Data Manual and Description

INnovative CULtural ToUrisM in European peripheries (INCULTUM)[1]

Sara Mitchell, Martin Hørlyk Kristensen, Maja Pedersen and Karol J. Borowiecki

Last updated: March 21, 2024

Abstract:

This data manual presents data collected from Tripadvisor during the EU-funded INCULTUM project. The collected variables are presented and summarised, and the data is validated though a simple comparison with official statistics.

Recommended citations:

*Borowiecki, Karol J., Pedersen, Maja U. and Palomeque, Marco (2024). *Putting the Periphery on the Map: Tourism Activity measured with Big Data.* Work in progress.

*Borowiecki, Karol J., Pedersen, Maja U., Mitchell, Sara B. and Alam Khan, Shahedul (2023). Navigating Cultural Landscapes: Approaches to Data Collection and Analysis in Tourism. In *Visiting the Margins: Innovative Cultural Tourism in European Peripheries*, Chapter 5. Routledge.

[1]**Acknowledgements**: The Project has received funding from the European Unions Horizon 2020 research and innovation programme under Grant Agreement no. 101004552.

Contents

1 Introduction 2

2 The data 4

3 Descriptive statistics 8

4 Validating Tripadvisor data 8

5 References 10

6 Tables 11

7 Figures 17

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. (2023b, 2024a,b) where the data it is described, validated and analyzed. The papers both present a visual and a formal validation of the data, before it is used to make inferences. A similar approach is also used in Borowiecki et al. (2023a), considering a different set of countries and including more languages collected from Tripadvisor.

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 the INCULTUM countries, 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 countries that are of relevance to the INCULTUM. 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). 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.[2] Attraction types are not mutually exclusive; that is, an attraction can be assigned to multiple attraction type

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

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 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 versus 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 321,055 attractions, 12,185,780 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 the 10 countries Albania, Denmark, France, Greece, Ireland, Italy, Portugal, Slovakia, Spain, Sweden. For each country reviews are collected in English. Furthermore for the countries Denmark, France, Italy, Portugal, Spain, and Sweden reviews are also collected in the native language. Table 1 lists the number of reviews posted on attractions in INCULTUM countries by language.

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.

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 activity and the Eu-

rostat 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 10 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 11 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 11) and the percentage-wise change in number of reviews left on Tripadvisor (i.e. Figure 10) 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.

5 References

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, 2023a.

Karol J. Borowiecki, Maja U. Pedersen, Sara B. Mitchell, and Shahedul Alam Khan. Final Findings Analysis Report INCULTUM. Deliverable D3.3, Visiting the Margins. INnovative CULtural ToUrisM in European peripheries, 2023b.

Karol J. Borowiecki, Maja U. Pedersen, Sara B. Mitchell, and Shahedul Alam Khan. Navigating cultural landscapes: Approaches to data collection and analysis in tourism. In *Visiting the Margins: Innovative Cultural Tourism in European Peripheries*, Routledge, chapter 5. 2024a.

Karol J. Borowiecki, Maja U. Pedersen, and Marco Palomeque. Putting on the map the periphery: Tourism activity measured with big data. Work in progress, 2024b.

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.

6 Tables

Table 1: Number of reviews by language

Language	Albania	Denmark	France	Greece	Ireland	Italy	Portugal	Slovakia	Spain	Sweden
English	20,261	157,227	1,196,886	882,169	894,179	1,842,324	678,371	40,878	1,774,871	141,302
Danish	0	87,409	0	0	0	0	0	0	0	0
French	0	0	3,067,122	0	0	0	0	0	0	0
Italian	0	0	0	0	0	849,703	0	0	0	0
Portuguese	0	0	0	0	0	0	445,914	0	0	0
Spanish	0	0	0	0	0	0	0	0	2,483,377	0
Swedish	0	0	0	0	0	0	0	0	0	107,164

Table 2: Number of observations in attraction dataset

Attraction Country	Attractions	Attraction Name	Rating	Reviews	Attraction Type	Location	Duration	Price	About
Albania	1,048	1,048	704	1,048	1,038	1,048	324	0	514
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
Greece	17,383	17,383	14,326	17,383	17,325	17,383	5,682	27	5,195
Ireland	9,361	9,361	7,838	9,361	9,343	9,361	3,199	18	3,846
Italy	122,477	122,477	87,986	122,477	122,274	122,477	48,619	186	10,584
Portugal	16,656	16,656	13,096	16,656	16,580	16,656	5,835	26	7,513
Slovakia	3,238	3,238	2,483	3,238	3,234	3,238	1,535	0	188
Spain	56,340	56,257	45,221	53,854	55,714	56,339	19,435	118	18,472
Sweden	7,505	7,505	5,985	7,505	7,495	7,505	2,532	8	2,231
Overall	321,055	320,972	242,220	318,569	319,913	321,054	119,767	439	57,475

