TaxDataDashboard

CS 4624
Multimedia, Hypertext, and Information Access

Report
May 9, 2023
Blacksburg, VA 24061

Instructor: Dr. Edward Fox
Client: Dr. Florian Zach
Authors: Courtney Anderson
Matthew Grillo
Namrata Inamti
Vishnu Sankaran

Contents

List of Figures

List of Tables

1 Abstract

The Internal Revenue Service (IRS) Form 990 is a form filed by nonprofit organizations. These forms collect information about the organization's finances, officers, contributors, and other information. These forms are required by law to be published. Starting in 2011, the IRS made the data available in an AWS bucket. As of 2020, the IRS also has published the data on its website.

In Spring 2022, Dr. Zach approached Dr. Fox with the TaxData project. He began working with groups of students to compile this data into a database and to begin analyzing the data. In the Spring 2022 semester, a group of students from the CS 4624: Multimedia, Hypertext, and Information Access capstone built a basic database and began to parse the data. In the Fall 2022 semester, one student took over the project and modified the database to be more effective.

This semester, we took over the project from where it was previously left off. Dr. Zach tasked us with creating visualizations that show correlations in the data. He wanted to be able to easily input a list of EINs and showcase a variety of graphs and charts for that list. These visualizations were then to be put on an interactive dashboard where all visualizations could be seen in one easy-to-access location.

The following report outlines the additional work we completed this semester through: Objectives, Deliverables, Requirements, Design, Implementation, Testing, User Manual, Developer's Manual, and Lessons Learned.

2 Introduction

2.1 Problem

The IRS has a specific form, Form 990, for non-profit, tax-exempt organizations to fill out each year with their data. Form 990 has variant and related forms that organizations fill out based on the size and type of the company. These types include Form 990, Form 990-EZ, Form 990-PF, Form 990-N, and Form 990-T. [2]

Once all of this data is collected it is published to the IRS website annually. However, the data is not easily accessible nor compiled for analysis. The main problem at hand is ensuring there is an easy way to analyze IRS data. This is currently difficult because of two factors. First, the IRS changes its forms each year. This means that the data from each year has various columns that may not directly correspond to the data from other years. Second, the dataset is large and difficult to work with. When the files are downloaded, there are various parts and information in different CSV files which makes pulling specific information difficult unless you know exactly where it is in the files.

2.2 Motivation

Our client, Dr. Florian Zach, is an Assistant Professor in the Howard Feiertag Department of Hospitality and Tourism Management, Pamplin College of Business, Virginia Tech, USA [3]. Dr. Zach is doing research on IRS data for non-profit organizations, specifically those related to Hospitality and Tourism Management.

His goal is for us to create a dashboard that allows him and his colleagues to easily view and analyze different aspects of the IRS data. He wants to look at data based on geographic location, such as the number of organizations per state each year, as well as visualizations showcasing revenue, expense, and asset data. Other issues Dr. Zach wants to look at are how gender plays a role in the compensation of those who run these organizations, and how successful these organizations are.

2.3 General Approach

For this project, we have been tasked with creating a dashboard for showcasing IRS tax data for non-profit, hospitality, and tourism management organizations. This is a continuation of the project done by two CS 4624 teams, one in Spring 2022 [4] and one in Fall 2022 [1]. The two previous groups worked mainly on getting all the data into a single database that can be easily queried. This goal was completed and we just needed to confirm all the data was up to date and that no new data needed to be added.

In order to build our dashboard we have decided to split our tasks into four main segments: visualization round one, gender identification, visualization round two, and dashboard building. In order to do this we will use Jupyter notebooks, SQLlite, and Python. Visualization round one will include building visualizations in a Jupyter notebook using the data provided by the previous teams. We will build maps and charts that will then be imported into our dashboard later on in the process. Gender identification will involve finding an online gender database that predicts gender based on a person's name, joining that with the names of the leaders of the desired organizations, and adding a new column to the database that predicts their gender. Visualization round two will include building visualizations that include this newly added gender information, and improving the interactivity of the visualizations from round one. Finally, building the dashboard will involve pulling together all the visualizations that we have built and making them accessible in one easy-to-use web application.

3 Requirements

In order to begin working on this project there were a few things we needed to gain access to. First, we needed the shared Google Drive. This Drive contains files from the past two semesters, including data, code, and reports. Next, we needed to access the Virginia Tech Gitlab hosted at code.vt.edu and to become maintainers of the git repository for the project. This contains the parsers created by the student from last semester and some of the visualizations that had been worked on in the past.

3.1 Deliverables

Through meetings between the client and the team, we decided on a few crucial deliverables that should be completed by the end of the semester.

1. An easily accessible dashboard that showcases all created visualizations. This dashboard should be interactive and should allow users to be able to explore the data on their own.

2. Visualizations that showcase the data in charts, maps, and other ways. This will specifically be the list of visualizations that Dr. Zach provided the team at the beginning of the semester.

3. Both the process we go through and the code that we write to create these deliverables should be well documented. Documentation is critical for Dr. Zach in using these visualizations and for any future teams that will work on this project.

4 Design

When it came to the design of our workflow, we took inspiration from guest lecturer Christian Johnson, whose development team utilizes AGILE methodology and two-week-long sprints to keep track of the progression of the project. We have designed the timeline of our project around two-week-long sprints, each with primary objectives to complete. The timeline can be found in Table 2 within Section 9.1. When it came to the design of our solution, we used the System Methodology approach, guided by the advice of Dr. Fox.

4.1 System Methodology

4.1.1 Users

We identified two types of users for our system: Dashboard Users and Database Users. The goal for dashboard users is to be able to view the IRS Form 990 data and visualizations that we have created. The goal for database users is to be able to visualize the data through custom visualizations and to query the database for specific pieces of information.

4.1.2 Tasks

The tasks that are discussed in this section are shown as diagrams in Figures 1, 2, and 3 found within Section 12.1. There is also a system workflow diagram in Figure 4 within Section 12.1 that summarizes the following tasks into one workflow that will be followed.

