***Aggregated Name***

A preliminary exploration of internet access in California: What factors contribute to people's access to the internet or related 3-C products and does the publicity efforts of California Public Utility Commission (CPUC) staff pay off?

***Project Description***

Having access to the internet, especially high speed internet, is vital to both individual development and country success. However, even in the contemporary United States, not everyone has access to it. Since the idea of digital equity has increasingly been rooted in people's heart, researchers in the telecommunication policy area have developed a series of theories, trying to understand what factors are hindering people from accessing the internet. An aggregated version of such theories is the resources and appropriation theory put forward by Van Dajk (2019), which puts forward a series of personal (such as age, race gender) and positional factors (such as education level, employment status) that may be in relation to the access to internet. Therefore, in order to empirically test this theory, the first research question of this analysis is whether age, gender, education level, employment status, and economic situation influence people's access to the internet or related 3-C products (especially high speed internet).

Promoting digital equity is not only the target of researchers, but also the aim of some government agencies. Lifeline is one important federal assistance program aiming at promoting universal access to telecommunication service, which provides direct discounts to eligible consumers' phone or internet service (USAC, 2022). However, Lifeline is long trapped by the low participation rate, though it seems to be meaningful (Burton & Mayo, 2007). According to discussion with CPUC staff, in order to let more people know about Lifeline, they have tried to set up tents around California's social security office for publicity purposes. Therefore, the second research question of this analysis will be to figure out whether their efforts have been transformed to influence.

***How to Run the Code***

Here is the link to the student's personal github project repository (https://github.com/TioHK/final_project_dsci510). In order to smoothly run the codes I provide, first, some files used in the running procedure should be downloaded first. Such data files are uploaded to my github respiratory and also listed in Table 1. These data files are assumed to be stored in a folder called data in the same directory as you run the code. If

not, the file name used in the code may need further adaptation. Similarly, some intermediary data files produced by the code will also be stored in the same folder. I also included the files needed in the data folder of my submission. Running the codes often need the support of a series of python libraries. The dependencies in use can be found in the requirement.txt file and also in Table 1.

*Table 1.Files to download and libraries to install*

| File Names | non-built in dependencies |
|---|---|
| 'ZIP_Code_Population_Weighted_Centroids.csv'<br>'Lifelineparticipants'<br>'data/tl_2020_06_puma10/tl_2020_06_puma10.shp'<br>'data/FO-Address-Open-Close-Times.xlsx' | requests==2.28.1<br>geopandas==0.12.1<br>pandas==1.5.1<br>numpy==1.23.4<br>matplotlib==3.6.2<br>dash==2.7.0<br>dash-core-components==2.0.0<br>dash-html-components==2.0.0<br>dash-table==5.0.0<br>plotly==5.11.0<br>geopy==2.3.0<br>seaborn==0.12.1<br>patsy==0.5.3 |
| Note. The file 'data/tl_2020_06_puma10/tl_2020_06_puma10.shp' is in a folder called 'data/tl_2020_06_puma10' with other files. However, you can not only download the single file because the files in that folder are interdependent. | |

After installing the libraries and files, the full version of code could be run by using the command `python code/main.py`.

However, for the purpose of making the running procedure more efficient. The separate parts of codes are also offered. They are in the folder of 'simple version of codes'. One of the folders is 'data_cleaning_and_collection.py', which takes a long time and can be skipped for efficiency purposes. And the other is 'data_cleaning_and_collection.py', which takes less time. The relevant running information is listed in table 2.

*Table 2. Relevant information about how to run sub-code files*

| Choices | If only want to replicate the data collection/cleaning process | If only want to replicate the data analysis/visualization process |
|---|---|---|
| File to prepare before running | 'data/FO-Address-Open-Close-Times.xlsx'<br>'data/ZIP_Code_Population_Weighted_Centroids.csv' | 'data/final_data_set_1.csv'<br>'data/tl_2020_06_puma10/tl_2020_06_puma10.shp'<br>'data/zipcenter.csv' |