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

Table 3: No. Obs. and Country

Attraction type	Albania	Denmark	France	Greece	Ireland	Italy	Portugal	Slovakia	Spain	Sweden	Overall
Boat Tours & Water Sports	96	210	3,838	2,380	410	3,748	1,626	55	3,968	431	16,762
Casinos & Gambling	2	23	194	16	60	73	15	14	104	28	529
Classes & Workshops	8	45	1,525	244	173	1,048	243	16	921	56	4,279
Concerts & Shows	14	201	1,774	197	270	2,081	205	37	1,136	176	6,091
Events	0	23	111	11	17	268	7	1	77	11	526
Food & Drink	20	94	3,325	266	79	3,704	335	46	1,191	46	9,106
Fun & Games	29	430	5,409	603	580	3,390	479	344	2,774	591	14,629
Museums	54	1,781	4,488	910	556	7,113	921	238	2,976	845	19,882
Nature & Parks	135	952	6,727	2,413	1,000	11,542	1,397	389	5,669	1,519	31,743
Nightlife	163	488	3,442	1,713	1,375	9,521	1,023	221	7,126	108	25,180
Outdoor Activities	160	637	7,514	4,389	1,065	10,111	2,417	173	7,548	1,196	35,210
Shopping	56	1,547	7,134	1,999	1,110	9,661	1,384	194	4,349	809	28,243
Sights & Landmarks	208	3,507	26,980	2,937	2,182	55,982	4,862	1,494	16,135	1,989	116,276
Spas & Wellness	43	147	3,737	557	354	2,651	478	113	2,372	219	10,671
Tours	335	695	12,102	4,897	1,781	13,841	5,040	262	10,495	1,189	50,637
Transportation	142	64	2,377	1,790	341	3,977	1,423	80	1,640	130	11,964
Traveler Resources	3	59	1,958	59	137	3,329	115	36	631	107	6,434
Water & Amusement Parks	3	15	162	40	14	199	23	19	87	67	629
Zoos & Aquariums	0	53	191	22	15	91	25	14	97	39	547
Other	4	148	1,414	124	84	1,586	236	27	2,996	235	6,854
NA	10	8	129	58	18	203	76	4	626	10	1,142
Total Attractions	1,048	8,977	78,070	17,383	9,361	122,477	16,656	3,238	56,340	7,505	321,055

Note: Attractions type categories are not mutually exclusive.

Table 4: Descriptive statistics (reviews)

Variable	Obs.	Mean	Std. Dev.	Min.	Max.
Rating	14,665,213	4.46	.91	0.00	8.00
Distance (km)	9,710,158	2,802.26	3,892.67	0.00	20,028.15
Word count	14,669,157	56.60	52.38	1.00	6,472.00
Words per sentence	14,669,157	16.55	11.10	0.00	3,619.00
Big words	14,669,157	20.85	8.69	0.00	100.00

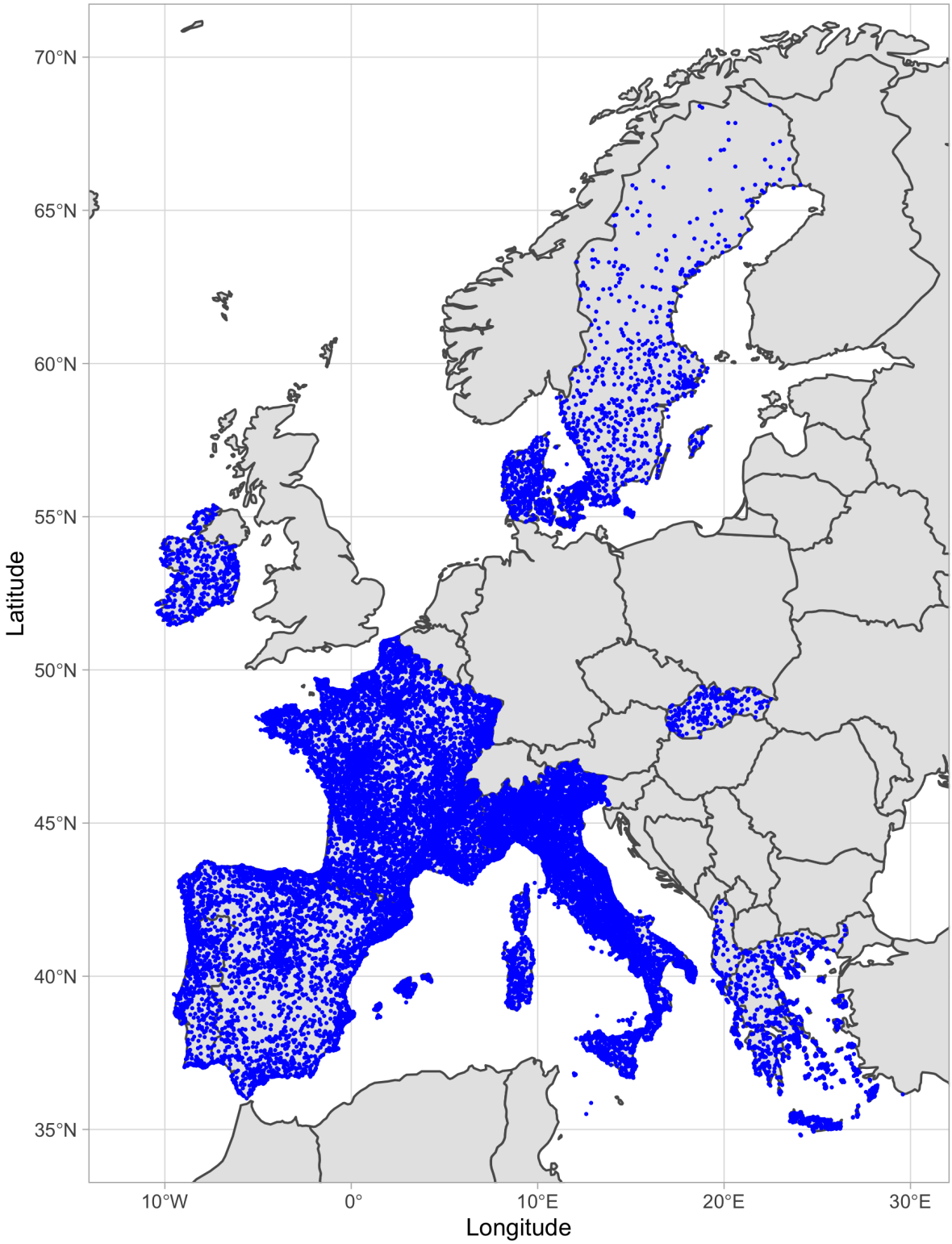
Table 5: Number of observations in review dataset

