**Report**

**Idea and Datasets**

This project started with the team looking at the main sources suggested by the project: the website [Kaggle](http://www.kaggle.com) and the website that consolidates all Australian Public data [Data.gov.au](http://www.Data.gov.au).

A first approach to Kaggle showed a dataset on Pet Ownership, which was a topic of interest for the whole team. At that stage, the plan was to cross tha information with geo-data from another sources to figure out the dog ownership patterns (breed, etc) for different areas. At that stage we also agreed that we could achieve a higher quality result in the allocated time by concentrating just on Victoria.

At that point we had found a website from the Department of Agriculture that had exactly what we needed: a list of Metro postcodes for Victoria. The information (essentially a list of four-digit numbers) was encoded in the website and not in a downloadable file. Immediately we recognised this as an opportunity to implement our knowledge of scraping. Emboldened by this finding, we started working on adding a third dataset with demographic information for each postcode. This could let us extract all sorts of interesting conclusions, not just about dog ownership, but also about the owners. We found an ideal source in the ATO, which publishes very detailed aggregate information and included that dataset to our project.

As promising as this looked, part of the group started exploring the Pet Ownership dataset and found that it was incomplete to the point of being unusable. It only contained a handful of postcodes, and in many cases information was missing.

While this was disheartening, we started bouncing off ideas and realised that the information we had was very powerful and complete. We had a very detailed dataset from the aTO (all 148 columns of it) and a list of how the authorities split “Metro” and “Country” Victoria. We knew this would let us implement interesting data slices, but we were just short of an interesting objective.

While discussing the columns in the ATO dataset, we found one that stroke as as very different to all other: “Charitable Donations”. We considered possible applications and realised that some work on this field could be useful for charities seeking contributors, by analising how likely they are to donate.

After discussing this idea we became really invested in the project, as it seems like a very useful outcome in the right hands.

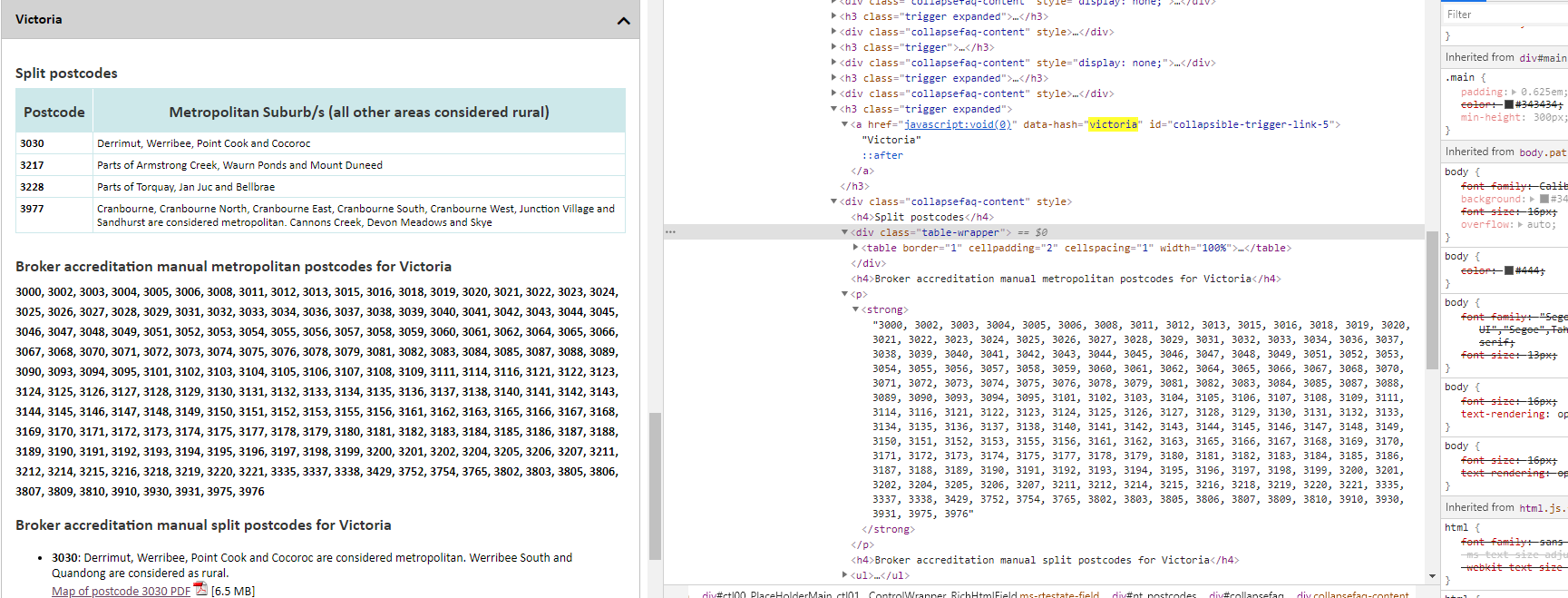
To make this information easier to present, we found another component: a csv file with all Australian postcodes and their corresponding suburb, published by Matthew Proctor ([link](https://www.matthewproctor.com/full_australian_postcodes_vic)). This also contains latitude and longitude, which could be useful for visualisations. Time permitting, this is a functionality we would love to implement.

