U6614: Research Proposal - Analysis of California‘s SASH program and drivers of adoption of residential solar systems among low-income communnities

Javier Baranda, Juan Farfan, Aparajita Rao

2023-03-27

Research question and policy background

Policy background

What do you hope to learn from the proposed analysis? What are the policy implications of your potential results?

The Single-Family Affordable Solar Homes (SASH) Program was a low-income solar program part of one of two initial programs under the California Solar Initiative (CSI). The program administrator of this project, GRID Alternatives (GRID), was the state-wide non-profit solar contractor. The SASH incentive was open to qualifying low-income homeowners within the service territories of the three major Investor Owned Utilities in California: Pacific Gas and Electric (PG&E), Southern California Edison (SCE), and San Diego Gas and Electric (SDG&E). The first phase of the program, SASH 1.0, analysed in this study and referred as SASH, accepted applicants across the state from 2009 to 2018 until the allocated program funding of USD 108M was exhausted.

The program offered an upfront subsidy for the installation of single-house residential solar systems ($3 per watt of capacity installed) for eligible homeowners, identified as those living in “affordable housing” single-family homes and with a household income of 80% below their county median income. This subsidy often accounted for a substantial part of cost of the solar system installation for the beneficiaries. In order to access the subsidy, eligible households had to Under the original SASH program, GRID Alternatives installed systems on over 5,200 homes in California between 2009 and 2018. In addition to providing incentives, the SASH program also aimed to support energy efficiency through workforce development, green jobs training opportunities, and low-income community engagement. We aim to utilize the SASH database of solar to validate whether the program was truly comprehensive.

The SASH program is a policy intervention that aims to provide low-income households with affordable access to solar energy. However, SASH and similar programs, considering only income as qualifying criteria, may not be reaching equally marginalized communities due to the existence of additional barriers for residential solar adoption, such as geographic location, language, education or demographic composition. This research will aim to investigate the effectiveness of the SASH program in increasing solar adoption among the low-income communities targeted and also explore additional barriers to adoption of residential solar systems beyond household income.

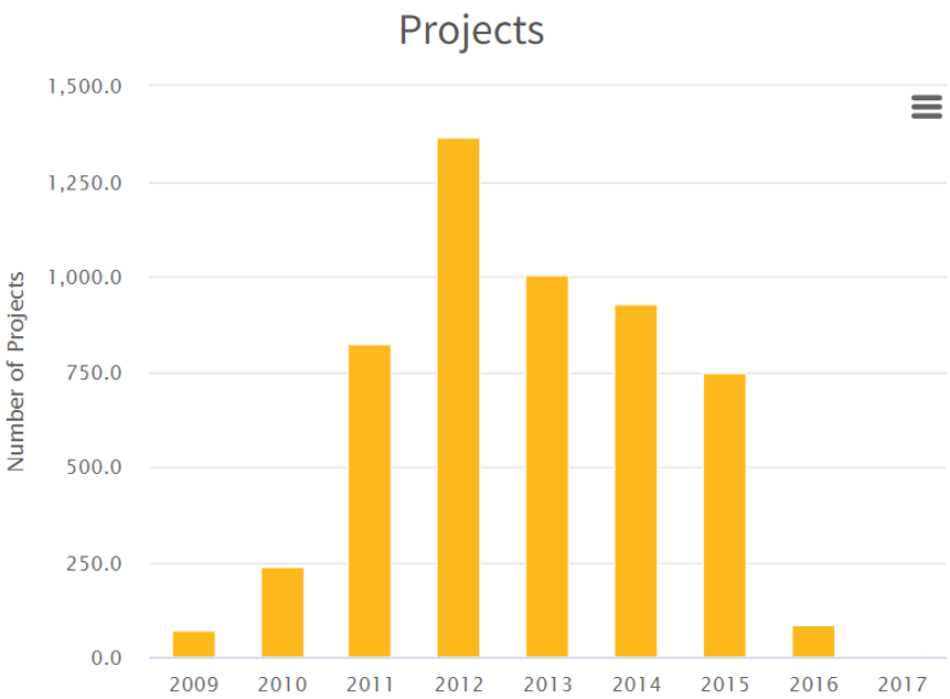


Figure 1: Distribution of SASH installations between 2009 and 2016

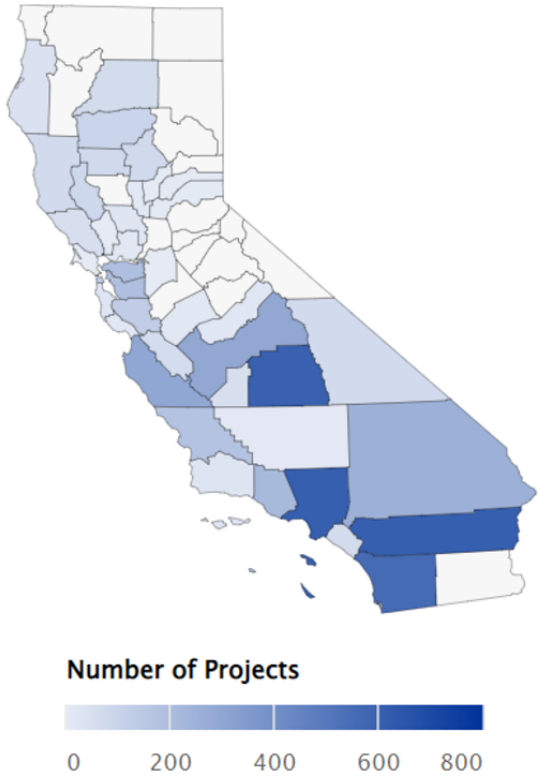


Figure 2: Distribution of SASH installations across counties in California

Research objective and policy implications

What policy context should we know? Make sure we understand what your "treatment" entails, and what factors may have contributed to the treatment (in the absence of random assignment)? Describe the nature of your policy/treatment variation.

What are the (potential) mechanisms linking your treatment or policy-relevant variable(s) of interest to outcomes? It may seem obvious to you, but make sure we know why might you expect to see any relationship between your policy/treatment variables and outcomes.

This research has a dual objective. Firstly, it aims to assess the effectiveness of the SASH program in reaching the low-income population that it was targeted to. This cans serve as an assessment of the performance of the SASH program and identify potential issues during its roll out that can inform the development of similar programs in the future. For this first objective of the research, the treatment considered in the study is the eligibility to receive the SASH subsidy for the installation of residential solar systems, which depended exclusively on household income (households situated under 80% of county median income). Therefore, the "treatment" is uniform across California state during the program period, and is defined by the percentage of eligible households per zip code area and year.

The second objective of this research is to understand additional drivers and barriers that affect the adoption of solar residential systems beyond household income, particularly within marginalized communities. For instance, in order to access the SASH subsidy, households had to go through an application and administrative process with their local utility. For this purpose, we will use both information on the installations of residential solar systems under the SASH program as well as wider data on the total number of residential solar systems in California during the same study period, provided by UC Berkeley Tracking The Sun (TTS) initiative.

This research aims to inform the design of potential solar subsidy programs that consider the multiple dimensions of differential solar adoption by households beyond income. While solar technologies are becoming increasingly affordable over the last years, additional barriers exist for marginalized and vulnerable communities to access residential solar systems, which has resulted in a growing ownership demographic gap. Since technologies like residential solar systems have the potential to alleviate energy poverty and environmental burdens often concentrated in marginalized communities, effectively promoting residential solar adoption among these groups continues to be a major policy focus in California, with subsequent programs developed after SASH.

Data Description

Describe your data sources, how was your input data generated, the unit of observation and population represented by your sample(s).

The 1. SASH Data: * Description: A database of solar systems installed under the SASH incentive program in California from 2008 to 2018. * Availability: Publicly available online through California Distributed Generation Statistics. * Unit of Observation: Individual system data grouped at zip code area-year, with the main variable being the number of systems installed under the SASH program. * Size: Including 5264 subsidized systems installed in California. * Population Represented: Subsidized residential systems installed under the SASH 1.0 scheme among low-income households, with 100% coverage of systems installed.

2. TTS Data:

- Description: A database of all solar systems installed in California from 2008 to 2018.
- Availability: Publicly available online through the Berkeley Tracking The Sun initiative.
- Unit of Observation: Individual system data grouped at zip code area-year, with the main variable being the number of total systems installed.
- Population Represented: All solar systems installed in California, with an estimated coverage of 77% of total systems installed.
- Size: Including over 707000 systems installed in California.
- Population Represented: Total residential systems installed in California, with estimated 77% coverage of systems installed.

3. 5-year ACS Data (Zip Code Level):

- Description: Socioeconomic data including population, racial composition, income distribution, language, and education information.
- Availability: Publicly accessible through tidycensus and the US Bureau of Statistics.
- Unit of Observation: Statistics calculated at zip code area-year level.
- Population Represented: Household characteristics at the zip code area level in California over the study period.

4. 5-year ACS Data (County Level):

- Description: Median area income statistics at the county level for California, used to calculate eligibility per zip code based on zip code level income distribution data.
- Unit of Observation: Statistics calculated at county-year.

Describe any key data restructuring or subsamples of the input data

Unit of observation: The unit of observation of the SASH and TTS datasets is each individual solar system installed in California. For our study, we are grouping and summarizing individual installations statistics (both under the SASH subsidy scheme and total installations represented by TTS) at the zip code - year level.

Outcome variable: Our outcome variable will be the installation rate, measured as the number of subsidized (SASH dataset) or total (TTS dataset) solar system installations per 10000 people at zip code - year level.

Treatment variable: The main policy/treatment variable to be explored for the SASH installation rate will be the eligibility share, or % of households at zip code - year level that qualify to receive the SASH subsidy. These are households which annual income is under 80% of the Area Median Income, the median household

income at county level on a given year. In order to obtain this eligibility share variable, we have extracted the number of households at zip code - year level that fall within the different household income "buckets" provided by the 5-year ACS. We will use these values to interpolate a function that represents the number of households in a given zip code - year that are under a certain income threshold. Then, we have obtained the median household income at county level for each year from ACS. The income threshold for the zip codes in each county - year is obtained as the 80% of the county median income. Through the interpolated function, we have identified the number of eligible households and the corresponding eligibility share for each zip code - year.

