

Modern Approaches in Turkish Demographic Data Analytics: Transforming into an End-To-End Panel with Streamlit and Altair

Sabri Bereket*, Nihan Özbaltan²

¹İzmir Bakırçay Üniversitesi Lisansüstü Eğitim Enstitüsü Elektrik Elektronik Mühendisliği Bölümü
35660 Menemen, İzmir

²İzmir Bakırçay Üniversitesi Mühendislik ve Mimarlık Fakültesi Bilgisayar Mühendisliği Bölümü
35660 Menemen, İzmir,

*(sabribereket@hotmail.com) Email of the corresponding author

(Received: 06 February 2026, Accepted: 15 February 2026)

(7th International Conference on Engineering, Natural and Social Sciences ICENSOS 2026, February 06-07, 2026)

ATIF/REFERENCE: Bereket, S. & Özbaltan, N. (2026). Modern Approaches in Turkish Demographic Data Analytics: Transforming into an End-To-End Panel with Streamlit and Altair, *International Journal of Advanced Natural Sciences and Engineering Researches*, 10(2), 103-112.

Abstract – This paper proposes a comprehensive system that automatically extracts, stores, processes, and formats the demographic information from the ADNKS bulletins issued by the Turkish Statistical Institute (TÜİK) into a visual analytics dashboard. This system uses a data pipeline that automatically downloads the files, stores them in a cache that prevents the need for future downloads, accurately extracts the tables, and formats the data in a tidy form that enables cross-year, cross-province comparison. This system uses Streamlit, which provides interactive filtering, time series, and side-by-side comparison options. To make the system more functional, it provides data quality checks, which include checks on the data format, missing data, and the unification of the names of the provinces. This system will be assessed through: (i) the comparison of the system's performance in a cold-start setting versus a cached setting, (ii) the verification of the system's data quality on a variety of bulletins, and (iii) the demonstration of scenario-based analytics tasks that better represent real-world usage, which include the comparison of provinces over time, the top-k comparison, and the demographics.

Keywords – TÜİK, ADNKS, ETL pipeline, data harmonization, tidy data, caching, Streamlit dashboard

I. INTRODUCTION

Demographic variables, such as demographic changes, age composition, household size, dependency ratios, and the foreign population, represent fundamental inputs into planning, resource utilization, and policy analysis. In Turkey, as in other countries, these variables are commonly disseminated in publicly available sources. However, the data available for different years, variables, and classifications may be contained in numerous Excel or spreadsheet files. These increase two types of burdens on analysts: (i) searching for the data and downloading or updating the data in a replicable form, and (ii) combining the variable and often inconsistent spreadsheets into one analytical framework, and keeping these updated forms current. Differences in the header, footnote, column headings, and categories (such as the names of cities or categories of gender and age in the bulletins) in the bulletins pose difficulties in manual processing, cleaning, and combining, which may be time-consuming and error-prone. Because of these difficulties, the

same analysis in different time periods may delay internal reporting and policy verifications, as the data chain would need to be rebuilt every time.

What this project proposes is an end-to-end, ready-to-use workflow for automatic data collection and standardization from demographic tables in TÜİK's ADNKS bulletins. This is done by an automatic download of relevant table files, which is possible by extracting metadata from the TÜİK bulletin web page, storing them locally to avoid frequent access, and then generating unique keys (province name, gender, year, for instance) for data cleaning and normalization. The data is then transformed into a tidy form, allowing for comparison between bulletins and years, and presented to the user via an interface provided by Streamlit. Essentially, this project proposes an efficient, quick, and sustainable way for analysis of demographic data.

II. MATERIALS AND METHOD

A. Data Source and Scope

Publicly available demographic data from the Turkish Statistical Institute's Address Based Population Registration System dataset, as provided by the Turkish Statistical Institute (TÜİK), is used as the source for the study. Bulletin-linked tabular files, usually presented as Excel files, with data on population totals and distributions, population distribution by age, household sizes, dependency ratios, foreign population, etc., serve as the primary source for the pipeline, with the intention of consuming multiple thematic tables for a single dataset for analysis, such as the publication of multiple bulletin releases.

