
Team Innovia

CS-251-1 INTRODUCTION TO DATA SCIENCE

A solid purple horizontal bar at the bottom of the slide.

Team Members

Team A

Kostiantyn Makrasnov

Michael Svetlichny

Spencer Gariano

A Team

Coby Pieros

Connor Weldy

Travis Herrera

Introduction


Client Background: Innovia Foundation

- Innovia is a Spokane based organization that raises money through donors to give grants to various projects and needs in the community, primarily in Eastern Washington.
- Innovia works to improve access to education, promote health and wellbeing, support arts and culture, create economic opportunity and enhance quality of life.



Analysis Topic

In October 2019, Innovia Foundation announced a new study of wealth in Eastern Washington and North Idaho demonstrating that \$42 billion is expected to transfer between generations over the next 10 years.

A white downward-pointing arrow on a blue background, connecting the first text box to the second.

Innovia aims to direct 5% of the wealth transfer back into community investments and programs

A white downward-pointing arrow on a blue background, connecting the second text box to the third.

Overall Analysis Topic: Utilize public datasets to identify high wealth individuals living in 10 counties in Eastern Washington and to also identify indicators of high net worth. Then if possible, predict likely donors.

Identifying High Wealth Individuals

Challenges

- Lots of Data Collection, Data Engineering, and Subsequent Dead Ends (more on this later)
- Open-endedness

Why is this interesting?

- The Application of DSM in Identifying High Wealth Individuals allows for the pinpointing of Potential Donors
- The Ability to Further Understand Indicators of Wealth

Methodology

General Analysis Approach



1. Collect Unclean Data



2. Filter Unneeded Rows



3. Aggregate Rows



4. Feature Engineering



5. Create Visualizations

6. Find Solutions

Predict Wealth

Predict Age

Predict Likely Donor



Datasets Attempted, and Challenges Encountered

- National Center for Charitable Statistics
(Source: <https://nccs-data.urban.org/index.php>)
- Washington State Voter Registration Database
(Source <https://www.sos.wa.gov/elections/vrdb/default.aspx>)
- DonorSearch Tool
(Source: <https://www.donorsearch.net/>)
- Whitworth Donation Database
(Source: Whitworth Dept of Institutional Advancement)
- Innovia's Donation Database
(Source: Innovia Foundation)



List of Datasets Used

- Washington Political Donations Dataset
(Source: <https://www.pdc.wa.gov/browse/open-data/contributions-candidates-and-political-committees>)
- Spokane Property Sales Dataset (Source: <https://gisdatacatalog-spokanecounty.opendata.arcgis.com/pages/treasurer-data>)
- Washington State Corporations Dataset
(Source: <https://www.sos.wa.gov/corps/alldata.aspx>)



Data Filtering

- N/A entity names were filtered out from all datasets
- Excluded all data that was not in the Eastern Washington area
(Ferry County, Steven County, Pend Oreille County, Lincoln County, Spokane County, Adams County, Whitman County, Columbia County, Garfield County, and Asotin County (Filtered by Zip-Code))
- For business owners, filtered out inactive organizations
- For political donations, filtered out organizations



Aggregating Rows

- **Identifying Rows:** Full Name, Address Line, City, State, Zip Code
- ***Dedupe*** allowed us to match rows across these columns even if the values were not completely the same
- **Challenge:** Need to provide training data unique to each data set (examples of duplicates/non-duplicates)
- **Challenge 2:** Although accurate, algorithm is $O(n^2)$ memory and time efficient*
- **Other:** Tableau Prep allowed also allowed for higher-level aggregation

Feature Engineering



- Summation of donation amounts
- Amount of individual donations
- Business age (year established – current year)

	contributor_name	contributor_address	contributor_city	contributor_state	sum	number
1	RON AND BARBRA WACHTER	X Hidden X	PULLMAN	WA	1650.00	10
2	ATTWOOD WAYNE		SPOKANE	WA	100.00	2
3	SIGLER SHIRLEY A		SPOKANE	WA	97.00	97
4	NYE WILLIAM E		SPOKANE	WA	171.50	159
5	BISHOP RYAN		SPOKANE	WA	100.00	2
6	BOB & DOREEN MORAN		USK	WA	563.00	13
7	HUGGINS KATHLEEN		SPOKANE	WA	72.00	2

