Introduction

Thank you for taking the time to read the technical documentation for the algorithm used to source a more expansive pipeline by New Profit, a venture philanthropy firm. In the following sections you will be provided with additional information on the purpose and intended use of the algorithm, an overview of the algorithm, and the scope of the remaining documentation.

Purpose and Intended Use

For more information on the purpose and intended use, please refer to the following report published by New Profit.

Link to Report (link pending)

Overview of the Algorithm

A rule-based algorithm was designed to identify the list of non profit organizations working in the field of economic mobility. Rule-based algorithms are ideal when the decision-making process can be clearly defined through explicit rules and a premium is placed on explainability of algorithmic outputs. These rules are presented in the Notebook Explanation section.

Scope of Documentation

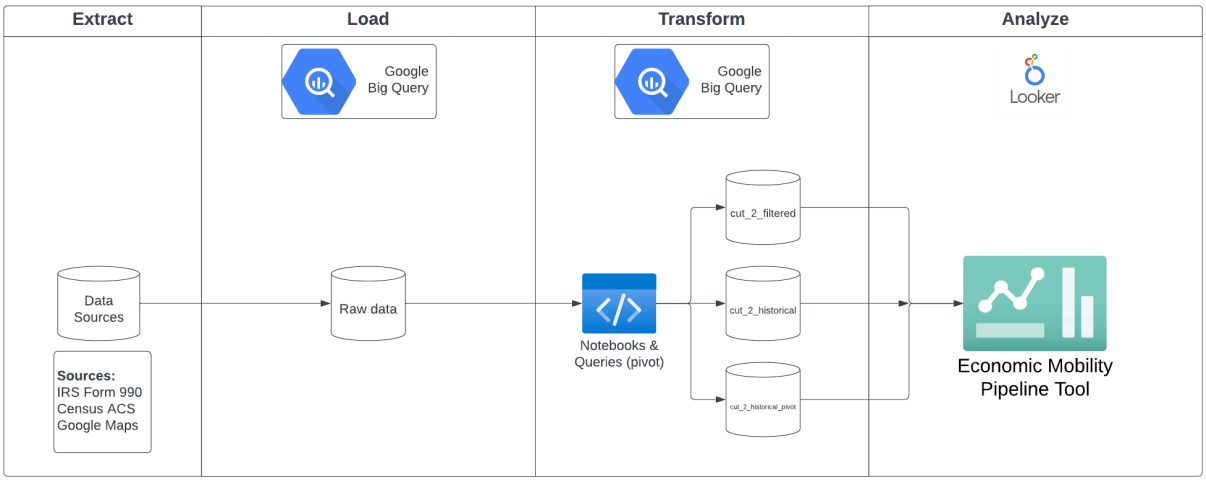
The scope of documentation provided here is designed for someone interested in executing a similar process for their own use case. As such, notebooks are provided with code and explanations.

Algorithm Description

Technical Architecture

The technical architecture for the algorithm is presented in the figure below. It follows a standard Extract, Transform, Load, and Analyze process. The extract step involves acquiring the necessary data. In this case, Form 990 filings in XML format[1], county-level data from the American Community Survey[2], and the Google Maps Geocoding API[3] used for converting addresses into latitude and longitude coordinates. This data is loaded into Google Big Query as part of the Google Cloud Platform where the transformation step is also applied. A series of notebooks are used to accomplish this and result in the creation of three separate tables for the Analyze Layer hosted on Google's Looker Studio. The purpose of this documentation is to make available the steps involved in taking the data sources listed and transforming them into a usable format. As such, the primary focus here will be on the Extract, Transform, and Load (ETL) steps outlined.

[1] Form 990 Series Downloads from the Internal Revenue Service
[2] American Community Survey Data from the Census Bureau
[3] Google Maps Geocoding API



Notebook Explanation

The steps involved in taking the Form 990 data and turning it into a usable tabular format is outlined in the following eight notebooks.

1. *Step 01 - Data Download*
2. *Step 02 - Decompression*
3. *Step 03 - Data Extraction*
4. *Step 04 - Base Data Loading*
5. *Step 05 - Reference Data Loading*
6. *Step 06 - Geocoding*
7. *Step 07 - Create Cut 1*
8. *Step 08 - Create Cut 2*

Step 01 - Data Download

Step 01 involves downloading the IRS Form 990 data. There is no straightforward way to load zip files directly to Google Big Query. Therefore, files are downloaded locally and then uploaded to Google Big Query. The whole process should take 30-40 minutes to complete and requires at least 30 GB of disk space locally.

Step 02 - Decompression

Step 02 involves managing the zipped files downloaded from the IRS. The files are compressed using the Deflate64 method. This format is not supported by the Google Cloud Platform suite of tools. It is recommended that after users download the zipped files locally, they should unzip them, and then re-zip them using a more common compression method. For this process, the tar gzip method was used (tar.gz). Zipped files can be uploaded to Google Big Query once they have been converted to a more common compression method.