The tasks for the dashboard users are as follows:

1. To analyze the IRS Form 990 data using pre-existing visualizations.
    (a) Begin running the dashboard.
    (b) Input a list of EINs for the desired data.
    (c) Select the desired visualization.
    (d) Input the desired year and state for the visualization.
    (e) View and analyze the data.

The tasks for the database users are as follows:

1. To create custom visualizations.
    (a) Create a new Jupyter notebook.
    (b) Perform a query on the database that extracts the desired data.
    (c) Use query data to create a dataframe.
    (d) Clean the data and ensure it is ready to be used.
    (e) Create the desired chart or graph.

2. To search the database for a specific piece of information.
   (a) Open the database in an SQLite server.
   (b) Perform a query on the database to find desired information.

5 Implementation

The main deliverable of this project is the Tax Data Dashboard. Our goal is to allow users to interact with and explore the data on their own. To accomplish this, we broke down our deliverable into multiple parts.

5.1 Getting new IRS nonprofit data and updating database

In order to have the dataset contain the most up-to-date information, you would have to:

1. Retrieve the most recent files containing the data from the IRS website
2. Make a backup of the current database
3. Modify the database update scripts from prior projects
4. Insert new data into the database using the update scripts
5. Validate that the database contains the full data set

Our team had started this process, and at step 1 we found out that there had not been any new data added, so we were able to skip this part.

5.2 Making the Visualizations

Our team used Jupyter Notebooks, Pandas, Matplotlib, and Plotly Express to get familiar with the data, create some visualizations, and prioritize which ones would be most useful in the dashboard. During this time our client, Dr. Zach, provided us with a list of visualizations for us to create. The visualizations that were created during this process can be seen in Section 12.2.

5.3 Gender Identification

Gender identification involved the addition of a column that indicates the officer's gender into the Schedule J tables in the database. This was done through a Python script that utilized a Python package called gender_guesser. This package is an open-source package that assigns a gender based on a first name. The detector takes in a string argument and returns either female, male, mostly_female, mostly_male, andy, or unknown. We parsed the database by querying the Schedule J tables one year at a time. Once the table was queried, a new table was added and then an insert query was performed for each unique first name that added the corresponding gender to that row of the table.

5.4 Making the Dashboard

Once the visualizations were created, we compiled them into one dashboard. We used Plotly Dash to create the dashboard. We found that Dash easily took

Plotly Express graphs and made them interactive. We were able to give each visualization its own page that can be viewed and modified with drop-down menus. Each page of the dashboard, as well as a short description of the page, can be found in Section 12.3.

6 Testing/Evaluation/Assessment

Due to the nature of our project, not much testing was needed to be conducted. After the first few weeks of the project, we had developed an understanding on the process that was needed to first obtain the relevant data, clean it, and visualize it. The visualizations themselves became the deliverables for the project. As such, evaluations and assessments were conducted by our client, Dr. Zach, on a bi-weekly basis. During these meetings we would show Dr. Zach the visualizations that had been created thus far in order to receive feedback on the design of them. He would offer ideas for improvements and if it was possible to modify them accordingly, we would go back and implement them before finalizing each visualization requested by the client.

7 User's Manual

7.1 Visualizations

Table 1 provides a list of all of the visualizations created along with a description of them and what file they can be found in. Users can use this table to find what they are looking for, and go directly to that file. Once the user accesses the Jupyter Notebook, there are instructions on how to edit the input EIN file and modify any variables that are meant to be interactive. The visualizations that are noted to take in a list of EINs, will read from a provided file to create visualizations only for the EINs provided. All of these visualizations are also included in the dashboard.

Visualization	Description	File
No. Organizations by State	Shows no. organizations from each state that filed taxes every year. User can select year of filing. Takes in a list of EINs.	NumOfOrgs_ByState_ByYear.ipynb
Organization Age Correlations	Shows the total revenue, assets, expenses, and liabilities vs. organization age. User can select variable. Takes in a list of EINs.	OrganizationAge_Visualizations.ipynb
Top 25 Employees	Shows the top 25 employees by total compensation, base compensation, bonuses, nontaxable benefits, retirement and deferred compensation, and other compensation. Indicates gender. User can select variable, state, and year. Takes in a list of EINs.	Top25_HighestPaid Employees.ipynb
Tax Exemption Status & Organization Form Type	Shows breakdown of filings and values by exemption status (501(c)(3), 501(c), 4947 (a)(1), 527) and organization type (corporation, association, trust, other). User can select variable.	TaxExemptionStatus_OrganizationForm_Visualizations.ipynb
Top Organizations	Shows the top 25 organizations by age, assets, benefits, expenses, liabilities, revenue, and salaries. User can select variable, state, and year.	TopOrgs_ByVariable.ipynb
Data Over Time	Shows the revenue, expenses, assets, liabilities, benefits, and salaries for an organization over time. User can select organizations. Takes in a list of EINs.	LineChart_ByVariable.ipynb
No. Organizations by State - Year Formation	Shows the number of organizations founded in each state each year. User can select year. Takes in a list of EINs.	Only in dashboard
Organization Compensation	Shows the ratio and percentage of organizations that offer club dues, discretionary spending accounts, first class or charter travel, housing allowance, tax indemnification, personal services, travel companions, business use of personal residence, and a written policy for these benefits. Takes in a list of EINs.	Only in dashboard

Table 1: Visualization Descriptions & Files

7.2 Dashboard

The dashboard can be run locally on any computer. There are a few prerequisites needed in order to make sure the dashboard can be run on a machine. The

following steps explain how to run the dashboard:

1. Clone the gitlab repository
2. Download the NPTR.db file from ARC
3. Place the NPTR.db file into the Dashboard folder
4. Place the EIN list file into the Dashboard folder
5. Open the folder in VSCode
6. Run the following commands in the terminal
    - pip install dash
    - pip install express
    - pip install sys
    - pip install sqlite3
    - pip install pandas
7. Run the command 'python create_csv.py *Input File*'. Here the Input File is the list of EINs you want to look as specifically.
8. Run the command 'python app.py'. This will start running the app.
9. Click on the http link in the terminal. This will open the dashboard in your browser.

Once the dashboard is running, the user should be able to look through all of the visualizations and use all of the available drop-down menus that make the dashboard interactive.

