*An Exploration of Voting Drivers Behind California Prop. 29 (Nov. 2022 Ballot)—Dialysis Clinic Requirements Initiative*

Authors: [Kasra Afzali](mailto:kasra@umich.edu) [Michael Light](mailto:mikalite@umich.edu) [Iris Lin](mailto:irisqlin@umich.edu)

`

Michael

* Finish Report first draft (incl. Visualizations, all sections)
* Will copy and paste choropleth code to secondary research section
* Adding R environment requirements
* Why R and why Quarto
* Move model analysis folder into exploratory data analysis folder

Kasra

* [x] Update README in EDA section
* [x] Clean up EDA Code to abide by rubric
* [x] Add an Appendix section extracting the visualization not in the final report
* [x] Create choropleth for stations per capital, slider for years 2018, 2020, and 2022
* [x] Write up stations per capita calculations and reasoning
* [x] Break out missing values heatmap into its own notebook and move to data visualization scripts folder
* [x] Save missing values heatmap image into data visualization folder

Iris

* + Delete pycache and icloud files?
* Run at the end
* Sources

Folders

1. Need to add requirements at end
2. Need to add in old documents
3. Readme done
4. Readme done, code run check done, WARNING FIX TODO
   1. CHHS: Warning download\_excel\_files pd concat dtypes mismatch
   2. Ballot: DeprecationWarning: `pl.count()` is deprecated. Please use `pl.len()` instead.
   3. print(df.group\_by("year").agg(pl.count()).sort("year"))
   4. CMS: may need some commenting
5. michael/kasta todo
6. Code run check done, readme done
7. Readme done, no code in file
8. Readme done,
   1. FutureWarning: use\_inf\_as\_na option is deprecated and will be removed in a future version. (pairplot visualization in Medicare Facility data prep notebook)

Last TODO:

* PEP 8 Convention (Michael)
* Repurpose 001 folder to hold requirements
* Final run at the end

Code Rubric (150 pts)

* Present a coherent, computational narrative of your data manipulation and analysis.
* Run without errors from top to bottom.
* Run without warnings from top to bottom.
* Include a mix of code and Markdown cells.
* Include comments in the code, whenever appropriate.
* Avoid generating excessive output (i.e., more than 10-15 lines per cell).

**Project Motivation**

Dialysis is a lifesaving treatment that removes waste from blood—acting as an artificial kidney for those with chronic kidney disease. Prop. 29, which failed to pass, was the third attempt to increase regulations on dialysis clinics in the state—preceded by Prop. 8 in 2018 and Prop. 23 in 2020. Proposition 29—Dialysis Clinic Requirements Initiative—aimed to establish increased regulations for both staffing and operations for the roughly 600 dialysis clinics in California, an estimated $3.5 billion industry. Notably, the proposed regulation required the presence of a physician or licensed practitioner during all treatment hours, increasing each clinic costs by several hundred thousand annually.

Proponents of the propositions claim that increased regulations improves patient safety and quality of care. Opponents however, argue that the increase in healthcare cost was unwarranted and would limit care coverage by overwhelming the clinic with costs and forcing them to close. ([source](https://www.latimes.com/california/story/2022-11-08/2022-california-election-prop-29-vote-dialysis-clinic-results)).

Our project explores trends in dialysis clinic access, quality of care, and ballot results in the state of California in recent years. Using publicly available data from the Center for Medicare and Medicaid Services and aggregated election results from California's Secretary of State, we’ve analyze associations between dialysis care and voting behaviors, specifically those related to recent statewide ballot initiatives designed to regulate California’s multibillion-dollar dialysis industry—including 2022’s Proposition 29, which failed to pass by a large margin. While private health insurers typically pay higher rates, medicare pays for dialysis treatment for the majority of people on dialysis in California ([source](https://lao.ca.gov/BallotAnalysis/Proposition?number=29&year=2022)) meaning most patient interactions with dialysis clinics would exist within the Medicare and Medicaid datasets.

This project uses a novel approach to study an area of public interest relevant not just to California but the entire country. Investigative reporters, patient advocacy groups, and labor organizations have spent significant resources over the past decade to raise public awareness of the dialysis industry and its need for regulation.

To our knowledge, this is the first project of its kind to explore the possible association between the way people vote and the quality of dialysis care they receive.

**Project Research Questions:**

The research questions this project set out to answer are broken into two categories: primary and secondary. Answering the primary question is the north star of the analysis; secondary questions are the first our analysis touches upon through further investigation of the data.

1. **Primary Research Question:** Is the quality of care of dialysis facilities correlated with the voting in favor or against dialysis industry regulation?
   1. The key assumptions:
      1. The relationship between Quality of Care and Voting Behavior is not confounded.
      2. A vote in favor of any of the three propositions can be interpreted the same way: as support for dialysis industry regulation.
   2. To test this relationship under the outlined assumptions, we approximate the quality of care using the following metrics:
      1. Five star rating
      2. Patient experience rating
      3. Facility Mortality Rate
      4. Number of available dialysis stations
      5. Staff rating
      6. Hospital readmission categorization (Worse than Expected, As Expected, Better than Expected)
      7. Profit/non-profit designation
      8. Parent company affiliation/independence
   3. Dialysis facilities in our data are categorized by:
      1. county
      2. city
      3. year
2. **Secondary and Tertiary Research Questions:** 
   1. What is the geographic coverage of dialysis clinics in California?
   2. Is there any correlation between organization structure (chain owned, profit vs non-profit) and the quality of care?

**Data Sources**

**Primary Data Sources**

Our project focuses on two primary sources of data: Center for Medicare and Medicaid Services (CMS) Quarterly Dialysis Facility Compare dataset from 2017 to 2024 and the California Secretary of State’s (SOS) Statement of Vote for Ballot Measures in November 2022, 2020, and 2018 ballots - when residents voted on propositions regarding dialysis clinic requirements.

CMS’s Quarterly Dialysis Facility Compare dataset includes star ratings for dialysis facilities, ranging from patient experience and quality of care metrics. It provides insights into how well these facilities perform in terms of patient satisfaction and clinical outcomes, with focus on critical areas like doctor-patient communication, hospitalization rates, and effective treatment methods. The patient experience star rating is derived from patient surveys. These patient experience surveys are administered twice a year from a random sample of patients receiving in-clinic treatment at a given dialysis facility. Facility ratings are calculated and reported using facility metrics that include factors like unplanned hospital readmissions, number of total and expected transfusions in a period, ratio of deaths to expected number of deaths, and waste removal from dialysis across patient types ([source](https://dialysisdata.org/sites/default/files/content/Methodology/Guide%20to%20QDFCC.pdf)).

The California Secretary of State Ballot Measures dataset features information describing the outcomes from the General Election in Nov. 2022, 2020, and 2018, across ballot measures for the counties and sub-counties in California. Sub-counties within these datasets include the congressional districts, state senate districts, state assembly districts, and cities for each county. These reporting levels allow for aggregation at multiple geographic levels for later in the analysis.

**Secondary Data Source**

The California Health and Human Services Speciality Care Clinic Complete Data Set from 2013 through 2023 was used as a secondary data source to extract more geographic data for each dialysis clinic, supplementing CMS’s Quarterly Dialysis Facility Compare dataset. Additional facility level features used include: senate district, congressional district, and latitude and longitude.