Difficulties Prior to Analysis

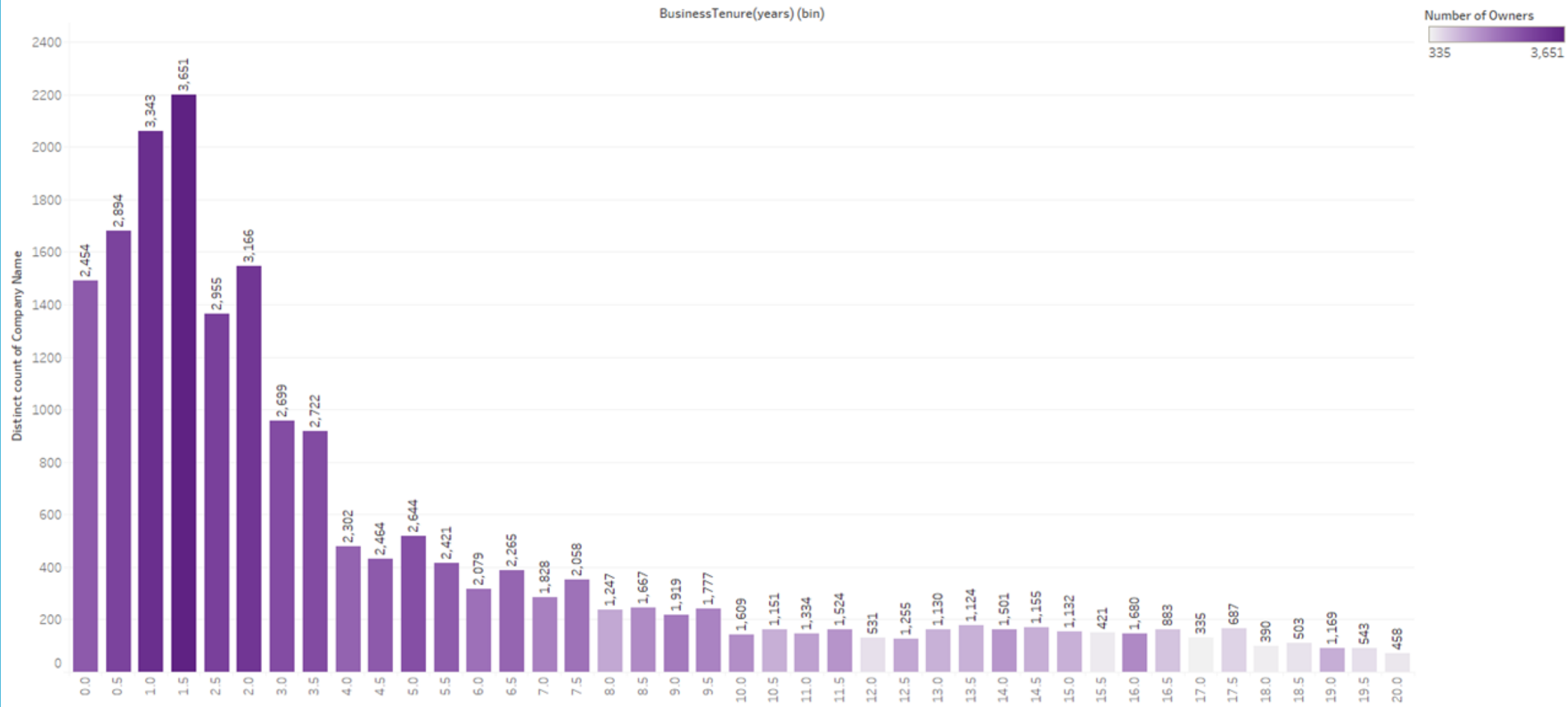
- In all datasets we analyzed, data was "noisy", meaning record values corresponding to a single individual wouldn't be perfectly the same
(i.e. "John Smith", "Smith John", "Smith, John", "John A. Smith", "John & Debbie Smith", etc.)

	contributor_name	contributor_address	contributor_city	contributor_state	sum	number
60835	3RD LEGISLATIVE DISTRICT DEMOCRATS	2422 E 9TH AVE	SPOKANE	WA	800.00	1
17620	3RD LEGISLATIVE DISTRICT DEMOCRATS	5603 W NORTHWEST BLVD	SPOKANE	WA	550.00	2
89725	3RD LEGISLATIVE DISTRICT DEMOCRATS	PO BOX 4963	SPOKANE	WA	150.00	1

- In order to successfully aggregate data, records would normally require to be exactly the same, this is where dedupe came in