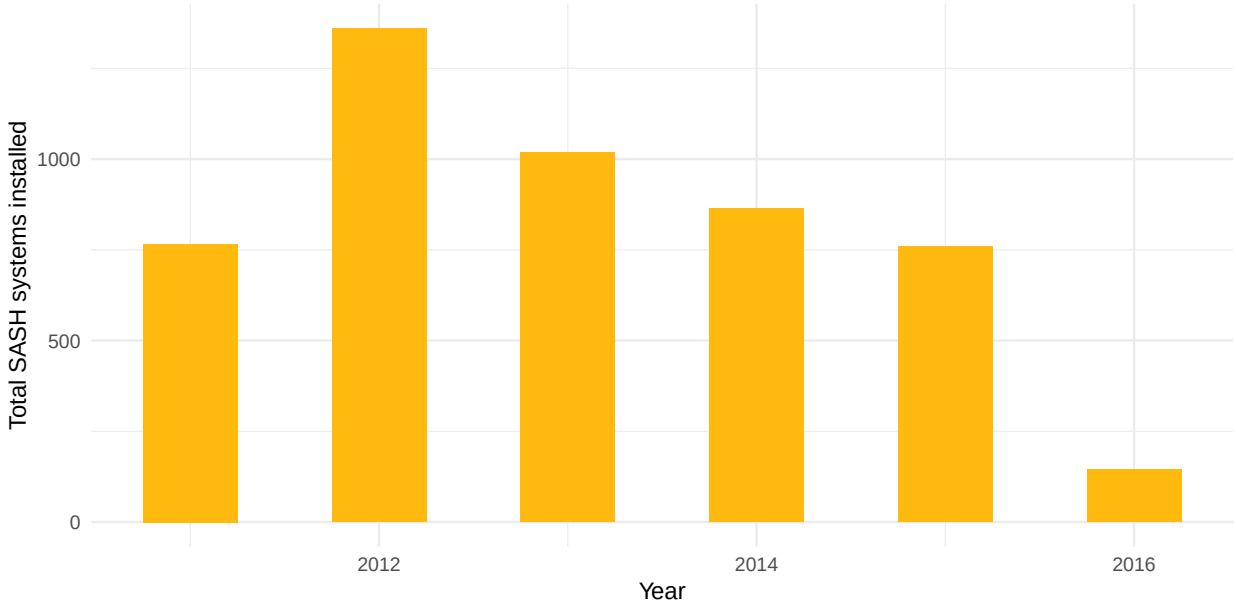
Control variables: A non exhaustive list of control variables can include the racial composition, educational achievement, English fluency, solar irradiation or share of houses suitable for solar installations at the zip code - year level. The socioeconomic control variables are obtained from the 5-year ACS and converted from count figures into shares of the total population by zip code - year.

Limitations of the data include that data disaggregated at zip code level does not seem to be available before the 2011-2015 5-year ACS survey. Thus, we are considering working with a subset of the SASH and TTS data running from 2011 to 2018 instead of 2008 to 2018, unless appropriate socioeconomic data from 2008-2011 is available. The most suitable datasets for obtaining solar irradiation and the share of houses suitable for solar installations are still being identified, but we don't expect it to affect the subset of data used for the study. Additionally, relying on the 5-year ACS survey for socioeconomic data limits the variation of socioeconomic and demographic variables over time in our study period.

Preliminary exploratory analysis

Show descriptive stats to summarize the distribution of your key X and Y variables, using tables and/or charts. One of the most important goals of this deliverable is to make sure you have enough sample variation for you to work with, so you should prioritize this preliminary EDA and report results, however preliminary.

The following bar plots show the number of SASH and total solar residential systems installed in California per year between 2011 and 2017. As we can appreciate, the number of subsidized SASH systems represents a very reduced share of the total systems installed.



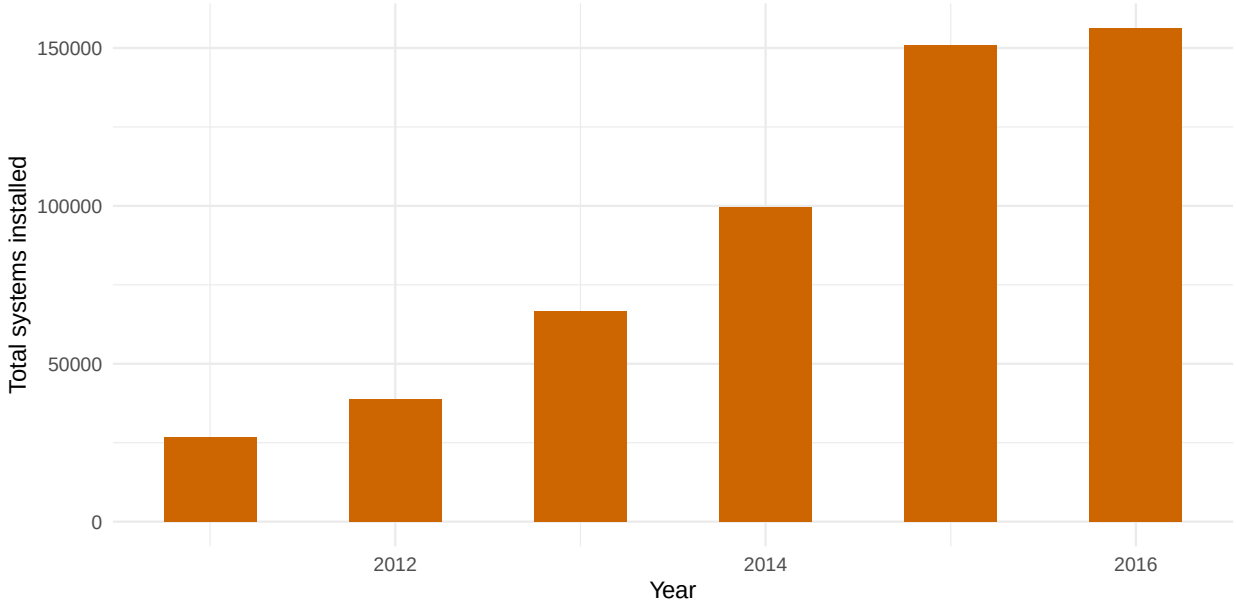
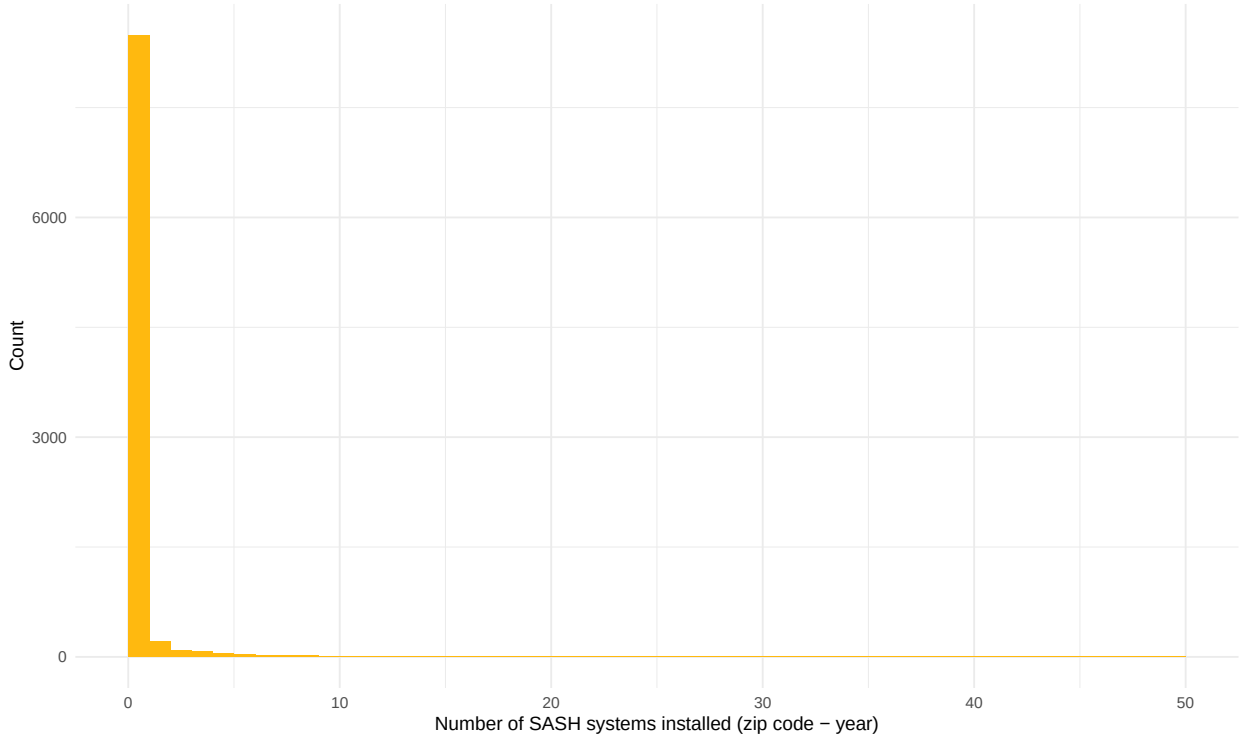


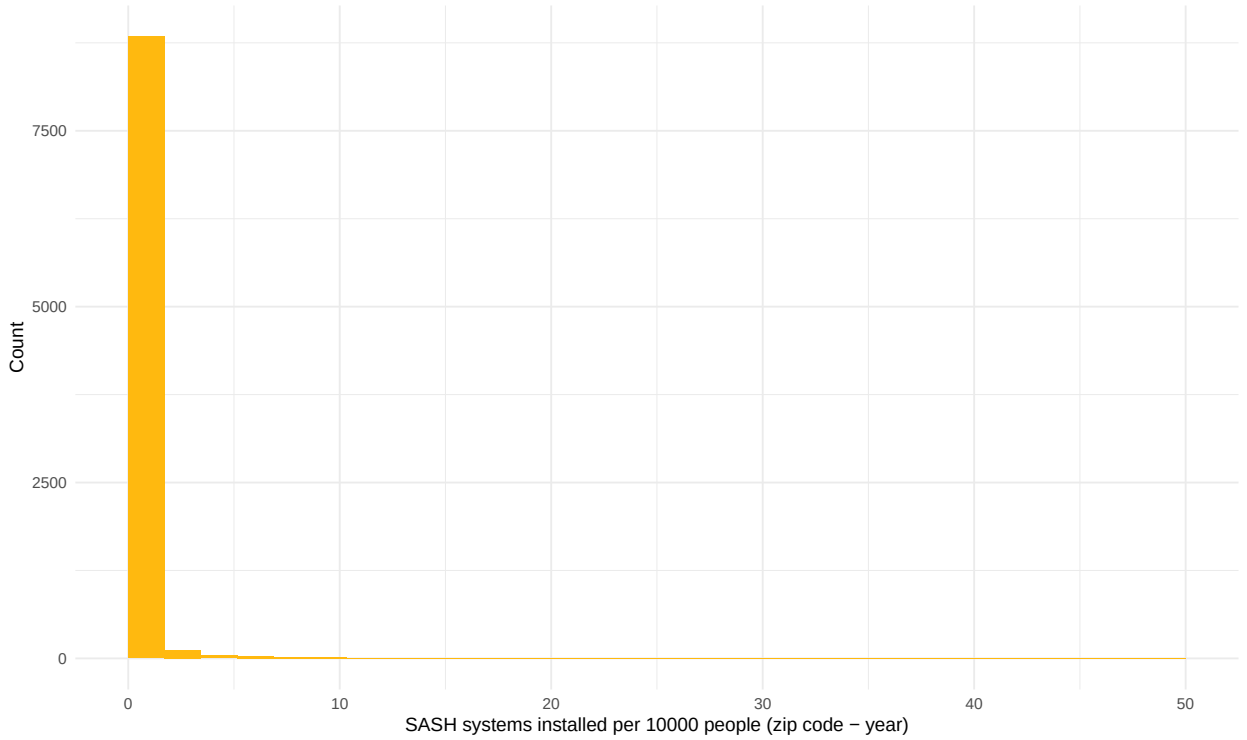
Table 1: Summary statistics at zip code - year level.

