Dean Fletcher

Analyzing Voter Turnout Patterns and Demographic Factors in Florida Elections

1. Project Motivation and Research Question

This project addresses how demographic, socioeconomic, and geographic factors influence voter turnout patterns across Florida counties and how these relationships evolved across recent election cycles (2016-2024). Understanding civic participation determinants is critical for election administrators improving access, civic organizations developing engagement strategies, and researchers studying democratic participation. Florida provides an ideal case study due to its demographic diversity, swing state status, and comprehensive county level data availability across both presidential and midterm elections.

The primary objective was creating a high-quality, analysis-ready dataset integrating election turnout with demographic and economic indicators using rigorous curation practices aligned with the Digital Curation Centre (DCC) Lifecycle Model. This project directly applies course concepts from data integration (M6), lifecycle models (M1), metadata documentation (M8), identifier systems (M9), and workflow reproducibility (M12).

2. Dataset Profile

The integrated dataset contains **335 county year observations** (67 Florida counties across 5 election years) with **16 variables** and **zero missing values**. Four primary data sources were integrated:

Florida Division of Elections (2016-2024): County level turnout data for five general elections. Variables: registered voters, votes cast, turnout percentage. Manual HTML extraction required due to lack of API access. Updated within 30 days post-certification.

U.S. Census Bureau ACS (2016-2020 5-year estimates): Demographic indicators including median household income (B19013), educational attainment (B15003), and total population (B01003). Rolling 5-year estimates provide reliable county level data through aggregated samples.

Bureau of Economic Analysis (2016-2023 annual data): Economic indicators from CAINC1 (personal income) and CAGDP2 (GDP). Year specific matching implemented following instructor feedback on temporal granularity. September 2024 release used, providing final 2023 estimates.

USDA Economic Research Service (2023 Rural-Urban Continuum Codes): Nine category geographic classification updated every 10 years. Applied uniformly across all years due to stable county classifications between updates.

Complete data dictionary and raw files available in project repository under data and documentation directories.

3. Data Curation Workflow

3.1 Workflow Overview and Lifecycle Mapping

The curation workflow systematically followed the **DCC Curation Lifecycle Model** (Higgins, 2008), implementing both sequential and continuous actions. The workflow comprised five major phases documented through modular Python scripts in the repository (scripts directory):

Phase 1 - Conceptualize and Plan: Literature review on turnout determinants, data source assessment, and proposal development specifying variables, geographic units, and temporal coverage (documented in /documentation/proposal.md).

Phase 2 - Create/Receive and Ingest: Manual extraction of election HTML tables to CSV format, programmatic Census downloads via data.census.gov API, and BEA bulk ZIP downloads. All raw data preserved in /data/raw with acquisition metadata. Encoding issues resolved during ingestion (BEA Latin-1 vs UTF-8) through explicit specification in clean_standardize.py.

Phase 3 - Appraise, Clean, and Transform: Quality assessment verified completeness (all 67 counties per year), accuracy (turnout calculations), and consistency (state totals matched). County name standardization implemented via lookup tables preserving originals. FIPS codes assigned as primary identifiers. Year-specific BEA temporal matching implemented through data_integration.py. One data quality issue identified: DeSoto County 2018 minor rounding discrepancy (53.7% vs 56.6%).

Phase 4 - Preserve and Store: Structured directory hierarchy (/data/raw, /data/processed, /scripts, /documentation) with Git version control. Original files preserved alongside processed versions. CSV format selected for maximum accessibility and interoperability.

Phase 5 - Access, Use, and Reuse: Comprehensive documentation through README reproduction instructions, inline code comments, and data dictionary. A public GitHub repository with clear licensing facilitates reuse by researchers, practitioners, and educators.

3.2 Data Integration Strategy (M6 Concepts)

Integration addressed three primary heterogeneities: **Format heterogeneity:** Varied encodings (Latin-1, UTF-8), structures (wide vs long format), and file types (HTML, CSV, ZIP). **Semantic heterogeneity:** County name variations ("St. Johns" vs "Saint Johns", "Miami-Dade" vs "Dade"). **Temporal heterogeneity:** Matching annual BEA data to election events while using appropriate Census aggregation periods.

