Lab Report

186.868 Visual Data Science 2024W

11912007 - Yahya Jabary

Code: `github.com/sueszli/uwaterloos-sunshines`

Contents

This project adheres to the traditional stages of a data science pipeline: Discover, Wrangle, Profile, Model, and Report. Each stage involves specific tasks and requirements, which are detailed in the subsequent chapters. The primary objective is to achieve a comprehensive understanding of the data and effectively communicate the insights through an interactive dashboard to a wider audience.

Significant emphasis is placed on the quality of the code and the reproducibility of the results. The `makefile` located in the root directory of the project repository includes commands for generating a `pip-compile`'d requirements file, ensuring an isolated virtual environment. Additionally, short scripts are available for execution within Docker or Conda environments.

Discover Stage

The tasks of this section are to:

- Discover your topic
  - Either select one of the suggested topics, or come up with your own idea. → Our custom topic was confirmed by the lecturer via email.
  - The topic should be possible to easily understand the data context, even if you are not an expert.
- Describe the two datasets you selected
  - Use at least two independent datasets.
  - The selected datasets must be multidimensional: they have to contain more than 5 variables.
  - The selected datasets must contain a sufficient amount of data rows.
- Keep the length to 1-2 A4 pages

Motivation and Context Salary transparency is a complex and often contentious issue in the workplace. While it can promote equity and reduce discrimination, it may also foster jealousy and resentment among employees. The approach to salary disclosure varies significantly between European countries and North America, particularly in the context of public institutions and academia.

In European countries, it is common practice to disclose compensation brackets for public officials, including employees at public universities. However, these disclosures typically do not reveal exact salaries or identifiable information, limiting the ability to study personal income inequality in detail.

North America takes a different approach, with many states and provinces enacting laws that mandate public institutions to disclose the salaries of top-earning employees. These individuals, often referred to as "Sunshines," are those earning more than $100,000 annually in salary, excluding additional benefits. The published lists, known as "Sunshine Lists," include the names of these employees, enabling more in-depth studies of personal income inequality.

The academic setting provides a particularly interesting context for studying income correlation. Academics typically publish their work openly and have established performance metrics, such as the total number of publications, citations, and h-index. This transparency in academic output creates a unique opportunity to examine the relationship between income and academic performance.

Given our personal connection to the University of Waterloo, we have chosen to focus their study on this institution's sunshine list. This decision provides a familiar and accessible dataset for analysis.

Datasets and Methodology The primary dataset for this study consists of the publicly available sunshine lists from the University of Waterloo. This dataset provides both tabular and time-series data, as it includes annual checkpoints. To enrich this information, we plan to join it with data from `csrankings.org`, which contains scholar IDs. These IDs can then be used to query additional resources such as Google Scholar and the Semantic Scholar API, providing more detailed information about the research output of individual employees.

By combining these datasets and API resources, the study aims to investigate the correlation between compensation and academic performance at the University of Waterloo. While specific research questions are yet to be formulated due to the exploratory nature of the approach, potential areas of inquiry include:

- The relationship between changes in compensation and role over time
- Correlation between academic performance metrics and salary
- Potential gender-based salary disparities, inferred through natural language processing techniques applied to employee names

We have selected the following data sources to support our study:

- University of Waterloo Salary disclosure for 2020-2023
- CS Rankings CSV based on the csrankings.org Github repository
- Google Scholar API
- Semantic Scholar API

Ethical Considerations and Data Handling While web scraping is legal in the European Union and the data used in this study is publicly available, we have committed ourselves to handling the information ethically. To avoid potential conflicts of interest and protect individual privacy, we will not visualize any identifiable information about specific employees or derive any conclusions that could harm individuals.

The focus of the study will be on detecting general trends and patterns rather than singling out individuals. All published data and insights will be aggregated to maintain anonymity. This approach ensures that the research can proceed without compromising ethical standards or potentially harming individuals, even in cases where significant discrepancies between pay and performance might be discovered. By adhering to these ethical guidelines, we aim to contribute valuable insights into the relationship between academic performance and compensation while respecting the privacy and dignity of the individuals whose data forms the basis of the study.

Wrangle Stage

The tasks of this section are to:

- Join the two or more datasets you selected into one big data table
  - How did you join the datasets? Which keys did you use to join the data? Did all keys match? Did you have to introduce new keys?
- Solve issues like formatting issues, missing data, faulty values, and non-matching keys.
  - Which data cleaning steps have been necessary? Did you experience data issues, and if so, which ones? How did you solve them? Did you use automated methods? Did you use visualization to inspect data issues?
- Visually show and explain the data quality of your dataset (for example, before and after cleaning steps). Come up with your own, creative, solution here.
- Use a charting library, not fully-featured applications.
- Keep the length to 3/4 to 1 A4 page.

Profile Stage

In this section we explore the data in detail, to completely understand its structure, and to discover any interesting patterns that can be found in there.

The tasks of this section are to:

- Find at least 3 informative insights in your dataset. For each one add a short text describing the insights plus one visualization.
- Use a charting library, not fully-featured applications.
- Keep the length to 3/4 to 1 A4 page per insight.

Model Stage

The tasks of this section are to:

- Build a model of the data you studied, to find answers to the questions you selected (possible models: linear regression, clustering, pca, anomaly detection, etc.)
- describe the modeling process
- Create one or more visualization(s) that describe the results of your model
  - How would you increase trust of your customers/colleagues in your modeling approach by using data visualization?
- Use a charting library, not fully-featured applications.
- Keep the length to 1-2 A4 pages.

Report Stage

The tasks of this section are to:

- Show your findings in an interactive dashboard to a broader audience. The findings may be related to the Model/Wrangle/Profile stages.
  - Use appropriate charts, visual encodings (e.g., color).
  - Use at least four different types of visualizations / charts.
  - Include interaction (e.g., filters, zoom, not a jupyter notebook), brushing & linking (changes in one view affect others, but not global filters)
- Use a library, not fully-featured applications.
- See examples: https://tuwel.tuwien.ac.at/mod/page/view.php?id=2433356