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
SASH installed systems per 10000 people	9135	0.37	4.07	0	0	0	239.73
Total installed systems per 10000 people	9135	29.1	52.65	0	4.1	37.41	2142.86
Share of SASH in total installations	7906	0.02	0.11	0	0	0	1
Median household income	8985	64193.17	28749.48	2499	43244	79188	250001
Hispanic share	9135	30.57	24.99	0	10.32	45.55	100
Share of non-fluent English	9135	9.46	11.04	0	1.34	14.26	78.54
Share with less than highschool degree	9118	16.25	15.01	0	4.98	23.2	100
Share with college degree	9118	30.41	20.9	0	14.38	43.24	100

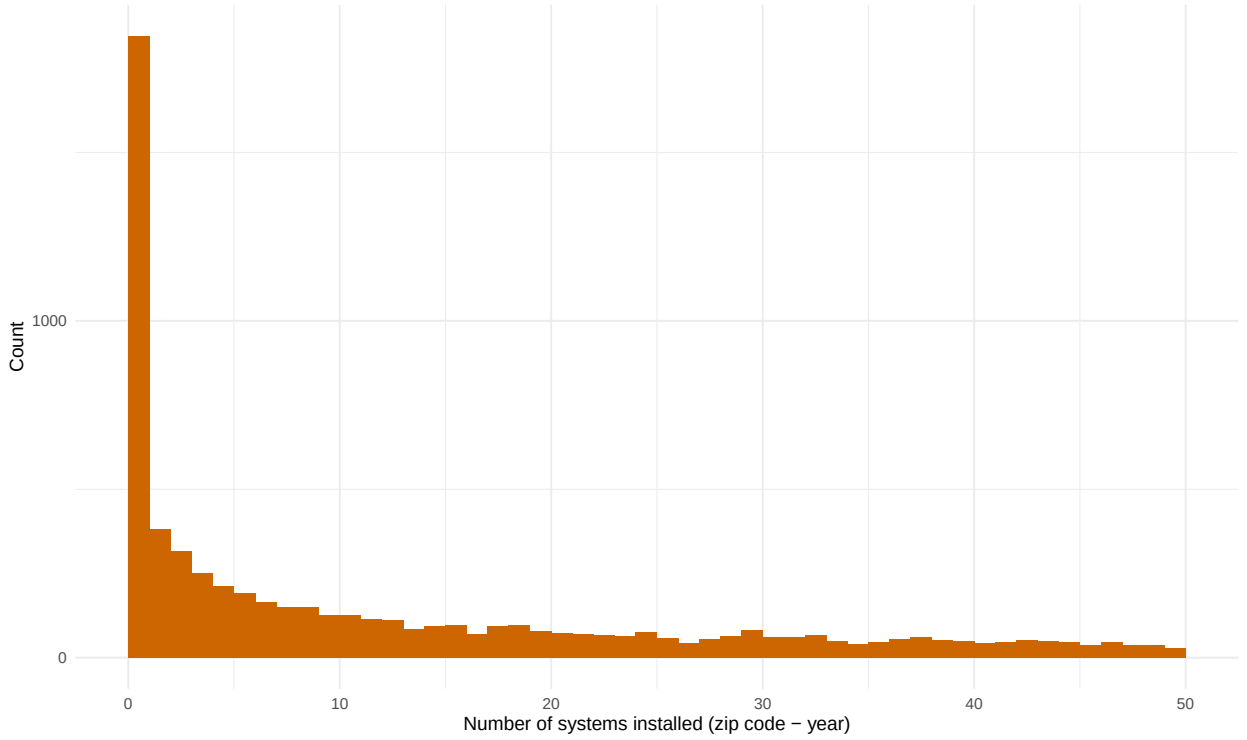
Table 1 includes the distribution of the proposed outcome variables (solar system installation rates), treatment variable (% of households eligible for susbsidy), and a non-exhaustive list of control variables including income, demographics, education, and language isolation.

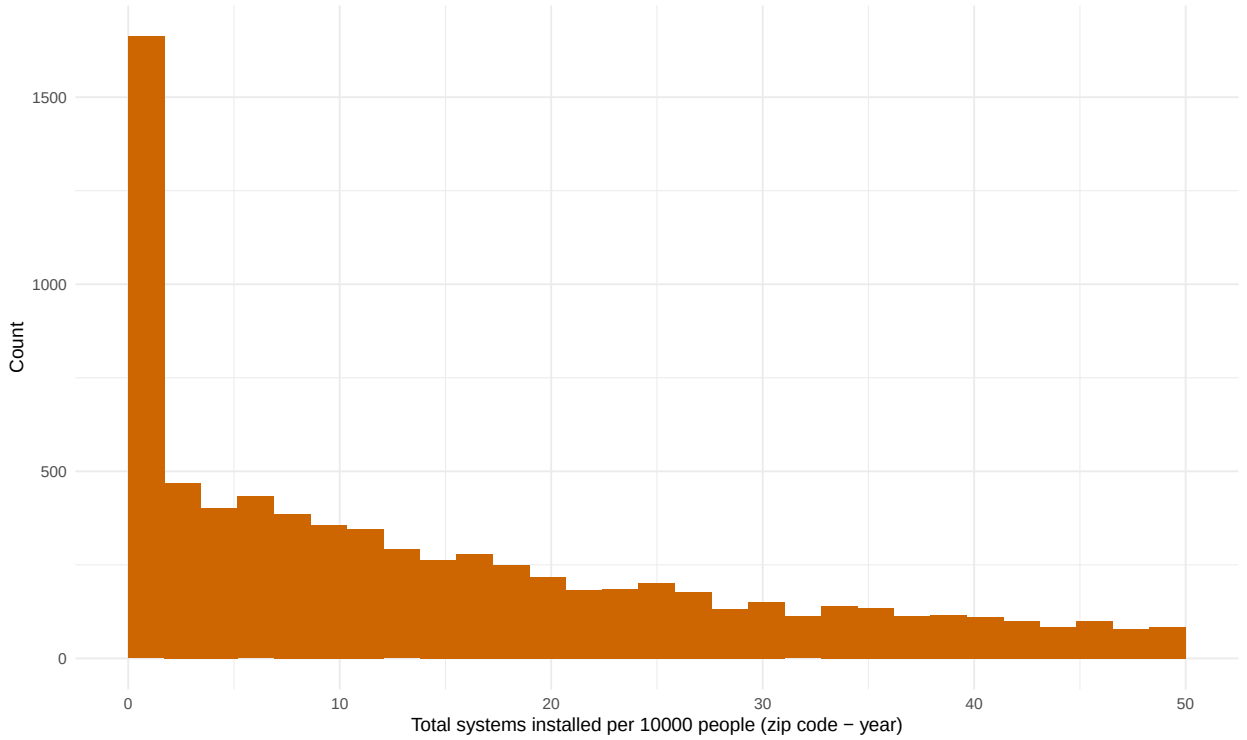
The following histograms represent the distribution of the number of SASH systems installed per zip code - year, with installation rates mostly under 1 per 10000 people.



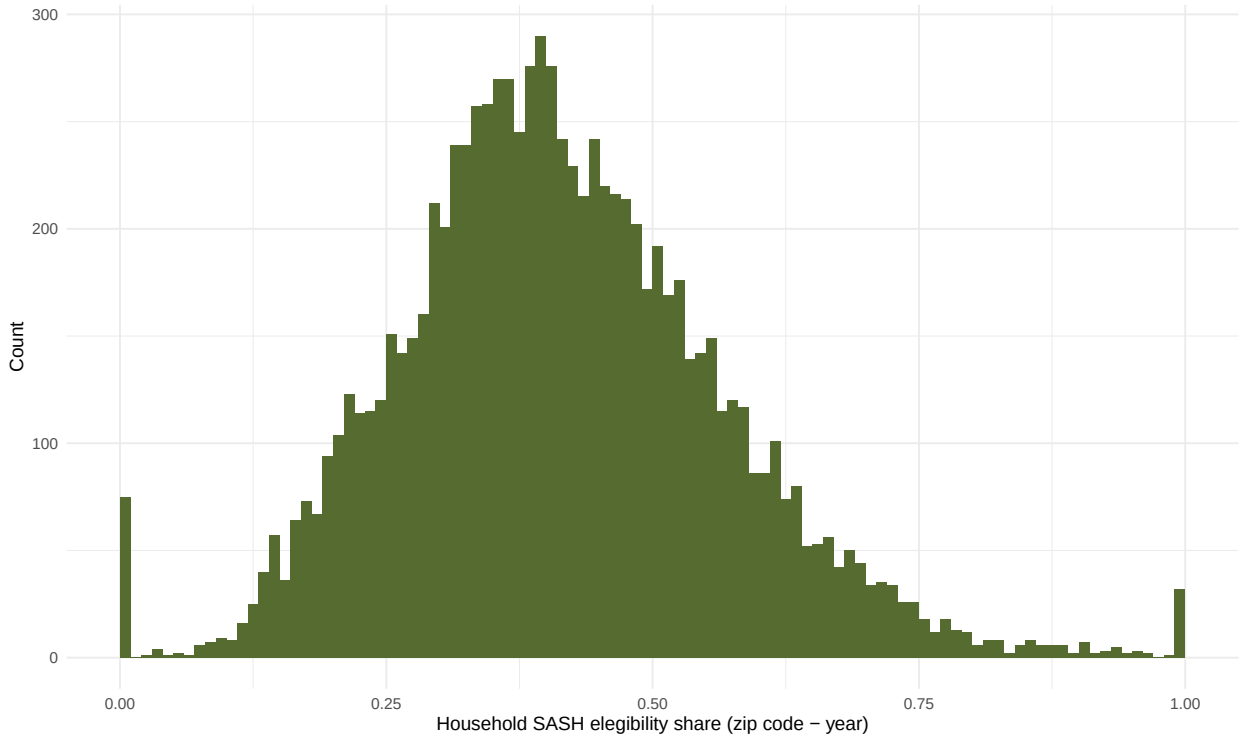


The following histograms represent the distribution of the number of total systems installed per zip code - year, with more spread installation rates.