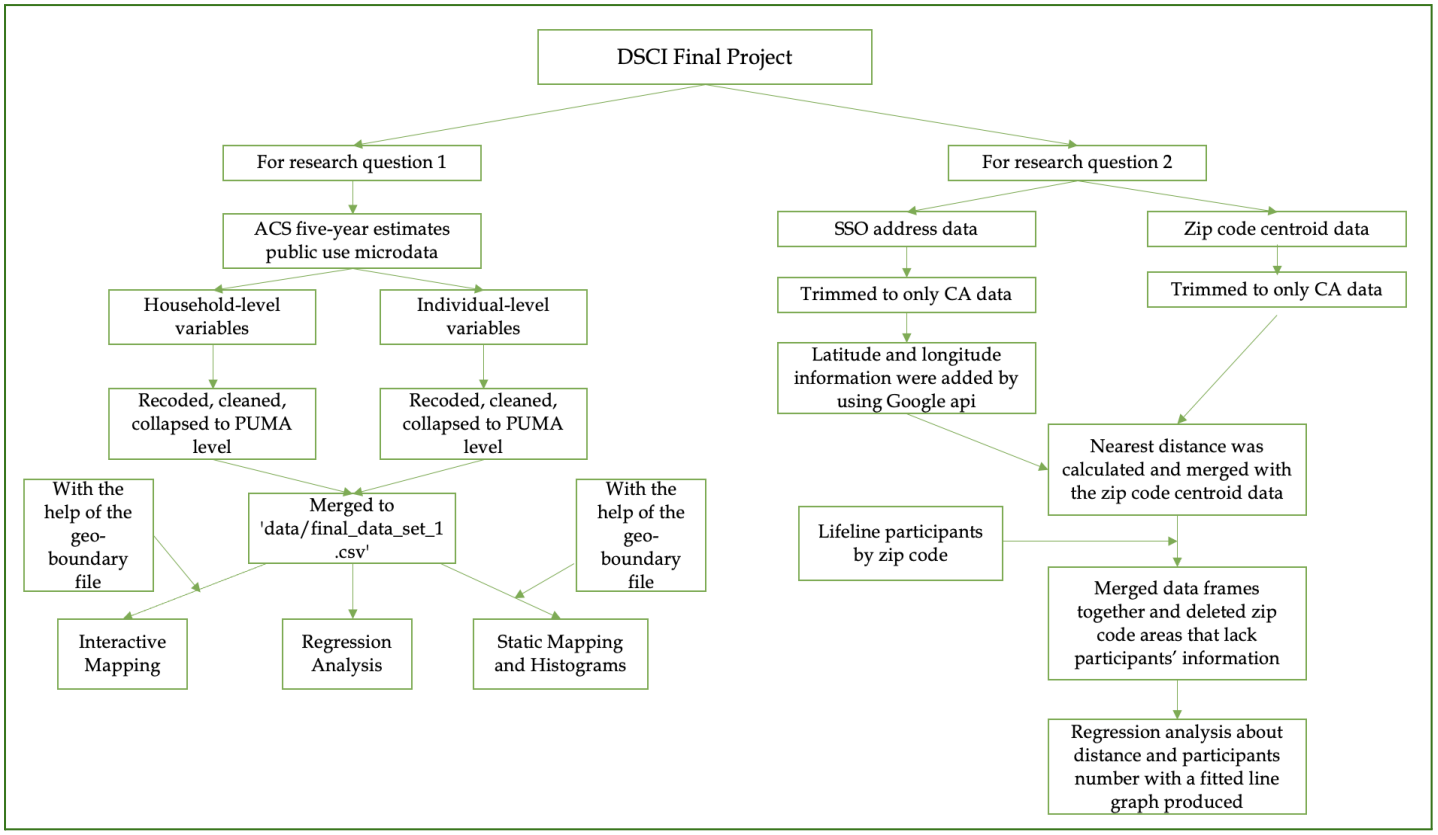
| | | |
|---|---|---|
| | | `'data/foinfo.csv'`<br><br>`'data/Lifelineparticipants.csv'` |
| Command line | `python code/separatecodes/cleaningandcollection.py` | `python code/separatecodes/visulizationandanalysis.py` |
| Expected final Output (intermediary output not included) | `'data/final_data_set_1.csv'`<br><br>`'data/foinfo.csv'`<br><br>`'data/zipcenter.csv'` | Three histograms,<br>Three static maps,<br>Four regression outputs,<br>two sets of results of VIFs,<br>an interactive map |
| Note 1. For both files, please also install the dependencies mentioned above (also in requirements.txt).<br>Note 2. The needed file will all be included in the data folder submitted. | | |

***Data Collection***

In order to answer the above-mentioned research questions, data from multiple sources have been collected. For the first research question, I have collected the 2020 release of American Community Survey (ACS) public use microdata. The data is collected through the census microdata api. The originally collected data include both household and individual level variables. Different levels of data have been stored into different files. Both of the initial data files contain around 1 million observations. After completing the initial data collection, the data has been recoded into different variables of interest. After that, both individual and household level data have been collapsed into Public Use Microdata Area (PUMA) level. And after collapsing to PUMA level, the datafiles have been merged into a single one for further analysis. And the PUMA level dataset contains 264 observations (i.e. PUMAs, See Appendix A for the final variables and how they are coded).

In order to give an initial exploration about the second research question, I use a zip code area's (zip code below) distance to the nearest Social Security Office (SSO) as a proxy for people's chance of having seen the publicity of CPUC. Therefore, the population weighted zip code centroid data has been collected from the U.S. Department of Housing and Urban Development (U.S. HUD) website and the SSOs' address data has been collected from the U.S. Social Security website. Since the original address data does not contain the geolocation information, Google Geocoding api is used for getting the longitude and latitude of each SSO. After that, the distance between each zip code centroid and the nearest SSO has been calculated with the help of `geo.py`. A simple data collection and analysis workflow has been included in Figure 1 for illustration purposes (see Appendix B for data source links).

