

Is North America more suitable for tourists than South America?

The aim of my MADE¹ project is to answer this question based on open data.

1. Data Sources

As of now, I am using two main data sources, as well as one extra data source for visualization purposes.

1.1. Data Source 1: INFORM Risk

The INFORM Risk Index² is provided by the Joint Research Centre of the European Commission. Its purpose is to evaluate the risk of humanitarian crises. The risk of humanitarian crises is obviously not the same thing as the individual risk for tourists, which I want to analyze. However, I have chosen the INFORM Risk dataset despite this lack in relevancy, because the data is of high quality in all other dimensions (accuracy, completeness, consistency and timeliness). Additionally, the index is not only open data but also open source, meaning that it is possible (although out of scope for this one-semester project) to adapt the methodology towards the risk profile of tourists.

The INFORM Risk Index is published under the Creative Commons Attribution 4.0 (CC-BY 4.0) license³. As such, it is possible to use the data without major limitations, provided that proper attribution is given. In addition to the attribution already present here, I will add additional notices (directly or as metadata) to derivatives like the final report or the resulting dataset.

1.2. Data Source 2: International Comparison Program

The International Comparison Program⁴ is managed by the World Bank on behalf of the United Nations Statistical Commission. One of its aims is, to produce “comparable price level indexes (PLIs) for participating economies”. As such, it was a natural choice for me, to use this dataset to compare price levels of different American countries. From my investigations, the dataset is of high quality regarding accuracy and consistency. However, the data is only complete for earlier years, so I limited myself to data for the year 2021.

The results of the International Comparison Program are also licensed under the CC-BY 4.0 license⁵. Consequently, I will take the same measures as with the INFORM Risk dataset to ensure proper license adherence.

1.3. Extra Data Source: Natural Earth

The Natural Earth dataset⁶ is a public domain map of the world, which is primarily intended for visualization purposes but a closer inspection shows that there is also lots of useful metadata attached to the dataset. The data and metadata is of exceptionally high quality.

As the Natural Earth dataset is in the public domain (CC0)⁷, the data can be used without limitations and no attribution is required, although of course possible.

2. Data Pipeline

My data pipeline is programmed in Python⁸ and based on the Dagster framework⁹. After comparison of different competing frameworks, I chose Dagster because it offers a good compromise between framework size and offered features.

2.1. Structure of the Pipeline

In a first step, the datasets are extracted from their respective source URLs and parsed into a Pandas¹⁰ DataFrame or GeoPandas¹¹ GeoDataFrame. Thanks to the (Geo)Pandas libraries and the Python standard libraries, this is possible with only few lines of code, regardless of the original

format (JSON, CSV, Shapefile). It is also easily possible, to perform simple reshaping and cleaning operations on the data (see Section 2.2).

After the largely unmodified datasets have been stored (see Section 2.3), sub-datasets are derived in cases where only part of the original data is relevant for the rest of the data pipeline. As such, a dataset `icp_metrics_2021` is extracted from the `icp_metrics` dataset and an `americas` dataset is extracted from the `countries` dataset.

All original datasets or one of their sub-datasets are eventually merged into a `combined_dataset`. This dataset can then be used to produce derivatives like visualizations and it will serve as the ground truth for the travel score calculation in the future.

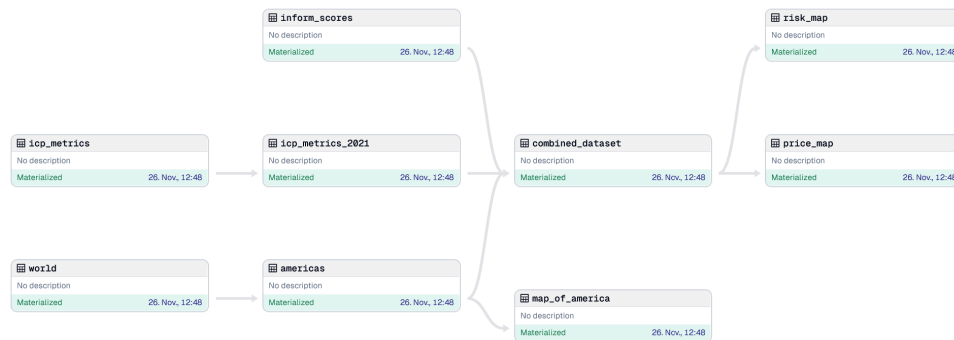


Figure 1: The pipeline structure as seen in the Dagster Web UI.

2.2. Reshaping and Cleaning

Depending on the specific dataset, different reshaping and cleaning operations are performed. All of them make use of the Pandas library, which provides as powerful toolset for such purposes. For example, multiple rows can be grouped together using `pivot` or unwanted metadata at the end of the file can be discarded with `drop`.