8 Developer's Manual

8.1 Database

The database at it stands contains 189 relations, most of which are partitions of the annual tax form 990, and form 990EZ data. It also includes a relation named "TagsMapping" which is used to convert XML paths into shorter, digestible attribute names appropriate to the data located at that XML path. Each attribute also has a description, which has been very helpful for drawing visualizations and comprehending the data that is stored in the database. All information regarding the "TagsMapping" relation is pulled from the Nonprofit Open Data Collective where it is kept up to date at `https://github.com/Nonprofit-Open-Data-Collective/irs990efile`.

More technically, the database is a sqlite3 relational database and can be interacted with easily through the sqlite3 Python package; this is also what is used in the parsing scripts and for our visualizations.

8.2 Visualizations

In order to get the data in a format that enables proper visualizations, compound SQL queries and external merging of the data is often necessary as a result of the design decisions of the previous capstone project team. For this, we found that the Pandas Python package is the best suited for viewing the dataset and making sure that the data is formatted correctly. Creating nice visuals for Dr. Zach to be able to interpret and include in his reports was a primary factor as well when creating visualizations. To accomplish this we utilized the matplotlib Python package as it allows for quick graphics to be created and configured to ours and Dr. Zach's liking.

9 Lessons Learned

9.1 Timeline/Schedule

Our timeline for the project can be seen in Table 2.

Week 1 & 2: 2/2-2/15	- Getting new IRS nonprofit data. - Updating Database with new data; ensure no duplicates.
Week 3 & 4: 2/16-3/1	- First round of Visualizations using Jupyter Notebook, Pandas, Matplotlib. - Take feedback from the client at the halfway meeting (2/19) and adjust visualizations.
Week 5 & 6: 3/2-3/15	- Gender identification.
Week 7 & 8: 3/16-3/30	- Adapt Visualizations to be interactive (filter data, etc.).
Week 9 - 12: 3/31-5/2	- Second round of visualizations and building the dashboard. - Demo

Table 2: Project Timeline

During the first two weeks of February, we were getting familiar with the pre-existing database. We spent time reading the reports and documentation of the previous teams, as well as meeting with Dr. Zach to learn more about the project. We gained access to a shared Google Drive that had information from the past two semesters, the ARC cluster that stores the database, and a git repository that had all the code made by previous students. After becoming familiar with the work that already existed, we went on to the IRS website to see if any new data had been added since the database had been updated in August 2022. We found out that there had not been any new data added and determined that the database was already up to date.

The next two weeks were spent playing around with the data and different ways of presenting it. We used Jupyter notebooks to explore correlations in the data, and start to build out figures that will eventually be in our dashboard. Dr. Zach asked for an extract of the data at this time, and we were able to provide CSV and JSON files for each of the tables that he wanted data from.

During the month of March we were able to complete the list of visualizations provided by Dr. Zach, as well as visualizations that we thought were interesting.

9.2 Problems

There were two main problems that occurred during the project.

1. The current database was comprised of 189 tables, each corresponding to a different part of a type of Form 990 and the year that the form was filed. This meant that each time the data needed to be aggregated a temporary data frame had to be created using Pandas. This took up extra time in our program runtime.

2. Duplicate EINs were found throughout the database. Many organizations had multiple filings for the same year due to a form revision.

9.3 Solutions

In order to solve these problems, the team came up with the following solutions:

- We created a template that would merge all of the necessary tables into one table at the beginning of each Jupyter Notebook.

- We found a way to query the database that removed the duplicate data.

We decided to find work-arounds for these problems in order to produce more results and ultimately make visualizations that Dr. Zach could use immediately.

9.4 Future Work

Work that we recommend be done includes:

1. Make the database more efficient. Remove the duplicate data that can be found for many of the EINs in each table. Combine all of the tables to no longer be by year, but just by form type.

2. Publish both the database and the dashboard to a server. This will be more efficient for developers when creating new visualizations and users when trying to run the dashboard.

10 Acknowledgements

We would like to thank Dr. Zach for giving us this project to work on and for collaborating with us to create the data visualizations and dashboard. We would also like to thank Dr. Fox for his guidance so far this semester.

11 References

Bibliography

[1] Kaleb Barker et al. *TaxData*. Virginia Tech CS4624 team term project report, Spring. 2022. URL: http://hdl.handle.net/10919/109974. (Accessed: 03.02.2023).

[2] IRS. *E-file for Charities and Non-Profits*. 2023. URL: https://www.irs.gov/e-file-providers/e-file-for-charities-and-non-profits. (Accessed: 03.02.2023).

[3] Florian Zach. Virginia Tech Pamplin College of Business directory entry for Professor Zach. 2023. URL: https://htm.pamplin.vt.edu/directory/zach.html. (Accessed: 03.02.2023).

[4] Maleehah Zafar. *TaxData2*. Virginia Tech CS4624 team term project report, Fall. 2022. URL: http://hdl.handle.net/10919/112875. (Accessed: 03.02.2023).

