**Report**

**Idea and Exploration**

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 the 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 [link](https://www.agriculture.gov.au/import/online-services/delivery-postcode/summary#victoria). 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 “Regional” Victoria. We knew this would let us implement interesting data slices, but we were just short of an interesting objective. It’s worth mentioning that for the purposes of this report we will utilise the words “country” and “regional” interchangeably, though the term we use in the datasets is always “Regional”.

While discussing the columns in the ATO dataset, we found one that caught our eye, as it was very different to all others: “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.

After some more exploration, we decided to test whether the list of postcodes from the Department of Agriculture was consistent with other sources. We looked at the Department of Immigration ([link](https://immi.homeaffairs.gov.au/visas/working-in-australia/skill-occupation-list/regional-postcodes%22)), which has a specific list of postcodes considered regional.

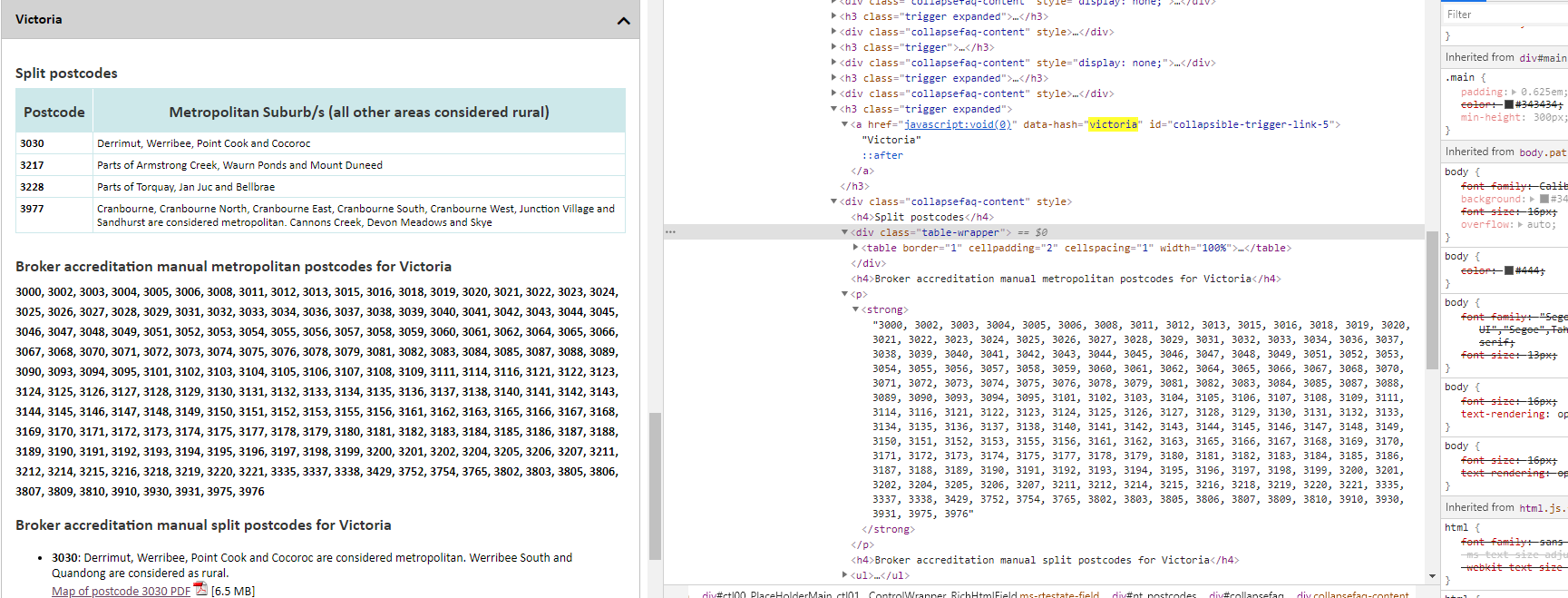
While we did some early work mashing both datasets, we found an issue that was unforeseen: certain postcodes were not considered in the Department of Agriculture as “Metro”, actually belong in Melbourne but for administrative reasons are not considered residential but where some donations are recorded. To avert this issue, we opted for using the Department of Immigration list, as it states regional

Discarded source: Postcode classification - 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.

