: No. attractions, by type

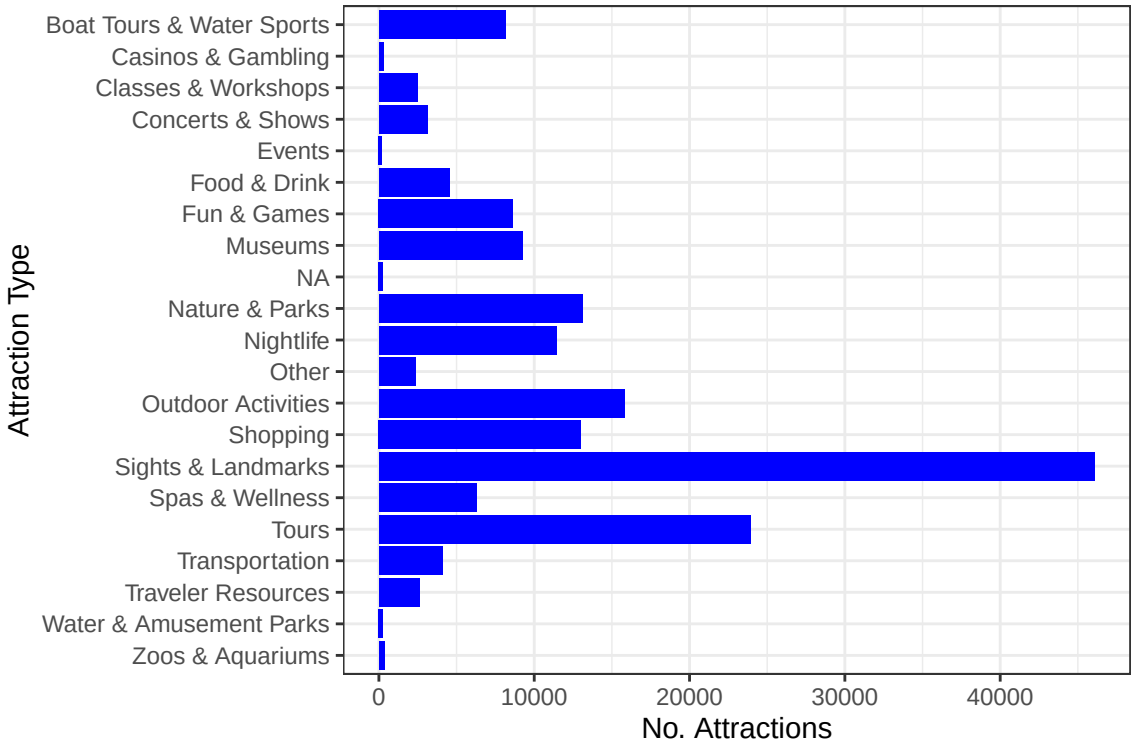


Figure 4: Denmark: No. attractions, by type