12 Appendices

12.1 Appendix A: Methodology Assignment

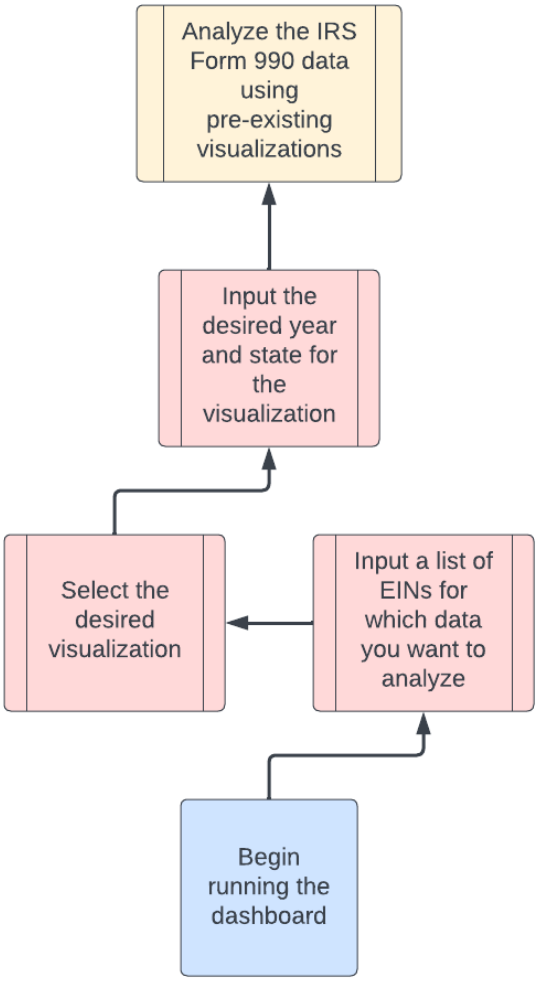

Figure 1: Dashboard User Task 1

The first task of a dashboard user is to be able to analyze the IRS Form 990 data using the visualizations we created this semester. Figure 1 shows the process a dashboard user would go through in order to successfully complete this task.

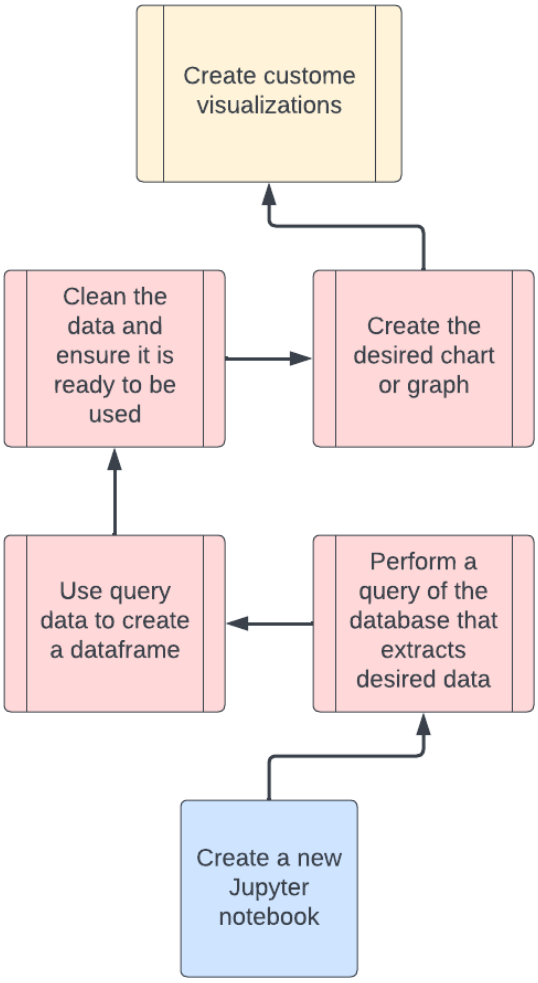

Figure 2: Database User Task 1

The first task of a database user is to create custom visualizations using the database. Figure 2 shows the process a database user would go through in order to successfully complete this task.

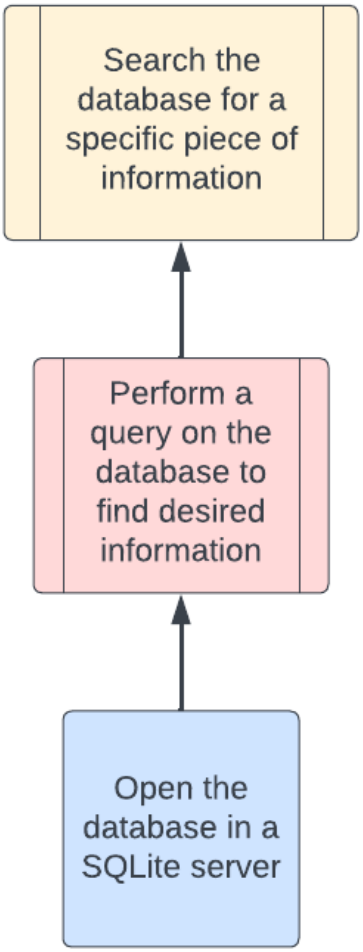

Figure 3: Database User Task 2

The second task of a database user is to be able to query the database for a specific piece of information. Figure 3 shows the process a database user would go through in order to successfully complete this task.

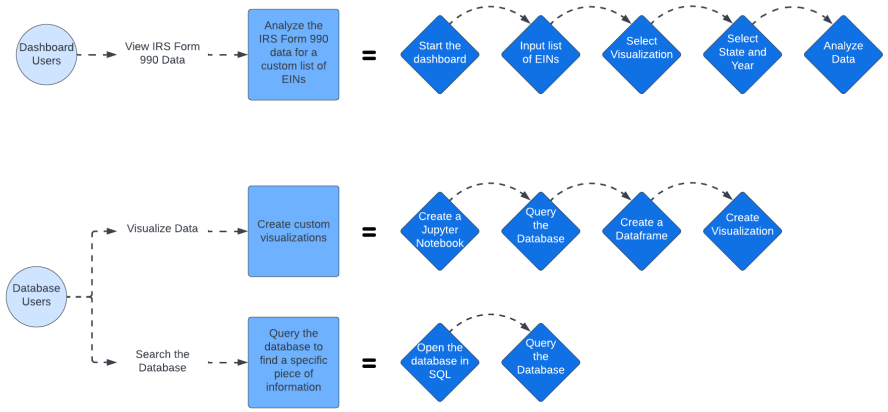

Figure 4: System Workflow

The system workflow can be summarized by the diagram shown in Figure 4. The circles represent the user types, the text are the goals, the squares are the tasks, and the diamonds are the subtasks.