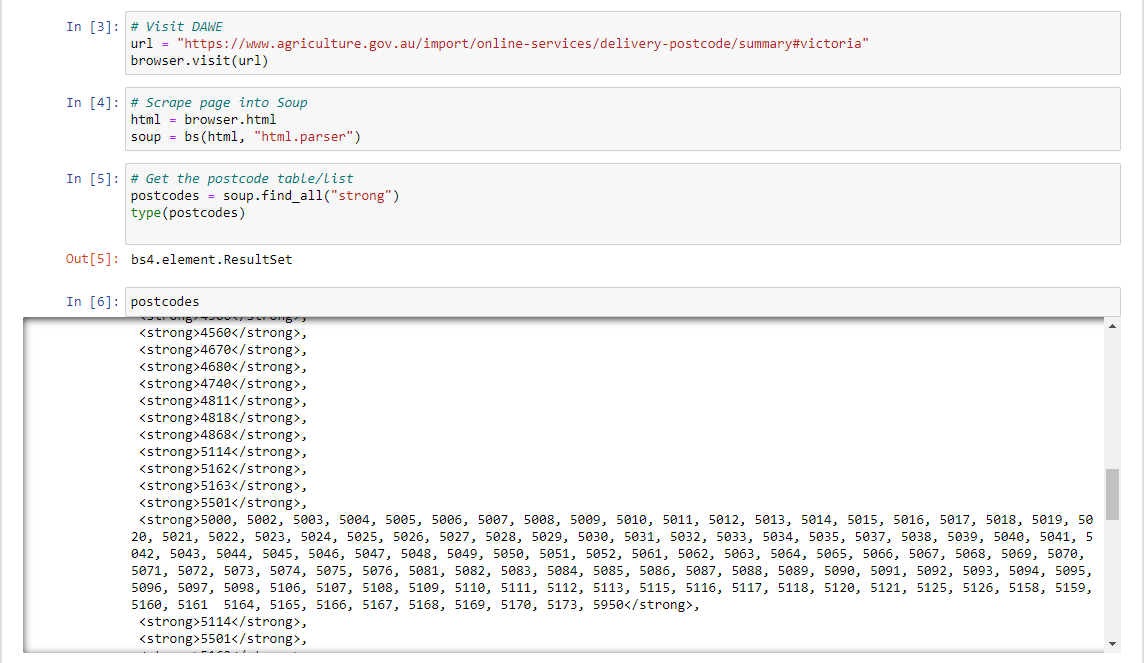
**Extraction**

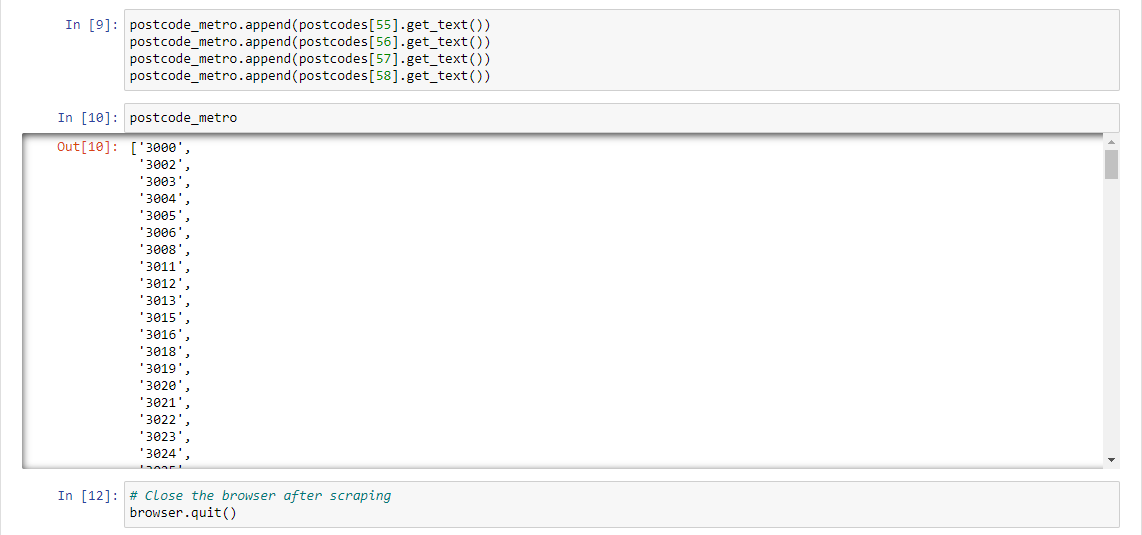
Final Sources:

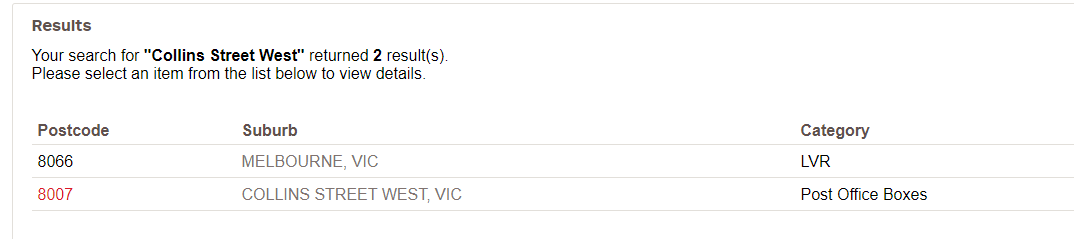
1. Designated Regional Area Postcodes- Department of Immigration ([link](https://immi.homeaffairs.gov.au/visas/working-in-australia/skill-occupation-list/regional-postcodes)),
2. Australian postcode dataset – offered by Matthew Proctor ( [website](https://www.matthewproctor.com/australian_postcodes) [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.



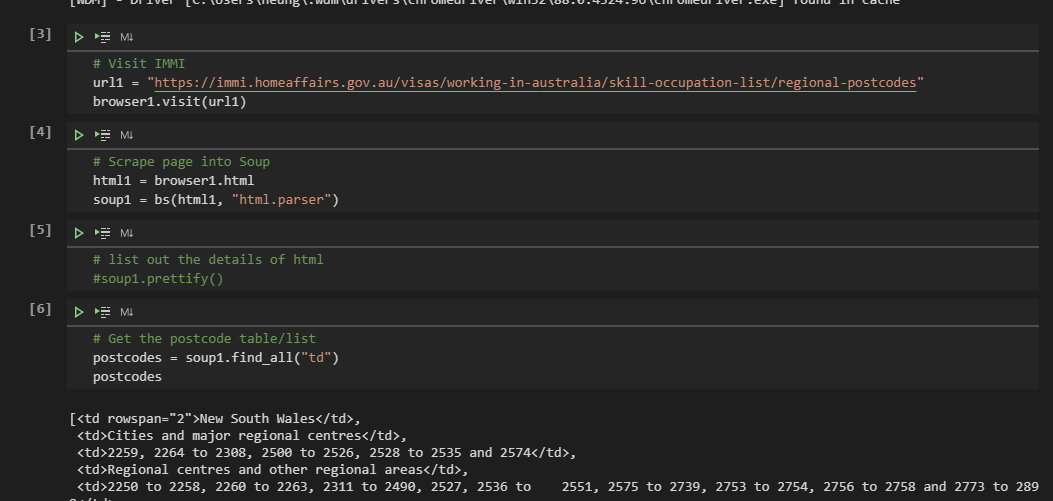
Through the use of BeautifulSoup on the Department of Agriculture website, we managed to get all “Metro” postcodes, having to do some additional cleansing to reach our final objective: a list of all Melbourne Metro codes.