Attraction country	Users	Attractions	Rating	Type of user	Date published	Date of experience	Country of attraction	Country of user	Distance
Albania	20,261	20,261	20,257	20,261	20,261	3,330	20,261	13,161	12,908
Denmark	244,636	244,636	244,536	244,636	244,636	30,087	244,636	178,351	177,275
France	4,264,008	4,264,008	4,262,623	4,264,008	4,264,008	640,766	4,264,008	3,066,076	3,033,029
Greece	882,169	882,169	881,962	882,169	882,169	246,480	882,169	602,689	581,624
Ireland	894,179	894,179	893,956	894,179	894,179	196,351	894,179	618,164	591,117
Italy	2,692,027	2,692,027	2,691,423	2,692,027	2,692,027	574,385	2,692,027	1,978,623	1,769,993
Portugal	1,124,285	1,124,285	1,123,735	1,124,285	1,124,285	274,553	1,124,285	631,316	595,193
Slovakia	40,878	40,878	40,815	40,878	40,878	6,685	40,878	27,541	26,975
Spain	4,258,248	4,088,552	4,257,476	3,577,585	4,258,248	2,057,190	4,258,248	2,987,465	2,809,404
Sweden	248,466	248,466	248,430	248,466	248,466	26,092	248,466	117,244	112,640
Total	14,669,157	14,499,461	14,665,213	13,988,494	14,669,157	4,055,919	14,669,157	10,219,966	9,710,158

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

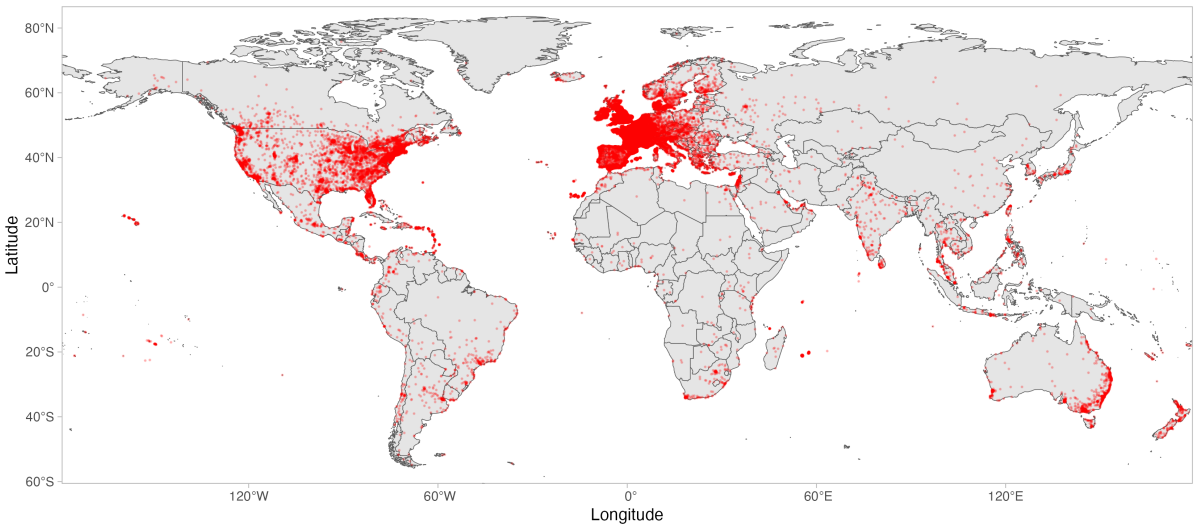
7 Figures

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 an attraction in one of the INCULTUM countries between 01.01.2016 and 01.04.2022