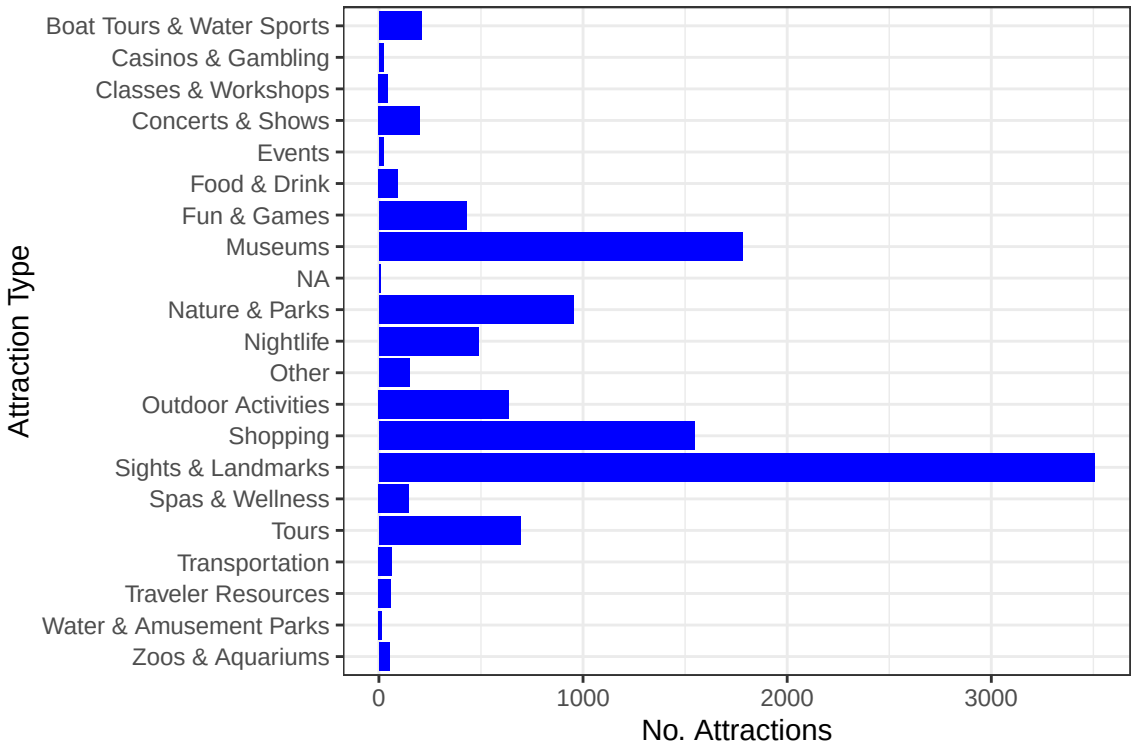


Figure 5: France: No. attractions, by type

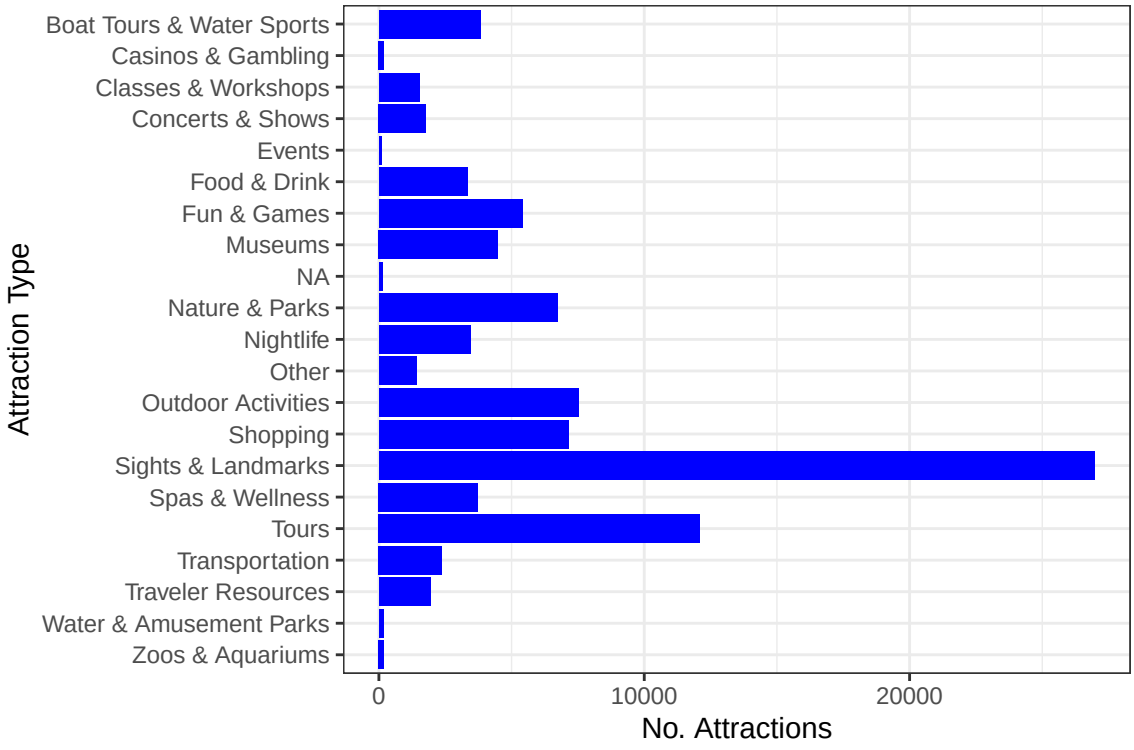