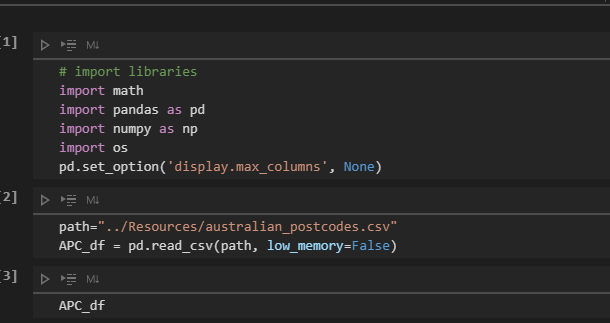


While working on this we made interesting findings: there are some postcodes in Melbourne that we have never heard of, such as 8066 (the postboxes in the Melbourne GPO) or 8010, that belongs to the University of Melbourne. 

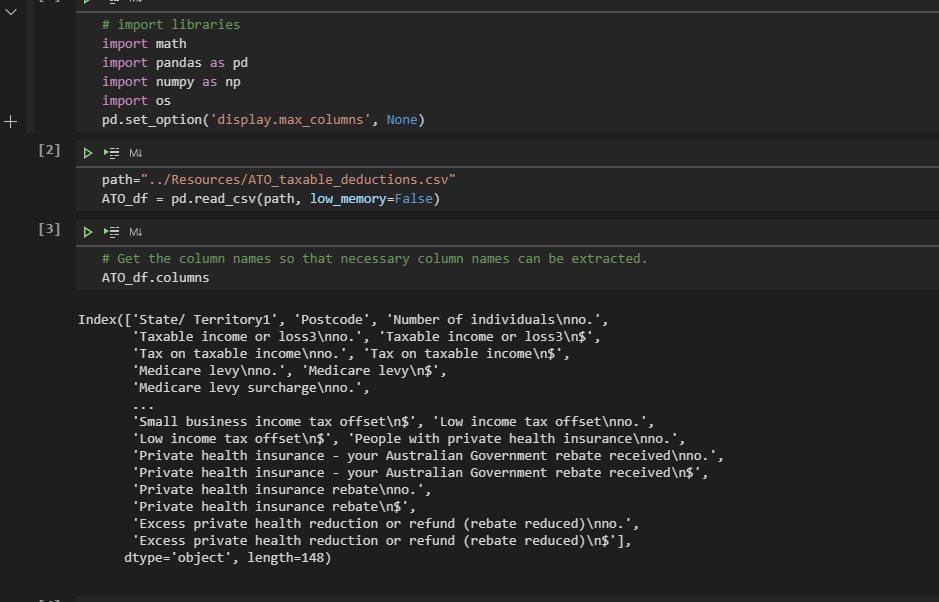
As explained on the Exploration section, this discovery suggested we discard the existing dataset and use the Department of Immigration one(Source 1), which involved starting our scraping from scratch.



Source 2 (Australian Postcodes) was a stragithforward import



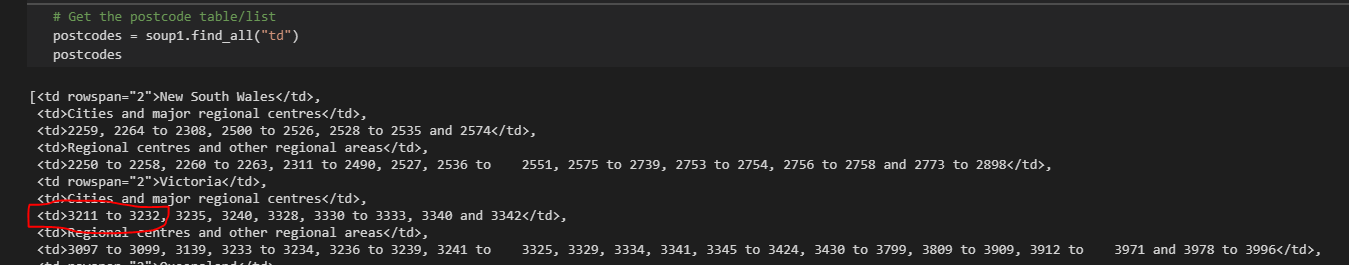
Source 3 (ATO file) was already in CSV, so importing it was not challenging