Step 03 - Data Extraction

Step 03 involves unzipping the zipped files and extracting their XML files. Here, the fields we are interested in extracting are specified. A storage container is specified for each year of Form 990 data we have. Processing a single year of data takes approximately three hours.

Step 04 - Base Data Loading

Step 04 consists of loading the extract base data used to identify the list of organizations that satisfy basic filing status, revenue, and expense thresholds.

Step 05 - Reference Data Loading

Step 05 involves loading key reference data used for conducting the down selection of organizations and enriching the resulting dataset. This includes the Google Geocoding API, loading in additional organization's identified by the Harvard Project on Workforce[4], and American Community Survey data.

Step 06 - Geocoding

Step 06 takes the addresses provided in the Form 990s and converts them into latitude and longitude points for visualization in maps and mapping to counties to acquire county-level statistics from the American Community Survey.

Step 07 - Create Cut 1

Step 07 is the first down selection step. It identifies organizations that satisfy the basic requirements outlined in Step 04.

Step 08 - Create Cut 2

Step 08 is the final step and is where most of the remainder of the use-case specific down selection rules are implemented. There are two important reference files necessary for conducting this process.

1. Bag of Words: a list of words and associated priority level for selecting organizations for inclusion in the final dataset
2. Bag of Words Remove: a list of words that remove an organization from being considered for the final dataset

Each list is provided in Appendix: Bag of Words[5]. Natural language processing is also applied at this step using spaCy's "en_core_web_sm" language model[6]. The process takes 2-3 hours to

[4] The Workforce Almanac Report: A System-Level View of U.S. Workforce Training Providers
[5] Appendix: Bag of Words
[6] spaCy's en_core_web_sm language model

complete. The words, phrases, and priority levels outlined are the result of a series of discussions with New Profit stakeholders. As such, the specific rules have limited value for alternative use cases. However, for those interested in adapting this logic for other use cases, simply modify the specific rules and bag of words lists to suit your particular scenario.

Down Selection Rules

Rule 1: organization does not mention any of the words listed in the bag of words remove list in either their name or mission statement

Rule 2: organization does not mention any non-US locations in their mission statement

Rule 3: organizations must meet total revenue and total expense criteria

Total expenses: 250,000 <= X <= 18,000,000

Total revenue: 250,000 <= Y <= 24,000,000

Using the "mission" attribute, organizations that meet the following conditions are passed through

Rule 4: organizations that mention one or more 1A priority words/phrase are passed through

OR

Rule 5: organizations that mention two or more 1B priority words/phrases are passed through

OR

Rule 6: organizations that mention three or more 1B or 1C priority words/phrases total are passed through (i.e. 1 1B and 2 1C would get through. So would 3 1C)

OR

Rule 7: organizations that mention six or more total words (1B, 1C, and 2)

AND

Rule 8: organizations that mention only one 1A priority word and no other priority words are removed from the dataset

After the list of organizations is generated, the final step consists of extracting historical data from previous filings as far back as those Form 990s filed in 2017. The resulting historical table presents a longitudinal view of each of the organizations based on the available data.

Limitations

As with any process, there are a few known limitations. In the sections below we will outline the known limitations and recommended mitigation strategies.

Limitations and Mitigations

Limitation	Mitigation
IRS Form 990 Data Delay: some estimates put the publication delay for Form 990 data at 1.5-2 years.	The IRS is working on the publication delay. Establishing protocols to acquire the data once published is the most straight forward approach. Additionally, users can explore alternative means to acquire some information. As an example, if quantitative time series data is available, then a forecasting model could be developed. Forecasting models are simply predictions, but could be helpful for planning purposes in the absence of actual data.
Organizations that don't file Form 990s: there are some organizations that do not file form 990s such as fiscally sponsored organizations.	At present, these organizations are considered out of scope. Acquiring information about them will require relying on traditional data collection approaches.
Organizations filing other types of Form 990s: there are different requirements involved in filing one Form 990 type over another.	The approach outlined here focused specifically on those organizations that file a standard Form 990. However, if a given use case involved other Form 990 types, then the notebooks could be modified accordingly. Note that doing so would require understanding of the structural differences between Form 990 types.

Conclusion

The algorithmic process outlined here represents an exciting step forward in making publicly available data more accessible and actionable. This work is intended to complement an already robust engagement process with leaders, organizations, and communities. The algorithm and its output cannot replace the critical role that humans play in advancing philanthropic work. If you have any questions, please contact the following.

Tessa Forshaw
tessa@peoplerocket.com

Jake Hale
jake.hale@peoplerocket.com