The following histogram represents the distribution of the % of households eligible for the subsidy at zip code - year level.



Empirical Strategy

Carefully describe the explanatory analysis you are planning on working towards. What is the policy variation you'll be exploiting in any regression analysis to come and potential threats to internal validity? Another way to think about this: what sort of comparisons will you be making with your planned regression analysis?

Our empirical strategy relies on constructing a zip code area-year panel and investigating the effect of variation in eligibility share on the SASH system installation rate. The strategy considers several potential confounding factors, such as population density, educational levels, racial composition, English fluency, solar irradiation, and depletion of funds for the policy intervention over time. We will use a fixed-effect strategy to mitigate some of these concerns.

The zip code and year fixed effects strategy will enable to control of variations in solar penetration across different areas and time periods. This will allow for a more accurate comparison between the penetration of solar systems among low-income communities and the overall trend of installations in the state.

The initial focus of the study is to assess how effective was the SASH program in reaching the low income communities that it was targeting, by analysing the penetration of solar systems among low-income communities (based on their eligibility share). This will be done throug a fixed effects strategy to mitigate time-variant confounding factors shared across all areas (e.g. decrease in solar system costs over time), as well as spatial-variant factors constant over time (e.g. specific demographic compositions at zip code level).

Secondarily, leveraging the wider trend of solar system installations in California over the study period available through the TTS dataset, the study aims to identify additional drivers of the adoption of solar home systems in California beyond household income. By doing so, we hope to inform better-targeted policies to accelerate solar adoption among marginalized communities, for instance including conditions that consider additional factors apart from income in order to qualify for future subsidy programs.

Outline key steps to prepare the data for analysis (data cleaning, recoding, merging, appending, aggregation, etc.).

In any data analysis project, it is important to clean and organize the data properly to make it usable for analysis. The first step in this process is to clean the initial datasets and select and relevant variables and a timeframe for the analysis. This ensures that the analysis is focused and efficient. In our, the TTS, SASH, and ACS datasets are used. From TTS and SASH datasets, variables measuring the installations are selected, while more technical variables are discarded. From the ACS dataset, relevant socioeconomic variables are selected to be used as controls for the study.

Once the relevant variables are identified, the next step is to recode them in a way that is useful. This often involves grouping variables at a certain level, such as the zip code - year level. This makes the data more manageable and allows for easy comparison across time and geographic regions. In our case, the installation rate of solar systems (defined as the number of installations per 10000 people) is calculated at the zip code - year level, which can be compared across different areas and years. The socioeconomic variables are also obtained at zip code - year level.

After the relevant variables are selected and recoded, the next step is to merge the different datasets at the zip code-year level. This allows for a more comprehensive analysis, as it combines information from various sources and can provide insights that would not be available if the datasets were analyzed separately. For example, by merging the TTS, SASH, and ACS datasets, it is possible to explore how factors such as income, education, and racial composition of a community affect the adoption of solar home systems.

Overall, the process of data cleaning, recoding, and merging is an essential part of the project. It ensures that the data is properly organized and that relevant information is included, which can lead to more accurate and meaningful insights.

Highlight key issues or limitations you need to address – be specific about how you plan to solve programming obstacles or fill critical data gaps!

To mitigate key issues or limitations, critical data gaps must be identified, such as the availability of zip code level information in the 5-year ACS before 2011. We will try to overcome this limitation by using the decennial ACS survey, or manually downloading data from ACS without tidycensus, in order to obtain the required data for earlier years. Furthermore, during the years 2017 and 2018 SASH systems were only installed in 5 different zip code areas due to the exhaustion of the program funding, with the large majority of the 5200 SASH installations happening until 2016. Therefore, the initial analysis considers the period from 2011 to 2016, where SASH installations were widely made across the state.

Another issue identified in early-stages of the project is the estimation of eligibility share for the SASH subsidy, based on the Area Median Income at county level published by the California Housing Board. Since these median income values for each year are not available in a digitized format (except for 2022), we have used the 5-year ACS dataset to obtain an estimate of the median income at county - year level, used then to calculate the share of eligible households for the subsidy at zip code - year level.

Regarding the data on solar irradiation and the share of buildings suitable for the installation of solar systems, we are aware that data for different years might not be available. Nevertheless, as these variables are unlikely to greatly vary during the period considered in the study, we will use an individual or average value for all years considered, which should still help to control for potential omitted variable bias in our study. We are still in the process of identifying the most suitable dataset to include solar irradiation data in our analysis.

Additionally, the project faces a potential challenge of small sample variation of SASH installations at the zip code - year level. To address this issue, as we advance with the project we may need to aggregate the data at the county - year level for the analysis of the effectiveness of the SASH program. We would return to the zip code level data only to assess additional drivers of solar adoption using the TTS dataset, which presents a larger sample variation.

By identifying these potential issues and developing specific solutions to address them, the project aims to ensure that the analysis is as robust and comprehensive as possible.

References

California Distributed Generation Statistics. Available at: https://www.californiadgstats.ca.gov/downloads/

Berkeley's Lab Tracking The Sun report. Available at: https://emp.lbl.gov/tracking-sun-tool

Knaide, J., Timmins, C., & Kimbrough, K. (2017). Low-Income Residential Solar: A SASH Evaluation.

Reames, T. (2020). Distributional disparities in residential rooftop solar potential and penetration in four cities in the United States.

Lukanov, B., Krieger, E. (2019). Distributed solar and environmental justice: Exploring the demographic and socio-economic trends of residential PV adoption in California.

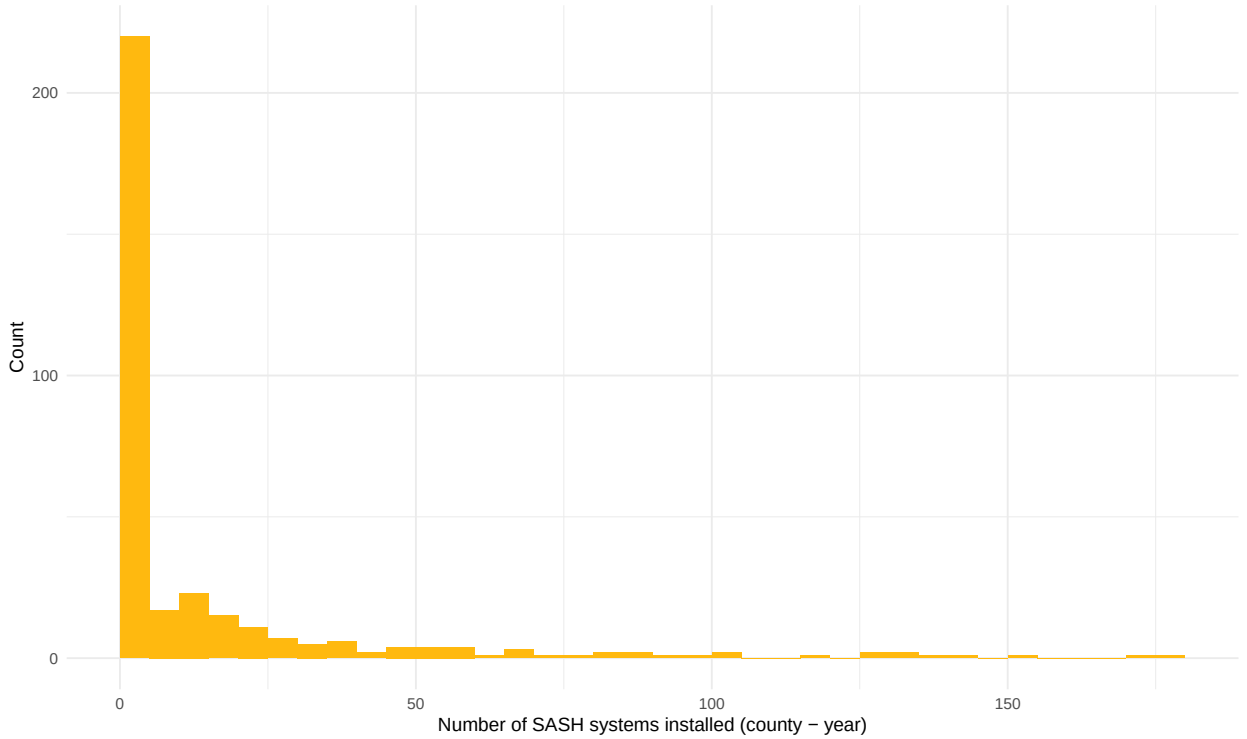
Table 2: Summary statistics at county - year level.