*Figure 1. Data Collection and Analysis Workflow*



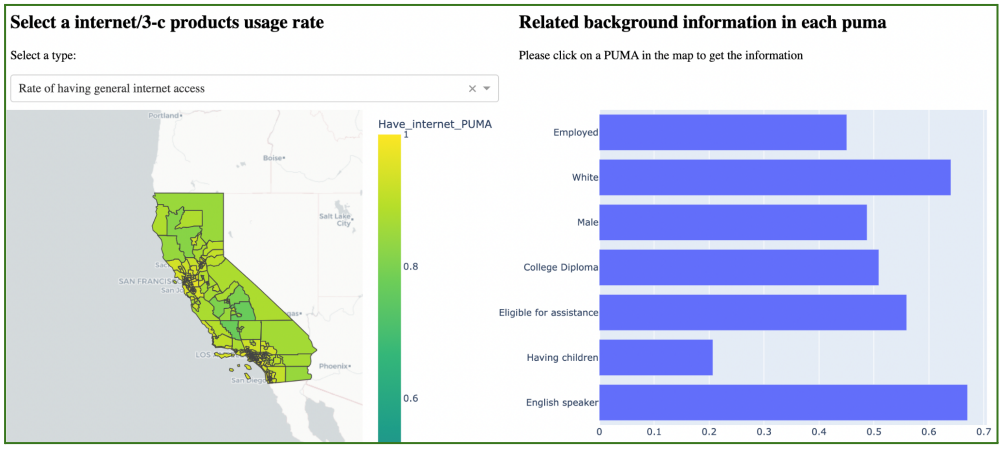

There's change about the data collection process compared to my initial plan. First, instead of collecting the 2017-2020 ACS data, only 2020 ACS data has been collected. Second, instead of collapsing data to zip code level, the data has been collapsed to the PUMA level, which is larger and leaves fewer amounts of observations. The plan changes are linked to some challenges I encountered during the data collection process. To begin with, the data structure of ACS public use microdata makes it meaningless to collect multiple years' data. For example, the 2020 dataset includes the data the U.S. Census Bureau collects from 2016-2020. Therefore, though there is a separate data file released for each year, many of the observations are overlapped, which makes it hard to see the trajectory of any variable. Moreover, after consulting the staff from the U.S. Census Bureau on a public slack channel, I realize that for data release before 2020, the microdata api may suffer from distinguishing individual and household, which makes it even harder to make use of pre-2020 data releases. Also, in order to protect observations' privacy, it is not able to get the information related to which zip code a person or household locates in. Though there are some crosswalks offered online to help transform PUMA level data to zip code level, how to use them is still in concern. For example, a crossfire may tell people that 30% of the households in one PUMA are in a certain zip code, it is still hard for you to identify which part of

households are in that zip code and what are the corresponding household characteristics. Except for the above mentioned challenges for research question 1, some challenges also emegres during the process of preparing the data to answer the second research question. Some of the zip codes that exist in the participants information file do not exist in the zip code centroid file, and there is no information about certain zip codes in the participants information file. Since I have no information about what the missing values mean (i.e. real missing or standing for 0) and why the zip codes do not perfectly match, I have to delete the zip codes that do not match or contain missing values.

*Analysis*

**For research Question 1** First, in order to better show the variable information (such as having high speed internet or not) of each PUMA for research question 1, I made an interactive map. A static presentation about the interactive map can be found in Figure 2 (and a gif version can be found in the Results folder). The left hand side map shows the rate of a PUMA's household level internet, high-speed internet, or 3-C products. By using the drop down menu, you can change the variable of interest. The right hand side includes two panels of horizontal bar charts (the second is not included in Figure 2 since the size), which plots some basic background information about each puma. The first panel shows "rate" related information and the other one shows money related information. By clicking the PUMA areas on map, people can change which PUMA's background information they want to be presented.

*Figure 2. Interactive map about internet access/background information related variables*




Second, a series of static figures are also produced by the codes. The first set of figures is about the distribution of the internet/high-speed internet/3-C products possession rates in different PUMAs. As can be seen from the figures, for both the three variables of interest, the

histogram shows a relatively left-tailed distribution, which indicates that most of the PUMAs have relatively high rates. However, it could be seen that compared to having internet/3-C products, the distribution of the rates of possession of high-speed internet is relatively scattered. This means that efforts still need to be made for promoting high-speed internet possession.