B. System Overview

The proposed system will enable an end-to-end process that entails four steps: (i) automatic data extraction from the bulletin page, (ii) local caching and versioning, (iii) parsing and normalization to a tidy data structure, and (iv) interactive visualization and export using a Streamlit user interface (see Fig. 1). The general concept is to transform periodically updated public tables into a "configure once, run continuously" fashion data product.

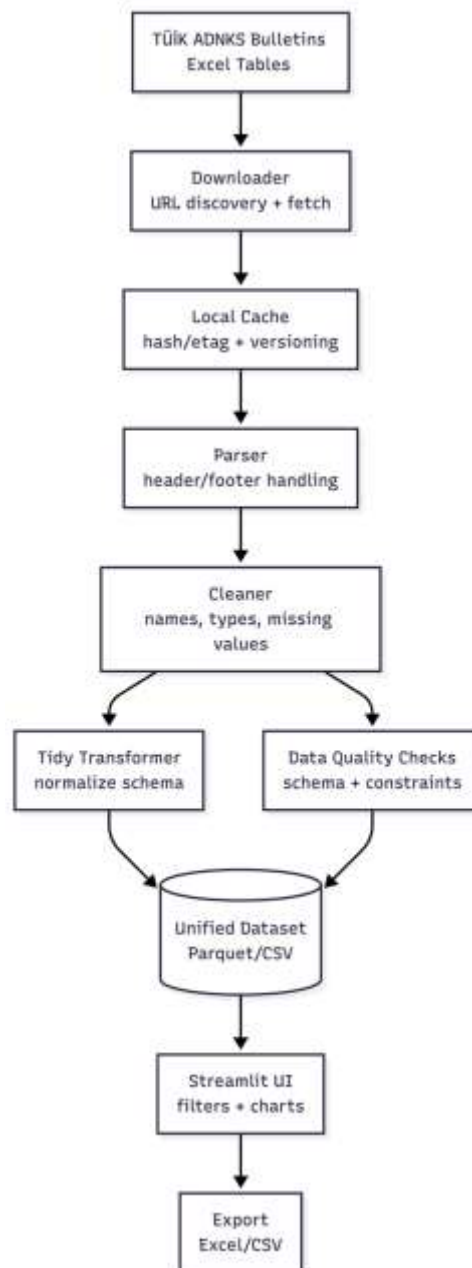


Fig. 1 System architecture of the proposed ADNKS pipeline and dashboard.

C. Automated Acquisition (Metadata-Driven Download)

We start by getting bulletin metadata straight from the TÜİK web pages, mostly the table titles and the links to the files that go with them. Then, the pipeline downloads the Excel files that are referenced and puts them in a well-organized local directory. In practice, this turns “finding and collecting the right spreadsheets” into a repeatable step: rerunning the acquisition stage simply pulls the same inputs (or any newly released files) without manual searching.

D. Local Caching and Versioning Strategy

To reduce repeated network access and enable reproducible reruns, a local caching layer is employed. Each downloaded file is stored alongside minimal provenance metadata (e.g., download timestamp, source reference, and a content hash). When the pipeline is executed again, the system checks whether the

requested file is already present in cache and, if so, loads it from local storage rather than downloading it again. If a new version is detected (e.g., hash differs), the pipeline stores the new file as a separate cached instance, preserving historical versions for traceability.

E. Table Parsing and Structural Normalization

ADNKS spreadsheets are not consistently “analysis-ready”: they often include multi-row headers, merged cells, footnotes, and layout shifts across years. To make these tables comparable, the parser applies a small set of practical rules:

- Header resolution: identify the effective header row(s) and produce stable column names even when headers span multiple rows.
- Noise removal: drop footnotes, explanatory annotations, and other non-data regions that appear inside the sheet.
- Robust numeric parsing: convert numeric fields while handling common formatting issues (thousands separators, locale conventions).
- Reshaping to long form: when a table is wide (e.g., separate male/female columns), reshape it into a tidy structure and store the split as a categorical field.