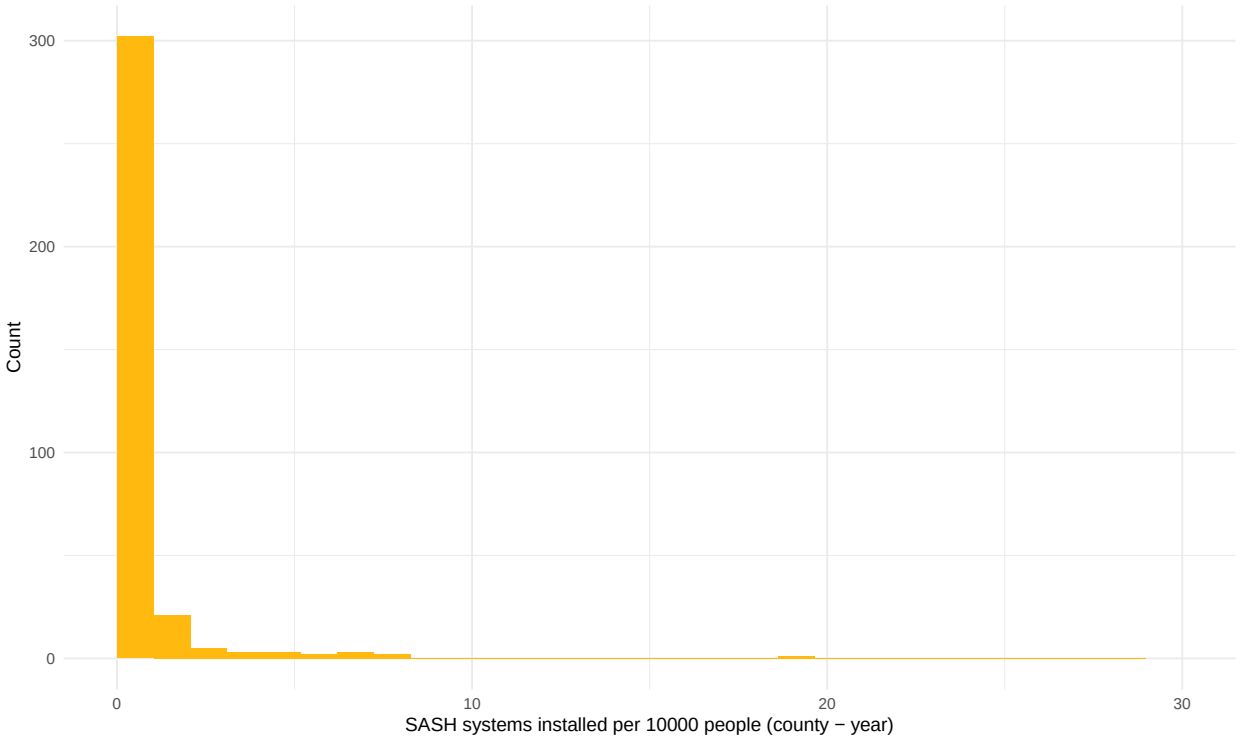
Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
SASH installed systems per 10000 people	342	0.51	1.58	0	0	0.34	19.59
Total installed systems per 10000 people	342	26.1	25.75	0	7.67	37.14	151.68
Share of SASH in total installations	331	0.03	0.07	0	0	0.02	0.47
Median household income	342	56186.79	14576.1	34974	44534.25	64956.5	101173
Hispanic share	342	29.34	17.29	6.74	14.68	42.61	82.7
Share of non-fluent English	342	8.83	6.23	0.49	3.54	12.95	29.5
Share with less than highschool degree	342	16.46	7.41	4.53	10.78	21.07	33.29
Share with college degree	342	25.35	11.3	10.47	16.83	31.9	59.19

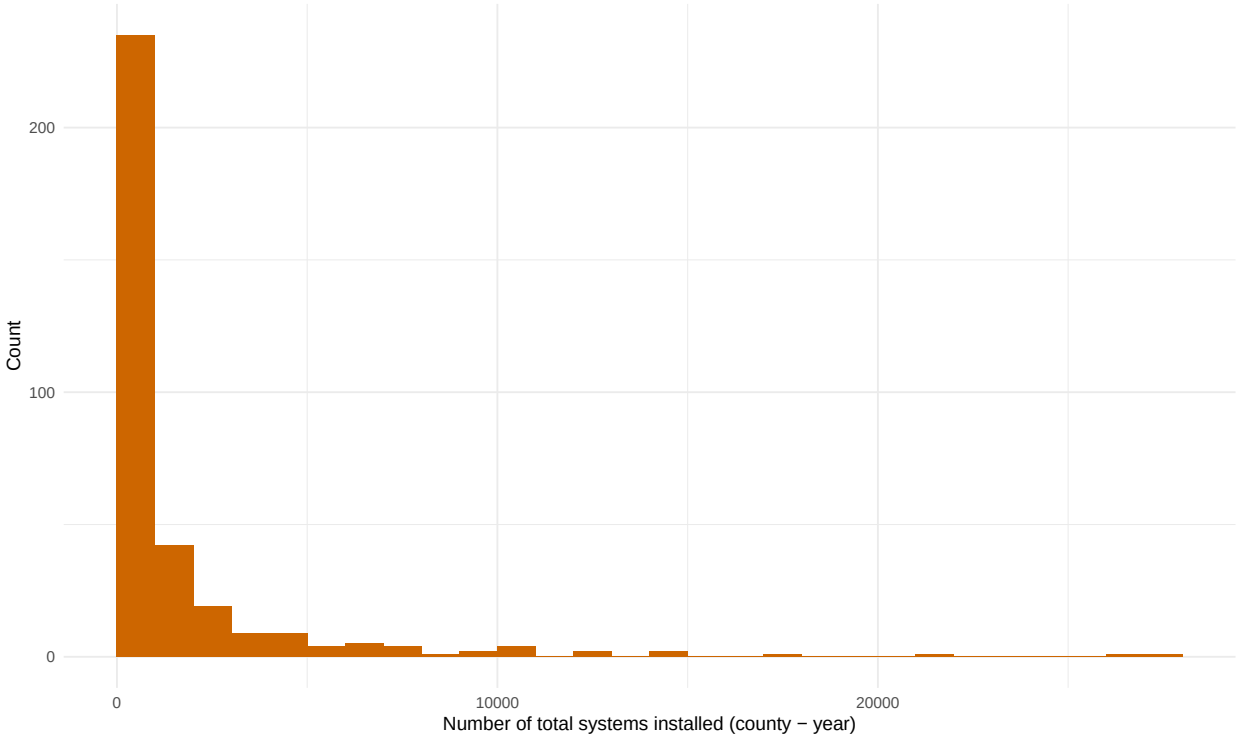
Appendix

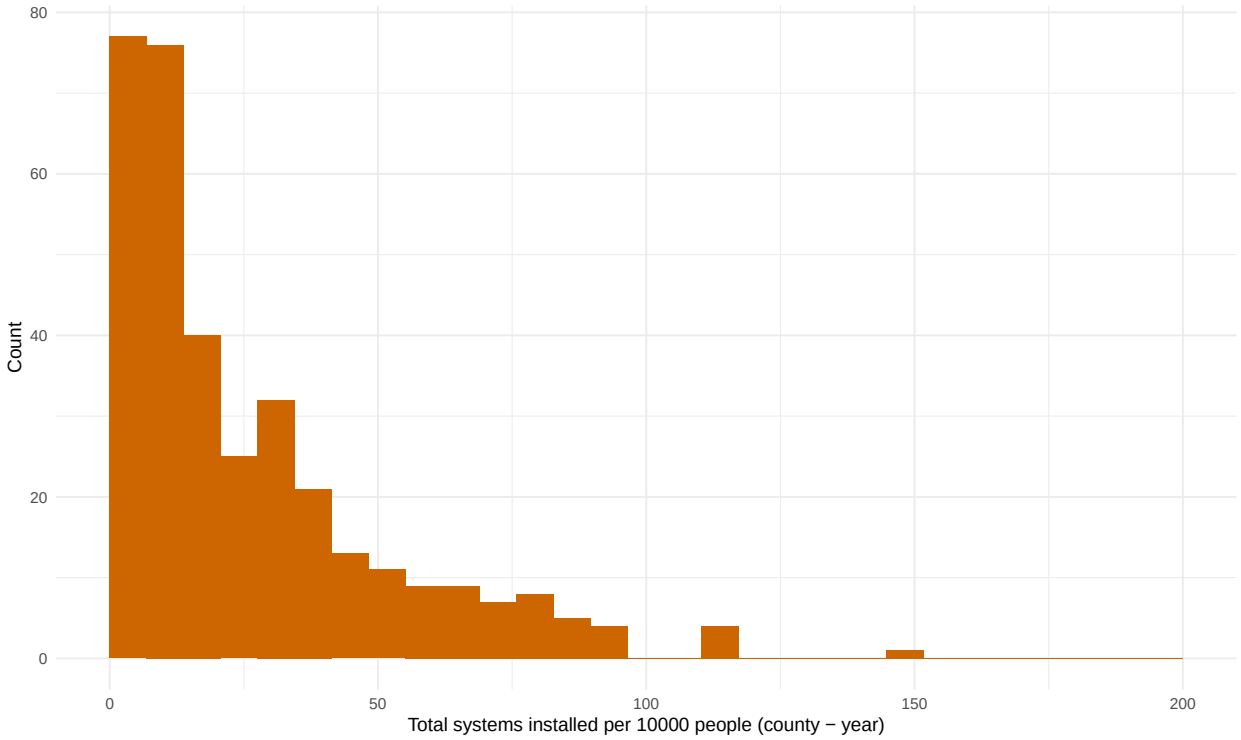
Aggregation at county - year level

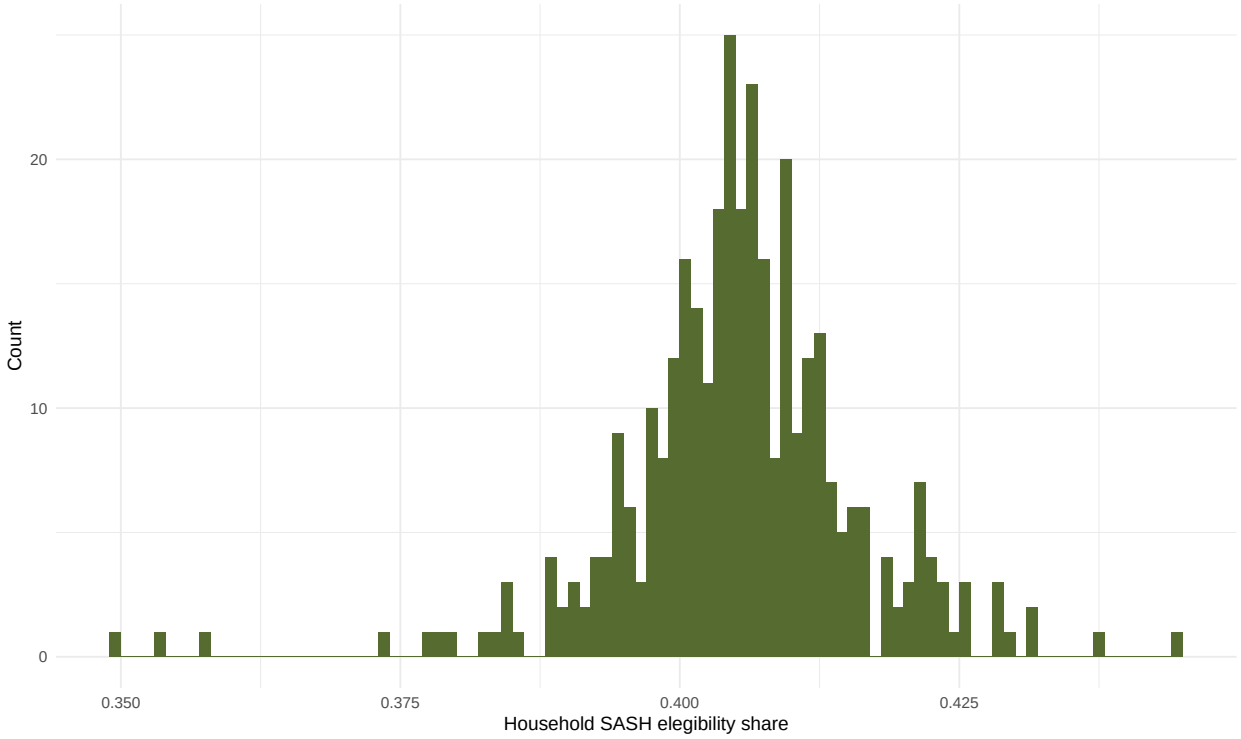
As mentioned in the Exploratory Analysis section, a potential challenge for the analysis is the small sample variation of SASH installations at the zip code - year level. This section includes the corresponding descriptive statistics and graphs when aggregating the SASH and TTS installations data at county - year level instead of the proposed zip code - year level.




300
200
100
0
Count
0
10
20
30
SASH systems installed per 10000 people (county − year)


200
100
0
Count
0
10000
20000
Number of total systems installed (county − year)


80
60
40
20
0
Count
0
50
100
150
200
Total systems installed per 10000 people (county − year)


20
10
0
Count
0.350
0.375
0.400
0.425
Household SASH elegibility share

Coding work to-date for importing, cleaning, recoding, restructuring and joining input data sources.

```
library("readxl")
library("writexl")
library(tidyverse)
library(dplyr)
library(data.table)
library(tidycensus)
library(weights)
library(vtable)
library(ggplot2)
library(ggbreak)
library(patchwork)

setwd("G:/My Drive/Spring 2023/Data Analysis in R/Final_Project/R Final Project")
getwd()

# ------------------------------------------------------------------------------
#    Initial data loading
# ------------------------------------------------------------------------------


# Filter data for SASH 1.0 project only from Low Income Incentive dataset
SASH_data <- read.csv("Data/LowIncome_Applications_Dataset_2023-02-16.csv") %>%
  filter(Low.Income.Program.Type == "SASH 1.0") %>%
  discard(~all(is.na(.) | . ==""))

str(SASH_data)

# Filter data for California 2008-2017 from TTS dataset
TTS_data_CA_2008_2017 <- read.csv("Data/TTS_LBNL_public_file_07-Sep-2022_all.csv") %>%
  mutate(installation_date = as.character(installation_date),
         zip_code = as.factor(zip_code)) %>%
  mutate(installation_year = substr(installation_date, 8, -1)) %>%
  filter(state == "CA", customer_segment == "RES") %>%
  filter(installation_year %in% c("2008", "2009", "2010", "2011", "2012", "2013",
                                  "2014", "2015", "2016", "2017"))

# Inspection
str(TTS_data)

summary(TTS_data$installation_date)
summary(TTS_data$zip_code)

# Save new datasets
write.csv(TTS_data_CA_2008_2017, "Data/TTS_data_CA_2008_2017.csv")
write.csv(SASH_data, "Data/SASH_data.csv")


## ------------------------------------------------------------------------------
## Get demographic data from the American Community Survey (via tidycensus)
## ------------------------------------------------------------------------------
```

```
# Download data from the US Census Bureau using the tidycensus package

# census_api_key("insert_census_key_here", install = TRUE)

# initialize new data frame to store ACS data from API requests
 CA_acs_zipcode_11_18 <- data_frame() # Zip code data only available from 2011 onwards in 5-year ACS

 v15 <- load_variables(2015, "acs5", cache = TRUE)

# #View(v15)

# Set up a for loop for each year of data to access
 for (i in 2011:2018) {

  # Query ACS data from the census API for each year 2010-2017
  acs <- get_acs(geography = "zip code tabulation area",
                 state = "CA",
                 variables = c(pop = "B01001_001",
                 white_pop = "B01001H_001",
                 hisp_pop = "B01001I_001",
                 asian_pop = "B02001_005",
                 black_pop = "B02001_003",
                 male_pop = "B01001_002",
                 med_age = "B01002_001",
                 nonfluent_english_pop = "B06007_005",
                 ed_LThighschool = "B23006_002",
                 ed_college = "B23006_023",
                 ed_total = "B23006_001",
                 medianinc = "B19013_001",
                 inc_under10 = "B19001_002",
                 inc_10to15 = "B19001_003",
                 inc_15to20 = "B19001_004",
                 inc_20to25 = "B19001_005",
                 inc_25to30 = "B19001_006",
                 inc_30to35 = "B19001_007",
                 inc_35to40 = "B19001_008",
                 inc_40to45 = "B19001_009",
                 inc_45to50 = "B19001_010",
                 inc_50to60 = "B19001_011",
                 inc_60to75 = "B19001_012",
                 inc_75to100 = "B19001_013",
                 inc_100to125 = "B19001_014",
                 inc_125to150 = "B19001_015",
                 inc_150to200 = "B19001_016",
                 inc_over200 = "B19001_017",
                 pov = "B17025_002"),
             year = i)

  # Transform data for later analysis and prep for join
    acs <- acs %>%
      select(-moe, -NAME) %>%
      pivot_wider(names_from = variable, values_from = estimate) %>%
      mutate(year = i,
```

```
            whiteshare = 100 * (white_pop/pop),
            hispshare = 100 * (hisp_pop/pop),
            asianshare = 100 * (asian_pop/pop),
            blackshare = 100 * (black_pop/pop),
            maleshare = 100 * (male_pop/pop),
            nonfluentshare = 100 * (nonfluent_english_pop/pop),
            LThighschoolshare = 100 * (ed_LThighschool/ed_total),
            collegeshare = 100 * (ed_college/ed_total),
            poverty_rate = 100 * (pov/pop),
            zip_code = as.character(GEOID)) %>%
     select(-GEOID, -ed_total)
   print(i)

  # Append each year of data to a combined dataset
  CA_acs_zipcode_11_18 <- CA_acs_zipcode_11_18 %>%
    bind_rows(acs)
 }

 # clean input data
 CA_acs_zipcode_11_18 <- CA_acs_zipcode_11_18 %>%
   mutate(zip_code = as.factor(str_sub(zip_code, -5, - 1)))

 # Save ACS data frame
 write.csv(CA_acs_zipcode_11_18, "Data/CA_acs_zipcode_11_17.csv")


# Get median income at county level - used as reference for SASH eligibility
CA_acs_countyAMI_11_18 <- data_frame() # Zip code data only available from 2011 onwards in 5-year ACS

v15 <- load_variables(2015, "acs5", cache = TRUE)

# #View(v15)

# Set up a for loop for each year of data to access
for (i in 2011:2018) {

  # Query ACS data from the census API for each year 2010-2017
  acs <- get_acs(geography = "county",
                 state = "CA",
                 variables = c(medianinc = "B19013_001"),
                 year = i)

  # Transform data for later analysis and prep for join
  acs <- acs %>%
    select(-moe) %>%
    pivot_wider(names_from = variable, values_from = estimate) %>%
    mutate(year = i,
           county = as.character(NAME),
           county_medianinc = medianinc) %>%
    select(-GEOID, -NAME, -medianinc)
  print(i)

  # Append each year of data to a combined dataset
```

```
  CA_acs_countyAMI_11_18 <- CA_acs_countyAMI_11_18 %>%
    bind_rows(acs)
}

# clean input data
CA_acs_countyAMI_11_18 <- CA_acs_countyAMI_11_18 %>%
  mutate(county = as.factor(str_sub(county, 1, - 20)))

# Save ACS data frame
write.csv(CA_acs_countyAMI_11_18, "Data/CA_acs_countyAMI_11_18.csv")

## ------------------------------------------------------------------------------
## Restructuring of installation datasets into zip code - year panel
## ------------------------------------------------------------------------------

# Read saved datasets
TTS_data_CA_2008_2017 <- read.csv("Data/TTS_data_CA_2008_2017.csv")
SASH_data <- read.csv("Data/SASH_data.csv")

# Get zip code - year summaries for TTS installations
TTS_data_zipyr <- TTS_data_CA_2008_2017 %>%
  mutate(zip_code = substr(zip_code, 1, 5)) %>%
  mutate(zip_code = as.factor(zip_code),
         year = as.factor(installation_year)) %>%
  group_by(zip_code, year) %>%
  summarise(total_sys_installed = n(),
            total_avg_size_DC = mean(system_size_DC))

summary(TTS_data_zipyr) # Average of 54 TTS installations per zip code

# Get zip code - year summaries for SASH installations
SASH_data_zipyr <- SASH_data %>%
  mutate(zip_code = substr(Host.Customer.Physical.Address.Zip.Code, 1, 5),
         year = substr(First.Completed.Date,
                       (nchar(First.Completed.Date)+1)-4,nchar(First.Completed.Date))) %>%
  mutate(zip_code = as.factor(zip_code),
         year = as.factor(year)) %>%
  group_by(zip_code, year) %>%
  summarise(SASH_sys_installed = n(),
            SASH_avg_size_DC = mean(Nameplate.Rating..KW.),
            SASH_avg_incentive = mean(Incentive.Amount),
            SASH_avg_syscost = mean(Total.System.Cost))

summary(SASH_data_zipyr) # Average of 4 SASH installations per zip code

# Join TTS and SASH installation data at zip code - year level
installations_data_zipyr <- SASH_data_zipyr %>%
  full_join(TTS_data_zipyr, by=c("zip_code","year")) %>%
  mutate(SASH_installations_share = SASH_sys_installed/total_sys_installed)

# Save joined dataset
write.csv(installations_data_zipyr, "Data/installations_data_zipyr.csv")
```

```r
# Inspection

summary(installations_data_zipyr$SASH_installations_share) # SASH installations represent 20% of total
installations_data_zipyr %>% filter(SASH_installations_share >1) %>% nrow() # 12 zip codes where SASH i

## ------------------------------------------------------------------------
## Merging datasets
## ------------------------------------------------------------------------

# Read SASH and TTS installations dataframe
 installations_data_zipyr <- read.csv("Data/installations_data_zipyr.csv") %>%
  select(-X)

 # Read ACS socioeconomic dataframe
 CA_acs_zipcode_11_18 <- read.csv("Data/CA_acs_zipcode_11_17.csv") %>%
  mutate(zip_code = as.factor(zip_code)) %>%
  select(-X)

 # Read list of counties for each zip code
 CA_county_zipcode <- read_excel("Data/CA_Zipcode_County_List.xlsx") %>%
   rename(zip_code = `IP Code`, county = County) %>%
   mutate(zip_code = as.factor(str_sub(zip_code, -5, - 1))) %>%
   select(-City, -Type)

 # Read Median Income per county from ACS (substitute for CA Housing Department data, not digitalized)
 CA_acs_countyAMI_11_18 <-  read.csv("Data/CA_acs_countyAMI_11_18.csv") %>%
