SIADS 593: Milestone I

Team Project Proposal

version 2022.07.27.1.CT

**Instructions:** please make a ***copy*** of this template file (do not edit original).

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

## 1. Team members

Please list your team members (2-3 max).

* Iris Lin (irisqlin)
* Kasra Afzali (kasra)
* Michael Light (mikalite)

## 2. Project summary

Summarize your proposed project in a few sentences.

#### What is your proposed project and why are you proposing it?

#### What are the question(s) you want to answer, or goal you want to achieve?

| Project Background:  Dialysis is a lifesaving treatment that removes waste from blood—acting as an artificial kidney for those with chronic kidney disease. Proposition 29—Dialysis Clinic Requirements Initiative—aimed to establish a set of regulations for both staffing and operations for the roughly 600 dialysis clinics in California. These regulations included adding more required staff for patient treatments, increasing reporting requirements for clinics, and prohibiting refusal of care based on a patient’s form of payment. Prop. 29, which failed to pass, was the third attempt to regulate dialysis clinics in the state—preceded by Prop. 8 in 2018 and Prop. 23 in 2020.  Project Objectives:  Our proposed project will explore geographic trends in dialysis clinic access, quality of care, and Medicare/Medicare spending in the state of California. Using publicly available data from the Center for Medicare and Medicaid Services and aggregated election results from California's Secretary of State, our goal is to 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.  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 sets out to answer are broken into two categories: primary and secondary/tertiary. Answering the primary question is the north star of the analysis; secondary/tertiary questions are the first draft/examples of the kinds of things we anticipate learning during the course of analysis.   1. **Primary Research Question:** Is the quality of care an individual dialysis patient receives associated with their vote in favor or against dialysis industry regulation? 2. **Secondary and Tertiary Research Questions:**     1. What factors (if any) influence the geographic distribution of dialysis clinics in California?    2. What is the association between clinic profit categorization and the quality of care they offer patients?    3. Do demographics, population profiles, and other voting behaviors—including historic voting behaviors—have an association with how people voted on Prop. 29?   *Sources:*   * *https://www.spur.org/voter-guide/2022-11/ca-prop-29-dialysis-clinic-requirements* |
| --- |

## 

## 3. Datasets

#### Describe one primary dataset and at least one secondary dataset. If other secondary datasets will be used please describe them as well.

#### The proposed datasets should exhibit different features/columns and/or different access methods, e.g., \*.csv file, \*.json file, API retrieval, web scraping, etc. Different time periods, for example, with the same features/columns is not considered a different dataset. Remember, the focus of the project in this Milestone course is to give you the opportunity to practice your data manipulation skills, so feel free to challenge yourself.

#### If you're unsure if your data sets are "different enough" describe the datasets and request a review via the *#siads593\_[semester]\_001\_project* Slack channel.

#### **Please note:** all proposed datasets ***MUST*** be publicly available to all members of the class (students, instructors, course support personnel, etc.). Use of proprietary datasets for this project is ***not*** permitted.

## 3.1 Primary dataset description

Describe your primary dataset. How is the data collected and how will you access it? Please share what features in the dataset are relevant to your topic. At a minimum, include the following information:

#### Short description (i.e., 1-3 sentences) of its key features

#### Estimated size (in records and/or bytes)

#### Location (give the URL or other access method)

#### Format (CSV, JSON, etc.)

#### Access method (download, web scraping, API, etc.)

| [**Dialysis Facilities: Patient Survey and Quality of Care Ratings Dataset**](https://data.cms.gov/provider-data/topics/dialysis-facilities)   * **Description:** This dataset includes star ratings for dialysis facilities based on patient experience and quality of care metrics. It provides insights into how well these facilities perform in terms of patient satisfaction and clinical outcomes, focusing on critical areas like doctor-patient communication, hospitalization rates, and effective treatment methods. The patient experience star rating is derived from surveys. * **Key Features:** Patient Survey Star Rating, Quality of Patient Care Ratingents. * **Size:** 7,577 rows * **Format:** CSV * **Access Method:** Download, API * **Link:** <https://data.cms.gov/provider-data/topics/dialysis-facilities> * When constrained to California only (736):   + 418 unique facilities   + 39 unique counties   + 251 unique towns     **Medicare Geographic Variation Dataset**   * **Description:** The Medicare Geographic Variation dataset provides insights into the demographic, spending, healthcare use, and quality indicators for the Original Medicare (Fee-For-Service) population. It highlights geographic differences in healthcare services across the United States, focusing on state and county levels. Spending is standardized to eliminate geographic payment differences, allowing for a more accurate comparison across regions. * **Key features:** Demographic indicators, spending indicators, healthcare utilization metrics, quality of care indicators * **Size:** 30,273 rows   + When constrained to CA 59 rows aggregated at the county level * **Format:** CSV * **Access:** Download, API * **Link:** <https://data.cms.gov/summary-statistics-on-use-and-payments/medicare-geographic-comparisons/medicare-geographic-variation-by-national-state-county> |
| --- |

## 3.2 Secondary dataset(s) description

Describe your secondary dataset(s). How is the data collected and how will you access it? Please share what features in the dataset(s) are relevant to your topic and describe the data types you expect. At a minimum, for each secondary dataset include the following information:

#### Short description (i.e., 1-3 sentences) of its key features

#### Estimated size (in records and/or bytes)

#### Location (give the URL or other access method)

#### Format (CSV, JSON, etc.)

#### Access method (download, web scraping, API, etc.)

| [**General Elections Statement of Vote, Ballot Measures (Nov. 2022)**](https://www.sos.ca.gov/elections/prior-elections/statewide-election-results/general-election-nov-8-2022/statement-vote)   * **Description:** The Ballot Measures dataset features 9 columns describing outcomes from the General Election from Nov. 2022 in California. These columns included the *County Name, Ballot Measure Name, Yes Votes,* and *No Votes*. The *County Name* column is a text column with the name of the California county. *Ballot Measure* name is a text column that includes the name of the proposition being voted on. The number of *Yes/No* columns are integer columns, detailing the number of Yes or No votes from occupants of each county specified. * **Size:** 407 Rows, 9 Columns * **Location:** <https://www.sos.ca.gov/elections/prior-elections/statewide-election-results/general-election-nov-8-2022/statement-vote> * **Format:** CSV * **Access Method:** Link   [**Voter Registration by County (Nov. 2022)**](https://elections.cdn.sos.ca.gov/sov/2022-general/sov/02-county-voter-reg-stats-by-county.pdf)   * **Description:** The Voter Registration by County dataset includes voter registration volumes for each of the 58 counties in California. Columns that we will be focusing on include *Total Registered*, number of people who registered to vote in each county, and columns representing each party - *Democratic, Republican, American Independent, Other, and No Party Preference*. All of the data in this dataset will be integers or floats. * **Size:** 61 Rows, 10 Columns * **Location:** <https://elections.cdn.sos.ca.gov/sov/2022-general/sov/03-voter-participation-stats-by-county.pdf> * **Format:** PDF * **Access Method:** Download and Convert to CSV   [**Census Demographics and Housing Characteristics (2020)**](https://data.census.gov/table/DECENNIALDP2020.DP1?g=040XX00US06,06$0500000&d=DEC%20Demographic%20Profile)   * **Description:** The General Profile and Housing Characteristics dataset includes Census information across all states and counties in the US from 2020. The main variables our analysis may use include those related to SEX AND AGE, RACE, and HOUSEHOLD types. More specifically, the total population volume who are within certain age bands, identify as certain genders and races. * **Size:** 61 Rows, 322 Columns * **Location:** <https://data.census.gov/table/DECENNIALDP2020.DP1?g=040XX00US06,06$0500000&d=DEC%20Demographic%20Profile> * **Format:** Excel * **Access Method:** Download   **Health Facility License Fee Reports:**   * **Description:**    + The California Department of Public Health (Department), Center for Health Care Quality (Center), Licensing and Certification Program is responsible for regulatory oversight of licensed health care facilities and healthcare professionals to provide safe, effective, and quality health care for all. This is an annual report on the Center’s periodic inspections and complaint investigations of healthcare facilities to oversee compliance with federal and state laws and regulations. These reports may be useful for analyzing state-level trends in dialysis clinic safety over time. * **Key features:**    + Complaint investigations of dialysis care facilities by year,   + Deficiencies cited at dialysis care facilities by year     - Deficiencies defined as:       * Substantiated allegations for violations of federal and/or state laws or regulations receive deficiencies that cite the violations of noncompliance. * **Size:**   + 1 report/year over the past 10 years   + ~900 kb/report   + ~50 pages/report * **Format:** PDF * **Access:** Download, pdf scraping * **Link:** <https://www.cdph.ca.gov/Programs/CHCQ/LCP/Pages/LicenseFeeReports.aspx> |
| --- |

## 3.3 [ YES ] Affirm: datasets are public.

Please write YES in the above box to confirm that your primary and secondary datasets are accessible and available to your classmates and the instructional team.

## 4. Cleaning and manipulation

Describe how you will need to manipulate your datasets: how will you handle missing or anomalous data? How will you join your primary and secondary datasets? What cleaning and manipulation challenges, if any, do you anticipate?

| **Primary Datasets:** The **Patient Survey and Quality of Care Ratings Dataset** will be merged with the **Medicare Geographic Variation Dataset**, with data filtered specifically for California. Key quality of care metrics, such as overall star ratings, nephrologist scores, linearized facility ratings, and rating responses, will be aggregated at the county level, either averaged or summed where appropriate. These metrics will provide valuable insights into patient satisfaction and facility performance across the state.  However, aggregating these metrics may pose challenges due to missing data, particularly in counties with fewer entries, which could skew comparisons. Approximately 43% of California entries (317 out of 736) lack patient survey data. To address this, a detailed analysis will be conducted to define a generative model for the missing data—for example, whether variables like facility location or parent organization are associated with missingness—and proceed with a principled approach for imputing missing values if appropriate.  By merging these datasets, we aim to explore the relationship between Medicare spending and quality of dialysis care, providing a comprehensive understanding of factors influencing patient outcomes and facility performance across California. **Secondary Datasets:** The **Ballot Measures**, **Voter Registration**, and **Census Demographics** datasets will undergo a cleaning process to ensure uniformity across the datasets. Though these columns are all in string format, the Ballot Measures county column refers to Alpine County, California as ‘Alpine’, while the Census Demographics dataset refers to the same county as ‘Alpine County, California’. Additionally, the Voter Registration dataset has stored their volumes as numbers formatted with commas which need to be removed. The largest challenge of the cleaning process may be converting the Voter Registration dataset from .pdf to .csv. After these cleaning steps, we will join the three secondary datasets on the county column.  To join the primary and secondary datasets, we will merge on geographic features, such as county relations, zip codes, and census blocks. |
| --- |

## 5. Analysis

Describe any analyses you plan to undertake. For each, please give the technique or approach and briefly explain what you expect to learn from it.