Analysis

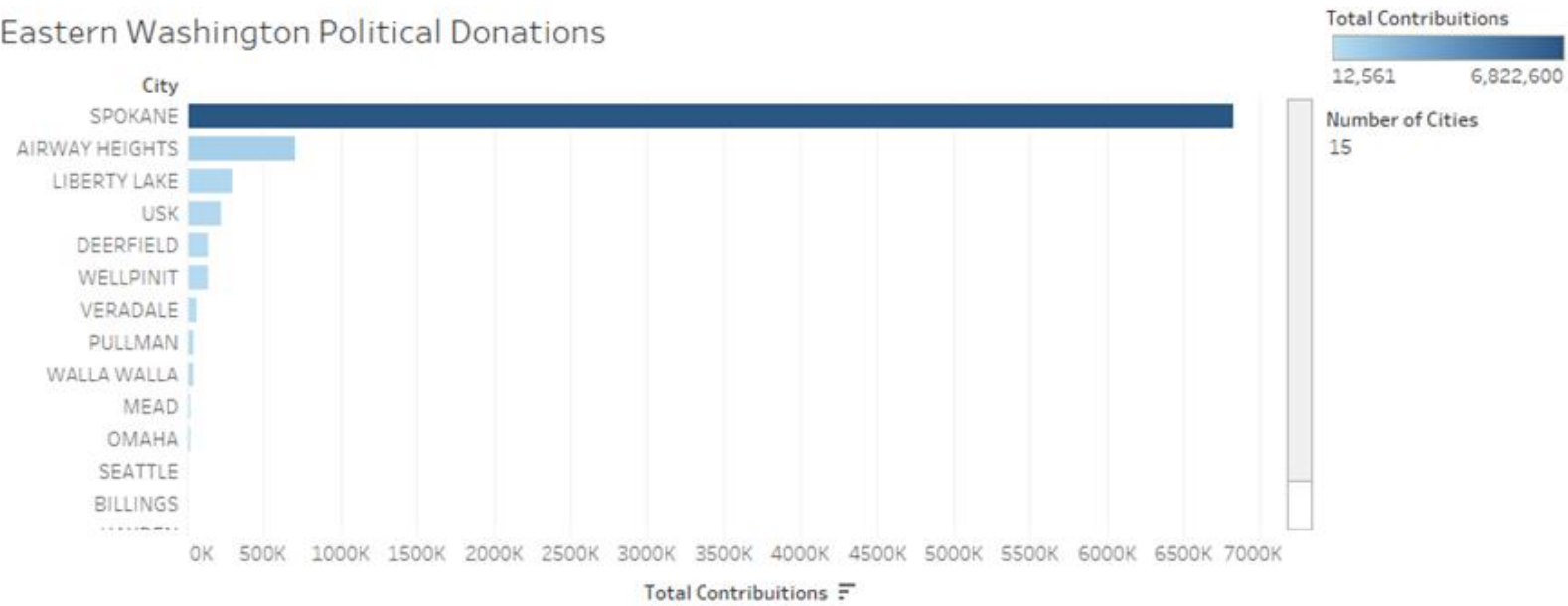
Spokane Companies. by Business Age



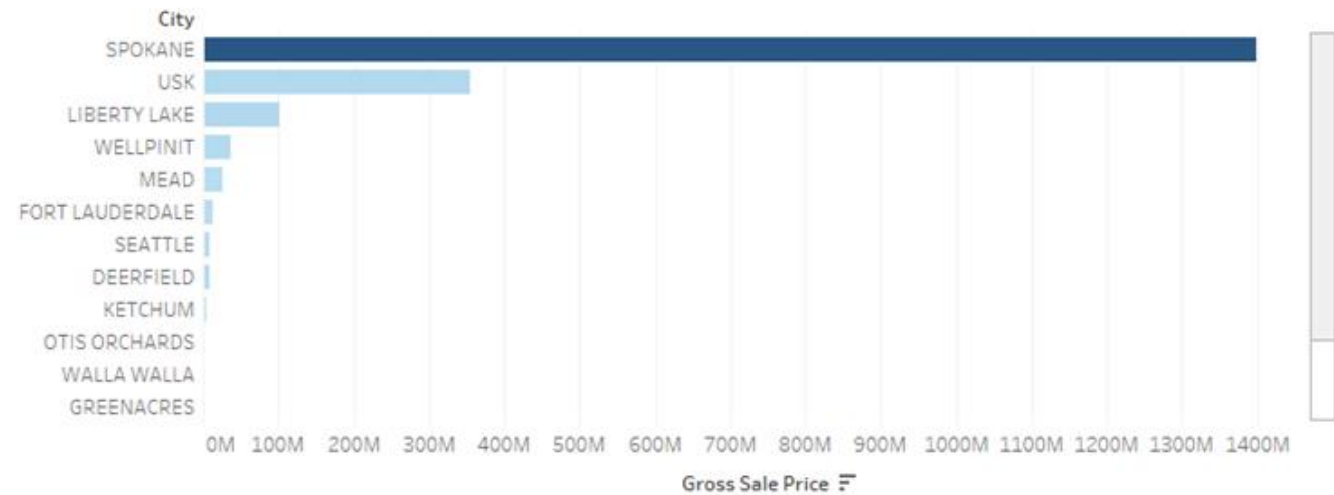
Distinct count of Company Name for each BusinessTenure(years) (bin). Color shows distinct count of Full Name. The marks are labeled by distinct count of Full Name. The view is filtered on BusinessTenure(years) (bin), which keeps 41 of 172 members. Source: Washington State Secretary of State Corporations Database