FIPS codes (five-digit federal identifiers) served as primary integration keys, providing stable, authoritative identifiers maintained by Census Bureau (M9: identifier systems). The integration workflow starts with standardizing county names and verifying FIPS assignments. Followed by extracting Florida counties (FIPS 12001-12133) from national files. From there we are able to implement temporal BEA matching for all years within the dataset. Generate a master dataset by performing left joins on FIPS codes. Then validation of the complete matching process. All integration logic documented in data_integration.py with comprehensive comments.

3.3 Temporal Matching Implementation

BEA Temporal Matching: Following instructor feedback emphasizing year-specific data utilization, economic indicators were matched to each election year rather than using a single year. This enabled longitudinal analysis revealing 46% per capita income growth from 2016 ($43,490) to 2024 ($63,477).

ACS Temporal Decision: Following instructor feedback regarding temporal granularity, ACS 5-year estimates (2016-2020) were retained rather than implementing year-by-year matching. This decision reflects fundamental methodological differences between ACS rolling aggregates and BEA point-in-time measurements. ACS 5-year estimates aggregate survey responses over five years, creating overlapping periods in consecutive releases. Additionally, demographic variables exhibit gradual secular trends rather than year-to-year volatility. The 2016-2020 estimates provide optimal statistical reliability while capturing demographic characteristics relatively stable across the study period.

3.4 Quality Assessment and Standards (M11)

Quality procedures implemented throughout curation included completeness checks (zero missing values verified), consistency validation (aggregated totals matched state figures), temporal checks (population/economic trends within plausible ranges), range validation (turnout 40-95%), and duplicate detection (FIPS-year unique identifiers). CSV format adheres to RFC 4180 standard. Variable naming follows snake_case convention for programmatic accessibility. FIPS codes conform to ANSI INCITS 38:2009 standard. All quality checks documented in /documentation/quality_report.md.

---

4. Ethical, Legal, and Policy Considerations (M2)

Public Data Usage: All data sources are publicly available aggregate statistics released by government agencies under open data policies. No individual level data or personally identifiable information is included, eliminating privacy concerns.

Data Provenance and Attribution: All sources properly attributed with complete citations. Original data preserved alongside processed versions to maintain provenance chain and enable verification.

Reproducibility Ethics: A public GitHub repository with comprehensive documentation enables replication and independent verification, supporting scientific integrity and transparency principles.

Potential Use Considerations: While the dataset analyzes civic participation patterns, users should recognize limitations in causal inference from observational data. The documentation explicitly notes that correlations do not imply causation and encourages responsible interpretation. Future research incorporating qualitative methods would provide richer understanding of mechanisms underlying observed patterns.

---

5. Data Models and Abstractions (M3-M5)

The project employs a **relational data model** with county-year observations as the unit of analysis. The conceptual model treats counties as geographic entities with time varying attributes (election outcomes, economic conditions) and time invariant characteristics (rural-urban classification). This aligns with the **Entity-Relationship** abstraction where counties are entities, elections are events, and demographic/economic indicators are attributes.

The **tidy data** abstraction (Wickham, 2014) guided data structuring that each variable forms a column, each observation forms a row, and each type of observational unit forms a table. The final dataset follows this principle with 335 rows (county-years) and 16 columns (variables), enabling statistical analysis.

FIPS codes function as **durable identifiers** following best practices for geographic data where they are unique (no ambiguity), stable (unchanged since county establishment), and authoritative (maintained by federal government). This contrasts with county names which exhibit variations across sources and time.

---

6. Key Findings and Analysis

6.1 Temporal Patterns

Presidential elections (2016: 74.98%, 2020: 77.52%, 2024: 80.39%) demonstrate 20+ percentage point higher turnout than midterms (2018: 63.06%, 2022: 57.11%), consistent with national patterns of differential engagement across election types.

6.2 Socioeconomic Relationships

The 2024 correlations show expected positive relationships with per capita income (+0.256) and education (+0.130). However, surprising negative correlations emerge for population size (-0.208) and GDP (-0.224), with larger counties exhibiting lower turnout despite higher incomes.

6.3 Urban-Rural Differential

The USDA classification integration explains the population turnout paradox. Small rural counties (e.g., Franklin: 94.1%) achieve significantly higher participation than large metropolitan areas (e.g., Miami-Dade: 69.2%) despite lower median incomes. The urban-rural turnout gap averaged 15 percentage points in 2024, warranting investigation into civic culture, community cohesion, and voting accessibility differences across geographic contexts.