Figure 6: Spain: No. attractions, by type

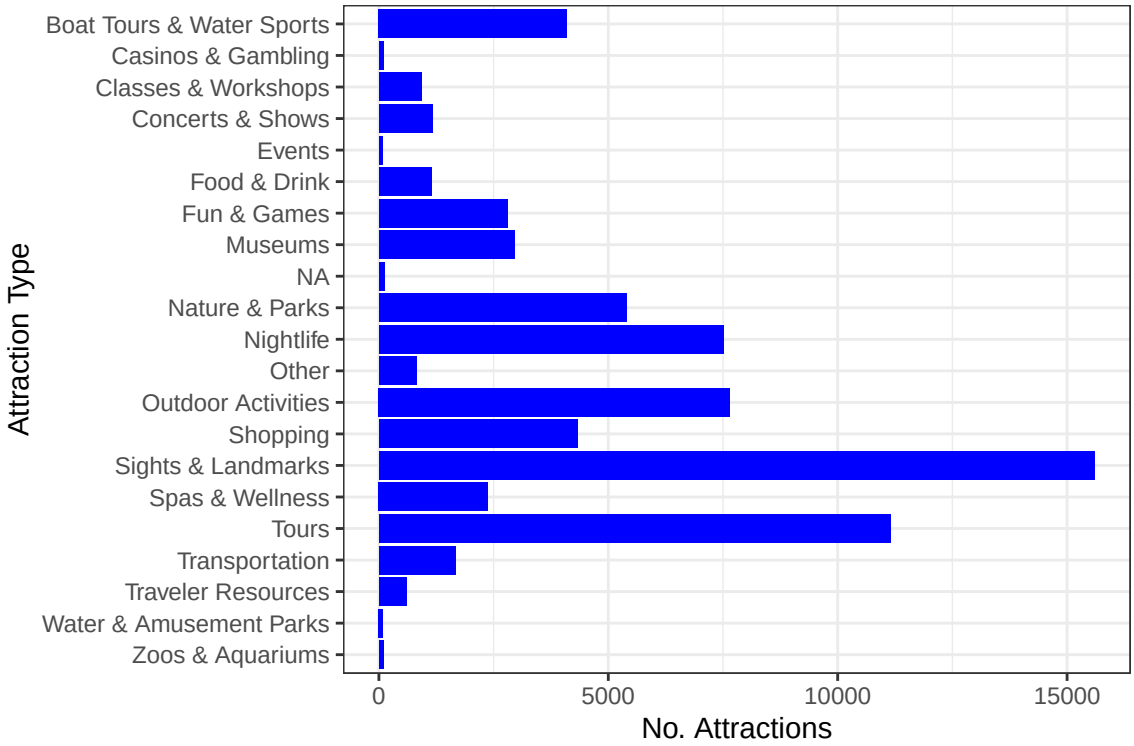
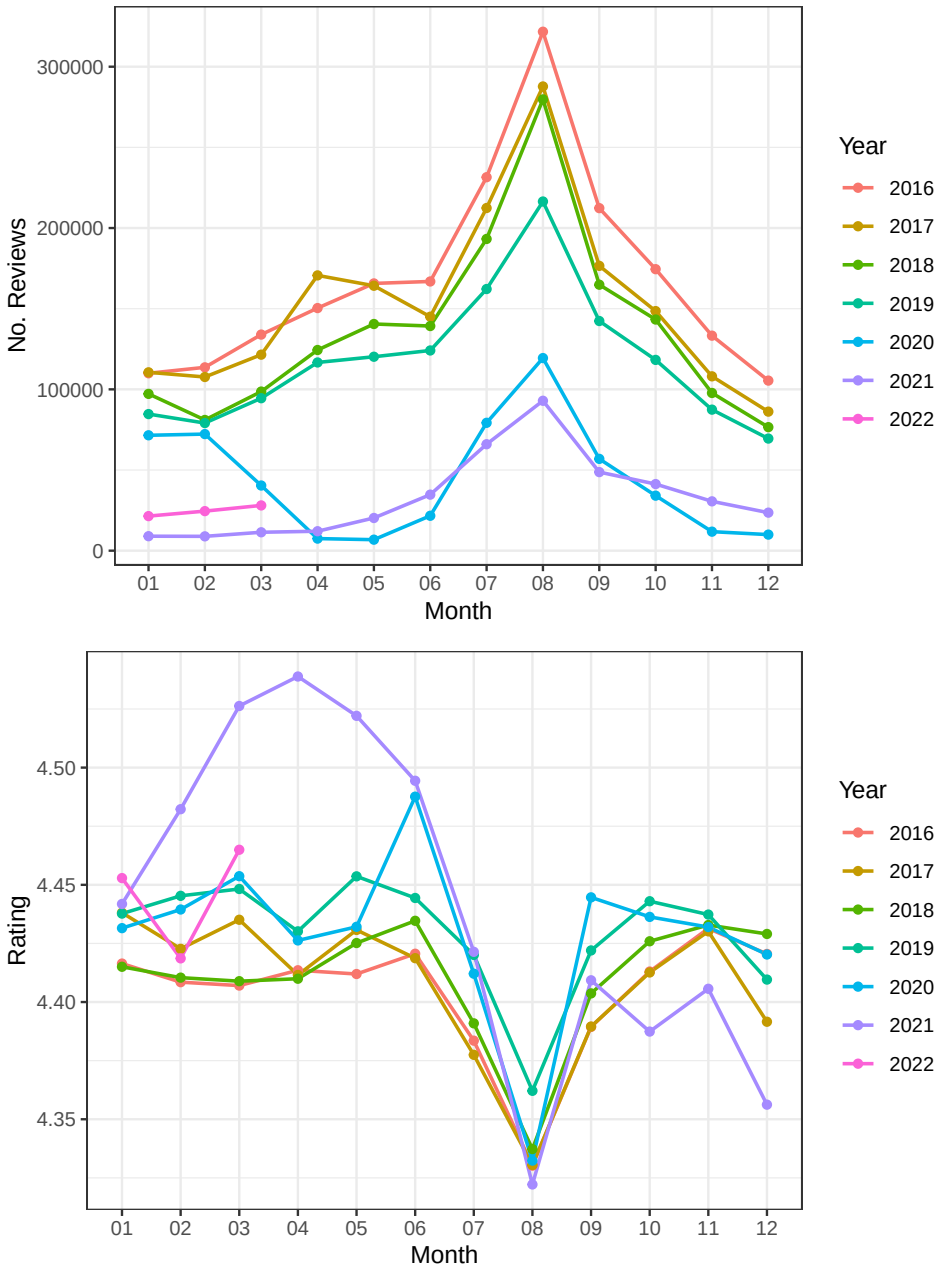