   select(-X)


 # Join installations and demographic data
 complete_data_zipyr <- installations_data_zipyr %>%
   mutate(zip_code = as.factor(zip_code)) %>%
   complete(zip_code,year, fill = list(SASH_sys_installed = 0, total_sys_installed = 0),
            explicit = FALSE) %>%
   left_join(CA_county_zipcode, by = "zip_code") %>%
   left_join(CA_acs_countyAMI_11_18, by=c("county","year")) %>%
   left_join(CA_acs_zipcode_11_18, by=c("zip_code","year")) %>%
   filter(year >= 2011 & year < 2017) %>%
   filter(zip_code != -1) %>%
   filter(!is.na(pop)) %>%
   #mutate(SASH_sys_installed = replace_na(SASH_sys_installed, 0)) %>%
   mutate(SASH_sys_installed = replace_na(SASH_sys_installed, 0),
          SASH_avg_size_DC = replace_na(SASH_avg_size_DC, 0),
          SASH_avg_incentive = replace_na(SASH_avg_incentive, 0),
          SASH_avg_syscost = replace_na(SASH_avg_syscost, 0),
          total_sys_installed = replace_na(total_sys_installed, 0),
          total_sys_installed = ifelse(
            total_sys_installed < SASH_sys_installed, SASH_sys_installed, total_sys_installed),
          total_avg_size_DC = replace_na(SASH_avg_size_DC, 0),
          SASH_installations_share = SASH_sys_installed/total_sys_installed,
          SASH_install_rate_10000 = SASH_sys_installed/pop*10000,
          total_install_rate_10000 = total_sys_installed/pop*10000,
          inc_threshold = 0.8*county_medianinc) %>%
```

```
  mutate(inc_under15 = inc_under10 + inc_10to15,
         inc_under20 = inc_under15 + inc_15to20,
         inc_under25 = inc_under20 + inc_20to25,
         inc_under30 = inc_under25 + inc_25to30,
         inc_under35 = inc_under30 + inc_30to35,
         inc_under40 = inc_under35 + inc_35to40,
         inc_under45 = inc_under40 + inc_40to45,
         inc_under50 = inc_under45 + inc_45to50,
         inc_under60 = inc_under50 + inc_50to60,
         inc_under75 = inc_under60 + inc_60to75,
         inc_under100 = inc_under75 + inc_75to100,
         inc_under125 = inc_under100 + inc_100to125,
         inc_under150 = inc_under125 + inc_125to150,
         inc_under200 = inc_under150 + inc_150to200,
         total_hh = inc_under200 + inc_over200,
         eligible_hh = NA) %>%
  select(-inc_10to15,-inc_15to20,-inc_20to25,-inc_25to30,-inc_30to35,-inc_35to40,
         -inc_40to45,-inc_45to50,-inc_50to60,-inc_60to75,-inc_75to100,-inc_100to125,
         -inc_125to150,-inc_150to200,
         -male_pop, -black_pop, -asian_pop, -white_pop, -hisp_pop, -nonfluent_english_pop) %>%
  relocate(year,county,zip_code) %>%
  arrange(year, county, zip_code)

# Interpolate number of people under 80% of zip code median income based on ACS income buckets
income_interpolation =  data.frame(15,20,25,30,35,40,45,50,60,75,100,125,150,200)

for (i in (1:nrow(complete_data_zipyr))) {
  hh_eligible = approx(x = income_interpolation, y = complete_data_zipyr[i,32:45],
                       xout=complete_data_zipyr[i,31]/1000, method = "linear")
  complete_data_zipyr[i,47] <- hh_eligible$y
}

# Calculate eligibility share in each zip code (% hh under 80% of median income)
complete_data_zipyr <- complete_data_zipyr %>%
  mutate(eligibilityshare = eligible_hh/ total_hh)

# Inspection
complete_data_zipyr %>% filter(is.na(eligible_hh))
complete_data_zipyr %>% filter(is.na(eligibilityshare))

# Remove 14 county observations with no or very small population
complete_data_zipyr <- complete_data_zipyr %>%
  filter(!is.na(eligibilityshare))

summary(complete_data_zipyr)
table(complete_data_zipyr$year, complete_data_zipyr$SASH_sys_installed)
table(complete_data_zipyr$year, complete_data_zipyr$total_sys_installed)

write.csv(complete_data_zipyr, "Data/complete_data_zipyr.csv")


## ------------------------------------------------------------------------
## Summary statistics
```

```
## ------------------------------------------------------------------------

complete_data_zipyr <- read.csv("Data/complete_data_zipyr.csv") %>%
  select(-X)

# Descriptive statistics

summary(complete_data_zipyr$SASH_install_rate_1000)

summary(complete_data_zipyr$total_install_rate_1000)

# Labels for summary statistics table
labs <- c('SASH installed systems per 10000 people',
          'Total installed systems per 10000 people',
          '% of households eligible for SASH program',
          'Share of SASH in total installations',
          'Median household income',
          'Hispanic share',
          'Share of non-fluent English',
          'Share with less than highschool degree',
          'Share with college degree')

# (Include in RMD) Summary statistics table with subset of variables
st(complete_data_zipyr,
   vars = c('SASH_install_rate_10000','total_install_rate_10000','eligibilityshare',
            'SASH_installations_share','medianinc','hispshare','nonfluentshare',
            'LThighschoolshare', 'collegeshare'),
   digits=2,
   labels=labs,
   title='Summary statistics at zip code - year level.')

#  Histogram for SASH installations by zip code - year
ggplot(data = complete_data_zipyr,
       aes(x = SASH_sys_installed)) +
  geom_histogram(fill="darkgoldenrod1", binwidth=1, boundary=0) +
  scale_y_break(c(300, 11000), ticklabels = NULL) +
  labs(x = "Number of SASH systems installed (zip code - year)", y = "Count") +
  scale_y_continuous(n.breaks = 4) +
  xlim(0,50) +
  theme_minimal()

# Histogram for SASH installation rate by zip code - year
ggplot(data = complete_data_zipyr,
       aes(x = SASH_install_rate_10000)) +
  geom_histogram(fill="darkgoldenrod1", bindwith =1, boundary=0) +
  # scale_y_break(c(300, 500), ticklabels = NULL) +
  labs(x = "SASH systems installed per 10000 people (zip code - year)", y = "Count") +
  xlim(0,50) +
  theme_minimal()


# Histogram for total installations by zip code - year
ggplot(data = complete_data_zipyr,
```

```
       aes(x = total_sys_installed)) +
  geom_histogram(fill="darkorange3", binwidth=1, boundary=0) +
  # scale_y_break(c(300, 500), ticklabels = NULL) +
  labs(x = "Number of systems installed (zip code - year)", y = "Count") +
  scale_y_continuous(n.breaks = 4) +
  xlim(0,50) +
  theme_minimal()

#  Histogram for total installation rate by zip code - year
ggplot(data = complete_data_zipyr,
       aes(x = total_install_rate_10000)) +
  geom_histogram(fill="darkorange3", bindwith=1, boundary=0) +
  # scale_y_break(c(300, 500), ticklabels = NULL) +
  labs(x = "Total systems installed per 10000 people (zip code - year)", y = "Count") +
  xlim(0,50) +
  theme_minimal()

# Histogram of eligibility share
ggplot(data = complete_data_zipyr,
       aes(x = eligibilityshare)) +
  geom_histogram(fill="darkolivegreen",binwidth=0.01, boundary=0) +
  # scale_y_break(c(300, 500), ticklabels = NULL) +
  labs(x = "Household SASH elegibility share (zip code - year)", y = "Count") +
  scale_y_continuous(n.breaks = 4) +
  theme_minimal()


# Calculate number of systems installed per year
data_zipyr_annualstat <- complete_data_zipyr %>%
  group_by(year) %>%
  summarise(annual_SASH_installations = sum(SASH_sys_installed),
            annual_total_installations = sum(total_sys_installed))


#  Bar plot for total SASH installations by year
ggplot(data = data_zipyr_annualstat,
       aes(x = year, y = annual_SASH_installations)) +
  geom_bar(stat = "identity", fill="darkgoldenrod1") +
  labs(x = "Year", y = "Total SASH systems installed") +
  theme_minimal()


#  Bar plot for total installations by year
ggplot(data = data_zipyr_annualstat,
       aes(x = year, y = annual_total_installations)) +
  geom_bar(stat = "identity", fill="darkorange3") +
  labs(x = "Year", y = "Total systems installed") +
  theme_minimal()



# Exploratory correlations
wtd.cor(complete_data_zipyr$SASH_install_rate_1000, complete_data_zipyr$eligibilityshare)
```

```
wtd.cor(complete_data_zipyr$total_install_rate_1000, complete_data_zipyr$eligibilityshare)

wtd.cor(complete_data_zipyr$total_install_rate_1000, complete_data_zipyr$poverty_rate)

wtd.cor(complete_data_zipyr$total_install_rate_1000, complete_data_zipyr$hispshare)


## ------------------------------------------------------------------------------
## Alternative: aggregation at county level - county data extraction from ACS
## ------------------------------------------------------------------------------


# Get median income and socioeconomic statistics at county level - used as reference for SASH eligibil
CA_acs_countystats_11_18 <- data_frame() # Zip code data only available from 2011 onwards in 5-year AC

v15 <- load_variables(2015, "acs5", cache = TRUE)

# #View(v15)

# Set up a for loop for each year of data to access
for (i in 2011:2018) {

  # Query ACS data from the census API for each year 2010-2017
  acs <- get_acs(geography = "county",
                 state = "CA",
                 variables = c(medianinc = "B19013_001",
                 pop = "B01001_001",
                 white_pop = "B01001H_001",
                 hisp_pop = "B01001I_001",
                 asian_pop = "B02001_005",
                 black_pop = "B02001_003",
                 male_pop = "B01001_002",
                 med_age = "B01002_001",
                 nonfluent_english_pop = "B06007_005",
                 ed_LThighschool = "B23006_002",
                 ed_college = "B23006_023",
