**Innovia Documentation Deliverable**

Michael Svetlichny, Spencer Gariano, Kostiantyn Makrasnov, Travis Herrera, Coby Pieros, Connor Weldy

The following document contains several different datasets with relevant information about the datasets that a team from Whitworth’s Introduction to Data Science course utilized to find high net worth individuals who are potentially likely to donate. Below for each dataset is a source, a link to the website that contains the dataset, what potential indicators the dataset has, and a description of the dataset with possible uses.

Along with this we have a list of scripts with descriptions that we used to link, filter, and aggregate the various datasets.

***To access our final, combined datasets, please use the following link to our OneDrive:*** [***https://mywhitworth-my.sharepoint.com/:f:/g/personal/cpieros23\_my\_whitworth\_edu/Eu4\_23NVwJRMm51Y97uJwJ8BGW-3wPUW70VBjHBAIi3Nwg?e=oICMQ8***](https://mywhitworth-my.sharepoint.com/:f:/g/personal/cpieros23_my_whitworth_edu/Eu4_23NVwJRMm51Y97uJwJ8BGW-3wPUW70VBjHBAIi3Nwg?e=oICMQ8)

If you have questions or have trouble accessing this link, please reach out to anyone on our team.

**Datasets:**

**Used Datasets**:

**Source**: Washington State Secretary of State Corporations Database

Website: [Corporations Data Extract Download - WA Secretary of State](https://www.sos.wa.gov/corps/alldata.aspx)

Possible Indicators: Those who are owners (or part owners/shareholders) of businesses in Washington state.

Description: This dataset can be a source of finding business owners in Washington State. There are four main files that you can pull from this website. We pulled all 4 and cleaned and combined the dataset, while focusing on just the Spokane area for both memory and time purposes. We chose to keep just the relevant variables for better interpretation. For memory and time purposes, we focused just on Spokane county. Additionally, we calculated the length of the business’s tenure based on when it was incorporated. Along with this, we filtered out inactive corporations. In the final dataset, there is an entry for each owner, their contact information (email), their affiliated business and UBI number, address, corporation type, and business age, among other. This dataset will be a part of our final deliverable and can be accessed through our OneDrive link as “BusinessOwnersCompleted.csv”.

**Source**: Washington State Political Donation Set

Website: <https://www.pdc.wa.gov/browse/open-data/contributions-candidates-and-political-committees>

Possible Indicators: People who donate politically and who make the maximum donation and most frequent donations might indicate a high net worth and willingness to donate.

Description: This dataset contains the records of people who donated politically. Significant variables in this dataset include name of the contributor, address, zip code, and contribution size. This dataset was aggregated to account for duplicate donors. For memory and time purposes, we focused just on Spokane county. The cleaned version of this data set can be found on the OneDrive as “CleanedPoliticalDonationData.csv”. We also filtered out just retirees for selection of having higher likelihoods of estate transfer. This is titled “RetiredPoliticalData.csv”. Additionally, this dataset was joined with property sales data. This combined dataset will also be a part of our final deliverable and is titled “CombinedPoliticalDonationsPropertySales.csv”.

**Source**: Property Sales Dataset

Website: <https://gisdatacatalog-spokanecounty.opendata.arcgis.com/pages/treasurer-data>

Possible Indicators: Property ownership, gross sale price of a property with respected owner, multiple property ownership

Description: We first downloaded the “All Tables” zip file which contains several different datasets. From this, we used sale\_info.txt and taxpayer\_info.txt to link a person’s name, address, and gross sale price of the house. The documentation for the code linking these datasets and aggregating is found under the “Script” section of this document. We believe this could be a useful dataset to find individuals who own homes that have higher values. For memory and time purposes, we focused just on Spokane county. Also, this dataset could be utilized to determine individuals who own multiple homes. The cleaned data set can be accessed in the OneDrive link as “SpokanePropertySalesCleaned.csv” Additionally, this dataset was joined with political donations data. This combined dataset will be a part of our final deliverable as “CombinedPoliticalDonationsPropertySales.csv”.

**Some More Attempted Datasets**:

**Source**: National Center for Charitable Statistics

Website: <https://nccs-data.urban.org/index.php>

Possible Indicators: key employee at non-profit organizations

Description: NCCS data contains several information about non-profit organizations within the United States filed through the IRS via a 990 form. Most datasets from the NCCS website contain records of non-profit organizations. The dataset “digdata.officers2005b.csv” found at <https://nccs-data.urban.org/data.php?ds=digitizeddata>, contains a list of officers, directors, and key employees from the non-profits Part V of the Form 990. This rather large file (3.4 Gb, > 13 million records, 29 variables) contains all the officers from every non-profit in the nation that filed a 990 through the IRS. It might be possible to filter this file to find officers that live in Innovia’s 10 counties through other variables in the dataset. A description about specific fields in this dataset are found at <https://nccs-data.urban.org/dd2.php?close=1&form=DD+Officers+v2005b>. Another useful source for looking up specific nonprofit organizations is GuideStar: <https://www.guidestar.org/> .

**Source**: Washington State Voter Registration Database (VRDB)

Website: <https://www.sos.wa.gov/elections/vrdb/default.aspx>

Possible Indicators: long time voters - may indicate someone who values civic involvement.

Description: This dataset could be a great source for finding people who have been voting for a long time which might be an indicator of someone who is civically involved. However, in order to retrieve a downloadable link to this dataset one must agree to a statement which includes:

“(1) Any person who uses registered voter data furnished under RCW 29A.08.720 for the purpose of mailing or delivering any advertisement or offer for any property, establishment, organization, product, or service or for the purpose of mailing or delivering any solicitation for money, services, or anything of value is guilty of a class C felony punishable by imprisonment in a state correctional facility...”

We are not sure if Innovia’s intentions would classify under this, so we did not investigate this dataset further.

**Utility datasets:**

This data will come in useful when running the scripts outlined below, in some cases they are required for filtering, and deduplication operations.

washington\_cities\_towns\_by\_county.csv – all Washinton town and city county linkings

eastern\_washington\_cities\_towns.csv - filtered dataset of target cities, UPPERCASE format

**Scripts:**

Machine used for scripts was CPU: I7-10700K, RAM: 3200hz 16 gb, M.2 storage drive.

* Aggregation\_for\_taxpayer-sales.r
  + **Description:** script in r to combine datasets with some walkthrough. As well as the capability to clean and aggregate data sets.
  + **Uses:** Computation level: semi-advanced and could be used when there is need to combine/aggregate datasets. This could take several minutes to complete.
* dedupe\_political\_donations.py
  + **Description:** Find clusters of similar rows in the political donations dataset and then combine those rows.
  + **Uses:** Computation level: advanced and should only ever be used no more than once a year. As well as resource heavy and running the script could take up days, possibly weeks on advanced machines.
  + Produces confidence interval, and a ClusterID to each record cluster.
* dedupe\_sales\_and\_taxpayer\_name\_spokane.py
  + **Description:** Find clusters of similar rows in the political donations dataset and then combine those rows.
  + **Uses:** Computation level: advanced and should only ever be used no more than once a year. As well as resource heavy and running the script could take up to 6 or more hours on advanced machines.
  + Produces confidence interval, and a ClusterID to each record cluster.
* filter\_citites\_by\_county.py
  + **Description:** Filter to find the rows that have the county names in the variable list countyList. Then drop all rows that do not fit in that county, then finally showing the names of all cities that fall into those counties.
  + **Uses: C**omputation level: non-advanced and could be ran as needed. You could change the variable `countyList` to others as well if needed. curDataPath on line 3 is the line changed to be applied to other datasets.