Figure 3: Overall: No. attractions, by type

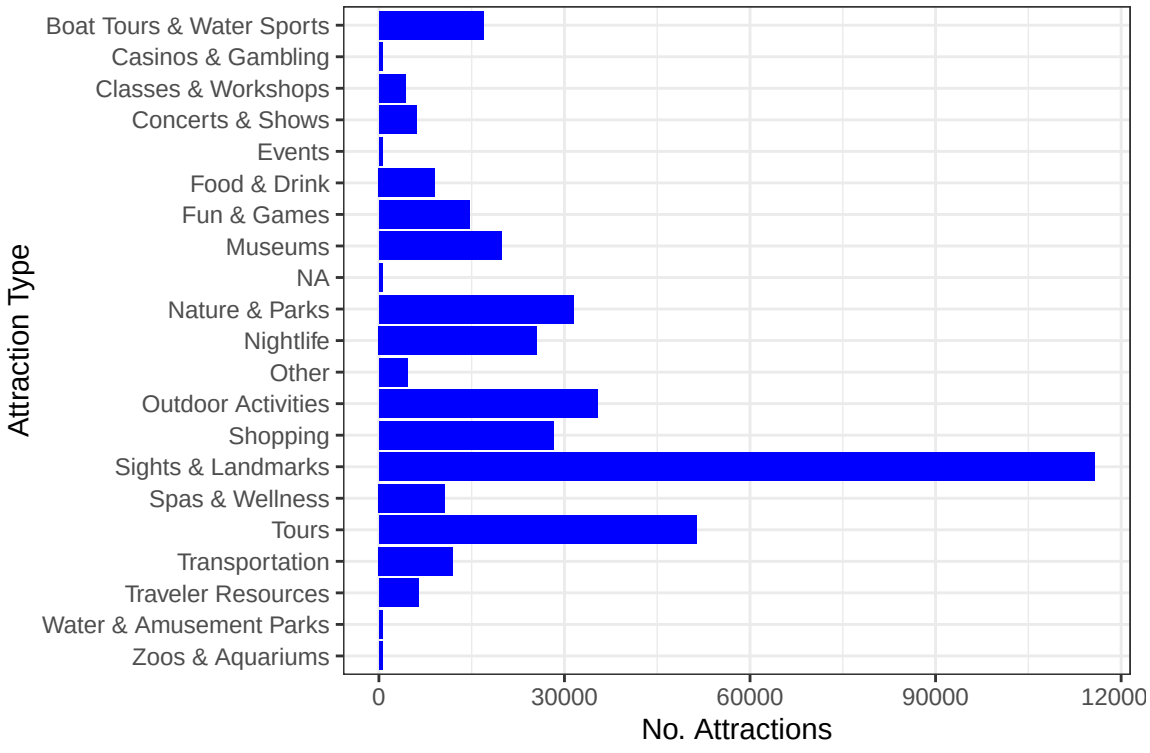


Figure 4: Aggregate tourism activity

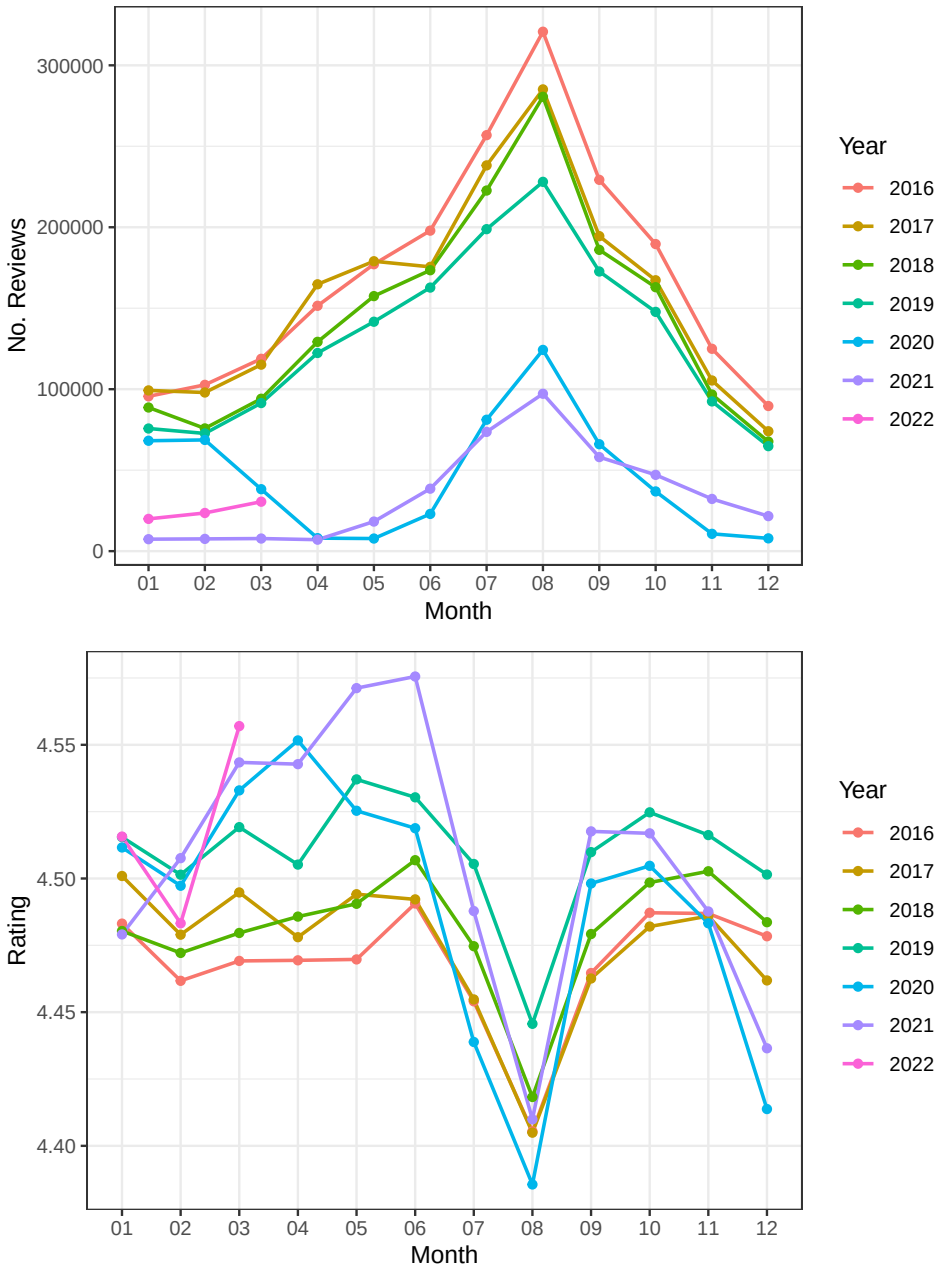


