Analysis of Granting Patterns of Foundations Across the Southwest

Final Report

Prepared By: Ankit Basu [934008673] and Archit Gupta [135001620]

Supervised By:

Dr. William A Brown
Professor, Director of the Center for Nonprofits and Philanthropy
Bush School of Government & Public Service

Dr. Lan Zhou
Associate Professor, Statistics
College of Arts and Sciences

Texas A&M University

Date: 11/20/2024

Introduction

The project, "Analysis of Granting Patterns of Foundations Across the Southwest", provides an in-depth analysis of how philanthropic institutions allocate their grants across the southwestern region of the United States, comprising Arizona, Arkansas, Colorado, New Mexico, Nevada, Oklahoma, and Texas. By analyzing grant-making patterns, the project uncovers trends in donations and grant distribution to identify disparities and patterns in donor behavior and recipient demographics.

This analysis is critical as it helps identify how grant funding correlates with regional economic indicators such as poverty levels, population size, and other demographic factors. It also reveals insights into the priorities of foundations and the effectiveness of their donations in addressing social challenges. The project focuses on visualizing these patterns through interactive dashboards, providing a user-friendly and intuitive tool for stakeholders to explore the trends in grant-making practices in the southwest region.

Objectives

The primary objectives of this project are to:

1. **Analyze grant sources:** Identify and categorize the origins of grants across the southwestern states. This includes mapping which states or regions contribute the most to philanthropy and which foundations are the largest donors.
2. **Examine utilization patterns:** Investigate how grants are distributed and utilized across different sectors and purposes. Analyzing the Grant Purpose will reveal key trends in how different types of causes or needs are being supported across the region.
3. **Identify recipient types:** Categorize and analyze the various types of grant recipients. This will help determine whether certain sectors (health, education, etc.) or organizations (non-profits, public institutions) are prioritized for grants.
4. **Assess population-wise distribution:** Evaluate the distribution of grants with population demographics. The objective is to see if the amount of funding corresponds to the needs of the population, such as underserved communities or regions with high poverty rates.
5. **Conduct geographic analysis:** Examine the geographic breakdown of grants, including the demographics of recipient regions. This analysis will highlight the concentration of grants in urban versus rural areas, offering insight into regional philanthropy patterns.
6. **Categorize donors:** Develop a classification system for donors and analyze their donation behaviors. By examining their total donations and repeat giving patterns, we can uncover trends in donor engagement and their commitment to certain causes or areas.
7. **Explore institutional relationships:** Investigate the relationships between recipient and donor institution types. This will allow us to see whether certain types of donors (e.g.,

foundations vs corporations) prefer to give to specific types of recipients, like educational or healthcare organizations.

Methodology

To achieve these objectives, we did the following:

- Collected and consolidated data from multiple sources, including IRS Form 990 XMLs and Census data. This includes identifying the keys in the data to join the data together to get a master data source.
- Developed and implemented data processing scripts to clean, organize, and transform the collected information. This was needed to be done as the data from the various sources were in different types and it needed to be consistent throughout.
- Conducted exploratory data analysis (EDA) to identify key variables and patterns. This involved Univariate, Bivariate, and Multivariate analysis.
- After EDA, by which we got a peek into how the data looks and what can be expected from the analysis, we moved on to the step of creating and designing the dashboards. This step dives much deeper into the relationship between various fields, gives us trends, and visualizes all this in the form of charts, graphs, and maps in addition to KPIs.
- The primary objective kept in mind during the dashboard design was the requirements laid out by our sponsor. The focus the dashboard keeps on is all the points Dr. Brown wanted to know more about and thus we deeply analyzed the same.
- Prepared a comprehensive final report summarizing key insights and trends in granting patterns across the Southwest. The report includes a few interesting patterns uncovered by the analysis. Also, a data dictionary has been created to aid in understanding the datasets and their sources.

Understanding IRS Forms

A crucial aspect of our data collection and analysis involves various IRS forms filed by nonprofit organizations.