*Figure 3. Histogram of the rate of having internet of PUMAs*

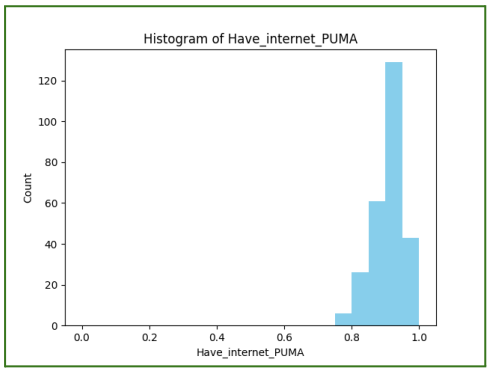



*Figure 4. Histogram of the rate of having 3-C products of PUMAs*

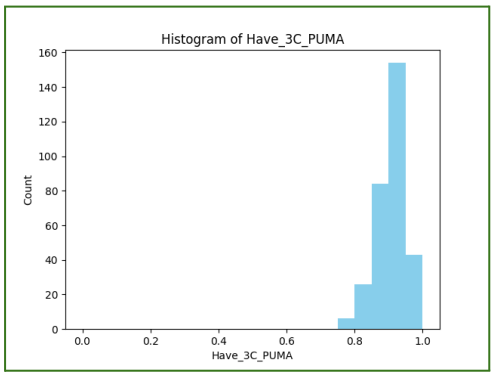
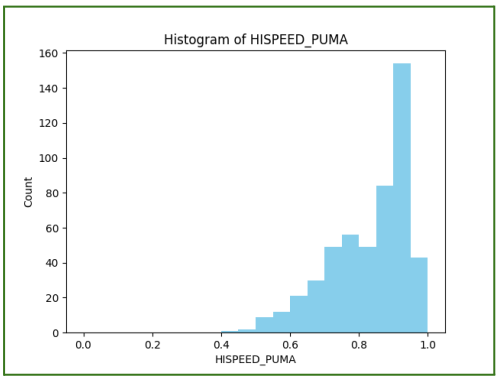



*Figure 5. Histogram of the rate of having high-speed internet of PUMAs*




The other set of figures shows the distribution of the above-mentioned rates in different PUMAs in the format of maps. The y-axis and x-axis are the latitude and longitudes. And the legend shows the corresponding color of each rate. As can be seen from the figure, there

is geographic divergence in terms of access. Generally speaking, near coast PUMAs are generally better in terms of internet/3-C products/high-speed internet access. Such figures can clearly show which PUMAs should be the focus for promoting access.

*Figure 6. Map of the rate of having internet of PUMAs*

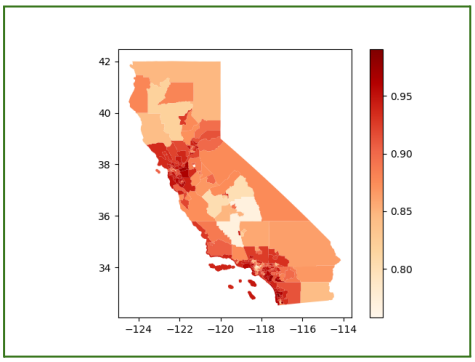



*Figure 7. Map of the rate of having high-speed internet of PUMAs*

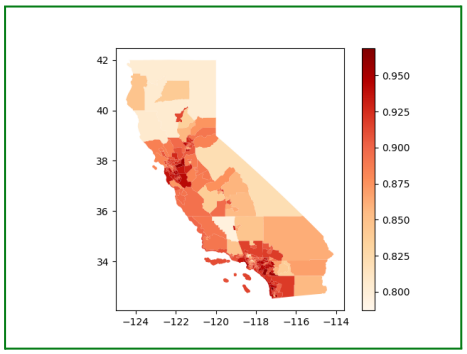
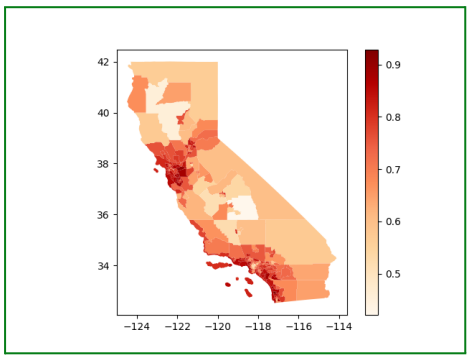



