**ETL Project Proposal – The Fundraiser Helper**

**Team Members:**

1. Franz Kiel
2. Linda Levy
3. Terry Tsang
4. Will Bobzin

**Project overview (200-300 words)**

* *We are developing a tool that will assist charitable organisations in planning their Victorian fundraising efforts.*
* *In essence, we are trying to determine what are the most “generous” postcodes, where charitable organisations would get better return from their fundraising efforts. To achieve this, we will group postcodes as “Metro” or “Country”, as their population compositions are vastly different. Metro has a considerable amount of population composed by travellers, tourists and short term residents that wouldn’t normally be expected to collaborate with charities.*
* *Once these are aggregated, we can find a “generosity index” that will let us know what is the expected donation per inhabitant for for each type of postcode.*
* *For each postcode, we can determine an “expected donation” by multiplying this index by the population, represented as taxpayers in the ATO. The “expected donation” will then be compared with the actual and this will let us know if a given postcode is overindexed (“generous”) or underindexed (“not generous”).*
* *Postcodes can then be ranked by generosity, with the most generous ones being an ideal location for fundraising activities.*
* *For additional clarity, we will utilise another dataset collected from the webthat advises what are the name of the suburbs under each postcode.*
* *This information will be consolidated in a dataframe, for simplicity of management and to enable future analysis by interested parties.*

**EXTRACT - Proposed data sources (minimum two data sources)**

* *We are starting with two sources of data:*

1. *Postcode clasification - Metropolitan vs Rural from Agriculture Department website (retrieved 23-FEB-21* [*link*](https://www.agriculture.gov.au/import/online-services/delivery-postcode/summary#victoria)*) This information will be scraped from the website.*
2. *Australian postcode dataset – offered by Matthew Proctor (* [*website*](https://www.matthewproctor.com/australian_postcodes)[*dataset*](https://www.matthewproctor.com/Content/postcodes/australian_postcodes.csv)*)*
3. *ATO census information for each postcode* [*link*](file:///C:\Users\znikr\Bootcamp\1.%09https:\data.gov.au\data\dataset\23b8c299-a85b-4fc0-a07d-5ed14e23a103\resource\ec5dba66-e3d1-47ed-b762-33b27d40484e\download\ts18individual06taxablestatusstatepostcode.xlsx)*. This is a cumbersome files, with multiple columns that are not relevant to our purposes.*

* *Based preliminary exploration, we expect the Agriculture website to provide a list of postcodes, which will form our “Metro” dataset. All others will be considered rural. From the ATO dataset we will extract population and charitable donations per postcode, which will be essential in creating the index.*
* *From the Australian postcode dataset we will retrieve a list of names for each postcode, which will help manage the outcomes.*

**TRANSFORM - Proposed clean-up and analysis**

* *We will utilise search functions to clasify postcodes as ‘Metro” and “Country”*
* *There will be a large amount of data dropped from the ATO census information datasheet, as most columns will not be relevant to this exercise and in order to minimise resource use. Columns will be renamed to make the dataframe more user friendly.*
* *We will use splinter and possible Beautiful Soup to scrape the data from the Agriculture website, that we will then implement to reflect the division between “Metro” and “Country”*
* *We will use Australian postcode dataset to fill in the suburbs under each postcode, by merging datasets and using a search function to complete the information. We believe this will make the use of the database more user friendly.*
* *We will perform preliminary explorations on all datasets, checking integrity and looking for incorrect/inconsistent and missing data. Basic statistical analysis will show outliers. We will also look for incomplete sources.*
* *We will manage the data using Jupyter Notebooks, with Pandas for dataframe management. We will implement the analytics using SQL, with Excel for very early analysis. We will perform joins to merge the data, particularly in relation to postcodes.*
* *After applying this transformations, we will have a streamlined, usable data source with no unnecessary columns. The first operaations will give us the tools to divide between “Metro” and “Country”.*

**LOAD - Data storage**

* *The final product will be a relational database, stored in PostgreSQL. This seems like the best option, as it will respect the integrity of each individual database, while maintaining the links that connect them. The key connection points will no doubt be postcodes, as they are a key in each connected database.*

**Potential limitations**

* *One of the limitations we currently foresee is that the available data only covers filed donations to registered charities disclosed to the ATO. While obviously there is a tax incentive to declare donations and deduct from taxes, these don’t cover gifts to individual or organisations that for whatever reason are not registered charities (eg. Sport clubs). While this is a weakness, we believe that it won’t distort the results of the exercise.*

#########################################################################

**ETL DIAGRAM**

It may be easier to explain the process in an ETL diagram (you don’t have to create a diagram but it may help your group to formulate a plan).

Most data engineers make sure to document the ETL processes for reference. One way to do this is to use ETL diagrams. This is usually drafted at the beginning of the project and finalised at the end of the project.

Here are a two examples of how an ETL diagram would look like:



