Semester Project:

Exploring the Impact of COVID-19 on CA WIC Dataset: Analysis and Visualization

Yijin Zhu

San José State University

INFM 203

Prateek Jain

May 7, 2023

Exploring the Impact of COVID-19 on CA WIC Dataset: Analysis and Visualization

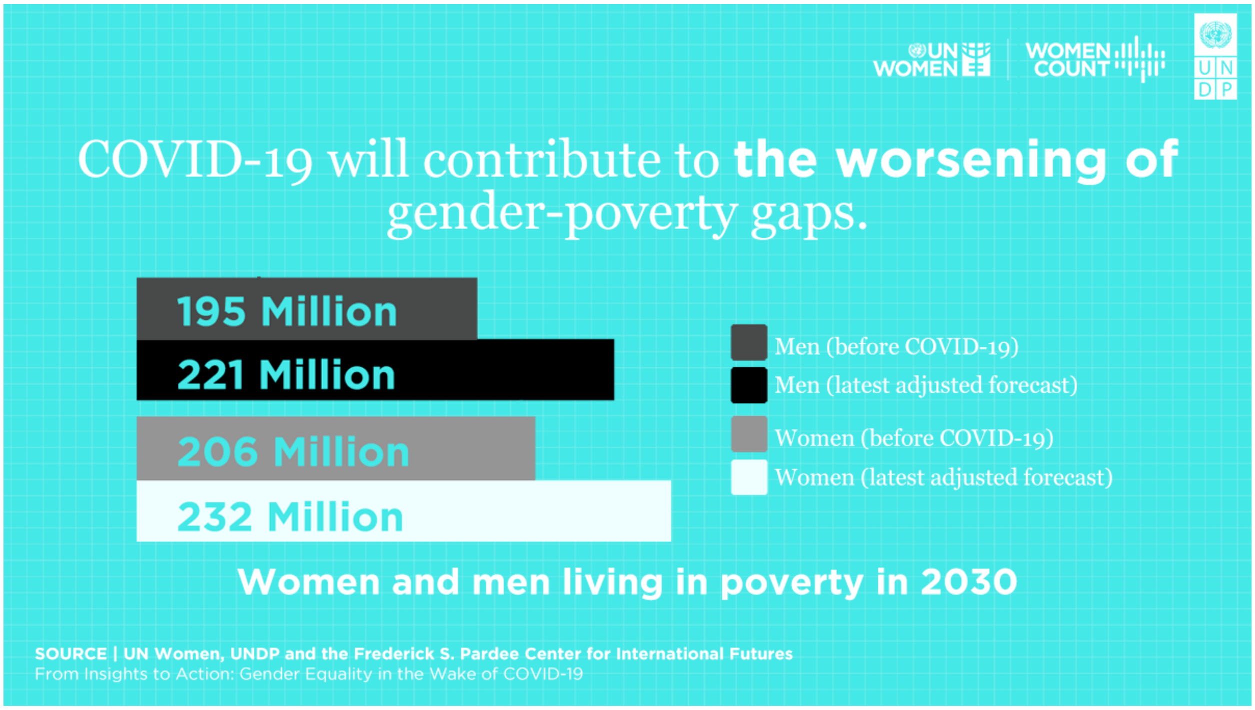
Project Motivation

Nowadays, as information and communication technologies continue to develop rapidly, big data analysis has achieved unprecedented growth. Big data is being widely used due to its applications in analysis and data-driven decision-making (Al-Barashdi & Al-Karousi, 2019). In the field of social welfare, big data can be a valuable asset, offering new and exciting opportunities for organizations of all sizes to improve and innovate their effort for helping the communities. Organizations can utilize multi-method analysis and big data interpretation to better understand their target audience's needs. This information can be used to improve and advance social services and develop interventions and programs that are more effective at meeting the needs of the community.

The economic crisis triggered by the COVID-19 pandemic has impacted the entire globe. While everyone is facing expected challenges, UN WOMEN (2020) points out that women are suffering the brunt of the economic and social fallout of COVID-19. Women who are poor and marginalized face an even higher risk of COVID-19 transmission and fatalities, loss of livelihood, and increased violence. The pandemic is excepted to widen the gender poverty gap in the long term (Figure 1). Statistics Specialists from UN WOMEN (2020) also mentioned that more women will be pushed into extreme poverty than men, especially the case among those aged 25 to 34, at the height of their productive and family formation period. Governments and social welfare organizations must take proactive steps to mitigate the negative economic impacts of COVID-19 on financially struggling women and their families.

Figure 1

Forecast of Women and Men Living in Poverty in 2030 (UN WOMEN, 2020)



SOURCE | UN Women, UNDP and the Frederick S. Pardee Center for International Futures
From Insights to Action: Gender Equality in the Wake of COVID-19

Problem Statement

The Women, Infants and Children (WIC) Program (Figure 2) is a federally funded health and nutrition program that provides assistance to pregnant women, new mothers, infants and children under age five. WIC helps California families by providing food benefits to individual participants, which can be applied to purchase healthy supplemental foods from about 4,000 WIC-authorized vendor stores throughout the State. WIC also provides nutritional education, breastfeeding support, healthcare referrals, and other community services. Participants must meet income guidelines and other criteria. According to the information on ca.gov, currently, 84 WIC agencies provide services monthly to approximately one million participants at over 500 sites in local communities throughout the State.