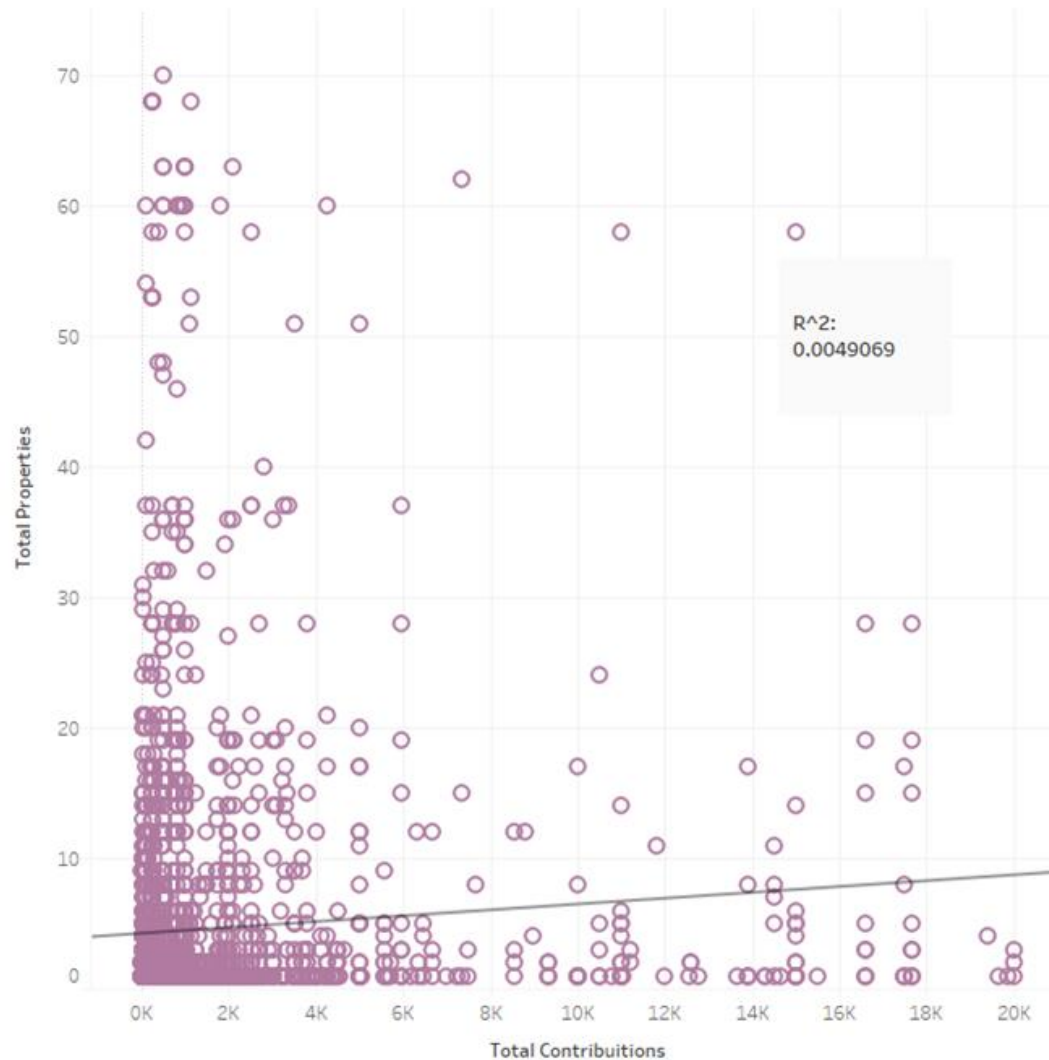
Eastern Washington Political Donations



Eastern Washington Total Property Value of Political Donors

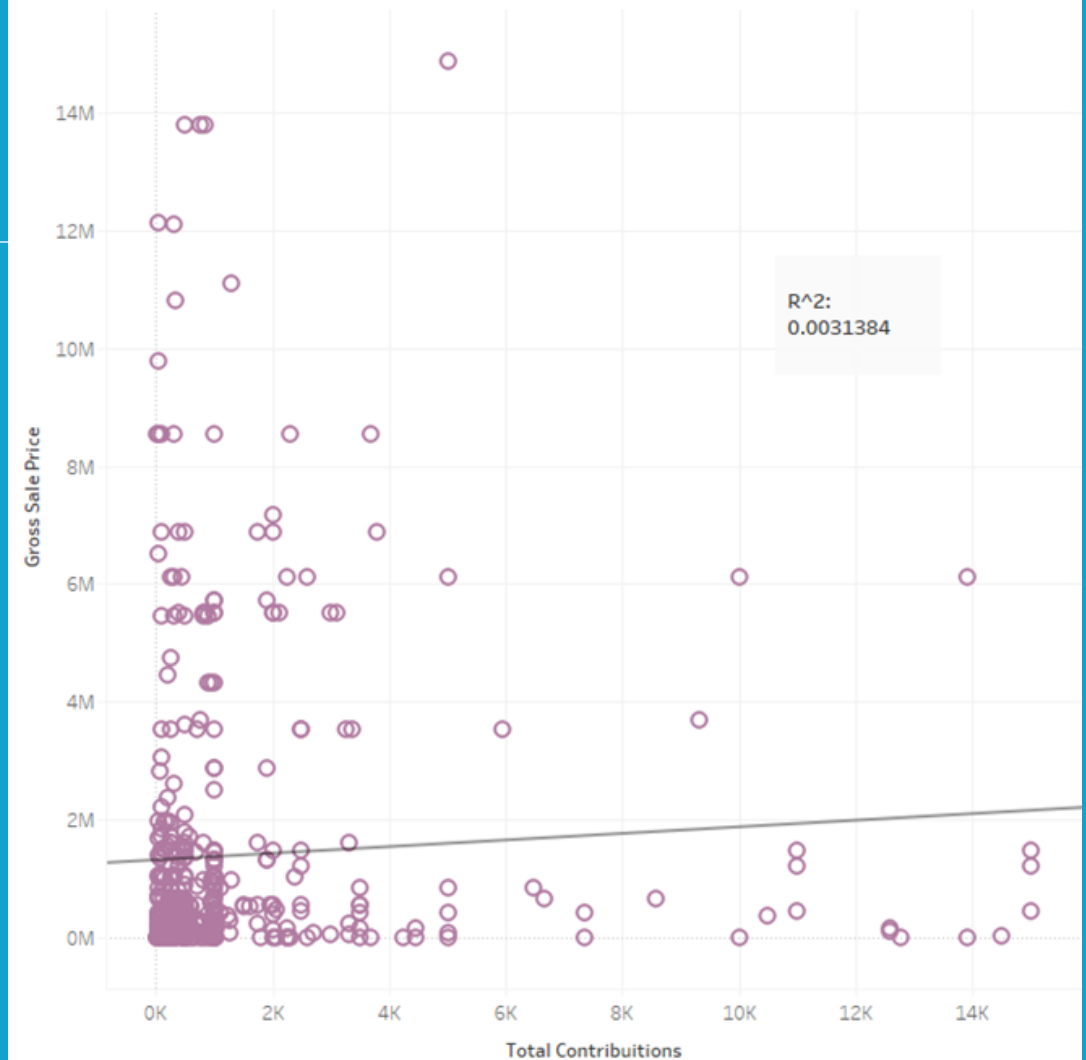