---

7. Challenges, Lessons Learned, and Next Steps

7.1 Technical Challenges

Data Format Heterogeneity: Integrating data from four agencies with varied formats required custom parsers for each source. The lack of standardized APIs for state election data necessitated manual HTML extraction, introducing potential for human error. Future work could explore web scraping automation.

Temporal Alignment Complexity: Balancing statistical reliability (5-year ACS aggregates) with temporal specificity (annual BEA data) required careful methodological consideration and justification. The project demonstrates that "more granular" is not always "better" when working with survey data.

Employment Variable Extraction: CAINC4 employment data extraction proved more complex than anticipated due to nested hierarchical structure, requiring additional parsing logic not completed within the project timeline. This highlights the importance of data profiling before committing to variable selection.

7.2 Methodological Insights

Identifier Importance: FIPS codes proved essential for integration, demonstrating the critical role of standardized identifiers in multi-source data projects. County name matching alone would have failed due to numerous variations.

Documentation Value: Comprehensive inline documentation in scripts facilitated iterative development and debugging. Future researchers can understand processing logic without reverse engineering code.

Preservation of Originals: Maintaining raw data alongside processed versions enabled validation and provided fallback options when encountering processing errors. This practice aligns with archival principles of preserving authentic sources.

7.3 Future Directions

Multivariate Analysis: The integrated dataset enables regression modeling to disentangle correlated factors (e.g., controlling for income when examining education effects). Hierarchical models could account for county-level clustering.

Temporal Trend Analysis: Year specific BEA data enables longitudinal analysis of how economic changes relate to turnout changes. Did 2020 pandemic economic disruptions affect 2020 participation?

Qualitative Extension: The urban-rural turnout differential warrants qualitative investigation through interviews with election officials and community members to understand mechanisms underlying observed patterns.

Methodological Extension: Incorporating 1-year ACS estimates for the three largest counties, where sample sizes support reliability, could enable more granular temporal analysis while maintaining 5-year estimates for smaller counties.

---

8. Connection to Course Concepts

This project synthesizes multiple course modules **Data Lifecycle** through systematic DCC model application, **Ethics and Policy** through public data usage and attribution practices, **Data Models** through relational modeling and tidy data principles, **Integration and Cleaning** through heterogeneity resolution and quality assessment, **Metadata and Documentation** through comprehensive data dictionary and documentation, **Identifiers** through FIPS code standardization, **Standards** through CSV and FIPS standards adherence, and **Workflow and Reproducibility** through modular scripts and version control.

The project demonstrates that effective data curation requires both technical skills (programming, data wrangling) and conceptual understanding (lifecycle models, quality assessment, ethical considerations). The emphasis on documentation, reproducibility, and transparent workflows reflects recognition that curated datasets are scholarly products requiring the same rigor as traditional research outputs.

9. Conclusion

This project successfully created a high quality, analysis ready dataset integrating Florida election turnout with temporally matched indicators across 335 county year observations. The curation process systematically implemented DCC Lifecycle principles through documented workflows emphasizing quality, reproducibility, and reusability. Key accomplishments include temporal BEA matching revealing 46% income growth, USDA integration explaining urban-rural turnout differentials, comprehensive quality assessment, and modular documented Python workflows published publicly. The dataset provides a foundation for studying democratic participation in diverse communities while demonstrating best practices in research data management.

Project Repository: https://github.com/Dfletchh/CS-598-Foundations-of-Data-Curation
Supplementary Materials: All scripts, data, documentation, and data dictionary available in repository

References

Digital Curation Centre. (2008). *DCC Curation Lifecycle Model*. Retrieved from https://www.dcc.ac.uk/guidance/curation-lifecycle-model

Higgins, S. (2008). The DCC curation lifecycle model. *International Journal of Digital Curation, 3*(1), 134-140.

U.S. Bureau of Economic Analysis. (2024). *Regional Economic Accounts*. U.S. Department of Commerce. https://www.bea.gov/data/economic-accounts/regional

U.S. Census Bureau. (2024). *American Community Survey*. https://www.census.gov/programs-surveys/acs/

U.S. Department of Agriculture Economic Research Service. (2023). *Rural-Urban Continuum Codes*. https://www.ers.usda.gov/data-products/rural-urban-continuum-codes/

Wickham, H. (2014). Tidy data. *Journal of Statistical Software, 59*(10), 1-23.

Florida Division of Elections. (2024). *Election Results and Statistics*. Florida Department of State. https://dos.myflorida.com/elections/data-statistics/