**Features of Interest**

| Data Source | File Information | Features |
| --- | --- | --- |
| Center for Medicare and Medicaid Services Dialysis Facilities | Type: .zip & .xlsx  Years: 2017-2024  Combined Size: 5,684 rows, 173 columns  Link: <https://data.cms.gov/provider-data/topics/dialysis-facilities> | * Facility Name * County * City * Facility Quality of Care Rating * Dialysis Center Staff Rating * Patient Experience Rating * Patient Readmission Category (Worse Than Expected, As Expected, Better Than Expected) * Mortality Rate * Number of Dialysis Stations * Licensee Type (Non-Profit, For-Profit.) * Organization Affiliation (Independent or Chain) |
| California Secretary of State November Ballot Measures | Type: .xlsx  Years: 2018, 2020, 2022  Combined Size: 7,122 rows, 7 columns  Link: <https://www.sos.ca.gov/elections/prior-elections/statewide-election-results> | * County * Congressional District * City * Ballot Proposition * Yes Vote Count for Proposition * No Vote Count for Proposition |
| California Health and Human Services Speciality Care Clinic Complete Data Set | Type: .xlsx  Years: 2013-2023  Combined Size: 6,605 rows, 143 columns  Link: <https://data.chhs.ca.gov/dataset/specialty-care-clinic-complete-data-set> | * Facility Name * County * Zip Code * Assembly District * Senate District * Congressional District * Census Tract * Latitude and Longitude |

**Data Manipulation Methods**

Our workflow was broken down into five stages:

* Data Collection
* Data Preparation
* Database Management
* Exploratory Data Analysis
* Statistical Analysis

**Data Collection and Preparation:**

The CMS Dialysis Facility Dataset was organized in .zip files, one per year. Within each .zip file existed several excel files grouped by variables. For the purposes of analysis, we were interested in only excel files relevant to describing a facility’s general information, facility ratings, and patient survey results. When importing and combining the data for this source, we chose to split the data in two separate parquet files at the facility level, one specifically for patient survey responses and one for facility ratings and measurements.

One challenge that we ran into when importing this dataset was the file naming conventions across the years. In 2021, the excel files were named in a different format than the other years. For instance, the patient survey related data included ‘ICH’, standing for In-Center Hemodialysis CAHPS Survey, the relevant file was named ‘59mq-zhts’ in 2021. To solve this issue, we included the file names into a list to select the exact files we needed rather than selecting all files with ‘IHC’ in the name.

In addition to expected occurrences of missing data (e.g. resulting from survey non-response), we also found that some variables of interest were missing seemingly due to administrative error. For example, the Center for Medicare Services ICHPS raw data files are missing many columns, including patient hospital readmission categorization and overall patient experience ratings, which are otherwise present historical ICHPS datasets. In these specific cases, a simple imputation was performed during the analysis step: we substituted 2018 ‘nan’ values with 2019 values at the facility level.

The SOS Ballot Data was imported by url to the excel file, one for each relevant proposition year - 2018, 2020, and 2022. To select the data relevant to the three proposition results needed for our analysis, we kept the columns that included ‘Kidney’ or ‘Dialysis’. Additionally, some manipulation was required for the geographic columns, renaming columns and backfilling rows, as the raw data included a multi-level index with sub-counties falling under its respective counties. The cleaned data for each year was merged into one final ballot data parquet with a column specifying the year and count and sub-county vote counts for each Dialysis Requirements Initiative proposition.

Similar to the challenge with the CMS Dialysis Facility dataset, there were inconsistencies in the SOS Ballot data between the years. In this case, in 2020 and 2022, county supervisorials were called ‘County Supervisorial’ but in 2018, the same geographic denomination was referred to as a ‘Supervisorial District’.

CHHS Speciality Care Clinic Complete Data Set was downloaded as excels for 2013 through 2023, one file per year. The main manipulations performed were on the files from years before 2018 to align them to the structure and naming convention of the files 2018 forward. First, we read the 2013 to 2017 data and 2018 to 2023 data into two separate data frames. Using a mapping dictionary, created by CHHS to map pre 2018 and post 2018 data columns, we renamed the 2013 to 2017 columns to their new respective names.

After mapping the columns, we checked to make sure that the same columns across the two files held the same file type, converting those that were not. Then merged the data frames to create a single parquet, outer joining to on columns that appear in both dataframes. Finally, we dropped rows with missing FAC\_NO, facility data. Most of these rows with empty FAC\_NO were due to formatting while others were removed to ensure we will be working with data attributed to a given facility.