*Figure 8. Map of the rate of having internet of PUMAs*




Third, to answer research question 1 in a deeper way, I run a set of ordinary least squares (OLS) regressions which use having internet/having 3-C products/having high speed internet as the dependent variables and a set of PUMA level household characteristics as independent variables. As the above figures show, the high-speed internet access gap is more severe than the other two, I will focus on interpreting high-speed internet related

regression outputs here. However, all the regression outputs are listed in the results folder. As Table 1 in appendix C shows, in the first round of regression output, the statsmodel gives out a reminder of the existence of multicollinearity. Therefore, I run codes to get the variance inflation factors (VIFs) to see which variables may be affected by the problem. Amone the VIFs, four of them crossed the commonly used rule of thumb (i.e. 5, Studenmund & Johnson, 2016). Since three of them are proxies of PUMA's economic development and the rate of employed people can also reflect this, I dropped those three variables (VALP_PUMA, HINCP_PUMA, Eligible_PUMA). The results of the second of regression outputs and VIF calculation are listed in Table 3 and 4. As can be seen from the figures, actually, the multicollinearity reminder has not been successfully eliminated. And in reality, the reminder can be dropped if I further drop the Aged_PUMA variable (the average age). However, the current included variables all have their theoretical meanings and after dropping the three variables, the remaining factors all have a VIF below five. Moreover, the major harm brought by multicollinearity is increasing the possibility of making type II error and the type II error is not so worrisome in our context of analysis. Therefore, based on the above-mentioned reasons, this second round of regression output is considered as the final version.

As indicated by the regression output, all the predictors are statistically significant predictors at at least 5 percent significance level. Among them, the percentage of English speakers (HHL_PUMA), percentage of college diploma holders (SCHL_PUMA), and percentage of employed people (WRK_PUMA) have positive estimated coefficients. This means that if a PUMA has more English speakers, college diploma holders, and workers, compared to the non-English speakers, non-holders, non-workers, the PUMA tends to have a higher high-speed internet access rate. And average age (Aged_PUMA) and percentage of males (SEX_PUMA) are negative predictors, which means if a PUMA has more aged people and male (compared to their young and female counterparts), the PUMA will have lower high-speed internet access rate. The shown results generally correspond to the resources and appropriations theory since young, working, educated and English-speaking people tend to have better social and economic resources (Van Dajk, 2019). However, the percentage of male variables generates an expected sign. It is possible that such a result is generated because of specific contextual reasons, which deserves further exploration.

