Summary of findings here (abstract?)

1. **Intro (Research Question, Data Sets Used)**
2. **Wrangling and Analysis Methods**

Because we use a variety of data sets, there were many data files we needed to process and collect all into one table. Thus, automating workflows in Python was key. In the initial wrangling steps we first examined files to study the structure of our data sets before attempting to read them in via Python and Pandas. One example issue we found was that the VCAMs data sets, we found that while most files had a regular tabular format, a few files had differing column names or extra columns that made it harder to use the same function. Some other issues included depression rate data in .pdf format, inconsistent area names (‘Greater Melbourne’ vs ‘Greater Melbourne *Area*’), and non-numeric data. Some examples of non-numeric data points found are ‘<5’ in the TSDR data, ‘NDP’ in the VCAMs data, and asterisks to replace missing values in the AEDC data set. Missing values were imputed as such in Python.

We used a combination of manual cleaning via Excel where it was trivial to do so or not worth automating, and Python for more automated cleaning, particularly for the VCAMs datasets. Picking out columns for depression rate and AEDC features were done manually, then processed further in Python together with the other data sets. Some other tasks done by hand include ensuring a consistent structure for VCAMs data, renaming files, and finding the index number of tables in the TSDR report.

Data sets were linked together by merging features over common keys – either LGAs or DHS Areas together with respective depression rates in young people for each region.

1. **Discussion of Results (Results, Significance and Value, Limitations etc.)**