To merge the four data frames on time period, geography, and facility names, data types and column names were standardized. First, the CMS facility rating dataset was merged with the CMS patient survey dataset. The CHHS facility measurements and ratings dataset was filtered to show only those related to dialysis clinics, then merged with the CMS facility and patient survey data. Reshaping the CMS and CHHS data by geographic level from the original facility level information, we merged these two aggregations of the data with the SOS Ballot Measures dataset.

After importing, cleaning, and manipulation, we were left with two parquet files, one with data aggregated at a city level and one with the data aggregated at an assembly district level.

**Database Management:**

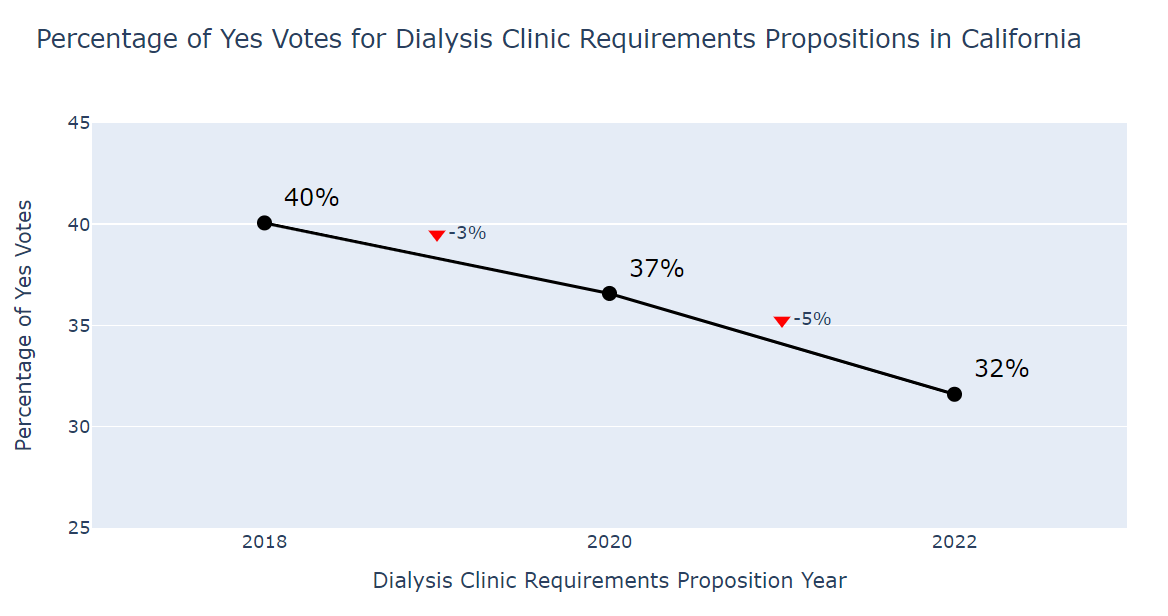
In addition to accessing our cleaned and manipulated datasets as parquet files, our project also created a custom-built relational database system implemented as a Python class. This database serves as a centralized repository for cleaned primary datasets as well as additional data that may be used for exploratory analysis, visualization, or future analysis. The design of our database management system prioritizes efficiency, flexibility, and ease of use.

Key features of our database are as follows. Please refer to the included database demo jupyter notebook to learn more about database functionalities.

1. Table structure:
   * Each table represents a distinct dataset
2. View structure:
   * Tables can have multiple views, which are subsets of table columns.
   * Views allow for focused access to specific data without altering the underlying table structure.
3. Dynamic View Creation:
   * The system supports the creation of new views on the fly.
   * Custom views can be defined based on the specific task
4. View Merging:
   * Multiple views can be combined to create new dataframes
   * This feature facilitates complex data analysis across different tables and views.
5. Conditional Querying:
   * The database supports querying based on user-defined conditions.
   * This allows for precise data retrieval and filtering capabilities.
6. Efficient Data Access:
   * The implementation ensures quick and reliable access to data across the entire database.
7. Code Quality and Maintainability:
   * The database class follows object-oriented, modular design and consistent naming conventions across the project
   * Improved refactoring, efficiency and adaptability, enhance readability and facilitate debugging.