Figure 5: Number of reviews over time, by country

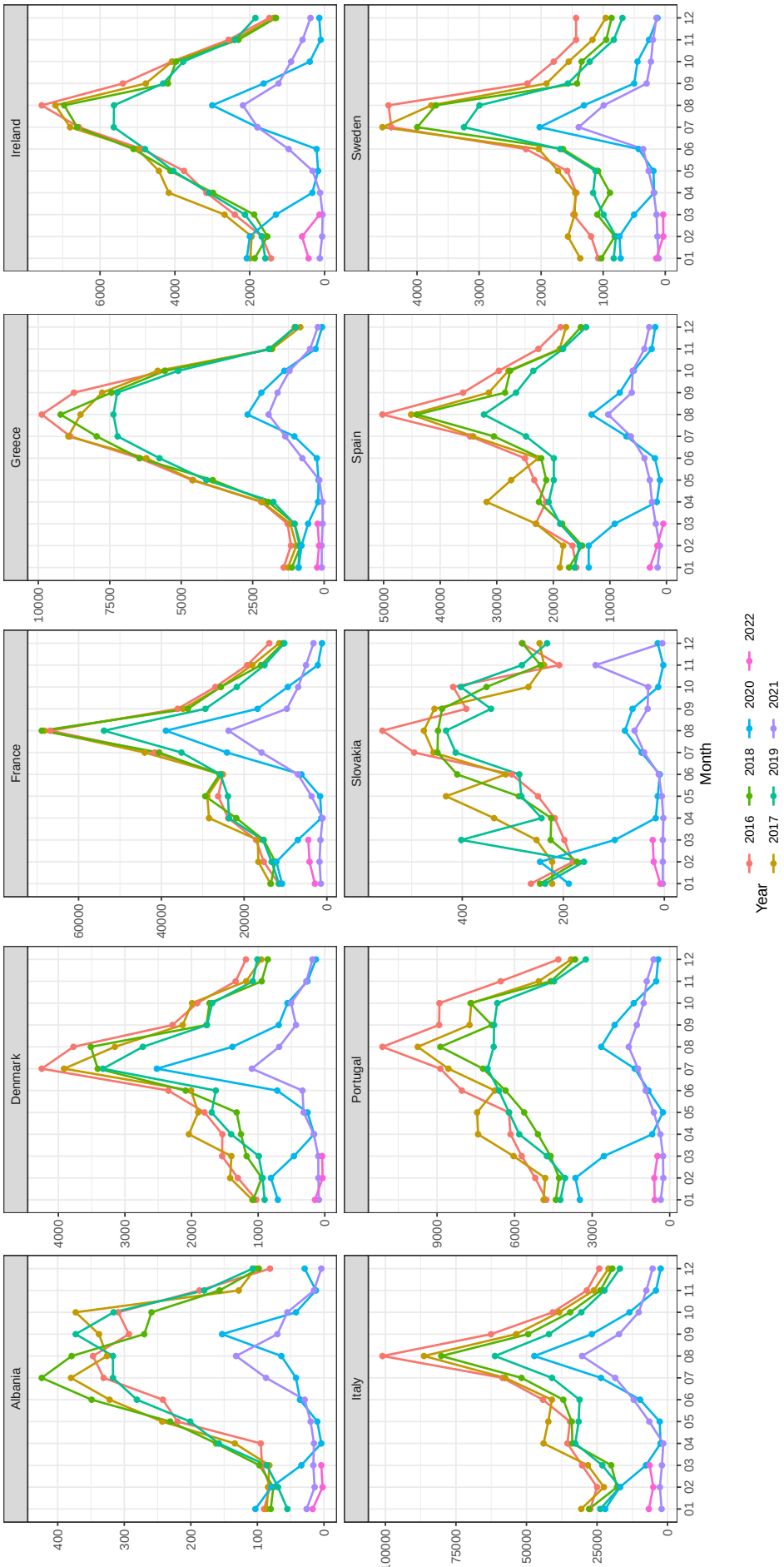


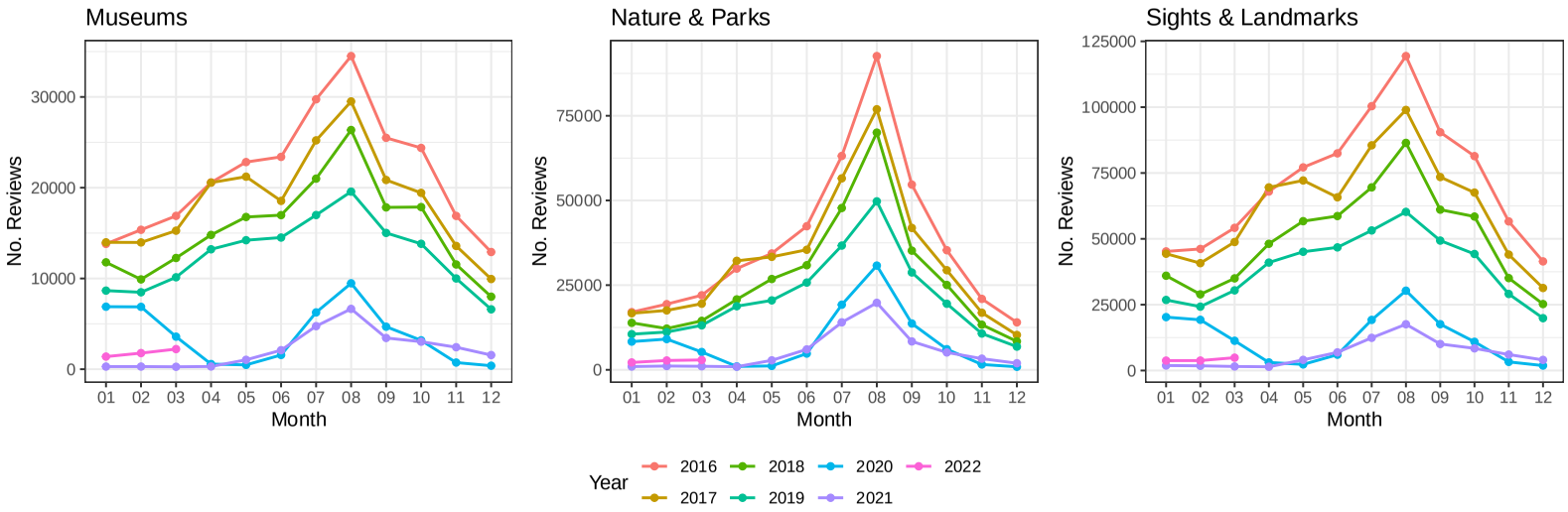
Figure 6: Average rating over time, by country