Figure 7: Aggregate tourism activity



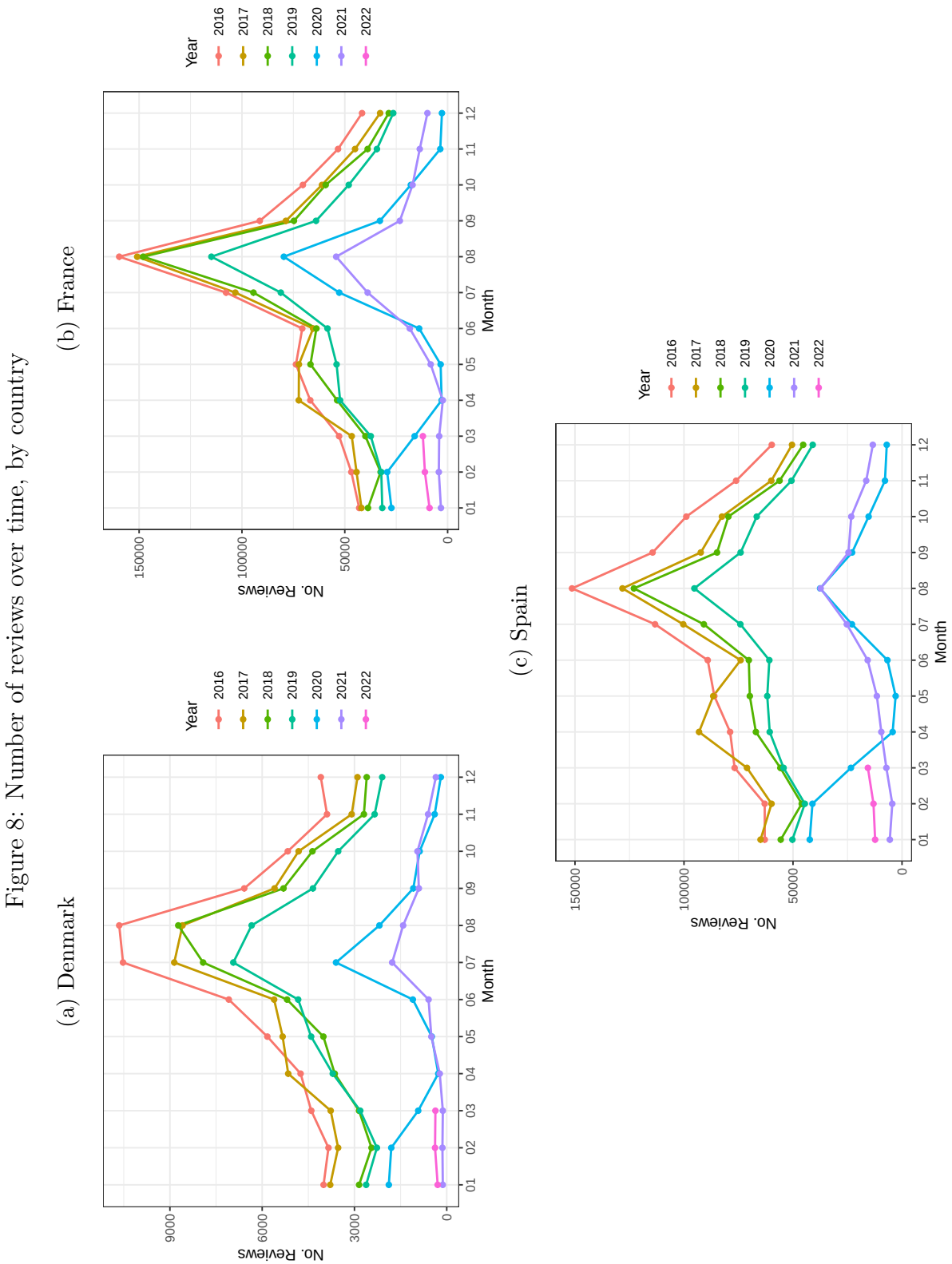


Figure 8: Number of reviews