LaborDE

Project Brief

FEBRUARY 14

Data Analyst: Nadia Ordonez Roman
CareerFoundry

Context

LaborDE, a prominent laboratory supplier based in Nordrhein-Westphalia, Germany, has excelled in providing diagnostic laboratories with cutting-edge equipment and materials, including a specialized department focused on cancer diagnostics products. In 2019, the German NGO **Open Knowledge Foundation Deutschland e.V.** and the British NGO **opencorporates** collaborated to release German Trade Register data via **OffeneRegister.de**, offering a comprehensive list of over 5 million companies to the public. Recognizing the strategic value of this data, LaborDE aims to leverage it for identifying potential laboratory customers, not only in their stronghold of Nordrhein-Westphalia but also for exploring new market opportunities in other Bundesländer. This initiative positions LaborDE at the forefront of the industry by adopting a data-driven approach to inform decision-making and foster growth.

Data Source

German companies dataset

The dataset, compiled by OpenCorporates primarily between June 2017 and January 2019, contains basic information on over 5 million German companies and their officers, sourced mainly from **Handelsregisterbekanntmachungen** and to a lesser extent, **Handelsregister** search results listings. OpenCorporates is generously sharing this dataset under an **open license**, aiming to demonstrate the advantages of releasing company information as open data. As a social enterprise committed to enhancing societal benefits, OpenCorporates actively advocates for open company data, collaborating with civil society, businesses, and governments to promote accessibility and transparency in critical information without financial barriers.
The database was downloaded as a **SQLite Datenbak** and later opened with Python using Jupyter notebooks. The original table named “company” was saved as “all_companies_de” in a pickle format, selecting relevant variables such as company names and addresses.

Bundesland German population dataset

This dataset was downloaded from the Kaggle community, however, the original data owner is the Statistisches Bundesamt, which is a federal authority of Germany responsible for collecting, processing, presenting, and analyzing statistical information concerning the topics of economy, society, and environment. This dataset contains the population, area, and population density numbers for all German towns and districts released on 31.12.2018. From these data, I will only retrieve the information at the Bundesland level. Also, I additionally retrieved GDP information from 2022 at the Bundesland level from the Statistische Ämter des Bundes und der Länder.

Regional German population dataset

I downloaded two files from Postleitzahlen Deutschland, containing postcode information in Germany. I later unified them by merging them using the postcode data as the key variable. These files were last updated on July 15, 2023, and are freely available under the Open Database License. Their source of raw data is linked to OpenStreetMap contributors. Population figures as the basis for calculations were originally extracted from the Statistical Offices of the Federal Government and the States.

Data Profile

In this project, there are two main datasets.

1. German company Bundesland
2. German company regional

German company Bundesland dataset

In this dataset, the German companies were grouped at the Bundesland level and merged with demographich, and financial atributes of the 16 Bundesländer of Germany.

Variable	Description	Time Variant / Invariant	Structured / Unstructured	Quantitative / Qualitative	Nominal /Ordinal /Discrete/ Continuous
bundesland_en	Bundesland company location - English version	variant	structured	qualitative	nominal
company_total	Total number of overall companies	variant	unstructured	quantitative	discrete
labor_total	Total number of potential laboratory customer ompanies	variant	unstructured	quantitative	discrete
cancer_total	Total number of potential laboratory customer ompanies, cancer-related	variant	unstructured	quantitative	discrete
plz_total	Total number of postcodes listed per Bundesland from the registered companies (from imputed postcodes)	variant	unstructured	quantitative	discrete
administration_unit_id	Bundesland id	invariant	structured	qualitative	ordinal
bundesland_de	Bundesland company location - German version	variant	structured	qualitative	nominal
area_sqkm	Bundesland area in square kilometers	invariant	unstructured	quantitative	continuos
population	number of people per Bundesland	variant	unstructured	quantitative	discrete
male	number of males per Bundesland	variant	unstructured	quantitative	discrete
female	number of females per Bundesland	variant	unstructured	quantitative	discrete
population_per_sq km	population per square kilometer per Bundesland	variant	unstructured	quantitative	continuos
gdp_mill_euros	GDP per Bundesland in million euros	variant	unstructured	quantitative	continuos