Final Datasets:

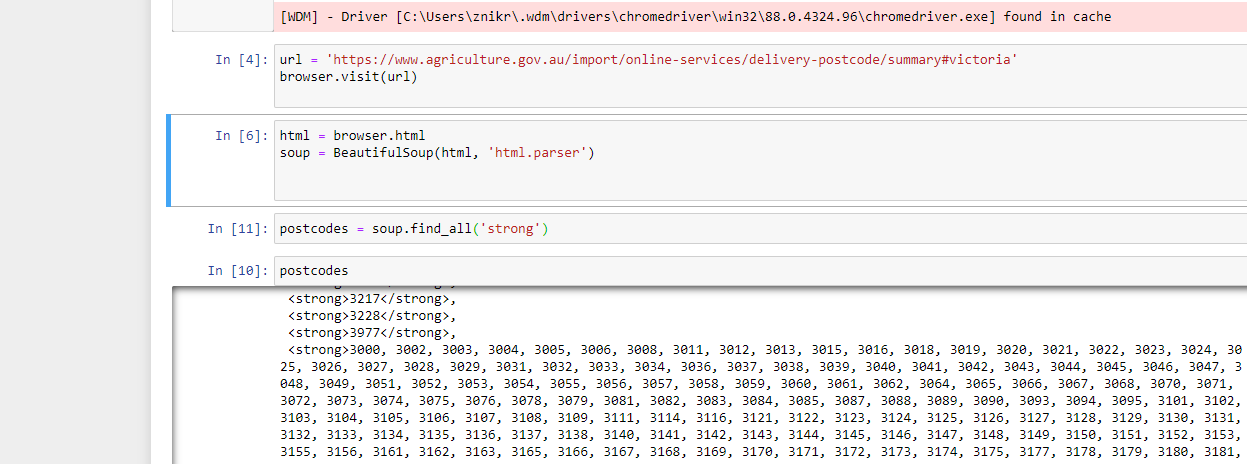
1. Postcode clasification - Metropolitan vs Rural from Agriculture Department website (retrieved 23-FEB-21 [link](https://www.agriculture.gov.au/import/online-services/delivery-postcode/summary#victoria)) This information will be scraped from the website.
2. Australian postcode dataset – offered by Matthew 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 information relates to individuals and aggregates information for the 2017-2018 tax year.

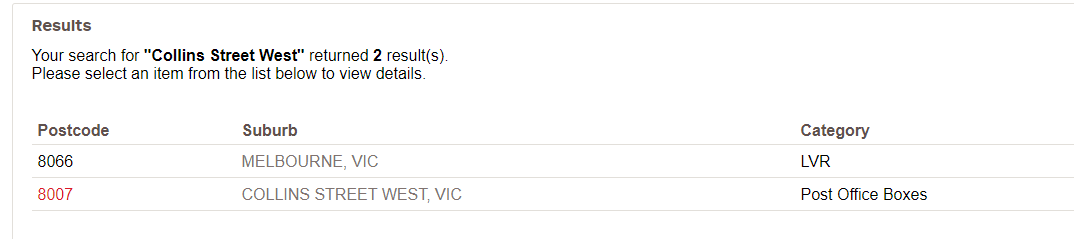
**Data cleansing**

Our first challenge was scraping the agriculture website (Source 1), which wasn’t as straightforward as expected, with the codes themselves embedded in the website text.

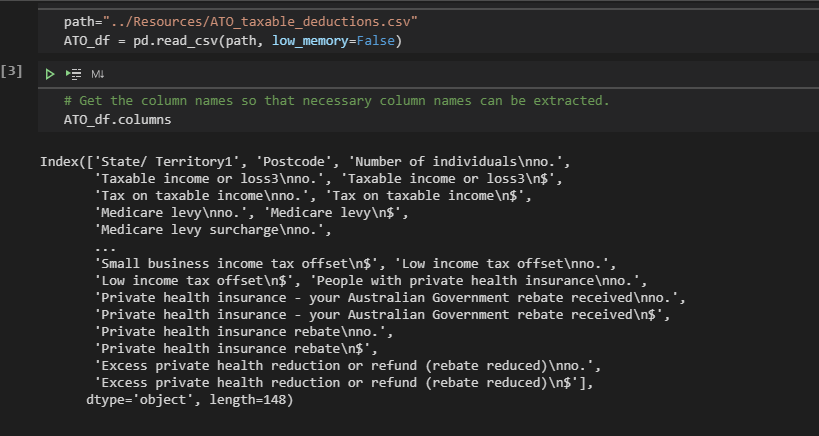


Through the implementation of BeautifulSoup we managed to get all postcodes, having to do some additional cleansing to reach our final objective: a list of all Melbourne Metro codes.