*Table 3. Regression output for giving high-speed internet access*

```
                            OLS Regression Results
==============================================================================
Dep. Variable:           HISPEED_PUMA   R-squared:                       0.590
Model:                            OLS   Adj. R-squared:                  0.582
Method:                 Least Squares   F-statistic:                     74.50
Date:                Tue, 13 Dec 2022   Prob (F-statistic):           3.85e-48
Time:                        22:19:37   Log-Likelihood:                360.76
No. Observations:                 265   AIC:                           -709.5
Df Residuals:                     259   BIC:                           -688.0
Df Model:                           5
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          1.0227      0.160      6.390      0.000       0.708       1.338
HHL_PUMA       0.0719      0.025      2.876      0.004       0.023       0.121
AGEP_PUMA     -0.0036      0.002     -2.033      0.043      -0.007      -0.000
SCHL_PUMA      0.4396      0.053      8.219      0.000       0.334       0.545
SEX_PUMA      -0.8565      0.271     -3.155      0.002      -1.391      -0.322
WRK_PUMA       0.3215      0.119      2.711      0.007       0.088       0.555
==============================================================================
Omnibus:                       13.115   Durbin-Watson:                   1.164
Prob(Omnibus):                  0.001   Jarque-Bera (JB):               13.770
Skew:                          -0.552   Prob(JB):                      0.00102
Kurtosis:                       3.173   Cond. No.                      3.04e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.04e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

*Table 4. VIFs of the included variables*

| | VIF | Variables |
|---|---|---|
| 0 | 1725.098066 | Intercept |
| 1 | 1.427323 | HHL_PUMA |
| 2 | 1.949243 | AGEP_PUMA |
| 3 | 3.625159 | SCHL_PUMA |
| 4 | 1.037491 | SEX_PUMA |
| 5 | 2.617522 | WRK_PUMA |

**For research Question 2** In order to answer question 2, a regression analysis has also been conducted. As can be seen from the output, the distance to the nearest social security office (SSO) is negatively associated with the Lifeline participants in a zip code. Controlling other factors as constant, a one km increase in the distance is associated with a 12 people decrease in terms of lifeline participation. This one-variable model explains 7.9% of the variance in lifeline participants. However, if we take a look at Figure 9, which is a scatter plot with the fitted regression line, there is actually no clear linear pattern. Therefore, we should be

cautious about the results. Further exploration is needed to see whether there is really an association or rather than a fitted line out of randomness or bias.

*Table 4. Regression output for the relationship between distance and Lifeline participants*

OLS Regression Results

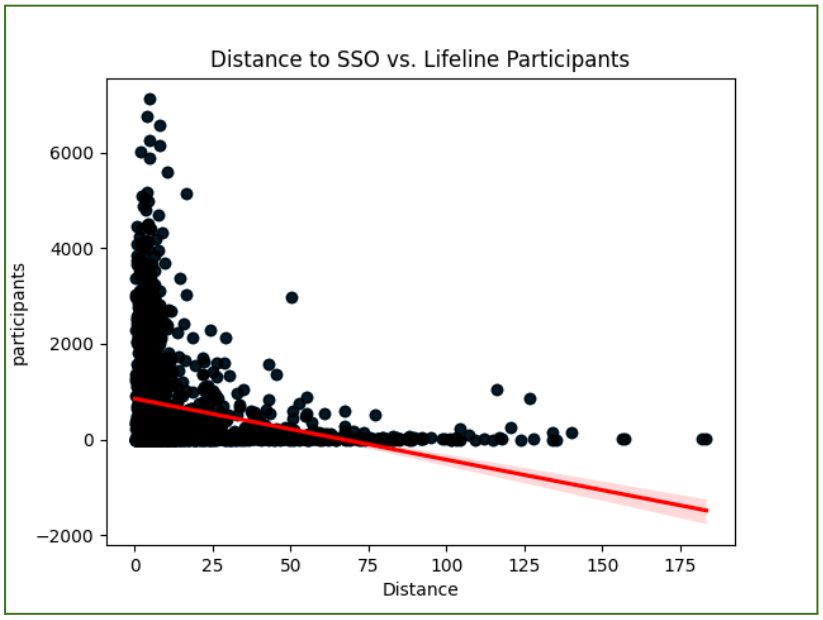
| | | | |
|---|---|---|---|
| Dep. Variable: | participants | R-squared: | 0.080 |
| Model: | OLS | Adj. R-squared: | 0.079 |
| Method: | Least Squares | F-statistic: | 169.0 |
| Date: | Tue, 13 Dec 2022 | Prob (F-statistic): | 4.15e-37 |
| Time: | 22:19:35 | Log-Likelihood: | -16216. |
| No. Observations: | 1954 | AIC: | 3.244e+04 |
| Df Residuals: | 1952 | BIC: | 3.245e+04 |
| Df Model: | 1 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| Intercept | 854.2339 | 27.900 | 30.618 | 0.000 | 799.517 | 908.951 |
| Distance | -12.7483 | 0.981 | -12.998 | 0.000 | -14.672 | -10.825 |

| | | | |
|---|---|---|---|
| Omnibus: | 963.711 | Durbin-Watson: | 1.492 |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 5436.285 |
| Skew: | 2.327 | Prob(JB): | 0.00 |
| Kurtosis: | 9.717 | Cond. No. | 36.1 |

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

*Figure 9. Scatter plot to show the relationship between distance vas lifeline participants with fitted line*




**Contributions** The present analysis and data visualization has its contributions. First, by running regressions to test question 1, this analysis makes its contribution to the empirical test of the resources appropriation theory, which can be the building block of future and

more rigorous tests of the theory. Moreover, the information provided by the regression output for the first research question can provide some information to practitioners in the telecommunication policy area about who are getting internet access/3-C products in California and who are not, which could help their daily work. Second, the current analysis makes a first step to test whether the publicity efforts of CPUC staff makes a positive contribution to Lifeline participation. Though the analysis is still far from precise, it at least provides some potential logics which may help further exploration. Third, the interactive map provided by this project can help people get the information about internet access and PUMA level social and economic related basic information in a quick, eye-catching, and interesting way. Such data visualization could be helpful for policy analysts' presentation, especially when they are presenting to the general public. Last but not least, the codes produced by the current project can be seen as a simple one-stop code shop to show telecommunication policy people how to use api, interactive data visualization, and run regression analysis in python. This may be helpful since at least USC wide, the common statistics tool for policy people is still Stata, which especially sucks in the visualization part.

**Changes and challenges** Firstly, the first research question has been changed compared to the initial plan. Initially, the first research question is initially designed to explore what factors are influencing Lifeline participation numbers in each zip code. Secondly, The analysis of the second research question has also been changed from the initial plan. Initially, the related variables grabbed from ACS will be used as control variables to better estimate the potential association with Lifeline participants. However, we are only estimating a single-variable regression now. Thirdly, the visualization plan has been changed. Initially, I would like to plot the trajectory of internet access in different PUMAs across years. However, currently, only the 2020 situation is shown. Such changes in plan are related to some challenges. To begin with, as I mentioned before, the Lifeline participation data provided by CPUC staff is on the zip code level. However, our influential factors grabbed from ACS are on the PUMA level. Such two levels are hard to merge with each other. Therefore, the research questions have to be changed and the ACS variables can not be used as controls for the second research question. Moreover, also as mentioned above, using different years' releases of ACS data will not really help to plot the trajectory since the structure of the data.

***Limitation and Future Research***

Though significant efforts have been put into formulating the project, it still bears several limitations. First, though the interactive map has achieved some basic interactive functionality, the dropdown menu still runs pretty slow. It may take more than ten seconds to load the changed map. Moreover, the visual effect of the interactive map is still not optimal. Second, there's still room for the static visualization to be better. For example, the figures can be combined with each other as a panel to save time for checking them one by one. Second, for the regression output for question 1, though I would like to explore the "influence" of the included variables, actually, by simply running OLS regression, I can only claim association between variables rather than causality. Moreover, I dropped some variables because of the multicollinearity concern. However, it should be acknowledged that after dropping variables, the sign of the coefficients of SCHL_PUMA and HHL_PUMA change. This may be the sign of the omitted variable bias, which should be noticed and acknowledged. Third, for the second research question, what I'm currently testing is the "intention to treatment effect" rather than the "treatment on the treated effect". This means that though the distance is used as a proxy for receiving the CPUC publicity, people are not guaranteed to see this. Therefore, in this way, I'm not targeting the real influence of the publicity itself. Moreover, similarly to the first research question, though I would like to test the effect of the CPUC publicity efforts, what I'm getting now is still the association rather than causal relationship. Also, for the second research question, the variable of participants is directly used in the regression. However, there's a chance that residents in each zip code are not equal. Therefore, it may be better to use the percentage of participants among the whole population rather than the number itself.