Here's an overview of the different forms relevant to our project:

Form 990

Form 990 is an annual information return required to be filed with the IRS by most organizations exempt from income tax under section 501(a), and certain political organizations and nonexempt charitable trusts. Key points about Form 990 include:

- Form 990 is filed by organizations with gross receipts greater than or equal to $200,000 or total assets greater than or equal to $500,000 at the end of the tax year.

- The form requires reporting on the organization's exempt and other activities, finances, governance, compliance with federal tax filings and requirements, and compensation paid to certain persons.
- It serves as a tool for transparency, allowing the public to view an organization's financial information and activities.

A sample form 990 and the more important fields are highlighted in the following image.
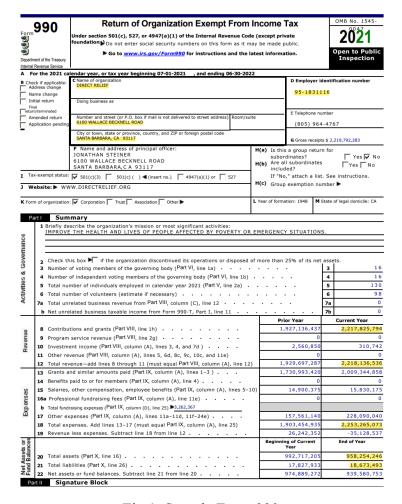

Fig 1: Sample Form 990

Form 990-PF

Form 990-PF is specifically for private foundations and serves multiple purposes:

- It's used to calculate the tax based on investment income.
- It reports charitable distributions and activities.
- Section 4947(a)(1) nonexempt charitable trusts serve as a substitute for Form 1041 when the trust has no taxable income.

Data Collection

For data collection, we use multiple publicly available datasets such as bulk XML Form 990/990-PFs from the IRS website (www.irs.gov), population data from the Census (www.census.gov), HUD Crosswalk data (www.huduser.gov), and NCSS Business Master File (nccs.urban.org).

Data Processing and Augmentation

We have developed a Python script for data collection from the IRS website and returned CSV files with the required fields. This script performs the following tasks:

1. **Bulk Download:** Downloads bulk IRS Form 990 XMLs in ZIP format, including forms like 990 and 990PF for different years from the IRS website. The website provides Form 990 XMLs from 2021 to 2023. For our analysis, we have chosen tax years 2020-2020.
2. **Iterates through all ZIP files and uncompress:** The script uses the package *zipfile* to uncompress the ZIP files and store the XML files in a separate directory. This step is extremely space-consuming and results in >65 GB of data.

   This step also can be executed manually where the ZIP extraction can be done manually and feed the XML files to the Python script for downstream processing.

```
ankitbasu@Ankit-PC:~/repository/TAMU/data-science-capstone$ ls -1A raw_data/ | wc -l
1951561
ankitbasu@Ankit-PC:~/repository/TAMU/data-science-capstone$ du -sh raw_data/
67G     raw_data/
```

Fig 2: Data Collection Statistics

3. **Reads XML files:** This step reads each XML file, understands their DOM (Document Object Model) structure, and extracts data points from the Form 990/990-PF XMLs.
4. **Data Augmentation:** As part of data augmentation, we gather additional data from sources like census, NCSS, and HUD crosswalks to generate the final datasets for further analysis.
5. **Generates CSV files:** Once the previous steps are successfully executed, the script generates CSV files for individual forms for further analysis.

We have successfully executed the scraper with Form 990 and 990-PFs to gather approximately 7,34,891 distinct grant records generated by the seven target states (with a huge computation overhead) in the tax years 2020-2022.

The data pipeline script can be found on GitHub.

Final Datasets

For our analysis, we have generated two datasets from the data collected from the sources above. The two datasets serve two separate purposes.

1. **Census Grants Dataset:** This dataset provides a county-level aggregation of the grants paid and received across different parts of the United States. To develop this dataset, we have taken all the counties in the US and then left-joined the grants and census data aggregated at a County Name, and Tax Year level.