over time, by country

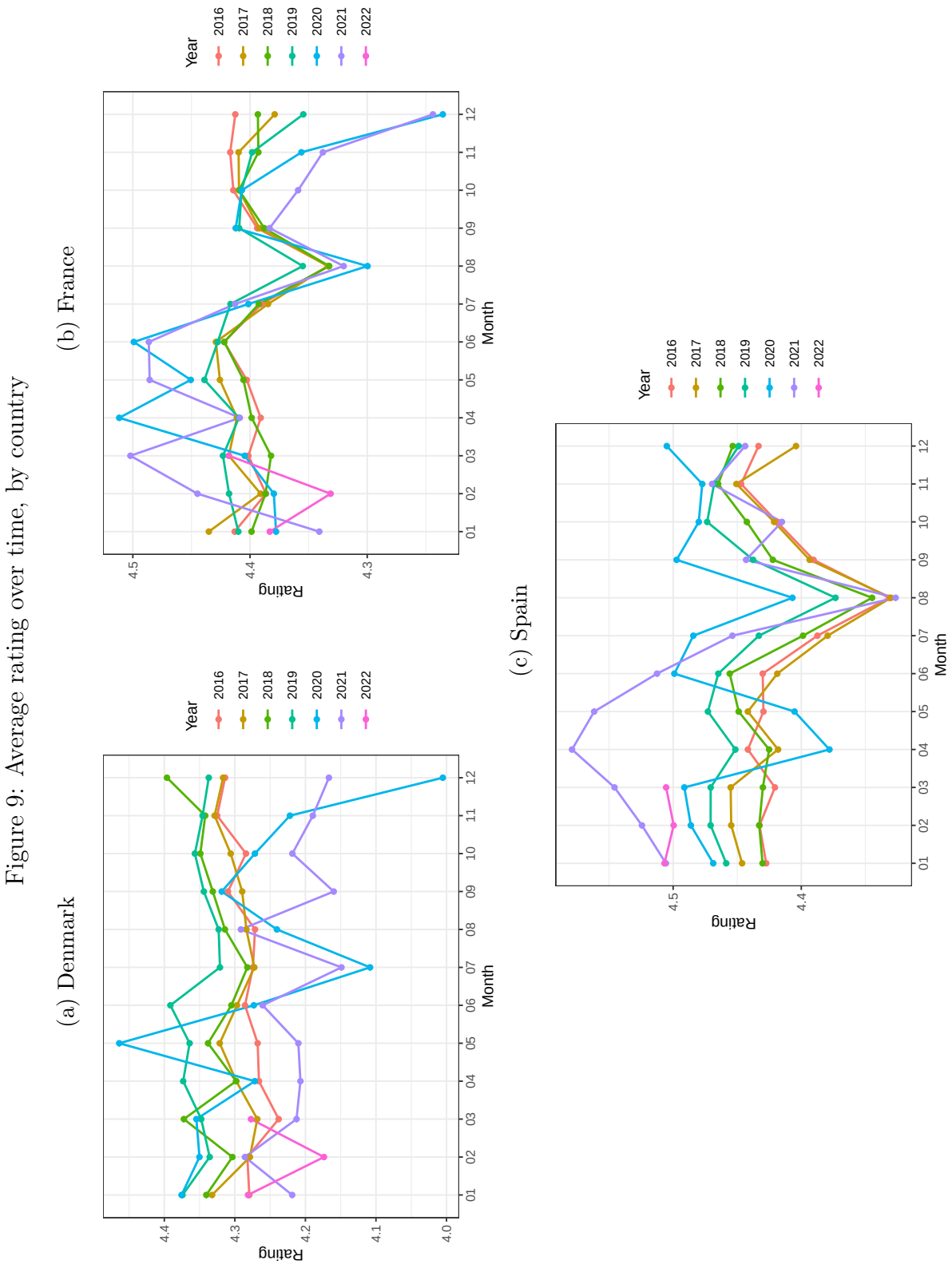
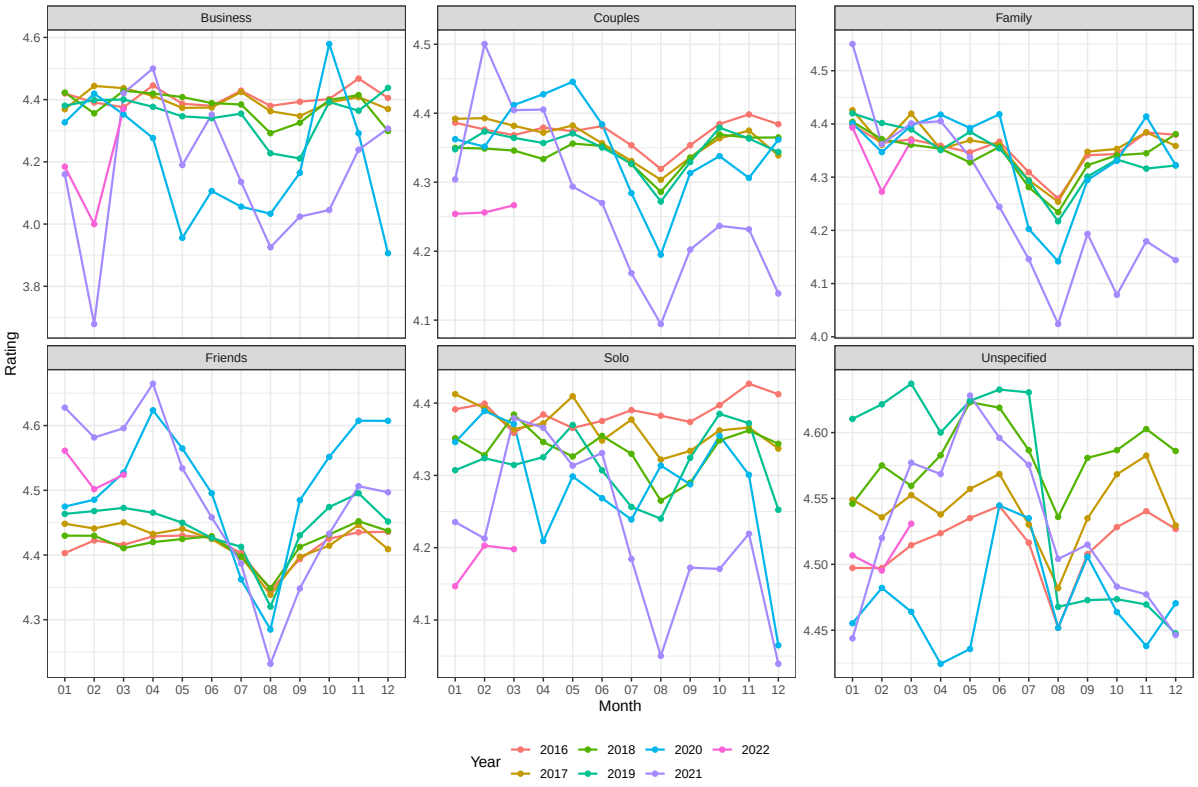