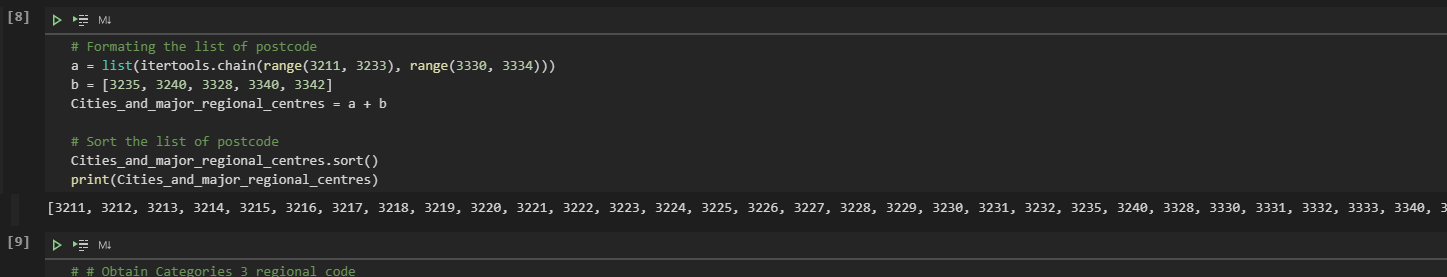
**Transformation**

**Data cleansing**

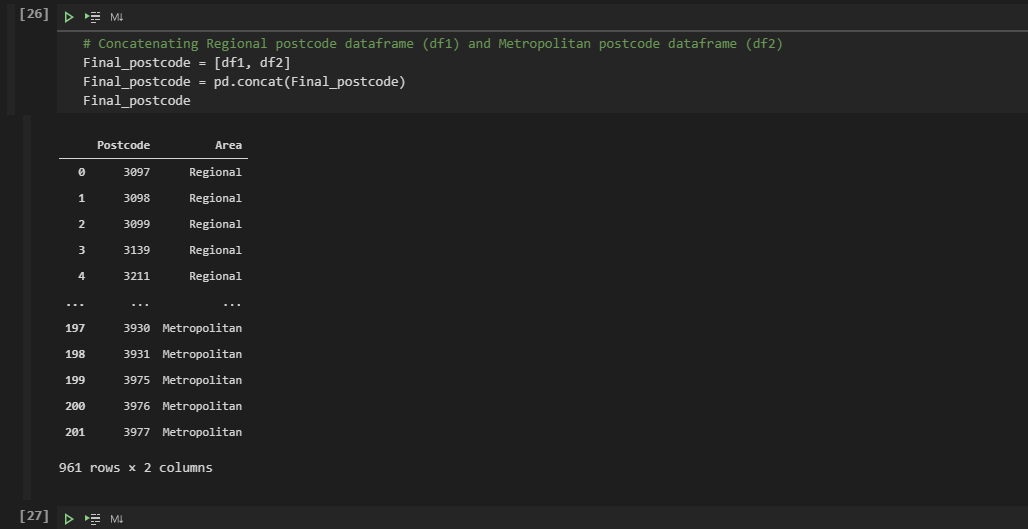
Our first challenge was cleaning the scraped data from the Department of immigration (Source 1), which wasn’t as straightforward as it might seem. Several of the postcodes are expressed as ranges, while others are listed:



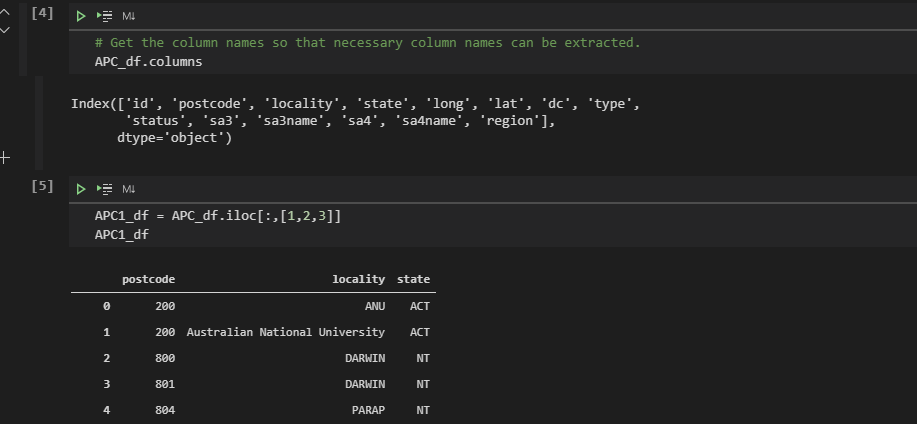
Usinge the *itertools*  library, we managed to fill in the missing items in each range:

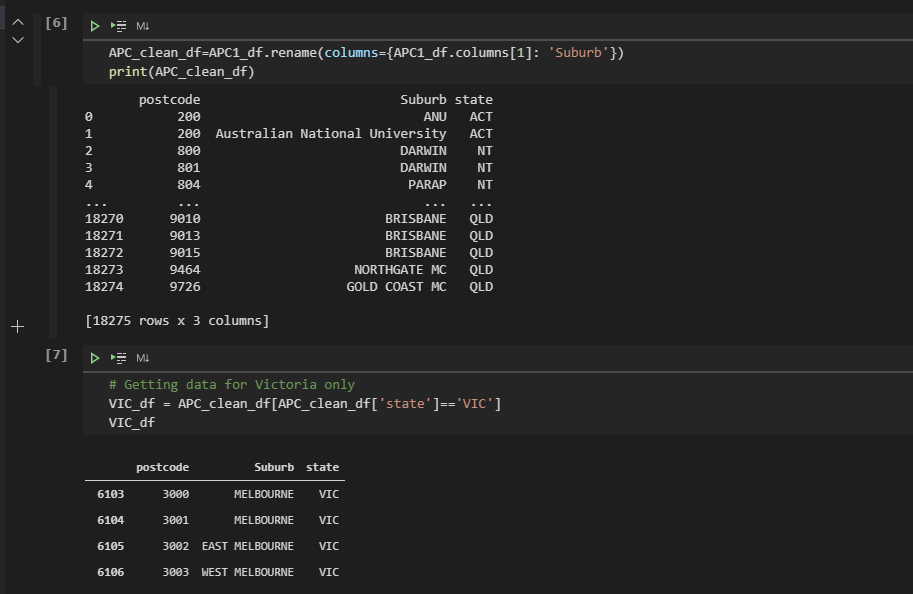


This allowed us to apply labels to each postcode, based on wether they are considered “regional” by the source or not. By default, those that aren’t were labelled as Metro.