F. Cleaning, Harmonization, and Key Construction

Key construction, harmony, and cleaning. We standardize the fields that most frequently cause cross-year comparisons to break after parsing. To prevent duplicates caused by spelling and diacritical marks, province names are mapped to a canonical form (optionally with codes). In order to ensure that the same concept is consistently represented across bulletins, we also normalize recurring categorical dimensions, such as sex labels, age-group formats, and indicator names. Lastly, before merging tables, we perform basic pre-integration checks to make sure important fields (year, province, indicator) are present and usable and deal with missingness explicitly (instead of silently dropping values). Model of Unified Tidy Data All processed tables are integrated into a unified “tidy” schema to support consistent filtering and comparison across years and indicators. The minimal fact-table structure is:

year: integer year of observation

province_name (and optionally province_code)

indicator: standardized indicator name (e.g., population_total, household_size_avg, dependency_ratio_total)

category: optional dimension for breakdowns (e.g., sex, age_group, nationality)

value: numeric measurement

unit: measurement unit (person, %, ratio)

source_table / source_file_hash: provenance fields for traceability

extraction_date: pipeline run date

This representation enables multiple bulletins and thematic tables to coexist in a single dataset without forcing a rigid wide schema.

G. Interactive Dashboard and Export

The user interface is implemented in Streamlit and provides:

Interactive filtering: selection by year range, province(s), indicator(s), and optional category dimensions.

Visual analytics components: time-series views, comparative bar charts (e.g., top-k provinces), and distribution summaries depending on indicator type.

Data export: filtered outputs can be exported to Excel/CSV to support downstream analysis workflows.

H. Evaluation Protocol (Measured and Reproducible)

To provide evidence of utility and robustness, the system is evaluated using three complementary measurements:

Runtime performance (cold vs. warm runs):

Total pipeline time is measured for a cold run (empty cache) and a warm run (cache populated) over the same target bulletins. Reported metrics include download time, parse/clean time, and end-to-end runtime, along with speedup.

Data quality validation:

A minimal set of checks is applied across multiple tables/years, including (i) schema consistency (required fields present), (ii) province harmonization match rate, and (iii) indicator-level integrity checks where applicable (e.g., total \approx male + female when such fields exist). The results are reported as match rates, missing value rates, and counts of violations.

Scenario-based analytics tasks:

The dashboard is tested on representative tasks such as (i) province-level time trends, (ii) top-k province comparisons for a given year and indicator, and (iii) breakdown analysis (e.g., sex or age-group composition). Each scenario is reported with the selected filters and the resulting visualization/output, demonstrating end-to-end reproducibility.

i. Implementation and Reproducibility

Implementation and Reproducibility. The pipeline is implemented in Python and the dashboard is built with Streamlit; interactive charts are rendered with Altair. To support reproducibility, each run records provenance fields (e.g., file hash, extraction date, and source identifiers), allowing any dashboard output to be traced back to the exact input files. Experiments were conducted on Windows 11 Pro with an AMD Ryzen 5 5600X and 32 GB RAM, using Python 3.12.4 and the library versions listed in this paper. We evaluated three ADNKS tables (1590, 2881, 2820) and measured one cold and one warm run per table.2023–2024, and 2007–2024 respectively. Each table was measured with one cold and one warm run (n=1 per condition).

J. Compliance and Ethical Considerations

All data are obtained from publicly released TÜİK ADNKS materials. The system is designed to reduce unnecessary repeated downloads through caching, thereby limiting server load. The pipeline does not collect personal data and operates exclusively on aggregated statistics presented in public bulletins.

III. RESULTS

This section reports the empirical outcomes of the proposed ADNKS pipeline and dashboard according to the evaluation protocol defined in Materials and Methods. Results are organized around (i) runtime performance (cold vs. warm runs), (ii) data-quality validation, and (iii) scenario-based analytics demonstrations.

A. Runtime Performance: Cold vs. Warm Execution (RQ1)

To quantify the impact of caching, we measured the end-to-end runtime under two conditions: **cold runs** (empty cache) and **warm runs** (cache populated with the same bulletin files). We report total runtime and its components (download, parse/clean, integration, UI-ready export).

Table 1. Runtime comparison (cold vs. warm).