Considering the limitations of the current analysis, I will do the following is having more time:

- Try to decrease the response time of the interactive map and use html/css related codes to enhance the visual appeals. Similarly, enhancing the quality of the static figures as well.
- Enhance the quality of the regression analysis by including more regressors, thinking more about function form, and better resolving issues like heteroskedasticity, multicollinearity, and omitted variable bias.
- Try to find a better way to quantify the publicity efforts for CPUC staff.
- Try to explore whether more advanced statistical models, such as regression discontinuity or instrumental variables, can be used to answer research questions and how they can be applied into the current research context.

- Try to explore how more advanced machine learning tools in the data science domain can better help to answer the current research questions. Such techniques include but are not limited to classification, decision tree, and Multi Layer Perceptron (basically, the supervised learning tools).
- Try to better modularize the codes and make it easier and clearer to be run by others.

Appendix A: Variables got from ACS data

| Variable Name | Variable Level and observation numbers | Variable Meaning | Coding Process |
|---|---|---|---|
| VALP_PUMA (in decimal) | On PUMA level; calculated from | The average property value for households | This is coded from the household level |

| | | | |
|---|---|---|---|
| | household level data | in each PUMA. | property value by taking average within each PUMA. |
| HHL_PUMA (in decimal) | On PUMA level; calculated from household level data | The percentage of English speaking households in a PUMA | Code all the non-English household language to 0 to make a binary variable on the household level, and then take average (since it is binary, it is equivalent to a percentage) |
| HINCP_PUMA | On PUMA level; calculated from household level data | Household year income | Code from household data by taking average. On the household level, negative values are transformed to 0 |
| Eligible_PUMA (in decimal) | On PUMA level; calculated from household level and person level data | The percentage of households that are eligible for several federal assistance programs (such as Medicare). This is also indicative for California Lifeline eligibility as a lower bound. | Several household and personal level federal assistance receiving status are used for coding this variable. Before collapsing to the PUMA level, the household is seen as an eligible household if the household or its members are receiving at least one assistance program. |
| AGEP_PUMA | On PUMA level; calculated from Personal level and person level data | The average age of people in a puma | This is coded from the person level data by taking average with PUMA. |
| SCHL_PUMA (in decimal) | On PUMA level; calculated from Personal level and person level data | The percentage of people having a education level on or beyond college | The original personal level education attainment variable is coded as a binary one (0=below college) After that, the average is taken. |
| SEX_PUMA (in decimal) | On PUMA level; calculated from Personal level and person level data | The percentage of males in a PUMA | Original data is coded as binary and then taking average PUMA wide |
| WRK_PUMA (in decimal) | On PUMA level; calculated from Personal level and | The percentage of people in a PUMA who has worked last | Original data is coded as binary and then taking average PUMA |

| | | | |
|---|---|---|---|
| | person level data | week (considered as employed) | wide |
| Have_internet_PUMA (in decimal) | On PUMA level; calculated from household level data | Percentage of household that having any kind of internet access | Original household level variable coded from several related variables (such as whether a household has satellite internet). After that, the average is taken. |
| Have_3C_PUMA (in decimal) | On PUMA level; calculated from household level data | Percentage of household that having smartphones or tablet | Original household level variable coded from several related variables (such as whether a household has smartphones). After that, the average is taken. |
| HISPEED_PUMA (in decimal) | On PUMA level; calculated from household level data | Percentage of household that having high speed internet access (not including cellular one) | Transform household level variable to a binary one and then take average |

Note 1. PUMA refers to public use microdata areas. It is designed by the U.S. Census Bureau for the purpose of data collection. PUMA also has a clear geographic boundary, which means it can be used for mapping. PUMA classification changes every 10 years. PUMA classification depends on many factors. Take an example here, Los Angeles County has been divided into more than 60 PUMAs.
Note 2. There are 264 observations (PUMAs) for each variable. No missing data exist on the PUMA level.
Note 3. Unless otherwise noted, household weight/personal weighting is considered during the coding process (weight is for generating population wide rather than sample wide estimation).
Note 4. Unless otherwise noted, all the missing values are treated as 0 originally by the U.S. Census Bureau.
Note 5. The general logic of data transformation is: transform data on the household or personal level→collapse to PUMA level by taking average

(US Census Bureau, 2022)

Appendix B: Data source information

| Research Question | Data Link |
|---|---|
| Research Question 1 | ACS microdata api<br>https://api.census.gov/data/2020/acs/acs5/pums?<br><br>ACS mapping used geographic boundary file<br>https://www2.census.gov/geo/tiger/TIGER2020/PUMA/ |
| Research Question 2 | Lifeline Participants Number<br>https://github.com/TioHK/final_project_dsci510/blob/master/data/Lifelineparticipants.csv<br><br>Social Security Office address:<br>https://www.ssa.gov/open/data/FO-RS-Address-Open-Close-Time-App-Devs.html<br><br>Google Geocoding api<br>https://developers.google.com/maps/documentation/geocoding/requests-geocoding<br><br>Centroid Information<br>https://hudgis-hud.opendata.arcgis.com/datasets/HUD::zip-code-population-weighted-centroids/explore?location=35.546609%2C-120.006125%2C3.59 |
| Note. For data files that are publicly available, the links to download them are listed. For non-public use data, the student's personal github respiratory link is listed. However, it should be noted that all the non-api files can be downloaded from the student's personal github respiratory. | |

Appendix C: First Round of regression outputs

*Table 1. Regression output for high-speed internet access*