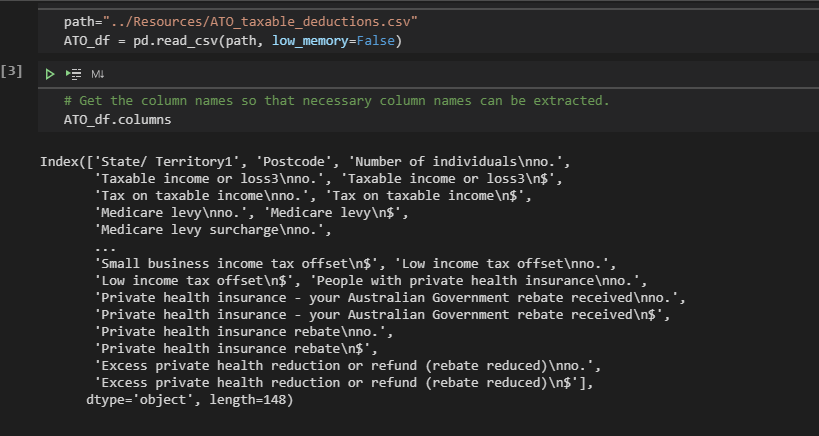


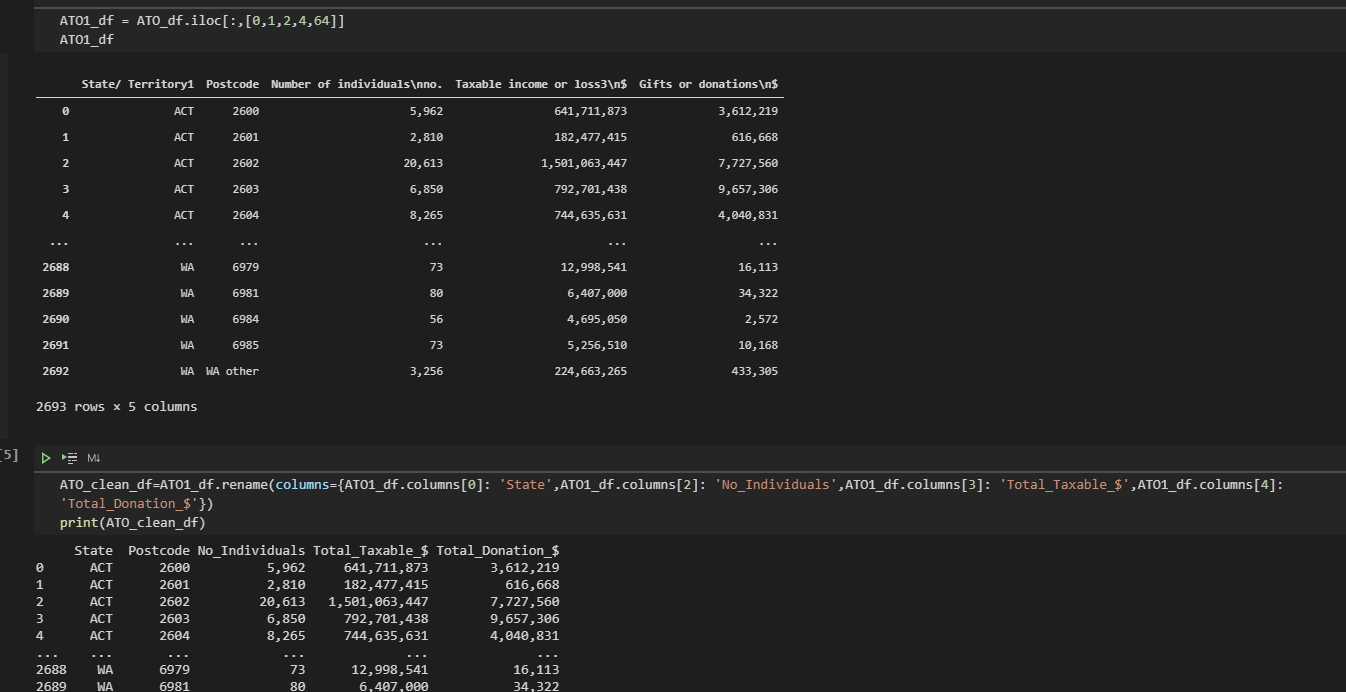
Source 2 presented a more modest challenge, being a smaller dataset with very clear columns. We dropped the columns that aren’t required:

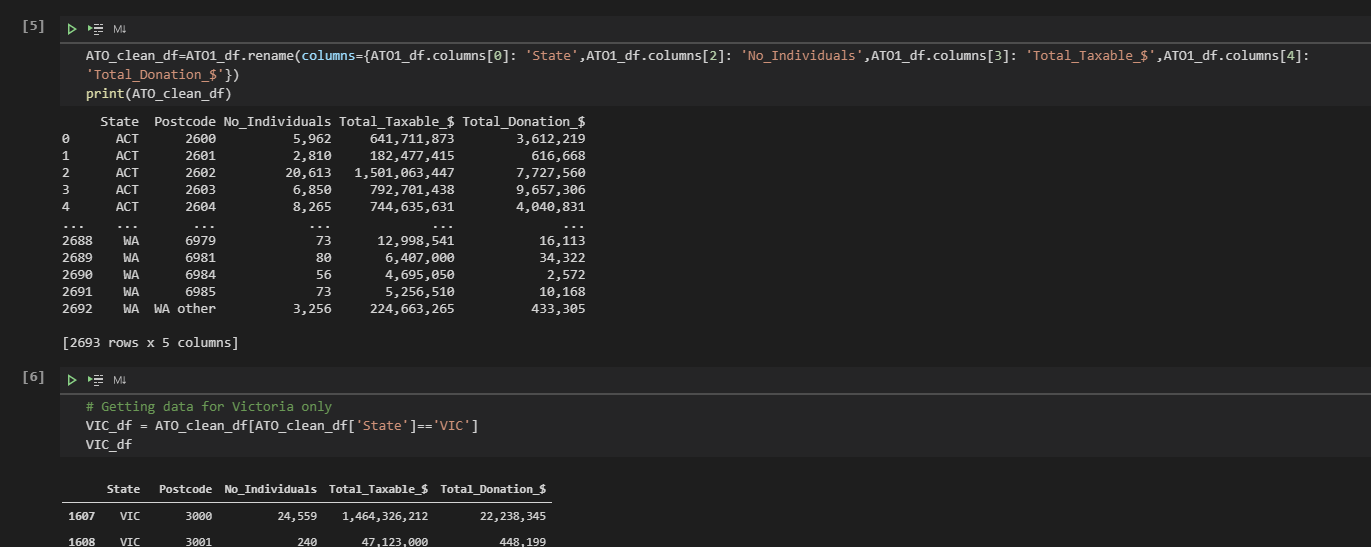


Additionally, we renamed columns and removed values for states other than VIC.

We had noticed during the exploration phase that the ATO file (Source 3) was way too large for our purposes, with 148 columns, most of which are not relevant to our analysis:

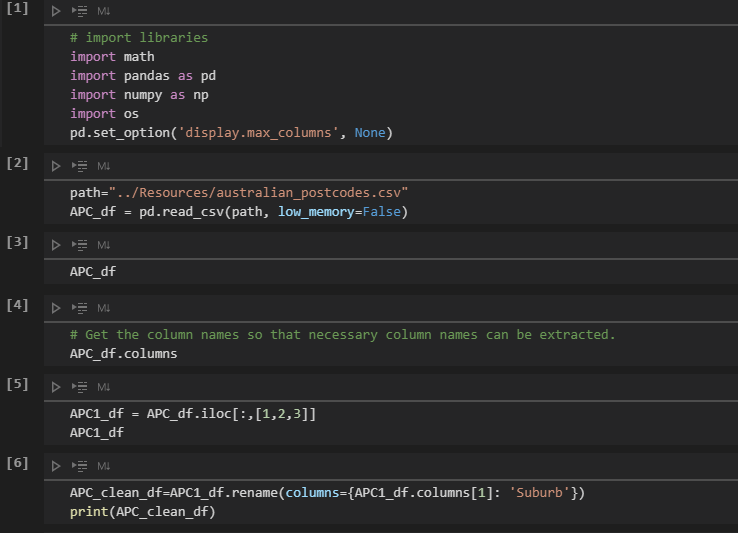


We dropped most of the columns to minimise resource usage and renamed the remaining ones using titles that are more convenient to manage:



Have finalise these transformations, we began working on the next step of this project: creating the framework to load the database using PostgreSQL.

LOADIN



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.