Total Contributions vs Total Properties Owned



Sources: Spokane County GIS Data Catalog, PDC

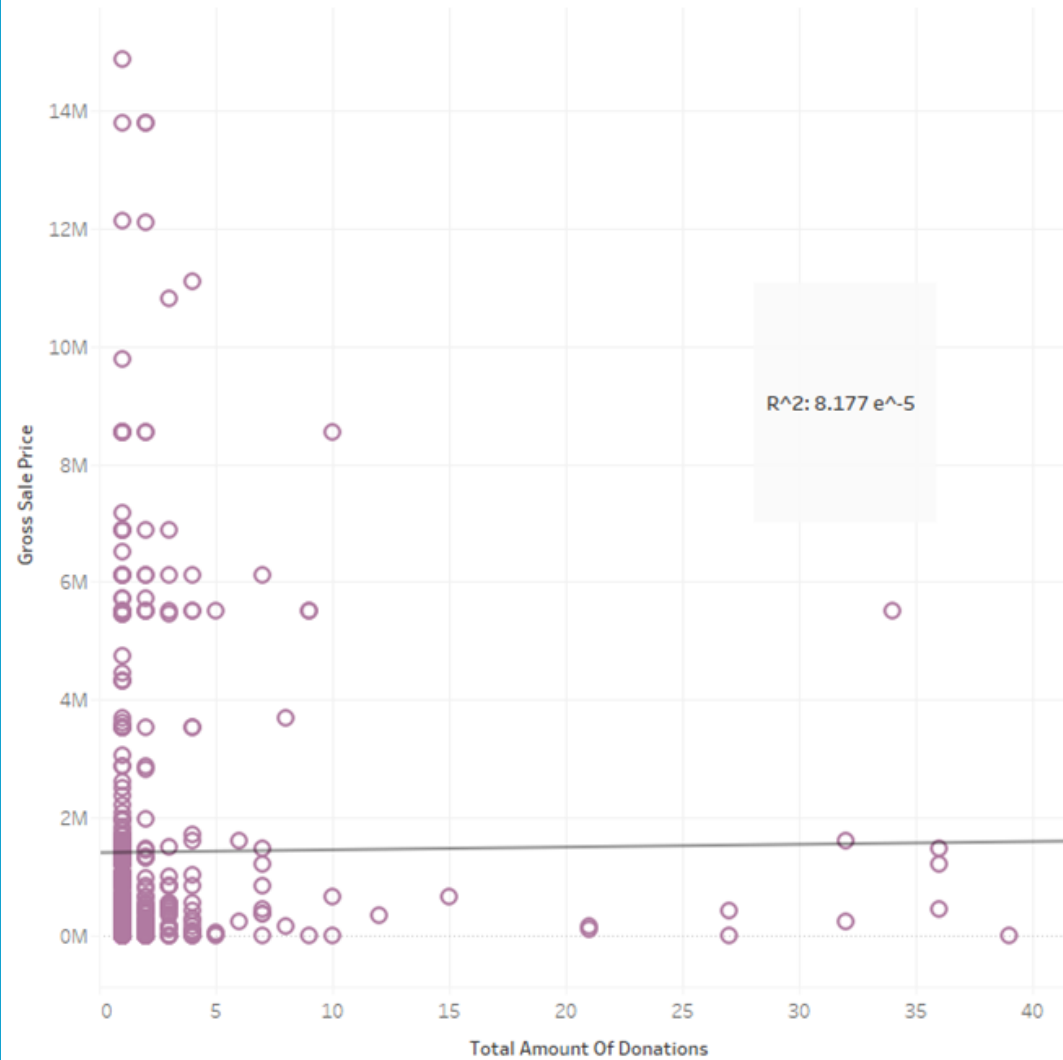
Total Contributions vs Gross Sales Price



