R Processing Pipeline – Adam Bernabeu

All the following codes can be found in /Dropbox/Elite_persistence/merging_work/code.

1. 1.Clean.R

Purpose: Clean raw multi-sheet census Excel files (1947–1996), standardize codes, and prepare year-specific cleaned datasets.

- **Inputs:** raw_data/*.xls (except 1966, 2006)
- **Outputs:** intermediate_data/cleaned_output_{YEAR}.xlsx
- **Key Points:** Filters to qism/markaz rows, horizontally merges sheets, fixes population mismatches, handles code format quirks (apostrophes, .0 endings, leading zeros). Special handling for 1986 and 1996 code columns.

2. 2.Merge.R

Purpose: Combine all cleaned census years into a long panel and attach 1947 codes for cross-year comparison.

- **Inputs:** intermediate_data/cleaned_output_{YEAR}.xlsx (including 1947)
- **Outputs:** final_data/merged_output_census.xlsx
- **Key Points:** Merges on Code_1996, drops extra columns, reorders variables, keeps code formatting safe for Excel.

3. 3.Final Merge Pop Land Parliament.R

Purpose: Merge land distribution data (1950–1961) with census panel; ready dataset for analysis and visualisation.

- **Inputs:** final_data/merged_output_census.xlsx, intermediate_data/land_dist_1950_1961.xlsx
- **Outputs:** final_data/final_merge_land_census.xlsx
- **Key Points:** Normalizes Code_1996 formats, adds "00" suffix for land data alignment, cleans final columns, writes auto-sized Excel.

4. Districts existing each year.py

Purpose: Create subset of districts present in all years since 1947.

- **Inputs:** final_data/final_merge_land_census.xlsx
- **Outputs:** final_data/districts_all_years_since_1947.xlsx
- **Key Points:** Trims codes to base level, checks complete year coverage.

R Visualisation and Analysis Tools

6. EgyptMapR — Reshape and District Count Plot

Purpose: Reshape dataset so each year's district names/codes appear as separate columns, and produce a bar plot of district counts by year.

- **Inputs:** final_data/final_merge_land_census.xlsx
- **Outputs:** final_data/reshaped_by_code_1996.xlsx (reshaped dataset), bar plot showing district count by year.
- **Key Points:** Cleans Code_1996, aligns years to 1996 base codes, merges year blocks horizontally.

7. RShiny Display — Interactive Dashboard

Purpose: Provide an interactive web app for exploring district-level demographic variables over time.

- **Inputs:** final_data/final_merge_land_census.xlsx, Egypt district shapefile (egy_admbnda_adm2_capmas_20170421.shp)
- **Outputs:** Interactive dashboard with:
    - Coloured map by year/variable
    - Aggregate trends across all districts
    - District-specific variable bar plots
    - District-specific trends over years
    - PNG export functions for all visuals
- **Key Points:** Dynamic variable filtering per year, clickable map to zoom on districts, consistent Code_1996 cleaning for shapefile join.

Workflow Order

1. **Clean** → 1.Clean.py
2. **Merge census years** → 2.Merge.py
3. **Merge with land data** → 3.Final Merge Pop Land Parliament.py
4. **Subset complete panel** → Districts existing each year.py
5. **Reshape & basic plots** → EgyptMapR script
6. **Interactive analysis** → RShiny Display script

General Notes:

- Code formatting of district IDs (Code_1996) is critical across all scripts.
- Apostrophes are used in Excel outputs to preserve leading zeros.
- Land merge logic assumes 4-digit base codes with a "00" suffix.
- You might have to change the directory for inputs in each code. The inputs are currently considered to be in ~/Dropbox/...
- R scripts assume final_merge_land_census.xlsx exists and has been cleaned by Python pipeline.