12.2 Appendix B: Visualizations

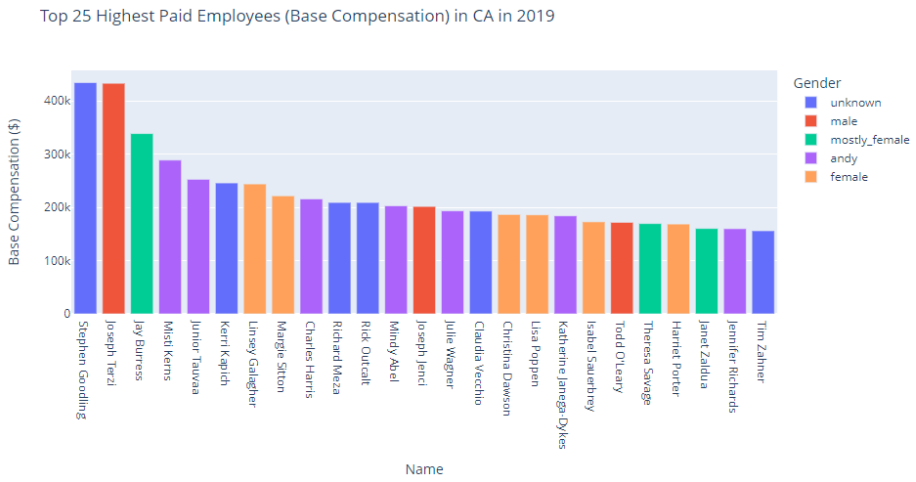

Figure 5: Top 25 Base Compensation CA 2019

Figure 5 shows the top 25 employees in California in the year 2019 for the base compensation variable. The bars are color-coded based on the gender of the employee.

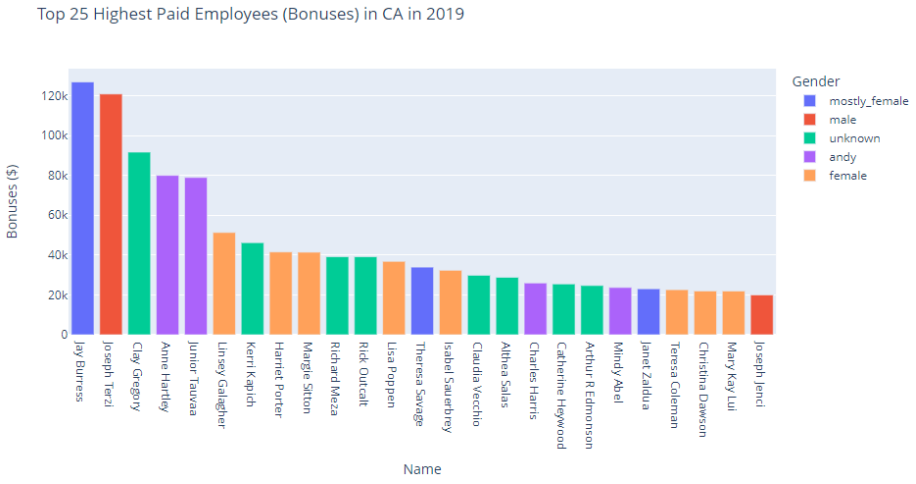

Figure 6: Top 25 Bonuses CA 2019

Figure 6 shows the top 25 employees in California in the year 2019 for the bonuses variable. The bars are color-coded based on the gender of the employee.

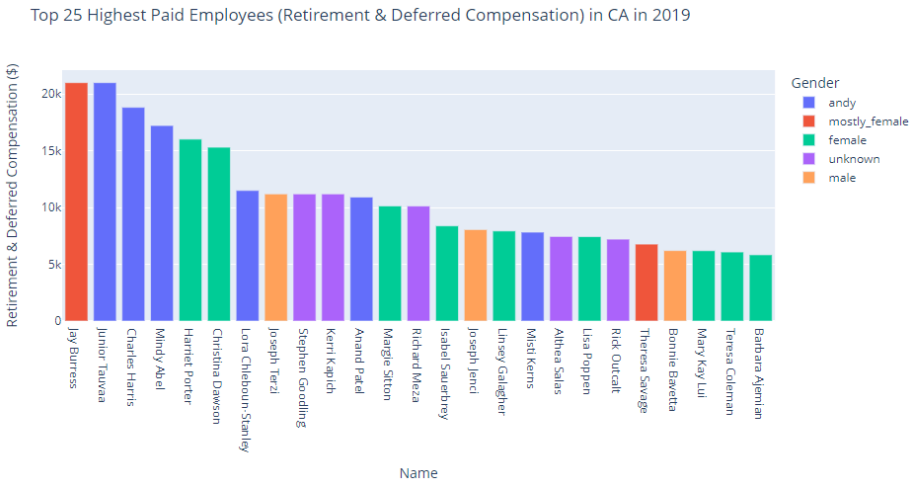

Figure 7: Top 25 Deferred Compensation CA 2019

Figure 7 shows the top 25 employees in California in the year 2019 for the deferred compensation variable. The bars are color-coded based on the gender of the employee.

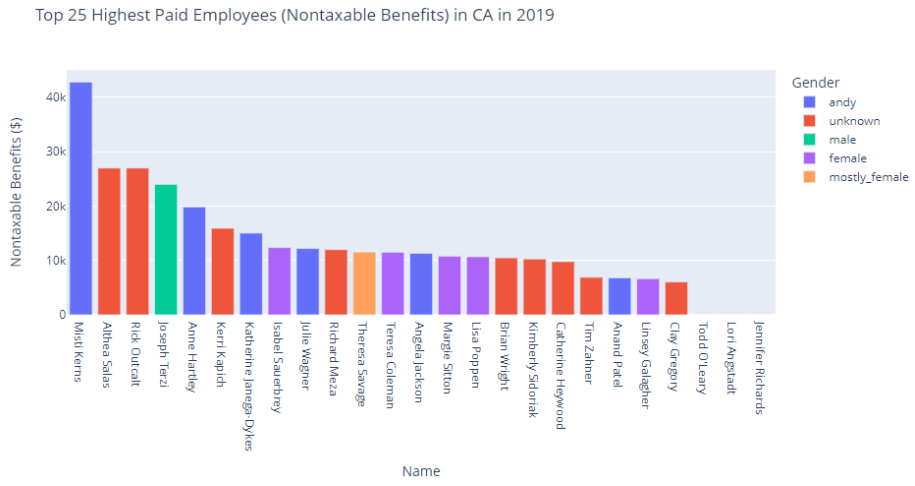

Figure 8: Top 25 Nontaxable Benefits CA 2019

Figure 8 shows the top 25 employees in California in the year 2019 for the nontaxable benefits variable. The bars are color-coded based on the gender of the employee.