Sources: Spokane County GIS Data Catalog, PDC

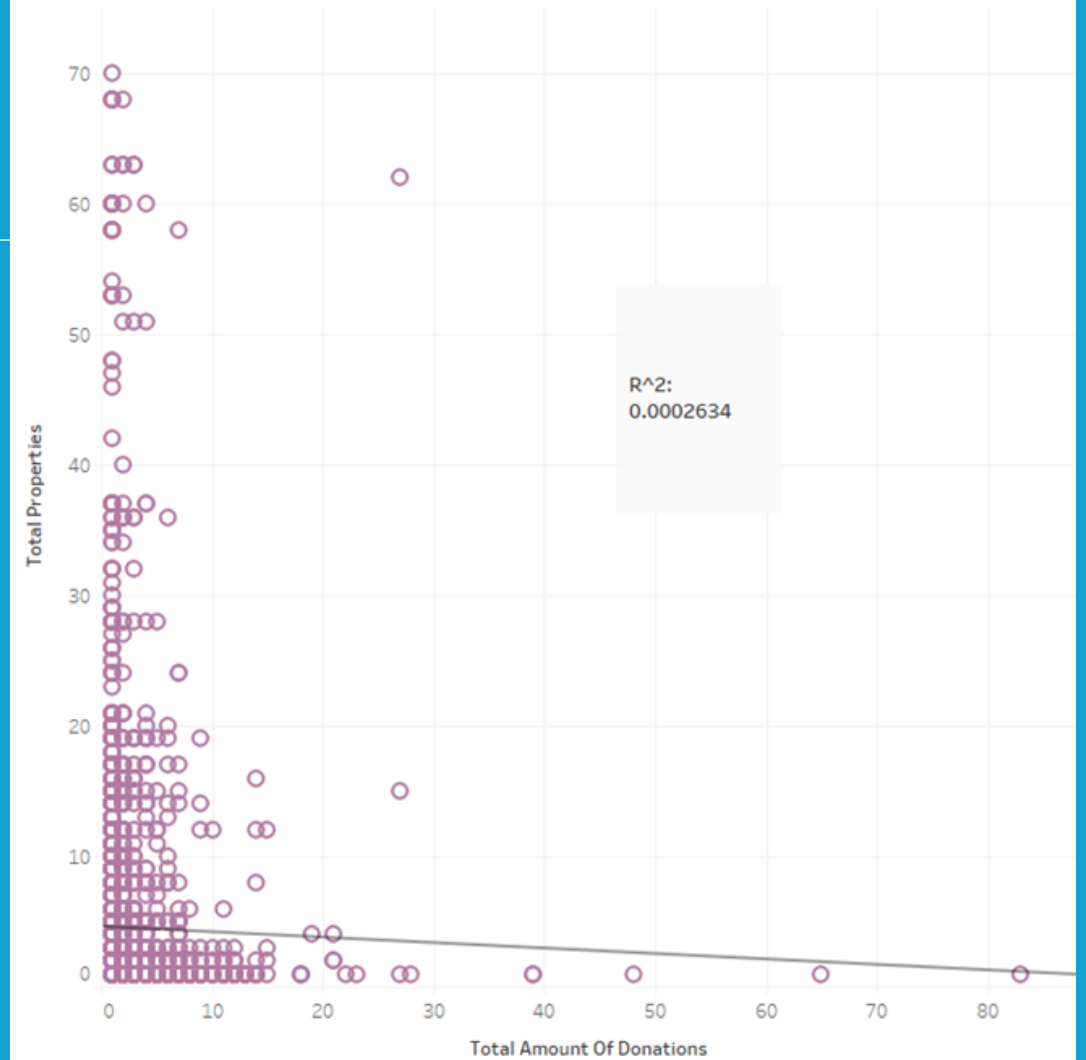


Total Donations vs Gross Sales Price



Sources: Spokane County GIS Data Catalog, PDC

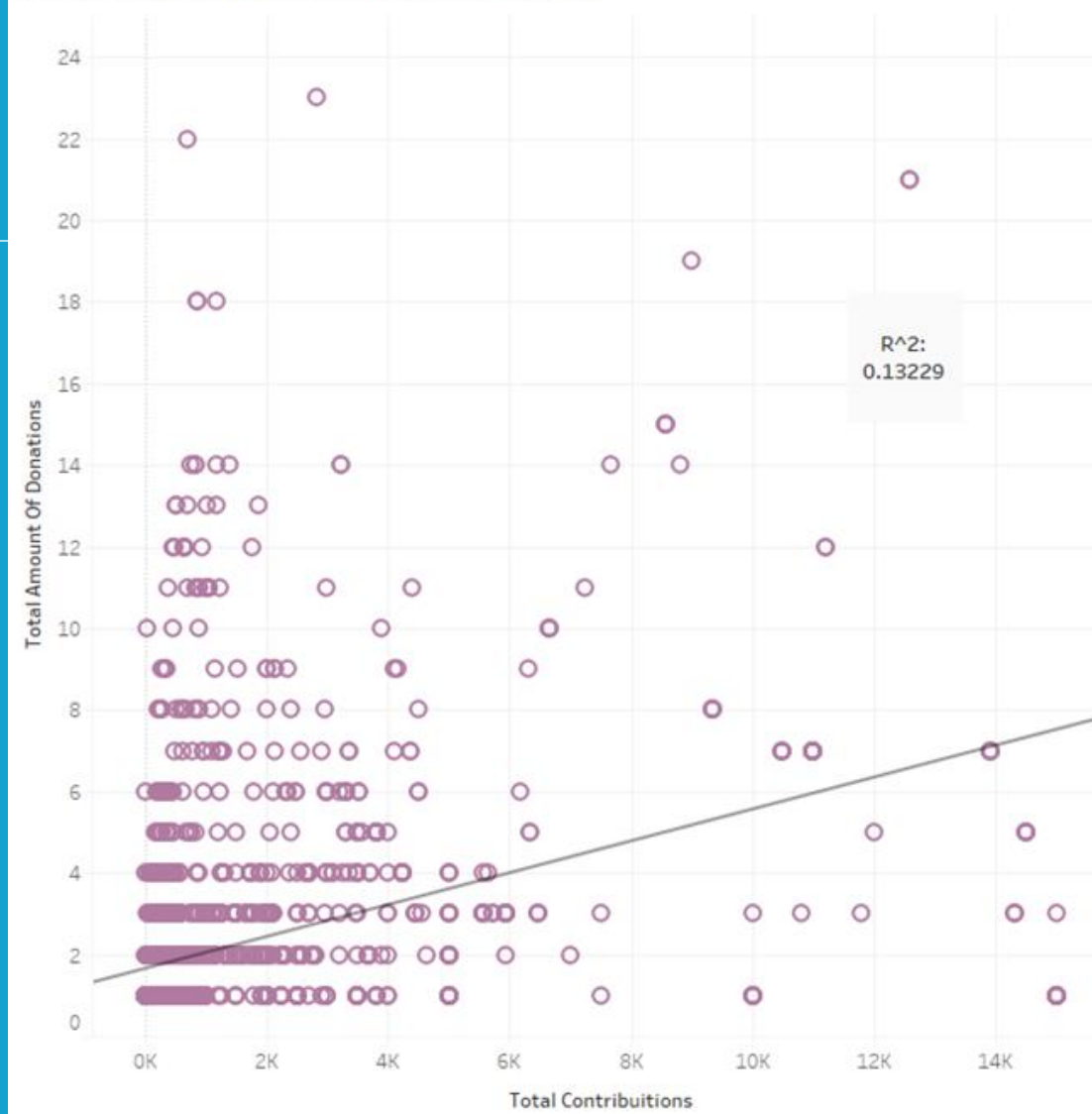
Total Donations vs Total Properties Owners



Sources: Spokane County GIS Data Catalog, PDC



Total Contributions vs Total Donations



Sources: Spokane County GIS Data Catalog, PDC



Deliverables to Innovia

1. In-depth documentation about accessed datasets
 - How to get updated data
 - Column descriptions
 - Usefulness of data in context
2. Scripts to filter, dedupe and aggregate raw data
3. Preprocessed datasets