| **Significant features we will explore throughout our analysis include:**   * Geographic location (county code, census block, latitude, longitude) * Medicare Expenditure (Total expenditure, expenditure per capita, beneficiaries covered) * Dialysis clinic rating (quality rating, patient survey rating ) * Voting record (y/n on prop.29, y/n on prop. 20, y/n on prop. 8) * Demographics (race, age, income, gender, health status, number/ratio of people with Chronic Kidney Disease   Using these features we we seek to answer the following questions:  **Primary Research Question:** Is the quality of care an individual dialysis patient receives associated with their vote in favor or against dialysis industry regulation?   * **Analysis Description:**   + We’ll use county and sub-county (census block) election results, voter demographics, and facility-level quality of dialysis care metrics to analyze the relationship between the quality of care provided to dialysis patients in a given geographic region and the way that region votes in favor or against increasing dialysis industry regulation. * **We expect to learn:**   + Whether there’s a relationship between the quality of care received by dialysis patients in a given geographic area and voter behavior therein.   + What is the relationship between quality of care received by dialysis patients and voter behavior re. regulation has changed over time (Props. 29, 23, 8 in 2022, 2020, and 2018 respectively) * **Examples of assumptions one might hold apriori:**    + People within a voting district receive dialysis within the same district   + Dialysis needs does not affect an individual’s likelihood of participation in the vote.   + People who voted in favor of Proposition 29 advocate for stricter safety, operational, and reporting standards in dialysis clinics, to improve the quality of care and protect patients at the cost of higher expenditure.   **Research Question:** What factors (if any) influence the geographic distribution of dialysis clinics in California?   * **Analysis description:**   + We’ll use county and sub-county (census block) voter demographics, county and sub-county dialysis clinic locations, and dialysis clinic metrics (quality of care, for-profit/non-profit incorporation status, treatment/operation costs, etc.) to analyze the geographic distribution of dialysis clinics across California. * **We expect to learn:**   + Where in California dialysis clinics are located   + What geographic, demographic, or other features of a given range (such as county or census block) are associated with a dialysis clinic being located therein.   + What/whether features of dialysis clinics are associated with their geographic location (incorporation status, patient quality of care, treatment cost., proportion of treatment costs covered by private insurance vs. Medicare, etc.) * **Examples of assumptions one might hold apriori:**    + Clinics are for profit and therefore will put clinics in areas where more dialysis patients would live.   + There’s some relationship between population demographics and clinic location.   **Research Question:** What is the association between clinic profit categorization and their quality of care for patients?   * **Analysis Description:** We will use the Dialysis Clinic profit categorization, quality rating, and patient survey rating to assess if there is a significant difference in quality rating and patient survey rating when the clinic’s profit categorization is either for profit or non-profit. This analysis may include looking at the mean and distribution of the rating features on a quantitative level and through visualizations like slab intervals, which function as more expressive versions of traditional visualizations like box/whisker plots that more effectively communicate uncertainty in quantitative variables like quality of care. * **What we expect to learn:**   + Whether for profit or non profit clinics have a different standard of care * **Assumptions and potential limitations:**   + Number of clinics under each categorization (for profit/non profit) is equal or can be sampled to balance each category   + Patient care can be laddered up to the clinic level - that groups of healthcare professionals and staff treating the patients are similar across an entire organization   **Research Question:** Do demographics, population profiles, and other voting behaviors—including historic voting behaviors—have an association with how people voted on Prop. 29?   * **Analysis Description:** Looking at the census demographic features, such as race, gender, and historical voting behaviors, we will try to understand which features, if any, contribute highly to their voting behavior on Prop. 29. To do this, we may build a regression or classification model and predict whether a certain county or smaller geographic population voted on Prop. 29 based on their demographic and voting profile. * **What we expect to learn:**   + If certain demographic or county profile features are more impactful to how people may vote on Prop. 29   **Potential challenges and limitations:**  One of the challenges we expect to face is with the quality of our primary dataset. It may be the case that clinic quality and patient survey rating are not representative of the quality of care offered by a given clinic to dialysis patients as survey data is highly subjective.  Another limitation that we face is the imbalance of our data. Of the population of voters, only a small subset interface regularly with dialysis clinics. Dialysis clinic access also varies geographically. It may be the case that a voter could reside in a county separate from where they receive dialysis. |
| --- |

## 6. Visualizations

Describe in 1-3 sentences at least **two** data visualizations that you plan to create. Include the chart type (e.g. bar chart, scatterplot, SPLOM, etc.) as well as the variables (features) you intend to plot.