Figure 2

California Department of Public Health (CDPH) and Women, Infants and Children (WIC) Logos (ca.gov, 2022)

To effectively analyze the impact of the COVID-19 pandemic on the WIC program in California, I can consider several variables, like redemption rates, and program participation. In this project, I will focus on utilizing big data analysis techniques to answer the following questions:

- How has WIC redemption changed in California over the past decade?
- Did the outbreak of COVID-19 affect WIC redemption statewide?
- If so, are there any similar trends in the data from big counties in CA?

Through the use of data analysis and visualization, I aim to gain valuable insights into the impact of the COVID-19 pandemic on the effectiveness of the WIC program. The results of big data analysis will help us examine the ability of the WIC program to support vulnerable populations during challenging times, so organizations and policymakers can be better informed for making decisions in the future.

Materials and Methodology

This project aims to analyze the impact of the COVID-19 pandemic on the WIC program by examining redemption datasets provided by the CA government. Multiple data analysis techniques, including Spark, Pandas, Matplotlib, and Seaborn will be utilized to gather, process, and represent data from the datasets to achieve this goal.

Data Collection

The datasets used in this project were obtained from the California Open Data Portal (https://data.ca.gov/). The primary dataset used in this project is from the title “California Women, Infants and Children Program Redemption by County”. This dataset provides information on the number of participants or families who redeemed their benefits and the corresponding dollar value of those redeemed benefits by the vendor county, starting from the calendar year 2010 and onwards. Also, one additional dataset, which contains the coordinates of all counties in California was used to support the analysis. These datasets were utilized to create a map to illustrate the analysis results and visualize the findings in a more straightforward but comprehensive way.

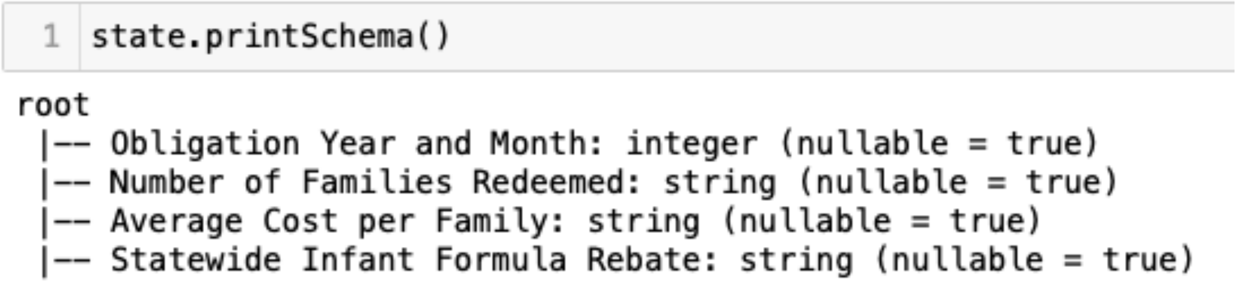
Data Cleaning and Preparation

In this section, I utilized Spark to read each CSV file into dataframes. Then, I used the filter function to split each dataframe into two parts: statewide data and county data. Following the filtering process, I aggregated each part of the data into two separate dataframes using the union function, one for the state and another for the counties. After performing a detailed examination of the dataframes, I decided to drop certain columns that contained redundant or non-informative data, such as Vendor County Code and Number of Food Instruments Redeemed.

I reconstructed the state dataframe by retaining only four columns: 'Obligation Year and Month', 'Number of Families Redeemed', 'Average Cost per Family', and 'Statewide Infant Formula Rebate'. This was done to simplify the dataframe and focus on the key variables that would be used for our analysis.

Figure 3

Example of state dataframe after reconstruction



```
state.printSchema()
```

```
root
 |-- Obligation Year and Month: integer (nullable = true)
 |-- Number of Families Redeemed: string (nullable = true)
 |-- Average Cost per Family: string (nullable = true)
 |-- Statewide Infant Formula Rebate: string (nullable = true)
```