German company regional dataset

Variable	Description	Time Variant / Invariant	Structured / Unstructured	Quantitative / Qualitative	Nominal /Ordinal /Discrete/ Continuous
plz_imputed	Company postcode imputed from ‘bundesland’	variant	structured	qualitative	nominal
company_total	Total number of overall companies per postcode location	variant	unstructured	quantitative	discrete
labor_total	Total number of potential laboratory customer ompanies per postcode location	variant	unstructured	quantitative	discrete
cancer_total	Total number of potential laboratory customer ompanies, cancer-related, per postcode location	variant	unstructured	quantitative	discrete
bundesland_en	Bundesland company location - English version	variant	structured	qualitative	nominal
plz_original	Company postcode as extracted from ‘registered_address’	variant	structured	qualitative	nominal
habitants	number of people per postcode location	variant	unstructured	quantitative	discrete
area_sqkm	Area in square kilometers associated to a postcode	variant	unstructured	quantitative	continuos
city	Company city location as extracted from ‘registered_address’, also imputed	variant	structured	qualitative	nominal
region	Company region location, also including imputed values	variant	structured	qualitative	nominal
bundesland_de	Bundesland company location - German version	variant	structured	qualitative	nominal

Business questions

To strengthen our market presence in NRW and expand our reach into other Bundesländer, stakeholders at LaborDE are committed to addressing key business questions. These inquiries will not only fortify our customer base but also strategically position us in the highly competitive landscape of laboratory diagnostics.

National Landscape - Bundesland:

- What is the nationwide distribution of potential customer companies specializing in laboratory diagnostics, and which Bundesländer has the highest concentration of potential customers? Rationale: Expanding our scope beyond NRW, this analysis aims to identify promising markets in other Bundesländer, providing valuable insights to inform our strategic expansion plans across Germany.
- What is the prevalence of companies offering diagnostic services specifically for cancer, and where are they situated in Germany? Rationale: Recognizing the demand for cancer-related diagnostic services is crucial for tailoring our offerings to meet specific healthcare needs, thereby positioning LaborDE as a specialized and sought-after service provider.

Regional Focus - NRW:

- What is the distribution of potential customer companies specializing in laboratory diagnostics across NRW, and which regions exhibit the highest concentration? Rationale: By understanding the geographical concentration of potential customer companies in NRW, we can tailor our outreach strategies to maximize impact in key cities, thereby enhancing our regional presence.
- What is the extent of the customer base served by laboratory diagnostic companies in NRW in terms of male and female population? Rationale: By quantifying the current customer base in NRW, we refine our service offerings and better meet the needs of our existing and potential customer companies serving these customers.

Data Limitations and Ethics

The German company dataset from OpenCorporates faces significant constraints. Terms of use from Bundesanzeiger and Unternehmensregister prohibit OpenCorporates from publishing data from these registers, rendering information on companies registered there unavailable. Additionally, the dataset is current only until January 2019, excluding companies established thereafter.

Addressing data ethics, the original German company dataset contains personal data subject to GDPR regulations. Therefore, stringent privacy measures and compliance with data protection laws are essential when managing and utilizing this dataset.

The simplicity of company details in the original file poses challenges in extracting precise laboratory diagnostics information. Potential biases in analysis arise from filtering and data imputations, especially when relying solely on company names for filtering potential customer companies, as these companies could operate under brand names unrelated to their business activities. The dataset's 58% of German companies with incomplete addresses led to postcode and region imputations based on frequently occurring values, introducing biases towards commonly listed postcodes and regions. The most reliable location attribute is at the Bundesland level, a crucial consideration in interpreting analysis results.