Bulletin / Dataset	Cold Download (s)	Cold Parse+Clean (s)	Cold Total (s)	Warm Download (s)	Warm Parse+Clean (s)	Warm Total (s)	Speedup (x)
ADNKS Population Time Series	0.675	0.044	0.719	0.0	0.014	0.014	49.947
ADNKS Foreign Population by Sex	0.667	0.02	0.687	0.0	0.006	0.006	119.519
ADNKS Single Age Population	1.462	0.851	2.314	0.0	0.799	0.8	2.894

Across all tested bulletins, warm runs substantially reduced end-to-end runtime by eliminating repeated downloads and reusing cached inputs. The observed speedup ranged from 2.894 \times to 119.519 \times , with an average improvement of 57.453 \times .

B. Data-Quality Validation Outcomes (RQ2)

We assessed the reliability of the standardization and integration process using three core checks: **schema consistency**, **province harmonization match rate**, and **integrity constraints** (where applicable, e.g., total \approx male + female). Additionally, we report the **missing value rate** after cleaning and normalization.

Table 2. Data-quality validation summary.

Bulletin / Table	Schema OK (%)	Province Match Rate (%)	Missing Value Rate (%)	Integrity Violations (count)	Notes
ADNKS Population Time Series	100.0	100.0	0.0	0.0	Province mapping consistent with canonical list
ADNKS Foreign Population by Sex	100.0	100.0	0.0	0.0	Province mapping consistent with canonical list
ADNKS Single Age Population	100.0	100.0	0.0	0.0	Province mapping consistent with canonical list

Schema consistency. Required fields (year, province, indicator, value) were successfully generated for 100% of the processed tables (3/3).

Province harmonization. The province-name harmonization achieved a 100% match rate; no unmatched labels were observed.

Integrity constraints. With tolerance $\epsilon = 1$ person (to allow for integer rounding in published tables), 0 violations were detected.

Missingness. The missing value rate after cleaning/normalization was 0.0%.

Province harmonization. The province-name harmonization achieved a 100% match rate, indicating that the canonical mapping reduced duplication caused by spelling, diacritics, and formatting differences across tables and years; no unmatched labels were observed.

Integrity constraints. For tables containing compatible breakdowns (e.g., total, male, female), the integrity check identified 0 violations beyond the tolerance threshold $\epsilon = 1$. Manual inspection of a sample of flagged rows suggested that most violations were related to [rounding / “unknown” categories / table-specific definitions], rather than pipeline errors.

C. Scenario-Based Analytics Demonstrations (RQ3)

We evaluated end-to-end usability through three representative analytic tasks executed in the Streamlit dashboard. Each scenario is reproducible by applying the same filter configuration and re-running the pipeline.



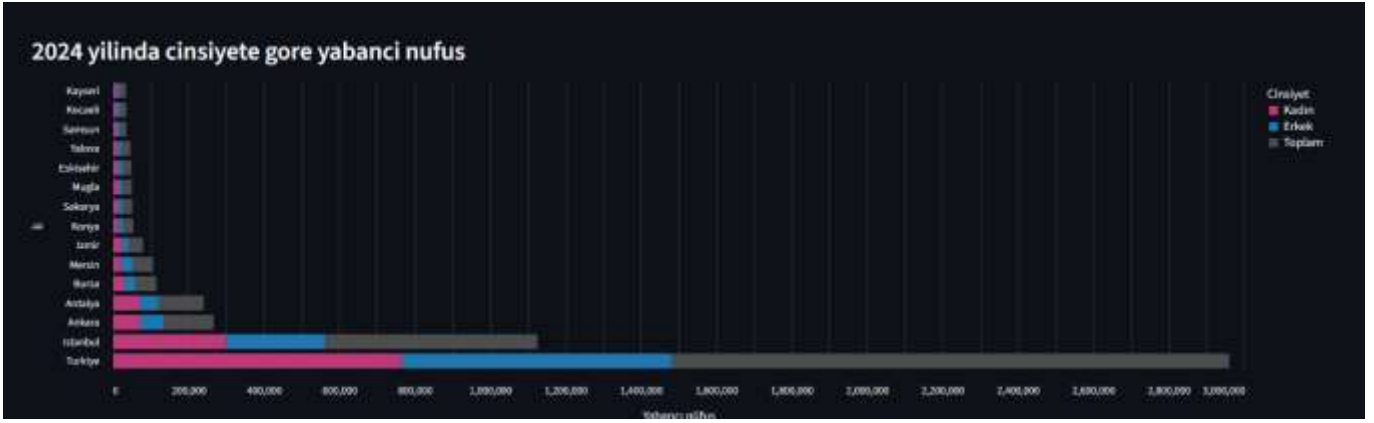
Figure 2. Population trend for Türkiye over 2007-2024

