Data Report

The Relationship Between Train Usage and Greenhouse Gas Emissions in EU Tourism Ecosystems

Methods of Advanced Data Engineering SS24

Presented by: Hend Abdo Hussien (23008093)

Contents

1. Research Question

How does the utilization of trains as mode of transportation within a tourism destination can influence the overall greenhouse gas intensity of the tourism sector within EU countries?

2. Data Sources

2.1. Datasets description

The chosen datasets are developed by European Commission Joint Research Centre. They provide tourism-relevant data and indicators collected from available, trusted sources concerning the tourism ecosystem on EU27 Member States aiming to characterize the tourism ecosystem at destination level (i.e., country), and track progress towards lower environmental impacts. The data sets can be described as follows:

1. Share of trips by train: This dataset provides information on the share of trips within a tourism destination that utilize trains as the mode of transportation. It gives insights into the preference for train travel which is relevant for understanding the transportation choices made by tourists. It is listed as one of the green pillars by EU Tourism Dashboard.

Source: Joint Research Centre Data Catalogue - UDP - Share of trips by train - European Commission (europa.eu)

2. Tourism GHG intensity: This dataset contains data on the greenhouse gas (GHG) intensity of the tourism sector, measured as GHG emissions per Million Euro of Gross Value Added (GVA) in the tourism sector. It helps in understanding the environmental impact of tourism activities within European Union countries.

Source: Joint Research Centre Data Catalogue - UDP - Tourism GHG intensity - European Commission (europa.eu)

3. Tourism Demand expressed by the number of nights spent in tourism destination: This dataset provides information on the total number of nights spent at tourist accommodation establishments in a destination (country or region) by both domestic and foreign tourists. It is provided by EU Tourism as a basic tourism descriptor to provide further context and characterization of the tourism activity of countries as it offers insights into the overall demand for tourism services in a particular destination. This measurement will help to assess the environmental impact of tourism activities within EU countries while accounting for the scale of tourism demand in each destination. Given that the high variance in tourism demand between destination countries can distort the GHG value.

Source: Joint Research Centre Data Catalogue - UDP - Nights spent - European Commission (europa.eu)

2.2. Data quality

The used datasets are structured tabular in csv format. The reliability, consistency and official nature of the Joint Research Centre of European Commission is are key in positively evaluating the accuracy of the data. More information about the Background and methodology used for the data collection and the relevance of the measurements can be found on the EU Tourism Dashboard.

The datasets used cover the time period 2019 to 2021 for all the EU27 Member States, Iceland, Norway and Switzerland. This time frame is chosen based on the availability and for maximizing the completeness of the data. However, some data was found missing from countries like Switzerland given that they were added to the data collection process recently. So the countries with missing values will be excluded from the dataset in the data transformation phase. It was worth noting that the pandemic time period is also included in the source data which might reflect anomalies given the

limited freedom of movement during the pandemic. While the data quality can be affected during this time period this limitation will be taken into account and discussed more in the section 4.

2.3. Data Profiling

Automated data profiling through Python scripts were used to analyzing the structure and content of datasets as part of the data exploitation phase (Figure 1). Distinct values of categorical fields where also explored to help in data transformation decisions. The two datasets Tourism GHG intensity and Nights spent in destination had unified structure, unlike the Share of Train Trips, which will required restructuring in the transformation phase.

Share of trips by train

#	Column	Non-Null Count
0	TERRITORY_ID	56 non-null
1	LEVEL_ID	56 non-null
2	NAME_HTML	56 non-null
3	UNIT	56 non-null
4	VERSIONS	56 non-null
5	2019	50 non-null
6	2020	50 non-null
7	2021	56 non-null
8	2022	50 non-null

Tourism GHG intensity

#	Column	Non-Null Count
0	VERSIONS	186 non-null
1	LEVEL_ID	186 non-null
2	TERRITORY_ID	186 non-null
3	NAME_HTML	186 non-null
4	YEAR	186 non-null
5	DATE	186 non-null
6	UNIT	186 non-null
7	VALUE	186 non-null