Figure 9: Average rating over time, by country

Figure 10: Number of reviews over time, by tourist type

Figure 11: Average rating over time, by tourist type



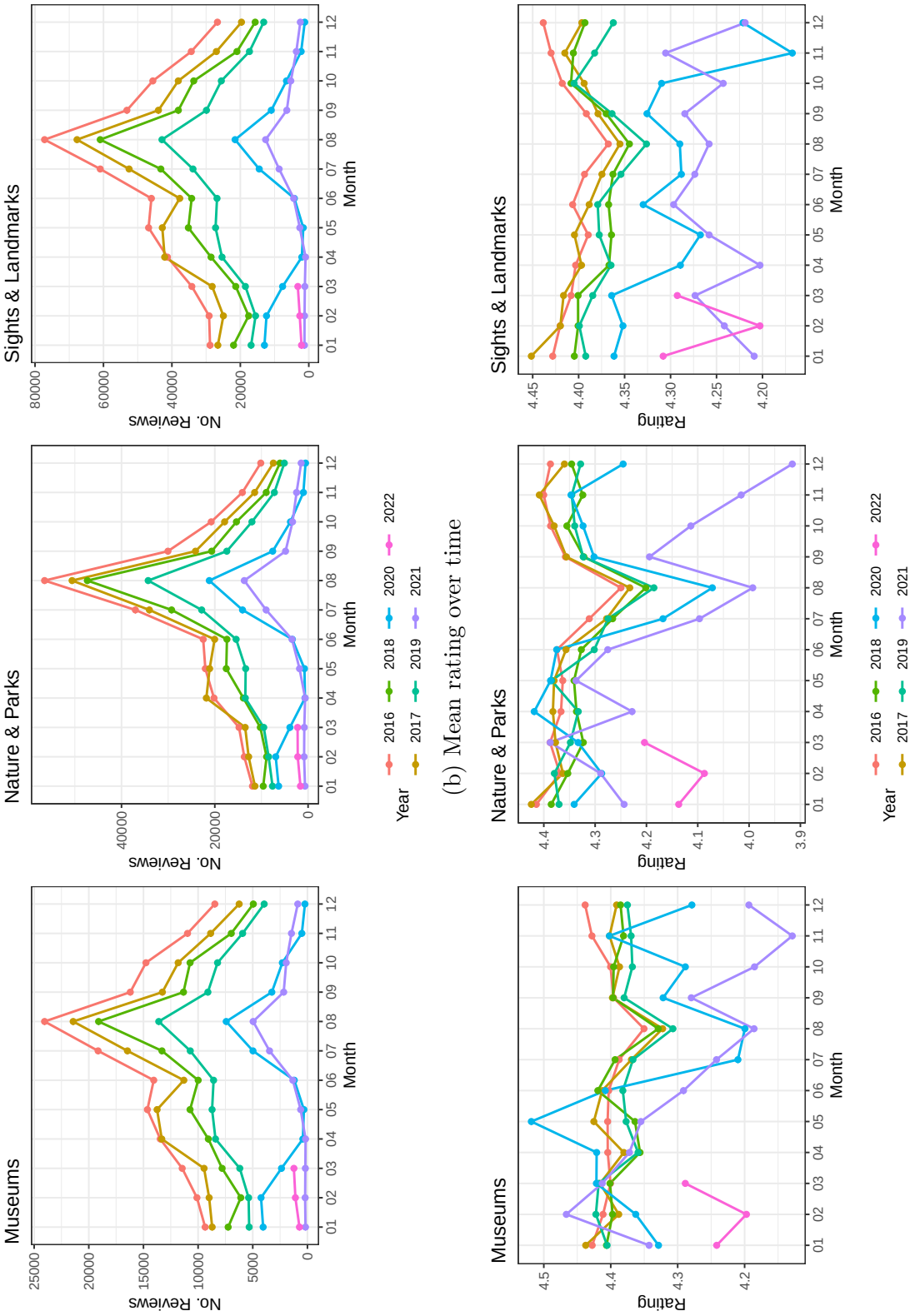
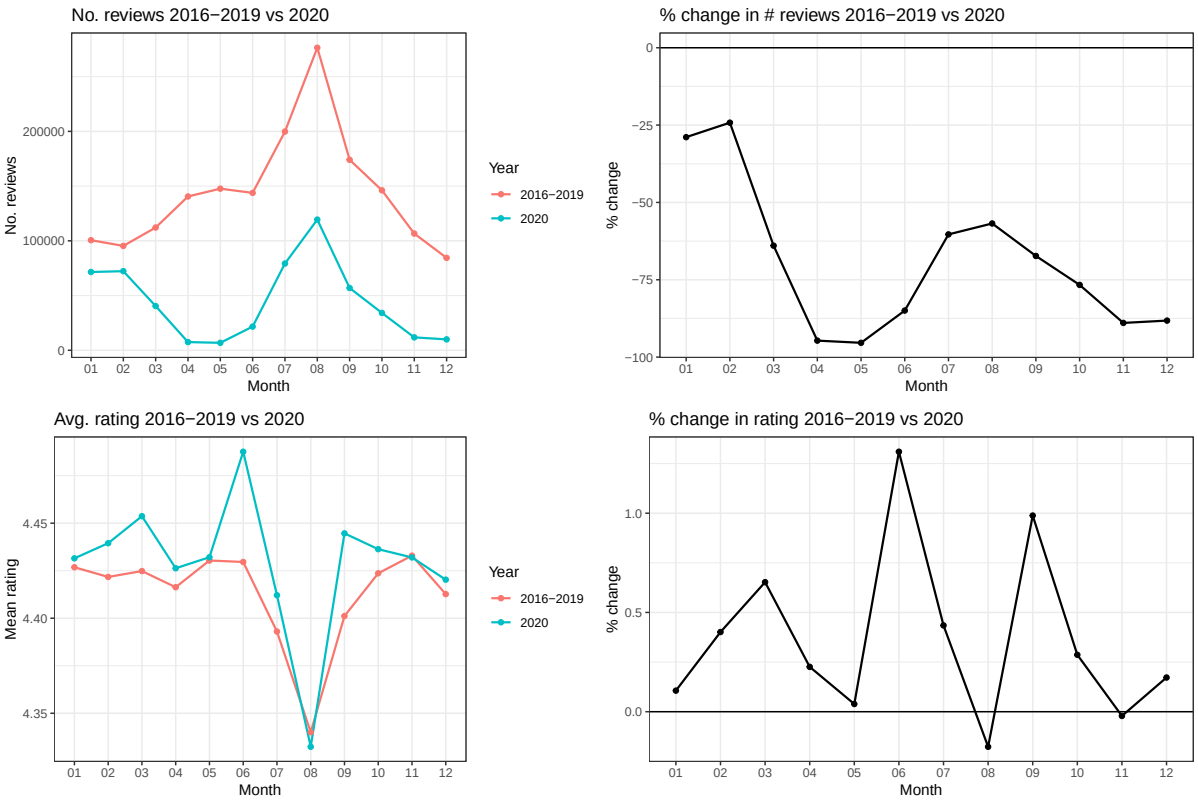


Figure 12: Tourism activity by attraction type

(a) No. reviews over time

(b) Mean rating over time

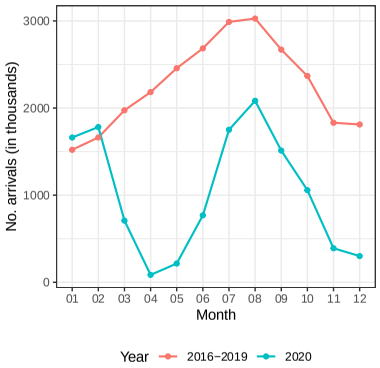
Figure 13: Tourism activity in Tripadvisor, 2016-2019 vs 2020



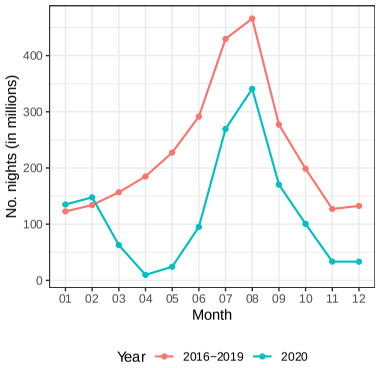
Source: Calculations based on Tripadvisor (2022) data.
Note: The figure illustrates the 2016-2019 mean and the 2020 level for Denmark, France, and Spain countries.