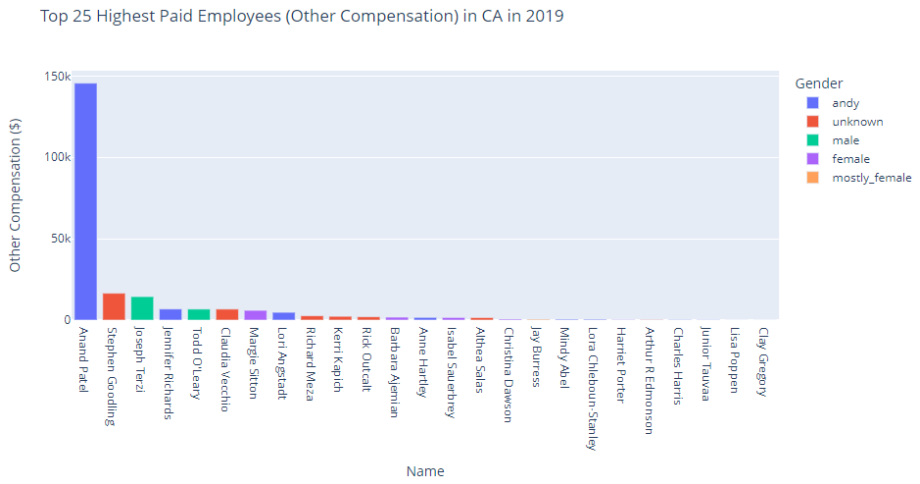

Figure 9: Top 25 Other Compensation CA 2019

Figure 9 shows the top 25 employees in California in the year 2019 for the other compensation variable. The bars are color-coded based on the gender of the employee.

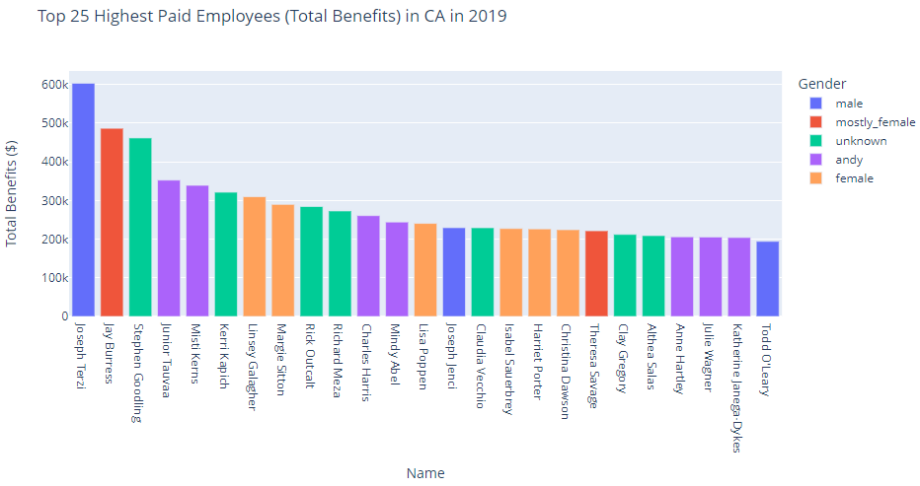

Figure 10: Top 25 Total Benefits CA 2019

Figure 10 shows the top 25 employees in California in the year 2019 for the total benefits variable. The bars are color-coded based on the gender of the employee.

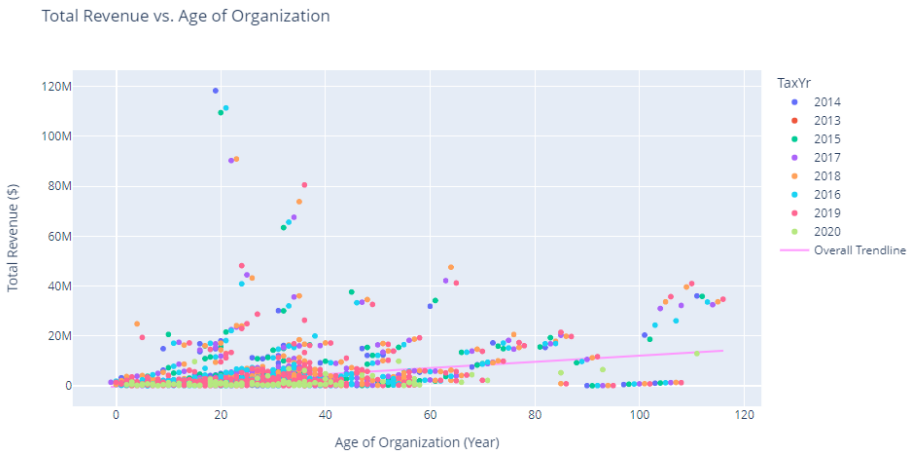

Figure 11: Total Revenue vs. Age Of Organization

Figure 11 shows a scatter plot with trend lines for each year taxes were filed comparing the total revenue of an organization to the organization's age.

12.3 Appendix C: Dashboard Pages

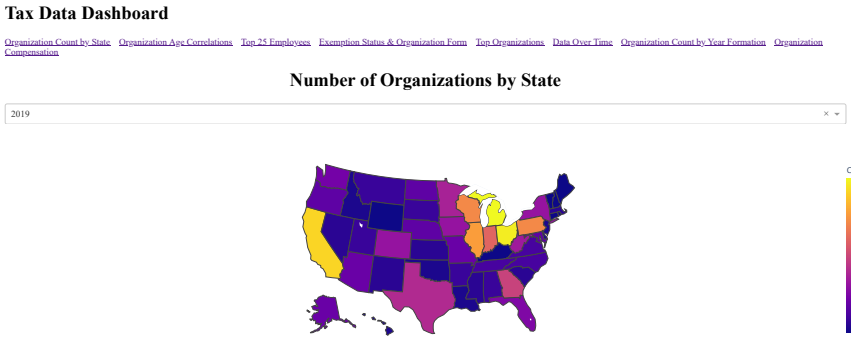

Figure 12: Organization Count Page

Figure 12 shows the organization count page of the dashboard. On this page there is a color map that shows the number of organizations filing taxes each year given a list of EINs. The drop-down menu allows the user to select what year they want to see data for.