**Analysis** (60 pts)

The Dialysis Clinic Requirements Initiative’s propositions have failed three times in California, with the percentage of ‘Yes’ votes to pass the initiative falling eight percent over the four years since it had been first proposed (see Figure 2). To understand more about one potential driver behind these voting behaviors – quality of care received from a dialysis clinic – we performed the following analysis.



*Figure 2. Percentage of ‘Yes’ Votes for Dialysis Clinic Requirements Initiatives (CA. 2018-2022)*

**Exploratory Data Analysis:**

A key feature of our EDA process was the implementation of interactive choropleth maps to encode geographical data across California's counties. To enhance the exploratory capabilities, we incorporated a dynamic slider functionality, allowing users to seamlessly transition between different features or access temporal data across years. This interactive approach enabled us to identify spatial patterns, temporal trends, and potential correlations within our dataset, providing a solid foundation for subsequent in-depth analysis.

**Primary Research Question**

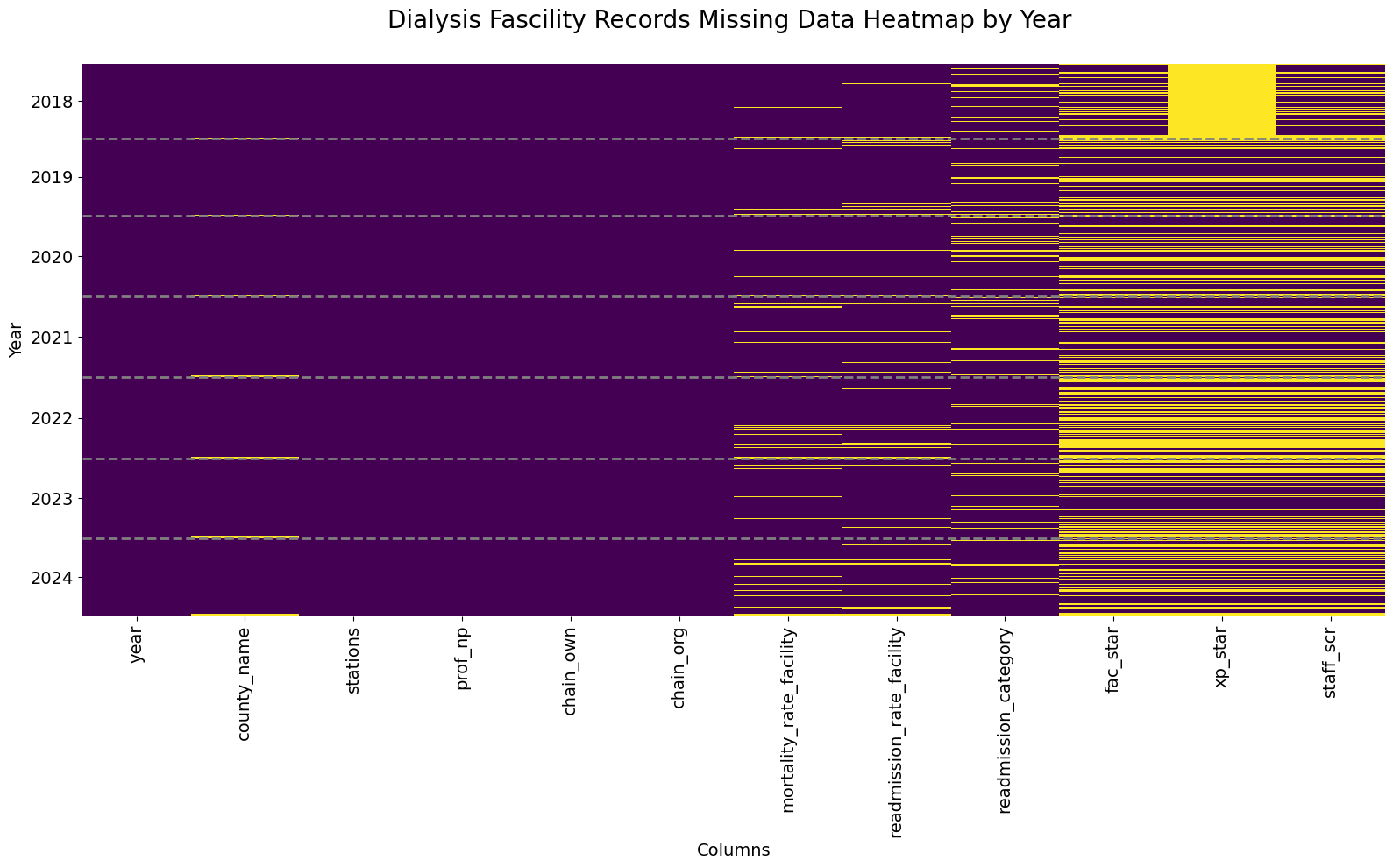
* Conditional Effects
  + What interesting relationships/insights did we find
  + What didn't work and why?
* Random Effects
  + What interesting relationships/insights did we find
  + What didn't work and why?
* Fitted Various Models between complete case analysis in the years that are relevant (2018, 2020, and 2022)
  + Found that the necessary sample size for the analysis is ~500