Figure 14: Tourism activity in Eurostat data, 2016-2019 mean vs 2020

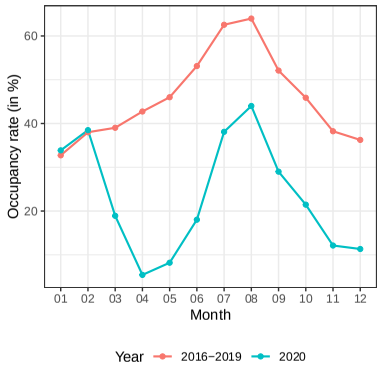
(a) No. tourism arrivals



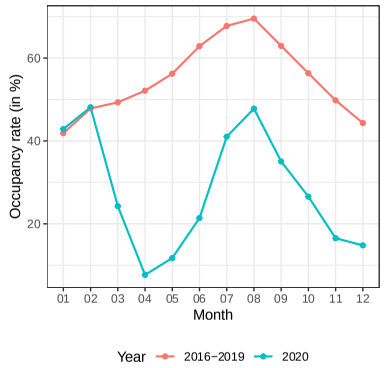
(b) No. tourism nights



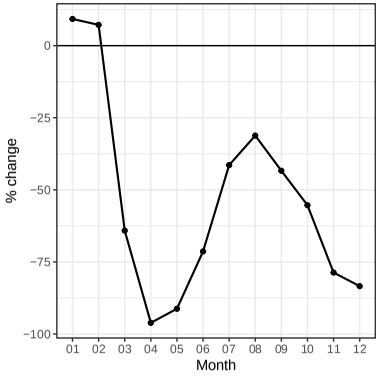
(c) Occupancy rate: Bed places



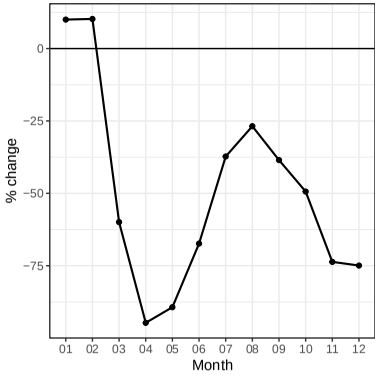
(d) Occupancy rate: Bedrooms



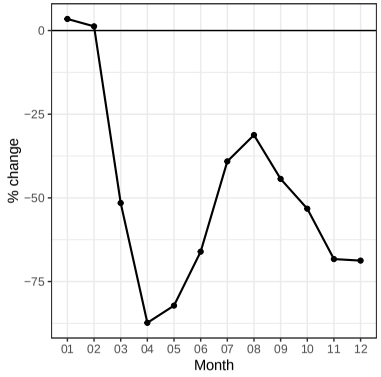
(e) % change in tourism arrivals



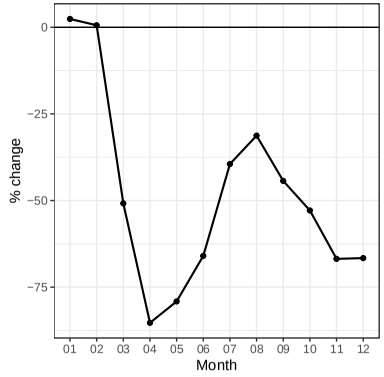
(f) % change in tourism nights



(g) % change in ccupancy rate: Bed places



(h) % change in occupancy rate: Bedrooms

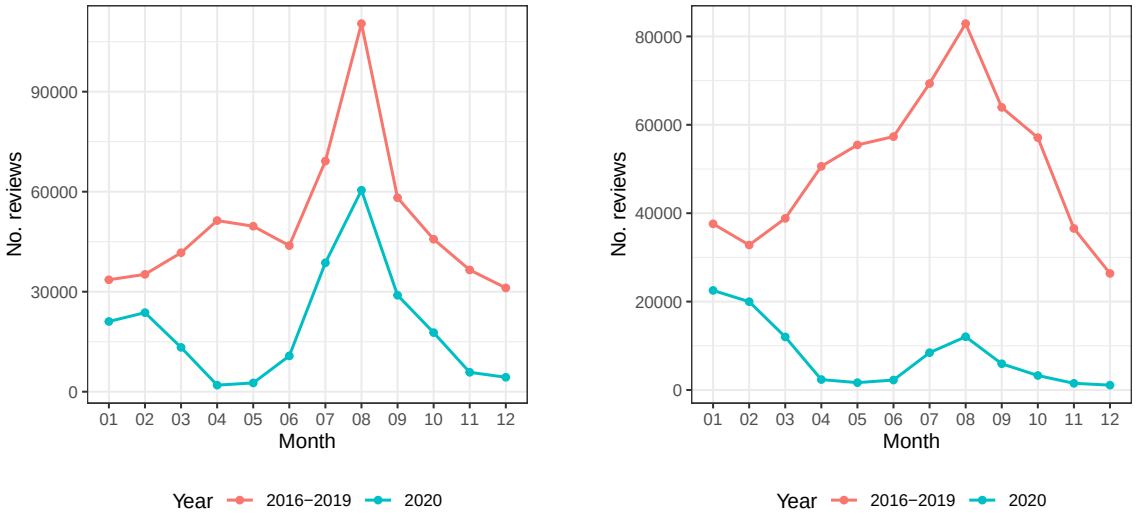


Source: Calculations based on Eurostat (2022) files TOUR_OCC_ARM, TOUR_OCC_NIM, and TOUR_OCC_MNOR.
Note: The figures in the top row illustrate the 2016-2019 mean and the 2020 level for EU-27 countries. The figures in the bottom row illustrate the percentage change from the 2016-2019 mean to the 2020 level.