2. **Form 990 Consolidated Dataset:** This dataset contains the individual grant information like grantmaker, recipient entities, their geographic locations, tax year, financials of the grantmaker for the particular tax year, etc. We gather the individual grant information from the Form 990/990-PF XMLs and join with the NCCS BMF and IRS BMF EO datasets for NTEE classifications for grantmakers and recipients.

An overview of the data dictionary is available as a google sheet here.

Exploratory Data Analysis (EDA)

We performed an in-depth Exploratory Data Analysis (EDA) to understand the structure and distribution of the data. This included checking for skewness in the distribution of key variables like grant amounts and recipient demographics, and applying transformations when necessary. We conducted univariate analysis to explore individual columns, followed by bivariate and multivariate analyses to uncover relationships between variables such as grant size, recipient location, and poverty levels. This helped us identify potential patterns and correlations for further analysis. This analysis also helped us understand what key data points we are missing in order to perform the required analysis and what other interesting patterns we could potentially find from the data about which we didn't think before.
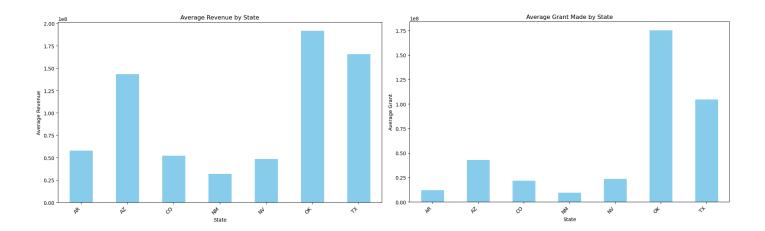
Fig 3: Average Revenue of Grant Makers by State; Fig 4: Average Grants Made by Businesses by State

Because of the EDA analysis done, we knew that this analysis is not true as the data is skewed for some states like Texas and Oklahoma, thus resulting in such a big difference in the average grants made by grantmakers in these states. Once the skewness of the data was removed, the analysis done using dashboards was much more reliable and actionable.

Data Analysis and Visualization

We have developed a Tableau Public (version 2024.3) dashboard to showcase the relationships and insights derived from our analysis. The dashboards are developed with interactivity in mind and different dashboards serve different purposes in our analysis. These visualizations are fed through CSV files and are best visualized on devices with a native resolution of 1920x1080 px (16:9) or above.
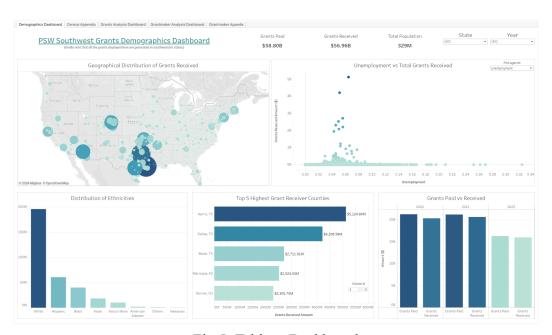

Fig 5: Tableau Dashboard

As part of our analysis, we have created the following visualizations that focus on different areas of analysis.

1. **Demographics Dashboard:** This dashboard focuses on the population-wise analysis of the grants aggregated at a county level. The dashboard provides Top-N counties served by southwestern states, state-wise population distribution, and shows grant density based on the geographic location of the recipients. Also, we have added interactive scatterplots to visualize patterns between grants received and various census parameters like total population, unemployment rate, median household income, poverty percentage, poverty

estimate, etc.

2. **Census Appendix:** Provides an easy at-a-glance appendix for county-level grants information over the years 2020-2022 that can be easily filtered by County names.

3. **Grants Analysis Dashboard:** This dashboard provides the analysis based on both grantmaker and recipient's perspective that can be controlled with a toggle. This dashboard provides grant breakdowns at NTEE classification and provides grants generated vs grants received geo-location relationships.