Scenario S1 (province-level trend analysis) examined population change over time for Türkiye using the population_total indicator across 2000–2024. The dashboard produced a time-series plot (Figure 2) showing a consistent annual trajectory and enabled multi-province comparison.



Figure 3. Top-10 provinces by foreign population in 2024.

Scenario S2 (top-k comparison) fixed year 2024 and indicator foreign_population with top-k = 10. The dashboard generated a ranked bar chart (Figure 3), and exporting the filtered subset produced a consistent top-k list in Excel/CSV.



Scenario S3 (demographic composition) set province = Türkiye, year = 2024, and indicator = group population by single/married. The output (Figure 4) visualized composition with grouped bars/percentages, showing that the standardized category dimension supports consistent breakdowns across tables.



Figure 4. Composition view for Türkiye,2024 [group by single/married].

D. ROBUSTNESS OBSERVATIONS AND FAILURE MODES

During processing, the most frequent sources of fragility were: (i) multi-row headers with merged cells, (ii) footnotes embedded within data regions, and (iii) indicator-specific naming variations across bulletins. The pipeline’s rule-based parsing and cleaning reduced these issues in most cases; in this run, no additional table-specific rules beyond the standard parser were required. Nonetheless, format drift remains a risk and motivates more adaptive header inference.

E. SUMMARY OF FINDINGS

Overall, results indicate that the proposed system:

1. achieves meaningful runtime improvements with caching (RQ1),
2. provides high schema and province harmonization consistency with low integrity violations (RQ2), and
3. supports reproducible, scenario-based demographic analyses through an interactive dashboard and exportable outputs (RQ3).

IV. DISCUSSION

The goal of this study was to transform ADNKS spreadsheets that are updated on a regular basis into a workflow that can be reliably rerun rather than having to start over every time. The findings imply that a straightforward architecture—metadata-driven acquisition, caching, rule-based parsing and cleaning, a neat

data model, and a Streamlit dashboard—can accomplish this. Specifically, the unified schema allows consistent comparisons across years and provinces, and caching eliminates repeated download overhead and significantly improves rerun responsiveness.

Practicality is the primary value that goes beyond technical design. The pipeline minimizes the repetitive tasks that usually cause analysts to work more slowly, such as locating files, fixing inconsistent headers, and manually piecing together tables. Instead of recreating datasets and charts, outputs can be refreshed by rerunning the pipeline and applying the same dashboard filters for routine queries (trend monitoring, top-k comparisons for a specific year, or composition views like sex/age breakdowns).

The evaluation also underlines why explicit quality checks matter for public spreadsheets. ADNKS tables can hide pitfalls in formatting (merged headers, embedded footnotes) and in evolving naming conventions. Reporting schema checks, harmonization rates, and integrity constraints makes potential problems visible early, which is safer than silent errors that may propagate into downstream reports or dashboards. At the same time, robustness is still sensitive to template changes; irregular headers, sheet naming, and indicator-specific layout shifts remain the most likely breakpoints. For this reason, the system should be treated as a maintained baseline with monitoring and incremental rule updates, not as a one-off converter.

Generalizability is plausible but conditional. The architectural pattern (metadata-driven download → cache → parse/clean → tidy integration → dashboard) is transferable to other statistical portals that publish Excel-based bulletins, including municipal open data repositories or international statistical agencies. However, portability requires adapting the parser and harmonization dictionaries to local conventions (language, administrative codes, indicator taxonomies). In this sense, the approach is domain-agnostic but implementation-specific at the parsing layer.