```
                            OLS Regression Results
==============================================================================
Dep. Variable:           HISPEED_PUMA   R-squared:                       0.778
Model:                            OLS   Adj. R-squared:                  0.771
Method:                 Least Squares   F-statistic:                     112.2
Date:                Tue, 13 Dec 2022   Prob (F-statistic):           3.57e-79
Time:                        22:19:37   Log-Likelihood:                442.14
No. Observations:                 265   AIC:                            -866.3
Df Residuals:                     256   BIC:                            -834.1
Df Model:                           8
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          1.7027      0.146     11.684      0.000       1.416       1.990
VALP_PUMA  -5.442e-08   4.04e-08     -1.346      0.179   -1.34e-07    2.52e-08
HHL_PUMA      -0.0842      0.024     -3.462      0.001      -0.132      -0.036
HINCP_PUMA  1.085e-06   3.43e-07      3.166      0.002     4.1e-07    1.76e-06
Eligible_PUMA -0.1859      0.018    -10.298      0.000      -0.221      -0.150
AGEP_PUMA     -0.0092      0.002     -5.552      0.000      -0.012      -0.006
SCHL_PUMA     -0.0512      0.060     -0.847      0.398      -0.170       0.068
SEX_PUMA      -0.9700      0.204     -4.756      0.000      -1.372      -0.568
WRK_PUMA       0.0336      0.100      0.336      0.737      -0.164       0.231
==============================================================================
Omnibus:                       19.159   Durbin-Watson:                   1.598
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               32.632
Skew:                          -0.429   Prob(JB):                     8.20e-08
Kurtosis:                       4.490   Cond. No.                     3.72e+07
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.72e+07. This might indicate that there are
strong multicollinearity or other numerical problems.
```

*Table 2. The VIFs of the included independent variables*

| | VIF | Variables |
|---|---|---|
| 0 | 2612.282221 | Intercept |
| 1 | 11.477610 | VALP_PUMA |
| 2 | 2.472116 | HHL_PUMA |
| 3 | 19.968830 | HINCP_PUMA |
| 4 | 9.125986 | Eligible_PUMA |
| 5 | 3.147443 | AGEP_PUMA |
| 6 | 8.449939 | SCHL_PUMA |
| 7 | 1.069548 | SEX_PUMA |
| 8 | 3.409657 | WRK_PUMA |

## References

Burton, M., Macher, J., & Mayo, J. W. (2007). Understanding participation in social programs:

Why don't households pick up the Lifeline? *The B.E. Journal of Economic Analysis & Policy*, *7*(1). https://doi.org/10.2202/1935-1682.1583

Studenmund, A. H. (2016). *Using econometrics: A practical guide*. Pearson.

USAC. (2022, September 26). Lifeline. Universal Service Administrative Company. Retrieved November 18, 2022, from https://www.usac.org/lifeline/

US Census Bureau. (2022). Public use microdata areas (pumas). Census.gov. Retrieved December 9, 2022, from https://www.census.gov/programs-surveys/geography/guidance/geo-areas/pumas.html