Figure 7: Number of reviews over time, by tourist type

Figure 8: Average rating over time, by tourist type

Figure 9: Tourism activity by attraction type

(a) No. reviews over time



(b) Mean rating over time

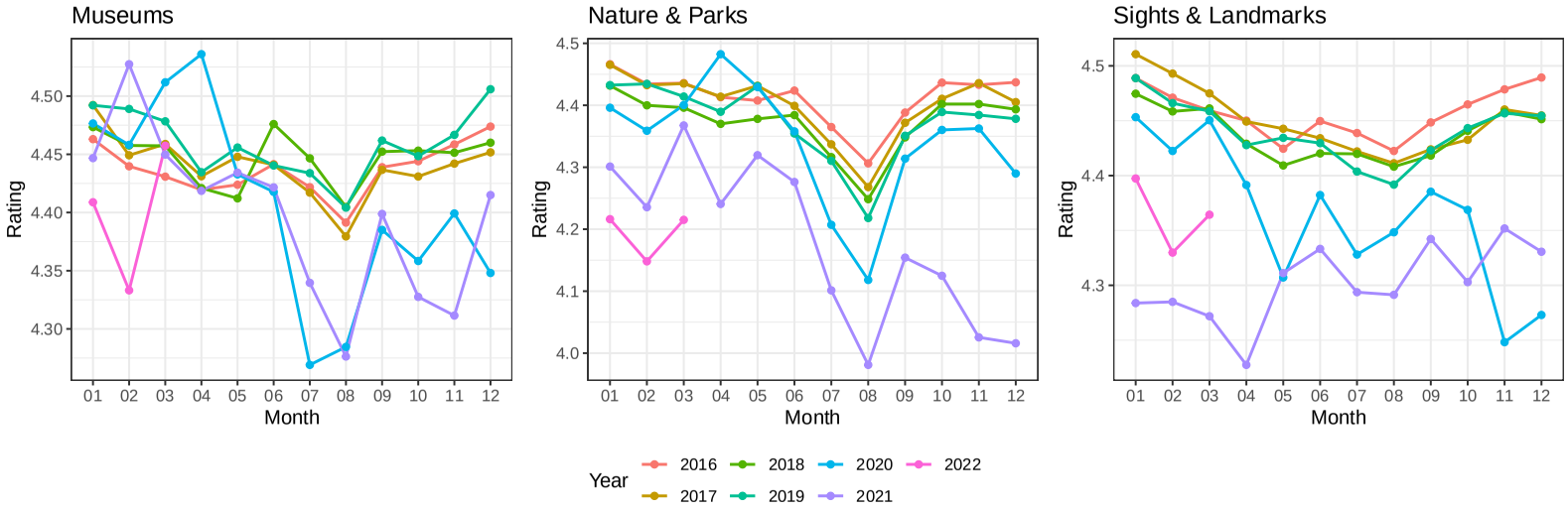
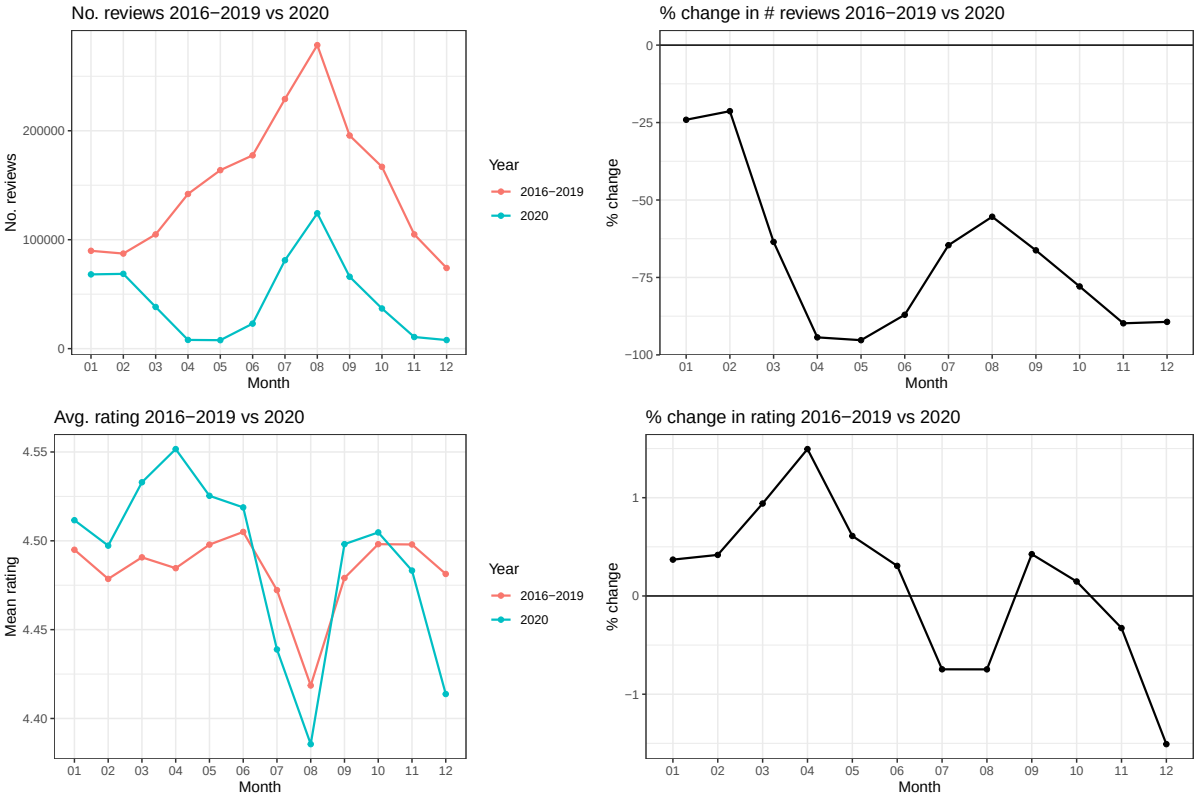


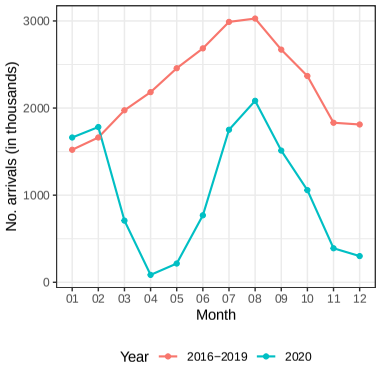
Figure 10: Tourism activity in Tripadvisor, 2016-2019 vs 2020



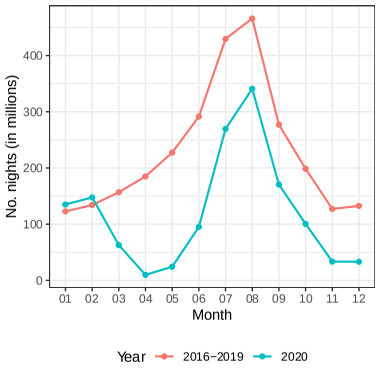
Source: Calculations based on Tripadvisor (2022) data.
Note: The figure illustrates the 2016-2019 mean and the 2020 level for INCULTUM countries.