Figure 15: Foreign vs domestic reviews in Tripadvisor, 2016-2019 mean vs 2020

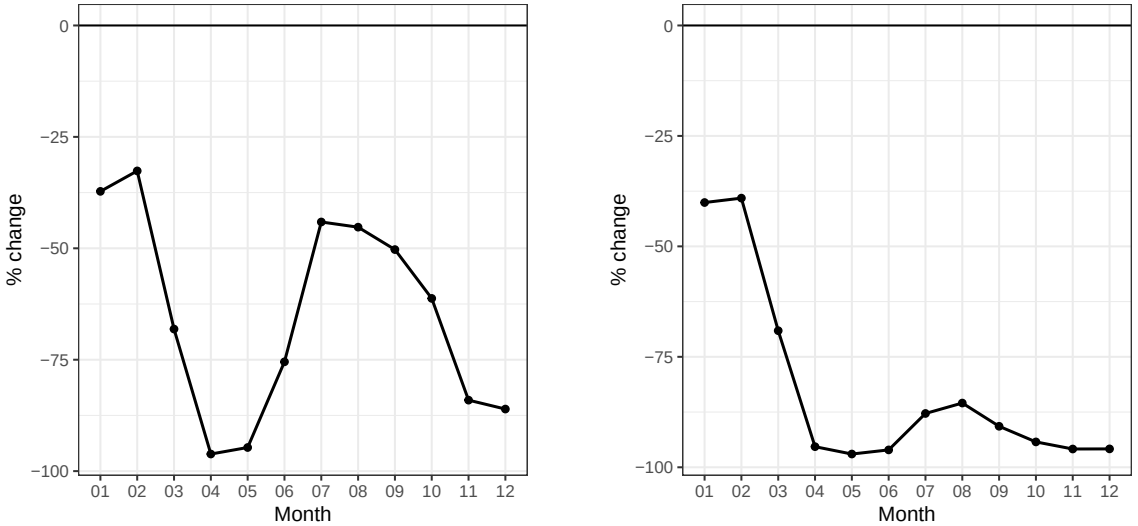
(a) # of domestic tourists

(b) # of foreign tourists



(c) % change in domestic tourists

(d) % change in foreign tourists

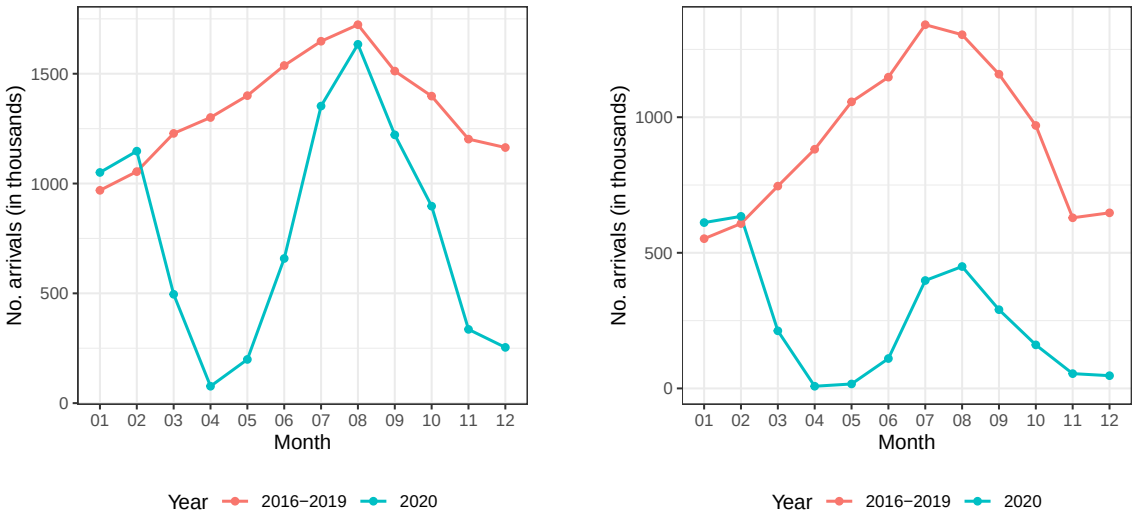


Source: Calculations based on Tripadvisor (2022) data.
Note: User and attraction location not available for all reviews.

Figure 16: Foreign vs domestic tourism arrivals in Eurostat, 2016-2019 mean vs 2020

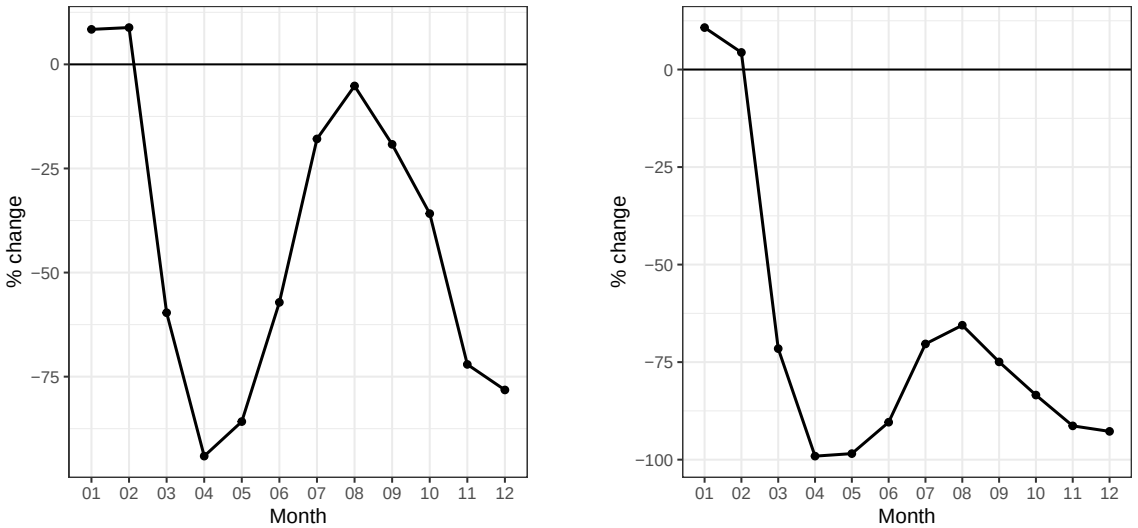
(a) # of domestic tourists

(b) # of foreign tourists



(c) % change in domestic tourists

(d) % change in foreign tourists



Source: Calculations based on data from Eurostat (2022) file TOUR_OCC_ARM.
Note: The figures in the top row illustrate the 2016-2019 mean and the 2020 level for EU-27 countries. The figures in the bottom row illustrate the percentage change from the 2016-2019 mean to the 2020 level.