**Secondary Research Questions**

In our exploration of dialysis facility records, we sought to understand the factors influencing the quality of care across different facility types and over time. Specifically, we want to investigate the geographic coverage of dialysis clinics in California. We began by examining the completeness of our dataset through a missing data heatmap. This revealed significant gaps in certain years and variables, particularly in earlier records.

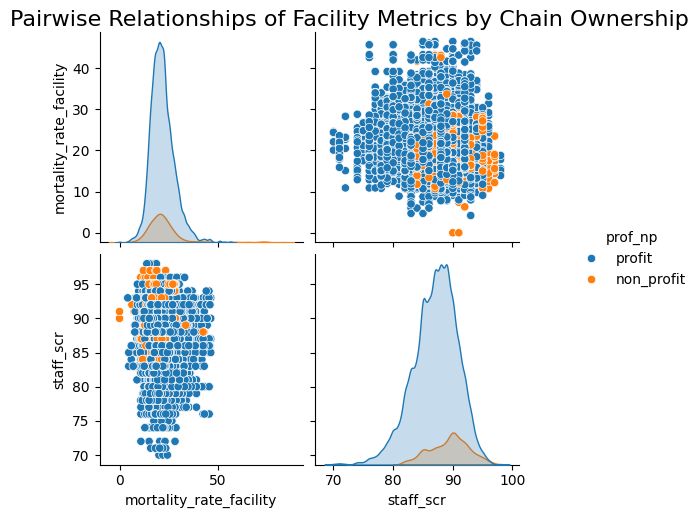
We began by examining the completeness of our dataset through a missing data heatmap. This revealed gaps in certain years and variables like patient’s experience rating , particularly in earlier records. These data inconsistencies suggested a need for cautious interpretation of long-term trends and highlighted the importance of focusing on more recent, complete data for robust analysis.

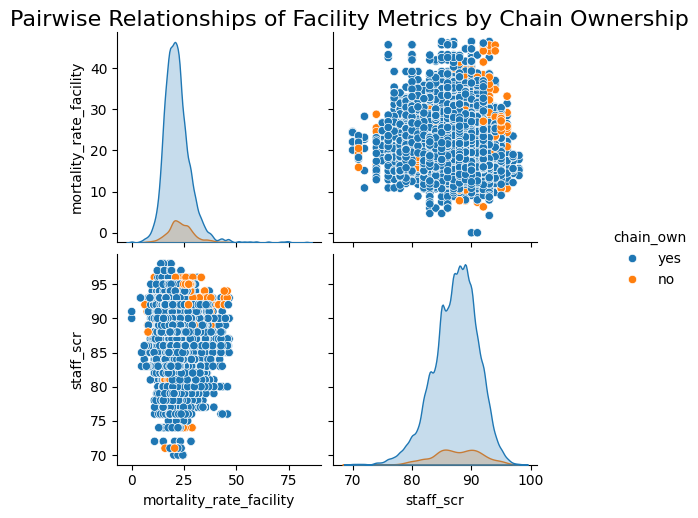
These data inconsistencies suggested a need for cautious interpretation of long-term trends and highlighted the importance of focusing on more recent, complete data for robust analysis.