Figure 11: Tourism activity in Eurostat data, 2016-2019 mean vs 2020

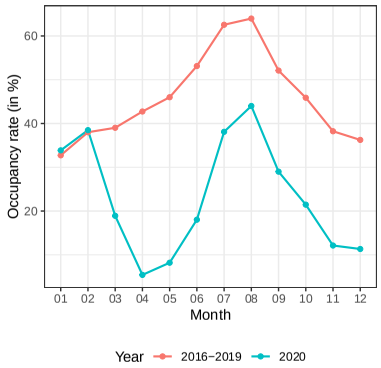
(a) No. tourism arrivals



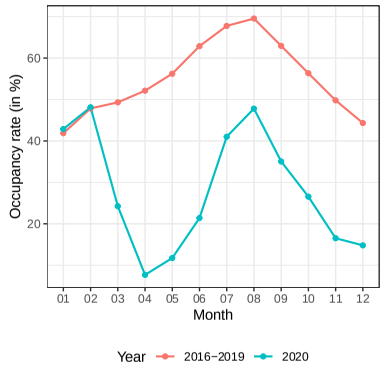
(b) No. tourism nights



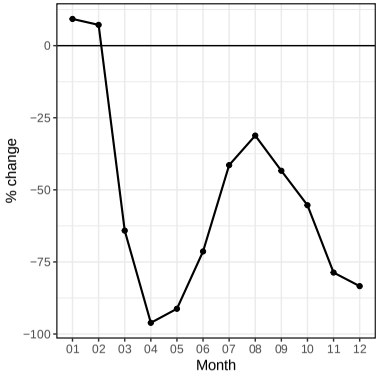
(c) Occupancy rate: Bed places



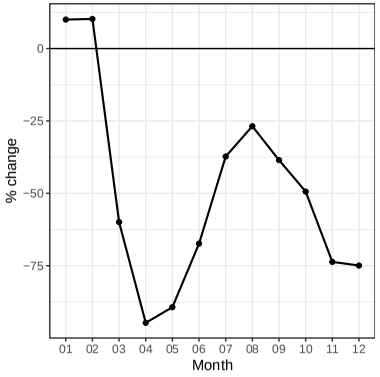
(d) Occupancy rate: Bedrooms



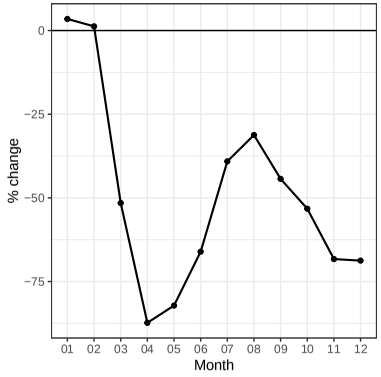
(e) % change in tourism arrivals



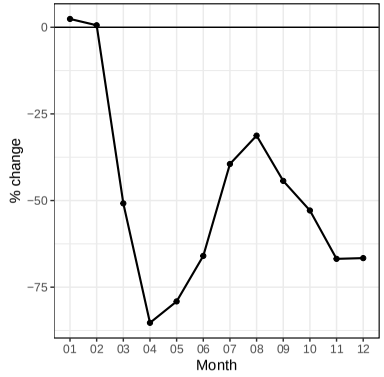
(f) % change in tourism nights



(g) % change in ccupancy rate: Bed places



(h) % change in occupancy rate: Bedrooms

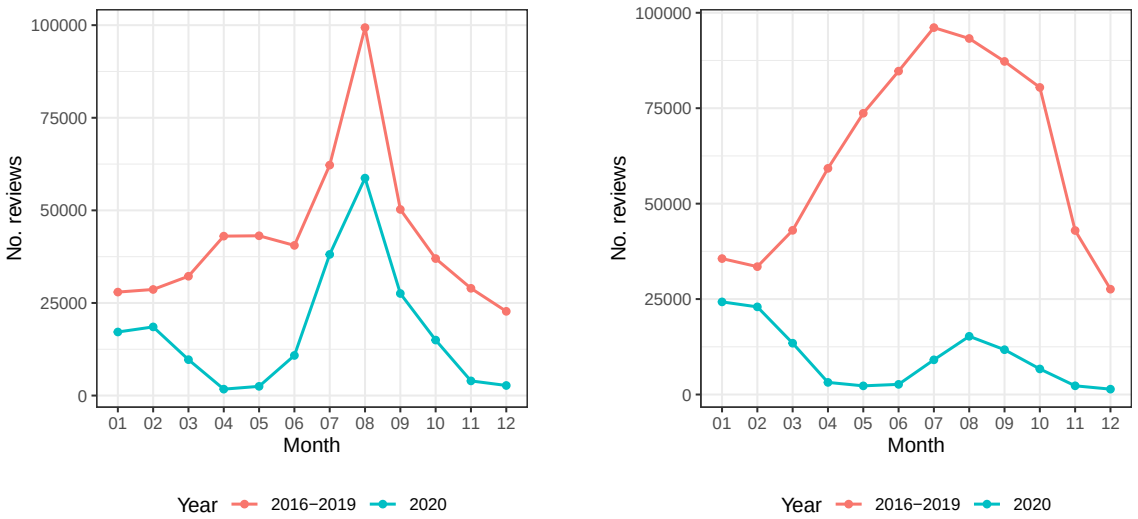


Source: Calculations based on Eurostat (2022) files TOUR_OCC_ARM, TOUR_OCC_NIM, and TOUR_OCC_MNOR.
Note: The figures in the top row illustrate the 2016-2019 mean and the 2020 level for EU-27 countries. The figures in the bottom row illustrate the percentage change from the 2016-2019 mean to the 2020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