2.3. File System

Between two pipeline stages, Dagster will automatically persist all in- and outputs (“assets”) to the file system. By default, the Python-specific Pickle format is used. However, for this project a custom IO-manager was employed to achieve asset storage in the form of SQLite¹² databases.

2.4. Encountered Problems

The main problem encountered was choosing the right dataset from different available versions. Both, European Commission and World Bank provide numerous different options to access their data in different formats. However, none of those options truly fits the requirements of this project.

In the spirit of rapid prototyping, this problem has been worked around for now by manually creating download links, that contain just the desired data in an easily processable format. However, this is neither the intended solution from the viewpoint of MADE, nor is it clear whether the ICP download link will continue to work over longer periods of time. Therefore, it might be a good idea to revisit this problem again towards the end of the project, in order to find a more permanent solution.

2.5. Meta-quality Measures

There are no automatic meta quality measures in place yet. However, it is easily possible to manually review the meta quality using Dagsters Web UI, by looking at the metadata which is added to each assets by the custom IO-manager. Additionally, the use of the SQLite format as discussed in Section 2.3 allows for straightforward inspection of the full datasets.

inform_scores

[View in Asset Catalog](#)

Latest materialization						
Run	Run 2930122e					View logs
Timestamp	26. Nov., 21:37					
num_records	191					
preview	Iso3	CC	HA	INFORM	VU	
	AFG	7.1	7.6	7.7	8.3	
	AGO	6.7	2.5	4.4	5.1	
	ALB	4.1	3.5	3.2	2.2	
	ARE	1.8	2.5	1.7	1	
	ARG	3.6	2.5	3	2.9	

Figure 2: Per-dataset metadata as seen in the Dagster Web UI.

3. Result and Limitations

The main output of my current data pipeline is the combined dataset described in Section 2.1. The output dataset inherits its nature from its ancestor datasets and is therefore of high quality regarding accuracy and completeness. The timeliness suffers a little, due to the ICP data being from the year 2021, while the consistency suffers as the INFORM data is taken from the most recent 2024 report. Regarding its relevancy, the dataset has yet to prove itself during the final report.

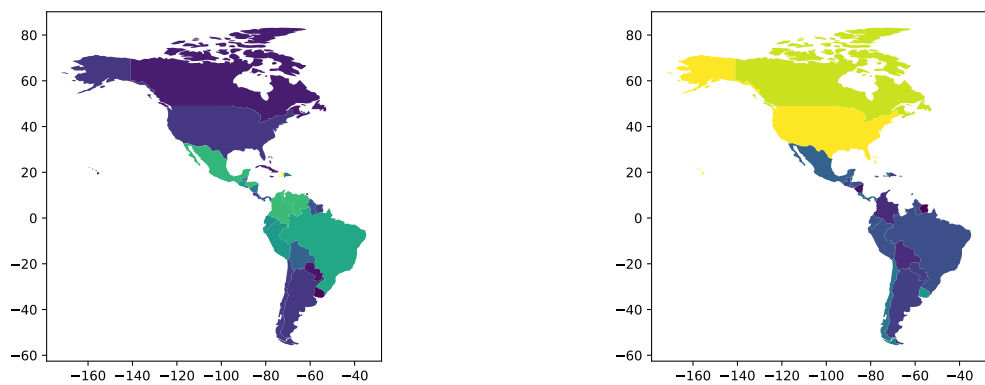


Figure 3: Visualizations of the combined dataset, indicators INFORM (left) and PX.WL (right).

As discussed in Section 2.3, the SQLite format was chosen for all assets. In practice, this format showed better compatibility with (Geo)Pandas than CSV while at the same time it is far more universal than the Pickle format.

4. License Glossary

CC-BY 4.0 – Creative Commons Attribution 4.0: <https://creativecommons.org/licenses/by/4.0/> 1

CC0 – public domain: <https://creativecommons.org/publicdomain/> 1

5. Bibliography

1. “Methods of Advanced Data Engineering,” <<https://oss.cs.fau.de/teaching/specific/made/>>
2. “INFORM Risk,” <<https://drmkc.jrc.ec.europa.eu/inform-index/INFORM-Risk>>
3. “INFORM Risk - Copyright,” <https://commission.europa.eu/legal-notice_en>
4. “International Comparison Program,” <<https://www.worldbank.org/en/programs/icp>>
5. “International Comparison Program - Details,” <<https://datacatalog.worldbank.org/search/dataset/0066092/International-Comparison-Program-2021>>
6. “Natural Earth,” <<https://www.naturalearthdata.com/>>
7. “Natural Earth - Terms of Use,” <<https://www.naturalearthdata.com/about/terms-of-use/>>
8. “Python,” <<https://www.python.org/>>
9. “Dagster,” <<https://dagster.io/>>
10. “Pandas,” <<https://pandas.pydata.org/>>
11. “GeoPandas,” <<https://geopandas.org/>>
12. “SQLite,” <<https://www.sqlite.org/>>