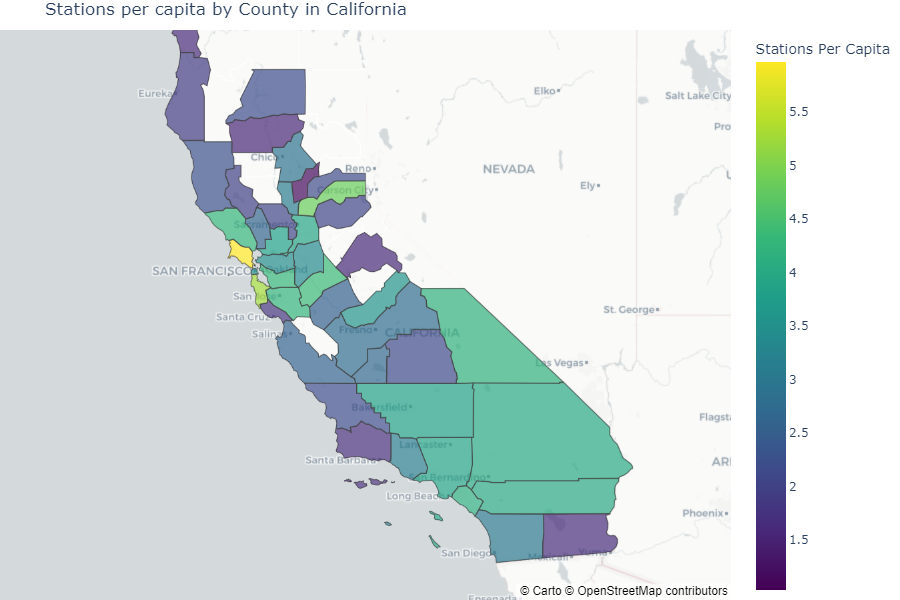


*Figure 1.Dialysis Facility Records Missing Data Heatmap (2018-2024): Variables sorted by missingness per year. Purple shows complete data, yellow indicates missing values*

Our pairwise analysis of facility metrics revealed complex relationships between facility characteristics and quality of care. We observed a slight negative correlation between facility size and mortality rates, suggesting potential benefits of larger facilities. Ownership structure played a role, with non-profit facilities generally showing better outcomes, though with considerable overlap with for-profit counterparts. Chain-owned facilities demonstrated more consistent metrics compared to independent facilities. While facility star ratings positively correlated with patient experience scores, the relationship was weaker than expected, indicating that patient satisfaction is influenced by factors beyond clinical outcomes. Over time, we noted a general improvement in mortality and readmission rates across all facility types, with chain-owned facilities showing more consistent progress. These findings underscore the multifaceted nature of dialysis care quality and the importance of considering various factors when evaluating facility performance and patient outcomes.







**Assumptions, Limitations, Ethical Considerations**

**Conclusion and Future Work**

1. **Future Research Questions:** 
   1. What is the geographic coverage of dialysis clinics in California?
   2. Do demographics, population profiles, and other voting behaviors—including historic voting behaviors—have an association with how people voted on Prop. 29?

**Statement of Work and Endnotes** (Can be on PG 11)

* Describes the contribution that each team member made to the project
* How we collaborated with each other

**Sources**

* [**https://lao.ca.gov/BallotAnalysis/Proposition?number=29&year=2022**](https://lao.ca.gov/BallotAnalysis/Proposition?number=29&year=2022)
* [**https://dialysisdata.org/sites/default/files/content/Methodology/Guide%20to%20QDFCC.pdf**](https://dialysisdata.org/sites/default/files/content/Methodology/Guide%20to%20QDFCC.pdf)