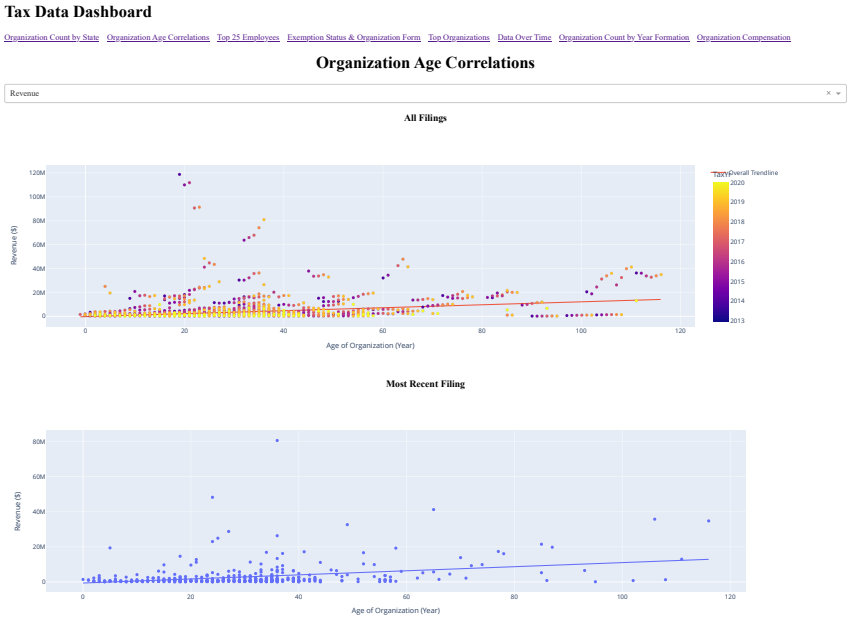

Figure 13: Organization Age Page

Figure 13 shows the organization age page of the dashboard. On this page there are two scatter plots. One shows the selected variable versus the age of an organization for all filings given a list of EINs. Another shows the same thing, but only for the most recent filing. The drop-down menu allows the user to select what variable they want to see data for.

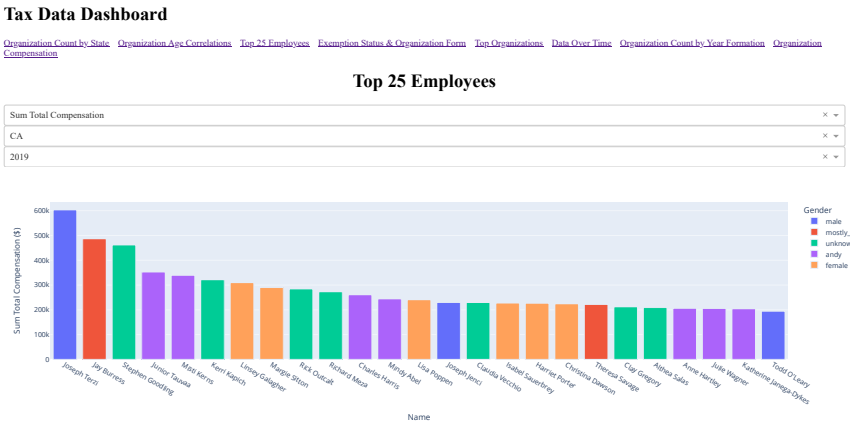

Figure 14: Top Employees Page

Figure 14 shows the top 25 employees page of the dashboard. On this page there is a bar graph that shows the top 25 employees for the selected variable, state, and year – given a list of EINs. The drop-down menus allow the user to select what variable, state, and year they want to see data for.

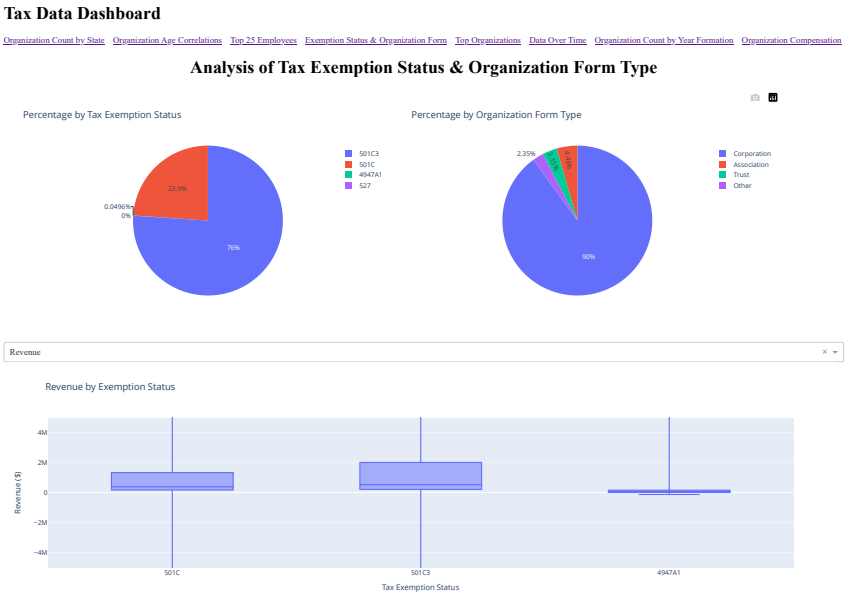

Figure 15: Exemption Status & Organization Count Page

Figure 15 shows the exemption status and organization form page of the dashboard. On this page there are two pie charts. One shows the breakdown of filings by exemption status and the other shows the breakdown of filings by organization type. The page also has two box plots. One shows the selected variable by exemption status, and the other shows the selected variable by organization type. The drop-down menu allows the user to select the variable they want to see data for.