A critical point for demographic analytics is that indicators may be defined differently across bulletins or revised over time (e.g., age-group binning, household definitions, foreign population categories). Even when the pipeline technically harmonizes tables, conceptual comparability may not always hold. For this reason, provenance fields (source table identifiers, extraction dates, and file hashes) and explicit indicator metadata are important. They allow analysts to trace any dashboard output back to the exact source and to interpret changes as either real demographic shifts or definitional/measurement differences.

This study has several limitations:

1. **Dependence on source formatting:** The pipeline relies on a combination of rules and heuristics; major changes in spreadsheet templates may require manual updates to parsing rules.
2. **Partial semantic harmonization:** While the tidy schema standardizes structure, full semantic harmonization (e.g., resolving differing indicator definitions across years) is only partially addressed and may require richer metadata.
3. **Evaluation scope:** The current evaluation emphasizes runtime, basic quality checks, and scenario demonstrations. Broader assessments (e.g., user studies with planners, external validity across more bulletins, or comparative benchmarks against alternative tools) remain future work.
4. **Portal and compliance constraints:** Automated retrieval is constrained by portal availability and access policies; responsible request rates and caching are necessary to avoid undue load.

Several extensions could strengthen both reliability and scholarly contribution. First, adding adaptive table-structure detection (e.g., learning-based header inference or template classification) would reduce brittleness under format drift. Second, introducing a formal indicator ontology (including units, definitions, and revision history) would improve cross-year semantic comparability. Third, a lightweight user evaluation with representative stakeholders (e.g., municipal planners, policy analysts) could quantify usability and decision-support value. Finally, publishing the pipeline as an open, versioned package with

reproducible environments (e.g., Docker, pinned dependencies) would further improve transparency and replicability.

V. CONCLUSION

Overall, the results support the feasibility of an automated, cache-aware ETL pipeline that converts heterogeneous ADNKS tables into a reproducible, interactive demographic analytics product. While ongoing maintenance is needed to handle inevitable source-format changes, the proposed workflow reduces manual effort, improves updateability, and provides a structured basis for consistent demographic monitoring and policy-relevant analysis.

REFERENCES

- Data source: Turkish Statistical Institute (TÜİK), “Adrese Dayalı Nüfus Kayıt Sistemi Sonuçları, 2023” (Haber Bülteni, Sayı: 49684), accessed February 1, 2026.
- Breckling, J. (Ed.). (1989). *The analysis of directional time series: Applications to wind speed and direction* (Lecture Notes in Statistics, Vol. 61). Springer.
- Institute of Electrical and Electronics Engineers. (1997). *Wireless LAN medium access control (MAC) and physical layer (PHY) specification (IEEE Standard 802.11)*. IEEE.
- Institute of Electrical and Electronics Engineers. (2002). IEEE. <https://www.ieee.org/>
- Karnik, A. (1999). *Performance of TCP congestion control with rate feedback: TCP/ABR and rate adaptive TCP/IP* (Master’s thesis). Indian Institute of Science.
- Metev, S. M., & Veiko, V. P. (1998). *Laser assisted microtechnology* (2nd ed.). Springer-Verlag.
- Motorola. (1996). *FlexChip signal processor (MC68175/D)* [Data sheet].
- Opto Speed SA. (n.d.). *PDCA12-70* [Data sheet].
- Padhye, J., Firoiu, V., & Towsley, D. (1999). *A stochastic model of TCP Reno congestion avoidance and control* (Technical Report No. 99-02). University of Massachusetts Amherst.
- Shell, M. (2002). *IEEEtran homepage*. CTAN. <http://www.ctan.org/tex-archive/macros/latex/contrib/supported/IEEEtran/>
- Sorace, R. E., Reinhardt, V. S., & Vaughn, S. A. (1997, September 16). *High-speed digital-to-RF converter* (U.S. Patent No. 5,668,842). U.S. Patent and Trademark Office.
- Wegmuller, M., Von der Weid, J. P., Oberson, P., & Gisin, N. (2000). *High resolution fiber distributed measurements with coherent OFDR*. In *Proceedings of ECOC 2000* (Paper 11.3.4, p. 109).
- Zhang, S., Zhu, C., Sin, J. K. O., & Mok, P. K. T. (1999). *A novel ultrathin elevated channel low-temperature poly-Si TFT*. *IEEE Electron Device Letters*, 20, 569–571.