| Analysis of our research questions will be supported by four primary types of visualization:   * Directed Acyclic Graphs:   + Used to represent the relationships our analysis assumes between variables and outcomes of interest. * Marginal Plots:   + Used to represent the effect of variables of interest on a given outcome when averaging over all other parameters in our analysis, such as the effect of clinic incorporation status on quality of care. * Prior and Posterior Density Distributions:   + Prior Predictive Plots: Slab interval plots used to represent the assumptions (Priors) and their implications on variables and outcomes of interest.   + Posterior Predictive Plots: slab interval plots used to represent the plausible range of values implied by our analysis for a given variable or outcome of interest, such as voter behavior. * Choropleth Maps:   + Used to represent geographic distributions of variables such as clinic location and voter demographics aggregated by county and sub-county.   We intend to plot the following variables:   * Medicare Expenditure (Total expenditure, expenditure per capita, beneficiaries covered) * Dialysis clinic rating (quality rating, patient survey rating ) * Voting record (y/n on prop.29, y/n on prop. 20, y/n on prop. 8) * Demographics (race, age, income, gender, health status, number/ratio of people with Chronic Kidney Disease |
| --- |

## 7. Ethical considerations

Does your choice of data raise any ethical issues? If so, briefly describe the concern and how you plan to mitigate it.

| When using voter data, one ethical issue could be a bias towards the majority—the population who has the physical and economic access to vote. In order to mitigate this ethical issue, we may need to hold out data from the majority population to balance the dataset that we use.  Another ethical issue that may arise is informed consent as we will be using Medicare/Medicaid datasets. Medicare/Medicaid qualified patients are typically those in lower income households, are pregnant, or have a disability. Informed consent of how data is collected, stored, and distributed may not be fully communicated. Nearly 10% of Medicare/Medicaid patients have a first language other than English and only 6% of them have proficient health literacy. While this ethical concern is hard to rectify, we can be proactive about the safety of the data that we collect and analyze to avoid any possible identification or misuse from the merging of multiple data sources.  *Sources:*   * [*https://www.medicaid.gov/medicaid/eligibility-policy/index.html*](https://www.medicaid.gov/medicaid/eligibility-policy/index.html) * [*https://www.cms.gov/blog/cms-releases-data-briefs-provide-key-medicaid-demographic-data-first-time*](https://www.cms.gov/blog/cms-releases-data-briefs-provide-key-medicaid-demographic-data-first-time) * [*https://hsrc.himmelfarb.gwu.edu/cgi/viewcontent.cgi?article=1173&context=sphhs\_policy\_facpubs*](https://hsrc.himmelfarb.gwu.edu/cgi/viewcontent.cgi?article=1173&context=sphhs_policy_facpubs) |
| --- |

## 8. Contributions

Indicate the contribution that each team member will make to the project.

| Assigned Roles:   * Lead and Facilitator: Michael Light   + As the Lead and Facilitator, Michael will make sure to clearly define and align project members to next steps in the project. During collaborative meetings, they will direct conversations in a forward direction while asking questions to group members to get input. * Notetaker and Devil’s Advocate: Kasra Afzali   + Kasra will take minutes during and develop agendas prior to each collaborative session. These agendas and minutes will ensure that questions and blockers that arise during the project are discussed as a group and addressed.   + As the Devil’s Advocate, Kasra will ask the tough questions, taking on the lens as a potential stakeholder or grader to the project. These will help us as a group delve deeper into the topic and take on challenging perspectives to our work. * Arbitrator and QA: Iris Lin   + As arbitrator, Iris will step in in times of conflict and try to create a compromise in ideas and opinions. They will try to see and explain both sides of the conversation to reach some agreement, or try to put a pin in the conversation for a later date.   + Iris will QA assignments before submission. During these checks, they will streamline the deliverable to create a cohesive sound. Additionally, these checks will make sure that all project tasks are addressed and missing pieces are filled. |
| --- |

### 

## Changelog

(2022.07.27.1.CT) Update for 593

(2021.07.24.1.AW) Adjust title, number sections, simplify section headings, edit text