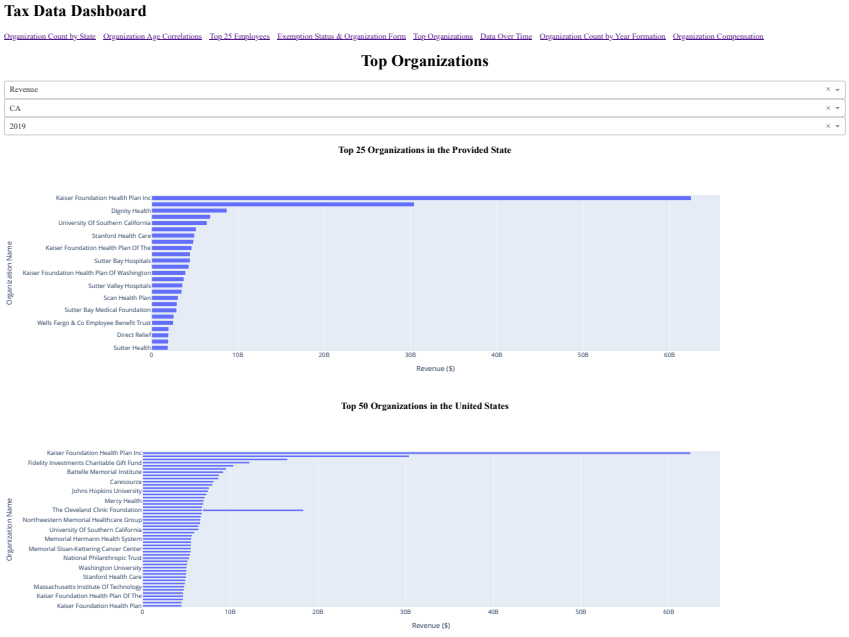

Figure 16: Top Organizations Page

Figure 16 shows the top organizations page of the dashboard. On this page there are two bar charts. One shows the top 25 organizations for the selected variable state and year given a list of EINs. The other shows the top 50 organizations across the United States for the selected variable and year, given a list of EINs. The drop-down menus allow the user to select the variable, state, and year they want to see data for.

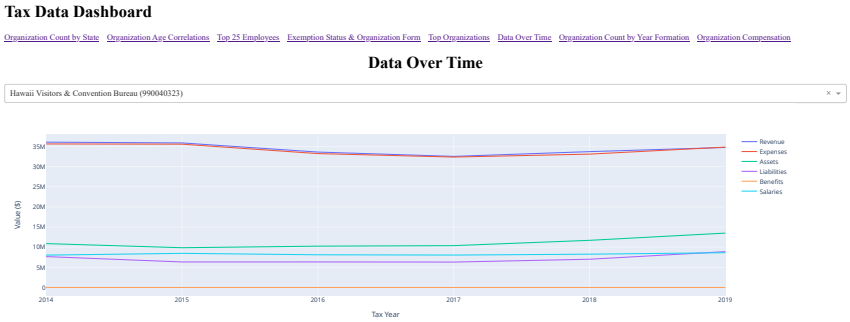

Figure 17: Data Over Time Page

Figure 17 shows the data over time page of the dashboard. On this page there is a line chart that shows the revenue, expenses, assets, liabilities, benefits, and salaries over time for the selected organization. The drop-down menus allow the user to select what organization they want to see data for.

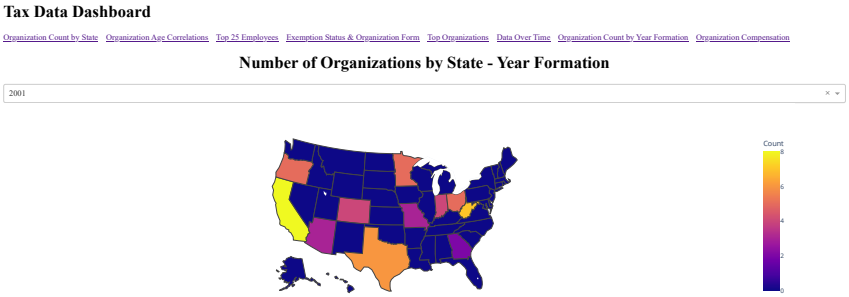

Figure 18: Organization Count by Year Formation Page

Figure 18 shows the organization count by year formation page of the dashboard. On this page there is a color map that shows the number of organizations founded in each state each year, given a list of EINs. The drop-down menu allows the user to select what year they want to see data for.

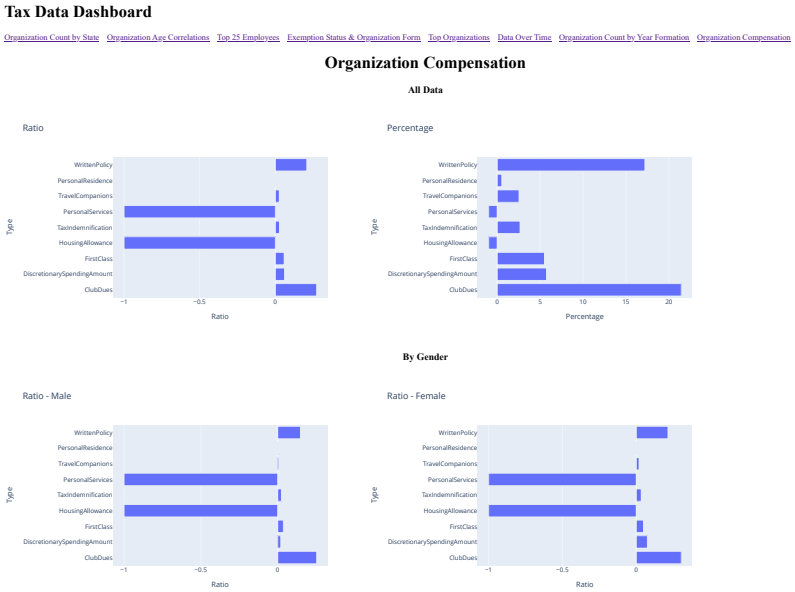

Figure 19: Organization Compensation Page

Figure 19 shows the organization compensation page of the dashboard. On this page there are six bar charts. The first two show the ratio and percentage of organizations that contribute to the different types of compensation, given a list of EINs. The bottom four show the same data broken down by gender.