Nights spent in tourism destination

#	Column	Non-Null Count
0	VERSIONS	248 non-null
1	LEVEL_ID	248 non-null
2	TERRITORY_ID	248 non-null
3	NAME_HTML	248 non-null
4	YEAR	248 non-null
5	DATE	248 non-null
6	UNIT	248 non-null
7	VALUE	248 non-null

Figure 1: Data Structure and the missing values of the used datasets

2.4. Data license

The data sources used in this project are provided by the European Commission and are subject to the European Commission Reuse and Copyright Notice. Anybody can directly and anonymously access the data, without being required to register or authenticate. It allows for the reuse of the data, provided that the source is acknowledged. For more information please check Joint Research Centre Data Catalogue - Use conditions - European Commission (europa.eu).
To fulfill the obligations of the license, it's planned to include a clear acknowledgment of the European Commission as the source of the data in any reports, publications, or presentations that include findings or insights derived from the data.

3. Data Pipeline

3.1. Components and used technology

Python was the primary technology used to implement this pipeline using data processing functions from Panda library, while the SQLite database management system was used for loading into the data sink . The stages used in the pipeline are listed below and illustrated (Figure 2):

- **Data Collection**: Data was collected from three CSV files using http links from the Joint Research Centre Data Catalogue. These sources included information on train trips, tourism greenhouse gas intensity, and nights spent at tourist accommodations.
- **Data Transformation**: the data underwent transformation steps to prepare it for analysis. It included filtering out specific rows based on criteria and transposing some data columns to fit the next pipeline blocks.
- **Integration**: the transformed datasets were combined using inner joins based on common keys such as territory ID and year. This integration allowed for the creation of a unified dataset for further analysis.
- **Loading**: Finally, the integrated dataset was loaded into an SQLite table to as a final sink which allows for retrieval and sharing of the dataset for future usage.

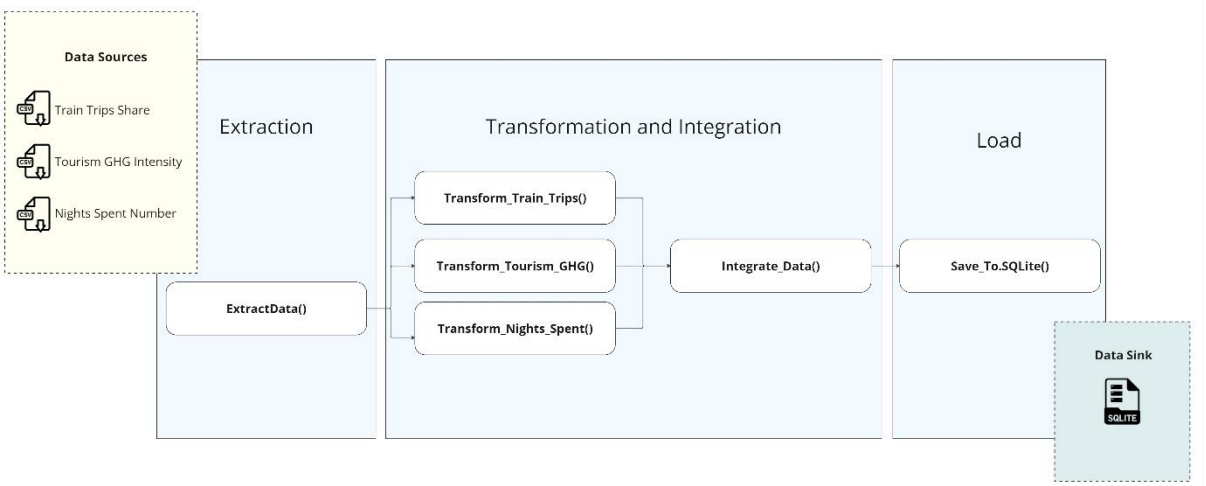

Figure 2: Data Pipeline Components

3.2. Data Transformation

The transformation process included mainly exclusion of unnecessary columns, filtering out data rows and transposing table structures:

- The datasets provided rows for older version of the data. The project is only interested in the most recent version from 2021. So the rows with older version were deleted and the column VERSION as well was eliminated.
- The datasets provided some rows for aggregated data for all the EU27 states together, which was irrelevant to the analysis, so these rows where filtered out and the column LEVEL was eliminated.
- The column for UNIT had a single value for each dataset. The unit of measure is fixed pro dataset and predefined. So the column UNIT was eliminated since it does not serve a function in the analysis
- The structure of dataset for Share of Train Trips had the values in column, one column for each year. This structure was inconsistent with the other two datasets where each years are used as key value which made more sense for filtering data based on year and integrating the three dataset. So this Share of Train Trips was restructured where the columns for years are transposed as rows.

4. Result and Limitations

4.1. Output Data Evaluation

The output data of the pipeline consists of a structured table (Figure 3) stored in an SQLite database. Each row of the table represents the indicators for a specific EU country pro a specific year from 2019 to 2021, along with corresponding the share of train trips, greenhouse gas (GHG) emissions value, and the number of nights spent at tourist accommodation establishments. Structured data is well suited for the analysis purpose given that there is a predefined set of attributes and data types for each column and its uniform structure making it easy to interpret and analyze.

	TERRITORY_ID	TERRITORY_NAME	YEAR	TRAIN_TRIPS_SHARE	GHG_VALUE	NIGHTS_SPENT_COUNT
0	AT	Austria	2019	0.16	29.0	127890554.0
1	AT	Austria	2020	0.13	47.0	79133399.0
2	AT	Austria	2021	0.14	54.0	66708839.0
3	BE	Belgium	2019	0.11	101.0	42512847.0
4	BE	Belgium	2020	0.09	156.0	20177486.0
5	BE	Belgium	2021	0.08	0.0	29220847.0
6	BG	Bulgaria	2019	0.03	72.0	27154791.0
7	BG	Bulgaria	2020	0.01	66.0	11968483.0
8	BG	Bulgaria	2021	0.02	49.0	17620268.0
9	CH	Switzerland	2019	0.31	48.0	56234630.0
10	CH	Switzerland	2020	0.23	0.0	38514354.0
11	CH	Switzerland	2021	0.26	0.0	45884488.0
12	CZ	Czechia	2019	0.08	31.0	57024767.0
13	CZ	Czechia	2020	0.07	51.0	31382494.0
14	CZ	Czechia	2021	0.05	53.0	31924242.0
15	DE	Germany	2019	0.23	45.0	436954848.0
16	DE	Germany	2020	0.18	62.0	260757872.0
17	DE	Germany	2021	0.22	61.0	266102654.0
18	DK	Denmark	2019	0.12	28.0	34325625.0
19	DK	Denmark	2020	0.11	41.0	23670530.0
20	DK	Denmark	2021	0.11	0.0	28556054.0

Figure 3: Sample of the output data table

4.2.Reflections and limitations

Reflecting on the output data structure and quality of the data pipeline, a potential enhancements to improve the analysis and interpretation of the results could be adding a derived data column for normalizing the GHG intensity per unit of demand aka Nights spent. A potential limitation would be that while the nights spent in destination is presented by the data provider as an expression of tourism demand and the share of train trips as an indicator of a sustainable practice in tourism destination, the higher granularity level where there is a distinction between domestic and foreign tourism activities could help when considering these factors, for example if a destination is has high prevalence of local touristic activities it may interfere with the representation of the train trips share.

Regarding the inclusion of the COVID-19 time period. The pandemic led to significant disruptions in travel and tourism globally, resulting in anomalies in the data due to restricted mobility and tourism activity during certain periods. While these anomalies may introduce challenges in interpreting the data, they also present an opportunity for insightful analysis. Given that COVID-19 pandemic affected the countries included in the dataset in relatively similar ways. So it is required to approach this analysis with caution and consider the unique circumstances and the contextual factors that may influence the observed results. For example, countries with stronger domestic tourism may be less affected compared to those who rely on international tourism.