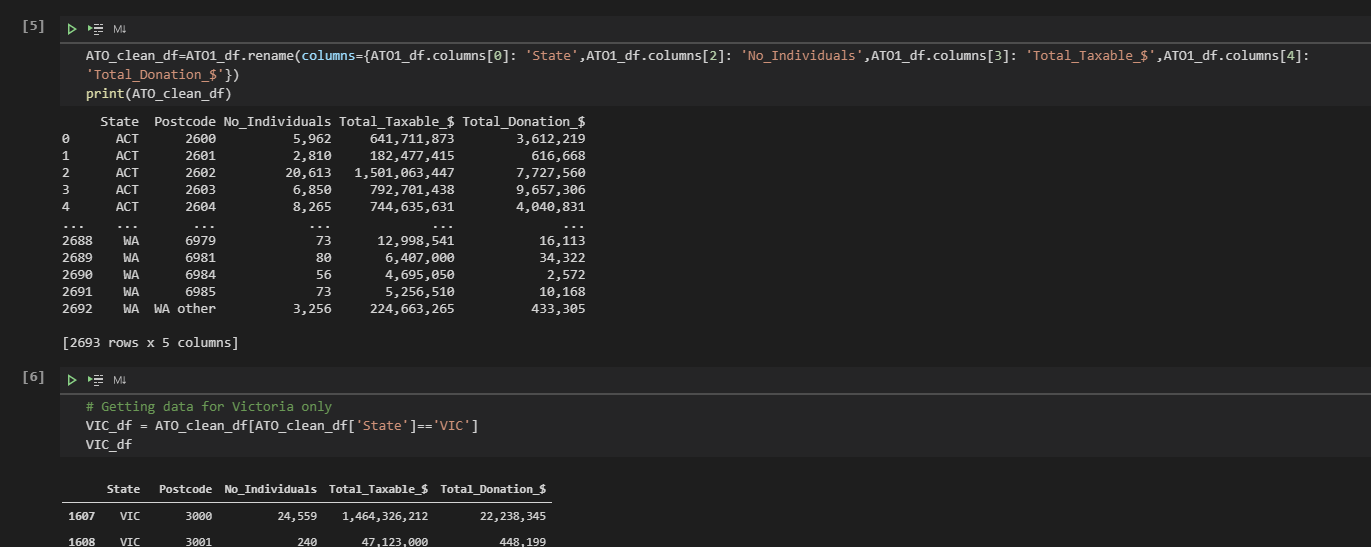


While working on this we made interesting findings in relation to postcodes: there are some postcodes in Melbourne that we have never heard of, such as 8066 (the postboxes in the Melbourne GPO) or 8010, that belongs to the University of Melbourne. These can be safely disregarded from our analysis, as we don’t expect any people to reside in these locations. 

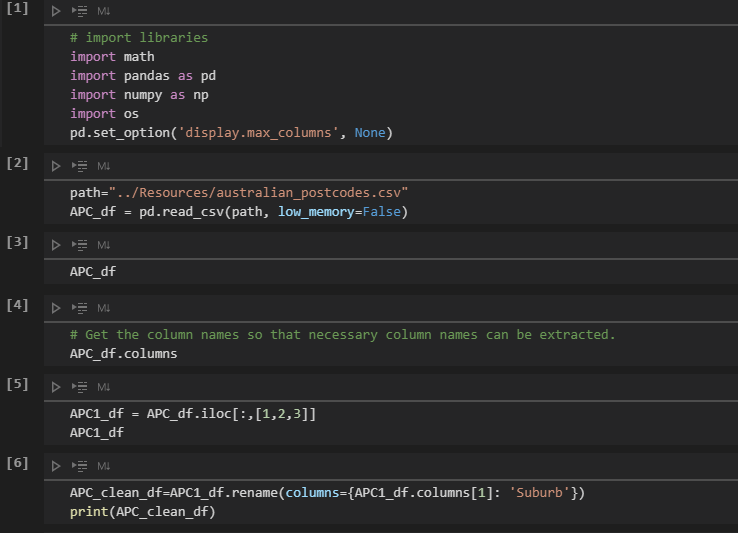
Exploring the ATO file (Source 3) we concluded that it was way too large for our purposes, with 143 columns, most of which are not relevant to our analysis.



We started work on dropping columns that are not relevant to our analysis, to minimise resources.



Source 2 presented a more modest challenge, being a smaller dataset with very clear columnes. We simply dropped the values for states other than VIC.



ADD PROCESS ON IMMIGRATION INFORMATION IF IMPLEMENTED

With this, the data is prepped and we are ready to load these datasets into a database and start working on our calculations.