Exploration of company information

Fabio Morea

2022-02-01

Contents

1 Scope and objectives

This notebook explores the datasets that will presumably underpin future research work for the PhD in Applied Data Science and Artificial Intelligence.

1.1 Background information:

Research, innovation and highly skilled people are considered to be important factors in economic and social development. Economic support policies often include funds to support research (for example with the creation of public research infrastructures), companies (for example with tenders to co-finance innovative projects) and the training of people with the necessary skills.

Area Science Park is a national research institution that manages a science and technology park located in Trieste (Italy). Its activities can be considered a public investment in support of research and innovation, for a value of approximately 20 million euros per year.

Currently Area is hosting 70 tenants (60 companies and 10 research centers) engaged in research activities in the fields of ICT, lifesceinces and materials. Their success (or lack of it) depends on a key - and often overlooked - asset: the community of over 1600 employees, researchers and entrepreneurs.

Area is interested in measuring the effectiveness and efficiency of its activities, focusing in particular on

- monitoring the economic performance of tenants,
- monitoring the community of skilled workers,
- comparing with similar groups, mainly at a regional or national scale, but also with respect to the science and technology parks in Austria and Slovenia.

To support research work, Area Science Park can provide some relevant datasets, cureated as a part of the project *innovation intelligence*. Innovation Intelligence aims to analyze information on companies in the Friuli Venezia Giulia region, using several data sources such as the chamber of commerce, the Regional Labor Market Observatory, a rating agency, as well as surveys on samples of companies.

1.2 Objectives of future research work - reserch questions

Research questions are currently defined on a general level:

- are tenant companies performing better than similar companies?
- how to measure similarity between two companies?
- how to exthed such measure to groups of companies?
- how to identify clusters or communities of companies?

The research questions above and the methodology outlined in this notebook are relevant also in other contexts, such ad sectoral cluster, public agencies supporting innovation and any kind of industrial area. The data set supports analysis focused on Friuli Venezia Giulia region, but can be extended to other regions (gathering relevant data from the Chamber of Commerce or from commercial data providers).

1.3 About this notebook

The notebook is divided in 6 sections: an introduction, a section for each dataset and a final section on potential future development.

1. Imprese_FVG
2. Bilanci_FVG
3. Rating_FVG
4. CO_FVG
5. Features: A basic example of sample feature selection, on a small subset, where each company is represented by 5 features
6. Further development: calculating the age of companies based on several dates, handling non metric features: defining a custimized similarity function to identify *similar* companies and estimate distances in a multi-dimensional space.

The notebook has been written using *R-Studio* and rendered with *boowdown* (https://bookdown.org/) package. Data data manipulation is based on *tidyverse* [https://www.tidyverse.org/], a data science library that includes *magrittr* (pipe operator %>%), *dplyr* (select, summarize...), *tibble* (a tidier version of the data.frame) and *ggplot2* (visualizations). A useful guide to tidverse is available online at the following address: [https://r4ds.had.co.nz/]

> TODO Some parts of the notebook are higlighted as “To Do”, to highlight potential improvements in analysis, code efficiency or need for further clarifications.

1.4 Data management plan

Raw data: The original data has been pre-pocessed by Area Science Park to fulfill the folloqing rquirements:

- encoded in UTF-8 cleaned from non-printable characters
- table columns are attributes (features, independent variables), renamed to be human- and machine-readable
- table rows are observations If you have multiple tables, they should include a column in the table that allows them to be linked
- splitted into several tables, created unique identifiers to connect the tables
- saved each table to separate .csv file with a hunam-readable name At this stage no attributes were removed or summarized. Raw data is available in local folder *data/raw*

Tidy data: This notebook explores all the attributes available in the raw data, and by merging, subsetting and transforming, produces a smaller, cleaner data set ready for further analysis. Tidy data is saved in local folder *data/tidy*

This notebook describes the process to create tidy data, and the meaning of each variable: - meaning, summary and visualizations ofeach attribute in tidy data - information about attributes that are not contained in the tidy data (basic meaning and reason why thy have not been included)

Updates: raw and tidy data may be updated periodically, since Area Science Park updates the raw data set twice a year; anyway the current version in based on June 2021 version and does not provide automatic updating scripts.