Figure 12: Foreign vs domestic reviews in Tripadvisor, 2016-2019 mean vs 2020

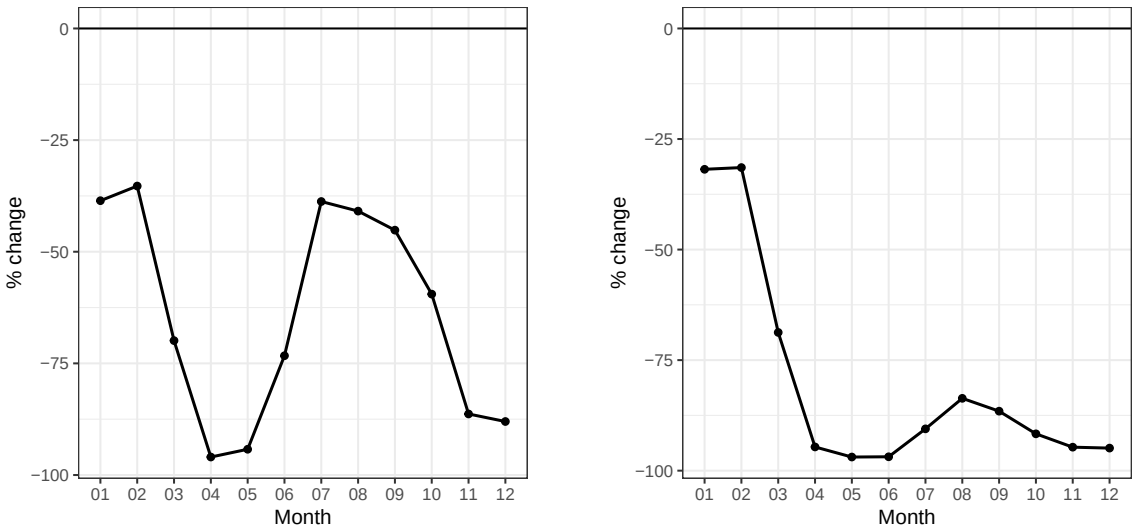
(a) # of domestic tourists

(b) # of foreign tourists



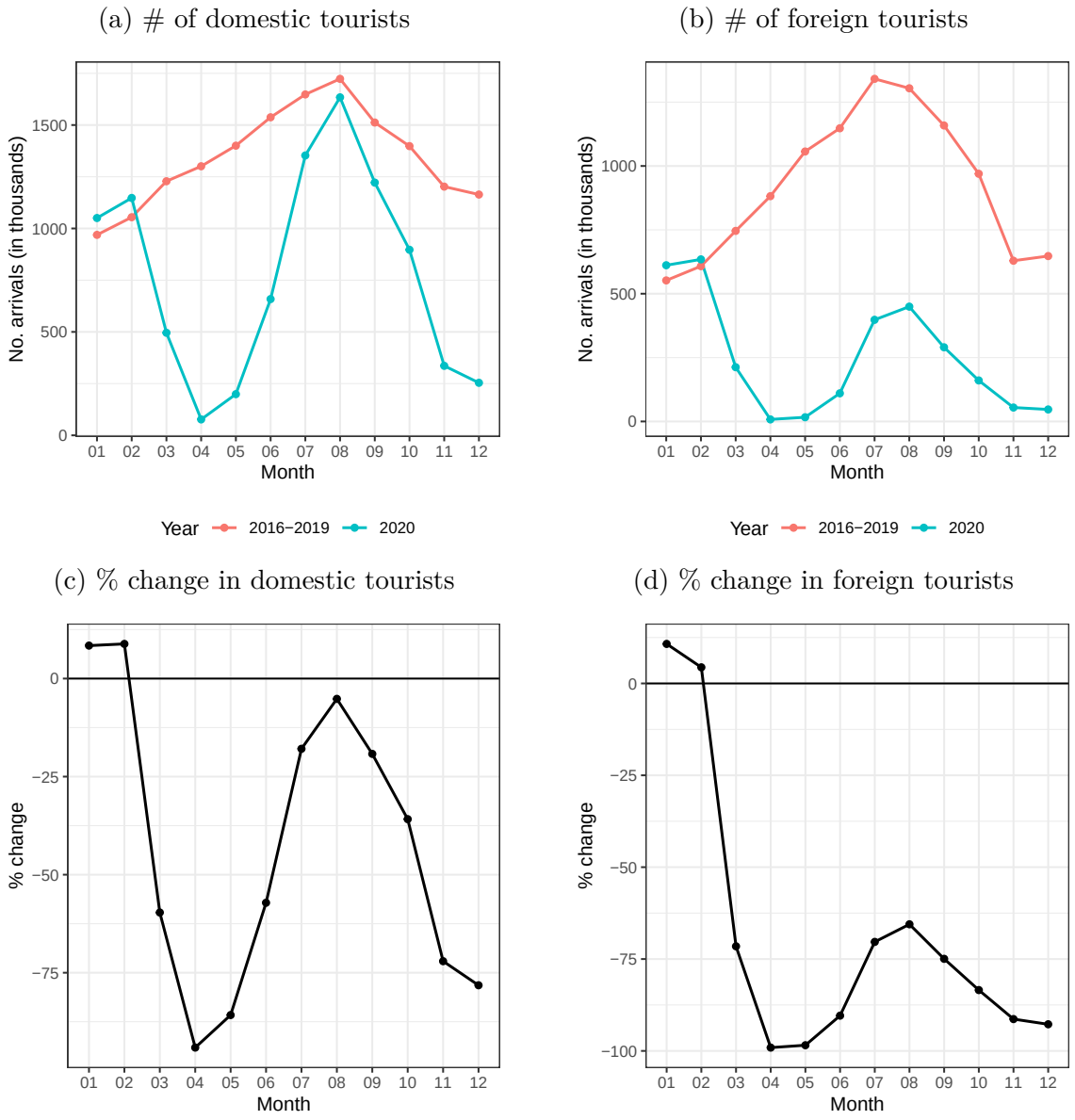
(c) % change in domestic tourists

(d) % change in foreign tourists



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

Figure 13: Foreign vs domestic tourism arrivals in Eurostat, 2016-2019 mean vs 2020



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.