                 ed_total = "B23006_001",
                 medianinc = "B19013_001",
                 inc_under10 = "B19001_002",
                 inc_10to15 = "B19001_003",
                 inc_15to20 = "B19001_004",
                 inc_20to25 = "B19001_005",
                 inc_25to30 = "B19001_006",
                 inc_30to35 = "B19001_007",
                 inc_35to40 = "B19001_008",
                 inc_40to45 = "B19001_009",
                 inc_45to50 = "B19001_010",
                 inc_50to60 = "B19001_011",
                 inc_60to75 = "B19001_012",
                 inc_75to100 = "B19001_013",
                 inc_100to125 = "B19001_014",
                 inc_125to150 = "B19001_015",
                 inc_150to200 = "B19001_016",
```

```
                inc_over200 = "B19001_017",
                pov = "B17025_002"),
            year = i)

 # Transform data for later analysis and prep for join
   acs <- acs %>%
     select(-moe) %>%
     pivot_wider(names_from = variable, values_from = estimate) %>%
     mutate(year = i,
            county = as.character(NAME),
            county_medianinc = medianinc,
            whiteshare = 100 * (white_pop/pop),
            hispshare = 100 * (hisp_pop/pop),
            asianshare = 100 * (asian_pop/pop),
            blackshare = 100 * (black_pop/pop),
            maleshare = 100 * (male_pop/pop),
            nonfluentshare = 100 * (nonfluent_english_pop/pop),
            LThighschoolshare = 100 * (ed_LThighschool/ed_total),
            collegeshare = 100 * (ed_college/ed_total),
            poverty_rate = 100 * (pov/pop)) %>%
     select(-GEOID, -NAME, -medianinc, -ed_total)
    print(i)

  # Append each year of data to a combined dataset
  CA_acs_countystats_11_18 <- CA_acs_countystats_11_18 %>%
    bind_rows(acs)
}

# clean input data
CA_acs_countystats_11_18 <- CA_acs_countystats_11_18 %>%
  mutate(county = as.factor(str_sub(county, 1, - 20)))

# Save ACS data frame
write.csv(CA_acs_countystats_11_18, "Data/CA_acs_countystats_11_18.csv")

## ------------------------------------------------------------------------
## Alternative: aggregation at county level - Merging datasets
## ------------------------------------------------------------------------


# Read SASH and TTS installations dataframe
installations_data_zipyr <- read.csv("Data/installations_data_zipyr.csv") %>%
  select(-X)

# Read ACS socioeconomic dataframe
CA_acs_zipcode_11_18 <- read.csv("Data/CA_acs_zipcode_11_17.csv") %>%
  mutate(zip_code = as.factor(zip_code)) %>%
  select(-X)

# Read list of counties for each zip code
CA_county_zipcode <- read_excel("Data/CA_Zipcode_County_List.xlsx") %>%
  rename(zip_code = `IP Code`, county = County) %>%
  mutate(zip_code = as.factor(str_sub(zip_code, -5, - 1))) %>%
```

```
  select(-City, -Type)

# Read Median Income per county from ACS (substitute for CA Housing Department data, not digitalized)
CA_acs_countystats_11_18 <-  read.csv("Data/CA_acs_countystats_11_18.csv") %>%
  select(-X)


# Join installations and demographic data
complete_data_countyr <- installations_data_zipyr %>%
  mutate(zip_code = as.factor(zip_code)) %>%
  left_join(CA_county_zipcode, by = "zip_code") %>%
  group_by(county,year) %>%
  summarise(SASH_sys_installed = sum(SASH_sys_installed, na.rm = TRUE),
            total_sys_installed = sum(total_sys_installed,  na.rm = TRUE)) %>%
  ungroup() %>%
  complete(county,year, fill = list(SASH_sys_installed = 0, total_sys_installed = 0),
           explicit = FALSE) %>%
  left_join(CA_acs_countystats_11_18, by=c("county","year")) %>%
  filter(year >= 2011 & year < 2017) %>%
  filter(!is.na(pop)) %>%
  #mutate(SASH_sys_installed = replace_na(SASH_sys_installed, 0)) %>%
  mutate(total_sys_installed = ifelse(
    total_sys_installed < SASH_sys_installed, SASH_sys_installed, total_sys_installed),
         SASH_installations_share = SASH_sys_installed/total_sys_installed,
         SASH_install_rate_10000 = SASH_sys_installed/pop*10000,
         total_install_rate_10000 = total_sys_installed/pop*10000,
         inc_threshold = 0.8*county_medianinc) %>%
  mutate(inc_under15 = inc_under10 + inc_10to15,
         inc_under20 = inc_under15 + inc_15to20,
         inc_under25 = inc_under20 + inc_20to25,
         inc_under30 = inc_under25 + inc_25to30,
         inc_under35 = inc_under30 + inc_30to35,
         inc_under40 = inc_under35 + inc_35to40,
         inc_under45 = inc_under40 + inc_40to45,
         inc_under50 = inc_under45 + inc_45to50,
         inc_under60 = inc_under50 + inc_50to60,
         inc_under75 = inc_under60 + inc_60to75,
         inc_under100 = inc_under75 + inc_75to100,
         inc_under125 = inc_under100 + inc_100to125,
         inc_under150 = inc_under125 + inc_125to150,
         inc_under200 = inc_under150 + inc_150to200,
         total_hh = inc_under200 + inc_over200,
         eligible_hh = NA) %>%
  select(-inc_10to15,-inc_15to20,-inc_20to25,-inc_25to30,-inc_30to35,-inc_35to40,
         -inc_40to45,-inc_45to50,-inc_50to60,-inc_60to75,-inc_75to100,-inc_100to125,
         -inc_125to150,-inc_150to200,
         -male_pop, -black_pop, -asian_pop, -white_pop, -hisp_pop, -nonfluent_english_pop) %>%
  relocate(year,county) %>%
  arrange(year, county)

# Interpolate number of people under 80% of zip code median income based on ACS income buckets
income_interpolation =  data.frame(15,20,25,30,35,40,45,50,60,75,100,125,150,200)
```

```r
for (i in (1:nrow(complete_data_countyr))) {
  hh_eligible = approx(x = income_interpolation, y = complete_data_countyr[i,26:39],
                       xout=complete_data_countyr[i,25]/1000, method = "linear")
  complete_data_countyr[i,41] <- hh_eligible$y
}

# Calculate eligibility share in each zip code (% hh under 80% of median income)
complete_data_countyr <- complete_data_countyr %>%
  mutate(eligibilityshare = eligible_hh/ total_hh)

# Inspection
complete_data_countyr %>% filter(is.na(eligible_hh))
complete_data_countyr %>% filter(is.na(eligibilityshare))

summary(complete_data_countyr)
write.csv(complete_data_countyr, "Data/complete_data_countyr.csv")


## ------------------------------------------------------------------------------
## Alternative: aggregation at county level - Summary statistics
## ------------------------------------------------------------------------------

complete_data_countyr <- read.csv("Data/complete_data_countyr.csv") %>%
  select(-X)

table(complete_data_countyr$county, complete_data_countyr$year)

# Descriptive statistics

summary(complete_data_countyr$SASH_install_rate_10000)

summary(complete_data_countyr$total_install_rate_10000)

# Labels for summary statistics table
labs <- c('SASH installed systems per 10000 people',
          'Total installed systems per 10000 people',
          '% of households eligible for SASH program',
          'Share of SASH in total installations',
          'Median household income',
          'Hispanic share',
          'Share of non-fluent English',
          'Share with less than highschool degree',
          'Share with college degree')

# Summary statistics table with subset of variables
st(complete_data_countyr,
   vars = c('SASH_install_rate_10000','total_install_rate_10000','eligibilityshare',
            'SASH_installations_share','county_medianinc','hispshare','nonfluentshare',
            'LThighschoolshare', 'collegeshare'),
   digits=2,
   labels=labs,
   title='Summary statistics at county - year level.')
```

```r
#  Histogram for SASH installations by zip code - year
ggplot(data = complete_data_countyr,
       aes(x = SASH_sys_installed)) +
  geom_histogram(fill="darkgoldenrod1", binwidth=5, boundary=0) +
  # scale_y_break(c(300, 8400), ticklabels = NULL) +
  labs(x = "Number of SASH systems installed (county - year)", y = "Count") +
  scale_y_continuous(n.breaks = 4) +
  theme_minimal()

# Histogram for SASH installation rate by zip code - year
ggplot(data = complete_data_countyr,
       aes(x = SASH_install_rate_10000)) +
  geom_histogram(fill="darkgoldenrod1", bindwith =1, boundary=0) +
  # scale_y_break(c(300, 500), ticklabels = NULL) +
  labs(x = "SASH systems installed per 10000 people (county - year)", y = "Count") +
  xlim(0,30) +
  theme_minimal()

#  Histogram for total installations by zip code - year
ggplot(data = complete_data_countyr,
       aes(x = total_sys_installed)) +
  geom_histogram(fill="darkorange3", binwidth=1000, boundary=0) +
  # scale_y_break(c(300, 500), ticklabels = NULL) +
  labs(x = "Number of total systems installed (county - year)", y = "Count") +
  scale_y_continuous(n.breaks = 4) +
  theme_minimal()

#  Histogram for total installation rate by zip code - year
ggplot(data = complete_data_countyr,
       aes(x = total_install_rate_10000)) +
  geom_histogram(fill="darkorange3", bindwith=1, boundary=0) +
  # scale_y_break(c(300, 500), ticklabels = NULL) +
  labs(x = "Total systems installed per 10000 people (county - year)", y = "Count") +
  xlim(0,200) +
  theme_minimal()

# Histogram of eligibility share
ggplot(data = complete_data_countyr,
       aes(x = eligibilityshare)) +
  geom_histogram(fill="darkolivegreen",binwidth=0.001, boundary=0) +
  # scale_y_break(c(300, 500), ticklabels = NULL) +
  labs(x = "Household SASH elegibility share", y = "Count") +
  scale_y_continuous(n.breaks = 4) +
  theme_minimal()
```