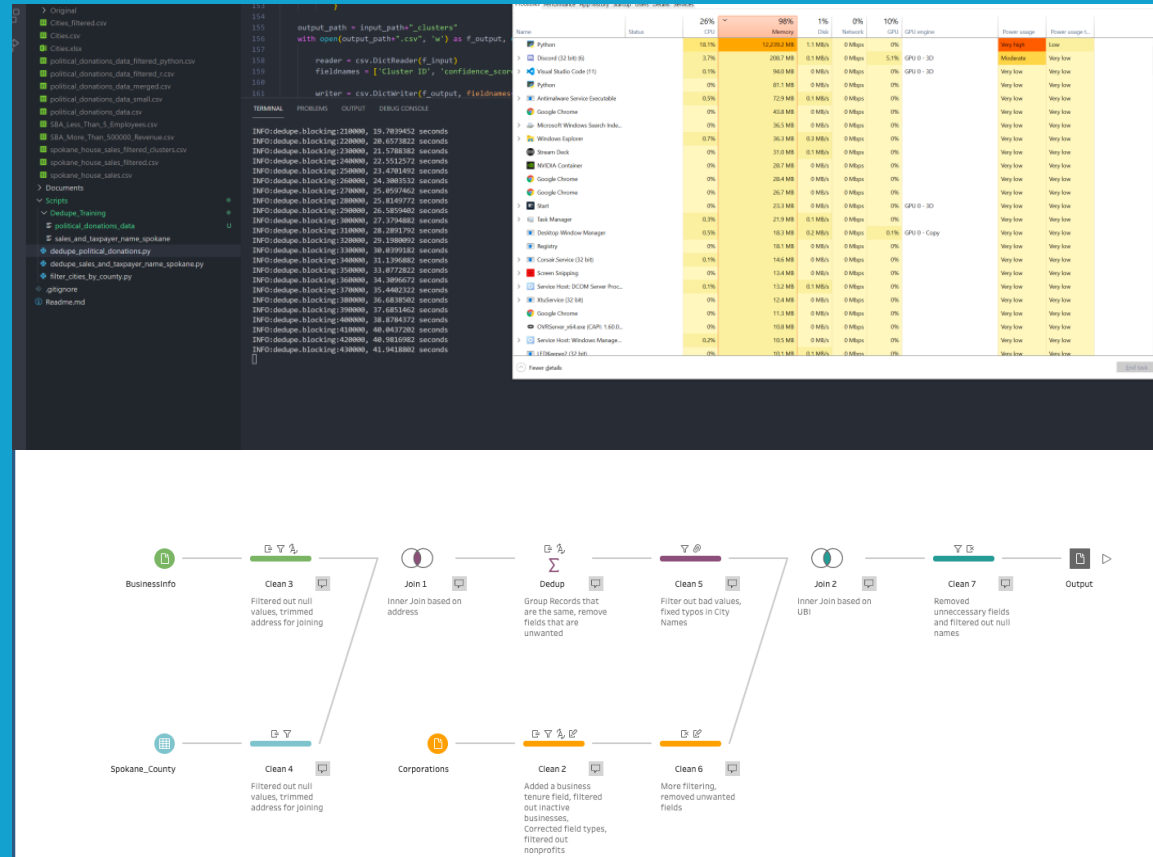


Resources Used

R and Python – Used for
Cleaning, Combining Datasets
& Analysis

Tableau – Used for
Visualization

Microsoft Access, Excel, and
Tableau Prep – Used for
Cleaning, Combining Datasets



Analysis Conclusions

- Feedback data from Innovia required to build a model that predicts likely donors
- Need results of outreach based on provided individuals
- No explicit data available for age and wealth of individuals from accessed datasets
- Only exploratory analysis was possible

Future Research

More Data Science
Techniques &
Predictive Modeling

Expansion of
Datasets Outside of
Spokane County

Important Lessons Learned

Like the DSM States, a Data Science Project Needs a Strong and Clear Description and Problem

Client Communication is just as Important as Analysis

Data Engineering is Complex and Difficult, as well as Resource and Time Heavy

Analysis can only be as good as the Data

References

Forest Gregg and Derek Eder. 2019. Dedupe. <<https://github.com/dedupeio/dedupe>>

Public Disclosure Commission. *Contributions to Candidates and Political Committees*. Washington, Data.WA.gov, May 2021. Website. 1 May 2021. <<https://www.pdc.wa.gov/browse/open-data/contributions-candidates-and-political-committees>>

Spokane County GIS Data Catalog. *Treasury Data*. Washington, Spokane County GIS, 2021. Website. 23 April 2021. <<https://gisdatacatalog-spokanecounty.opendata.arcgis.com/pages/treasurer-data>>

Washington Secretary of State. *Corporations Data Extract*. Washington, Washington State Secretary of State. Website. 5 May 2021. <<https://www.sos.wa.gov/corps/alldata.aspx>>

Questions?