4. **Grantmaker Analysis Dashboard:** Provides an overview of grantmakers and their statistics across different states and tax years. Top 10 grantmakers, Boxplots for different industries and sub-industries, donut charts for percentage analysis of grantmaker organization types and industry types. Also, provides the correlation between grants generated in MSA/non-MSAs and where the grants have been paid.

5. **Grantmaker Appendix:** Pivot for grantmakers across different regions and their grantmaking patterns along with financial details. This data can be filtered by EIN, Business Name, and State.

The dashboard is currently hosted on the Tableau public server and is [available here](#).

To refresh the dashboards, please install Tableau Public on your local desktop, download the .tbwx file from Tableau Public, go to the Data Source Tab, and select the updated files for both files before refreshing. This will refresh and update the dashboard with the latest data. The refreshed dashboard can be hosted on the Tableau Public server once refreshed.

Findings

As part of our data analysis, we have come across some interesting trends that are worth mentioning:

1. **Most of the grants are generated in MSA regions** (around 93%), and 89% of those grants end up in Metropolitan Statistical Areas. Implying that a very small percentage of grants are reaching out to micropolitan areas or they are not properly traced.
2. **Around 45.7% of the total grants are provided by the Public, Societal Benefit** industry, making it the biggest grantmaking industry in the Southwest.
3. **Most of the grants end up in the same geographic boundary**, both at the State, County, and CBSA levels.

4. The largest grantmakers in the Southwest region are **Texas (49%), followed by Colorado (13.8%) and Oklahoma (9.5%)**. While Nevada and New Mexico are the bottom two states at 6% and 1.3% of all grants.
5. The **total grant provided in 2022 was significantly less** than in the years 2020 and 2021.
6. Private Grantmaking Foundations provide **approximately 3 times more** funding than Private Independent Foundations, Public Foundations, and Community Foundations.

Further exploration of the dashboard and datasets will uncover more interesting patterns in the granting landscape of Southwestern states.

Challenges

While working on this project, we came across the following issues that created major blockers for our progress:

1. **Data Incompleteness:** The biggest challenge we faced was that the data sources available online or from previous projects were very limited and did not provide a major insight into the issue we intended to analyze. Our only option was to self-explore the data sources online and figure out a way to generate data sources from scratch.
2. **Absence of Categorical Data:** Collecting the data for categorization of the organizations and the purpose of the grant was particularly hard because either it was not publicly available, or available in the form of natural language text. Harnessing useful information from this was problematic as there was no particular category defined in the text, and our efforts of using LLMs from classification also went in vain. LLMs didn't work for this task because of the large dataset, and no domain-specific fine-tuning of the LLM.

Conclusion

The "Analysis of Granting Patterns of Foundations Across the Southwest" project is well executed, achieving the goal of publishing the dashboards to visualize key data points. Our comprehensive approach, combining IRS Form 990 data with Census information, provides a solid foundation for uncovering meaningful insights into the granting landscape of the Southwestern United States.

We were able to answer the questions laid out by Dr. Brown such as the granting pattern across different states, counties, and types of organizations. In addition to that, we also provided visualizations to see howthe trends look like for the years 2020, 2021, and 2022. Also, the users would be able to analyze the behavior of grants across the metropolitan and micropolitan areas

across the Southwest states. This visualization will enable stakeholders to explore the data and draw their conclusions, fostering a deeper understanding of the granting ecosystem in the Southwest.

We hope that this project will bring an impact on granting patterns by helping the user to make a data-driven and informed decision. The dashboard does highlight the skewness in nature of donations made with most of the money going to the metropolitan area and with this data-backed information, stakeholders could potentially focus on the smaller regions.

References

1. Tableau Public Servers ([link](#))
2. IRS Form 990 / 990-PF datasets ([link](#))
3. NCCS Business Master File ([link](#))
4. IRS EO Business Master File by States ([link](#))
5. NTEE V2 Codes ([Github](#))
6. HUD Crosswalks Datasets ([link](#))
7. Census Datasets ([link](#))
8. Guidestar/Candid ([link](#))