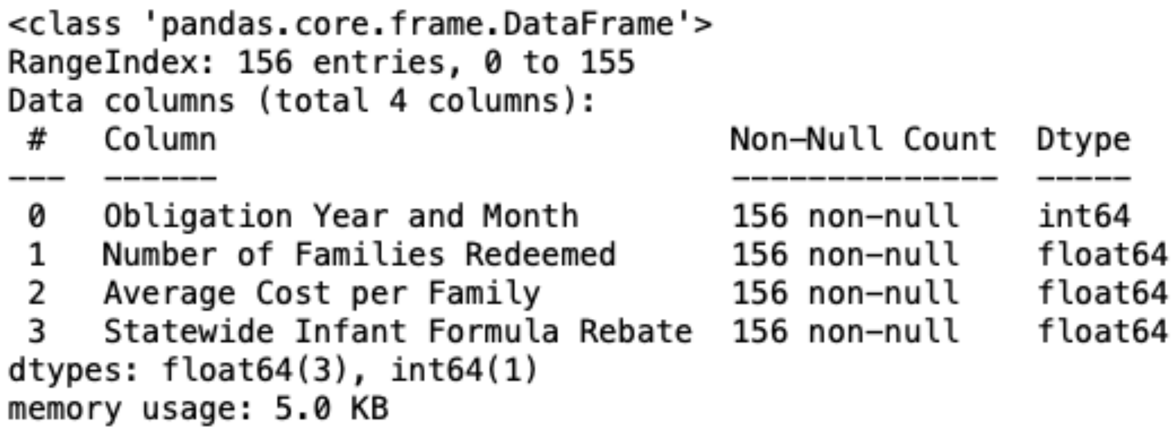
As the schema of state data shown in Figure 3, our dataframes still need some preparation before conducting any analysis or plotting, because:

- Some columns are currently stored as objects and need to be converted into numerical data types.
- I need to remove dollar signs and commas in certain columns before casting them into numerical types.
- The hyphen '-' is used to represent all missing values in the dataset, and I need to replace it with actual NaN values.

By performing these data-cleaning steps with the help of Pandas, such as converting data types and replacing missing values, I can ensure that our data is ready for analysis and visualization (Figure 4).

Figure 4

Example of state dataframe after data cleaning



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 156 entries, 0 to 155
Data columns (total 4 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   Obligation Year and Month       156 non-null    int64
 1   Number of Families Redeemed     156 non-null    float64
 2   Average Cost per Family         156 non-null    float64
 3   Statewide Infant Formula Rebate  156 non-null    float64
dtypes: float64(3), int64(1)
memory usage: 5.0 KB
```

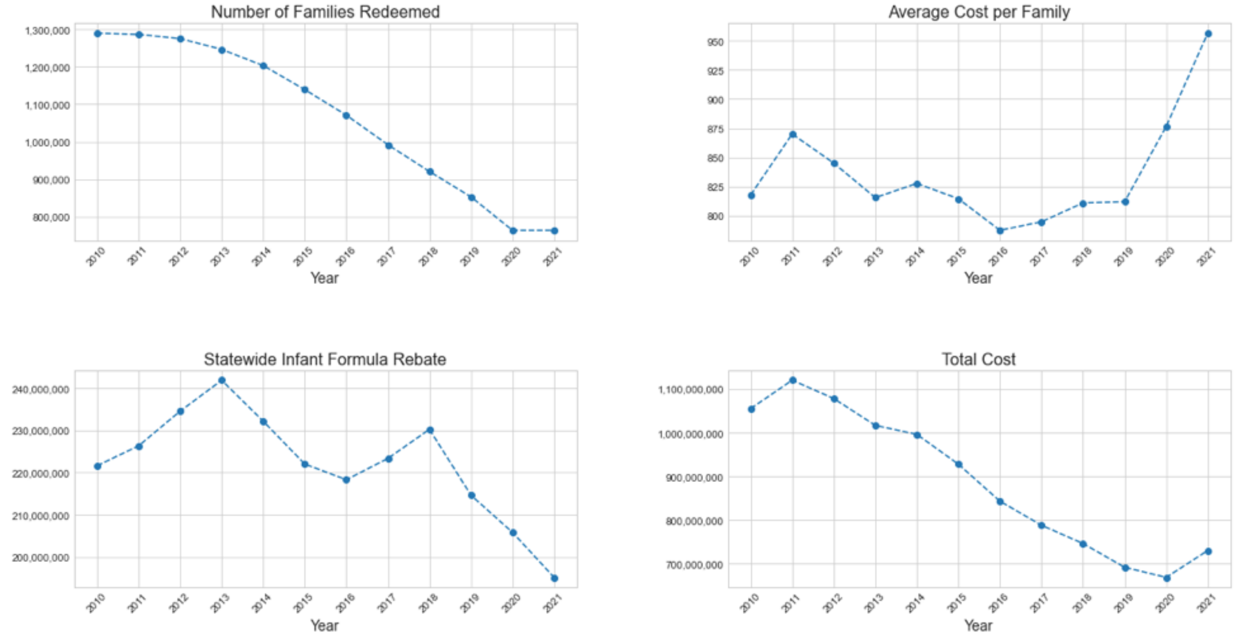
Data Analysis

In this section, I utilize Seaborn built on Matplotlib to analyze both state-level and county-level data. I begin by examining general redemption trends from 2010 to 2021 before creating a detailed redemption pattern from 2019 to 2021, which was a period that coincided with the breaking out of the COVID-19 pandemic and the subsequent statewide lockdown in California. The county-level dataset provides an opportunity to break down and integrate features that cannot be seen at the state level. In this part, I evaluate the relationships between various features, such as the number of participating families versus average cost, the number of redeemed families versus the COVID-19 timeline, and redemption versus location. The findings from our analysis are detailed in the following section.

Analysis results

Figure 5

CA WIC: General redemption trends from 2010-2021



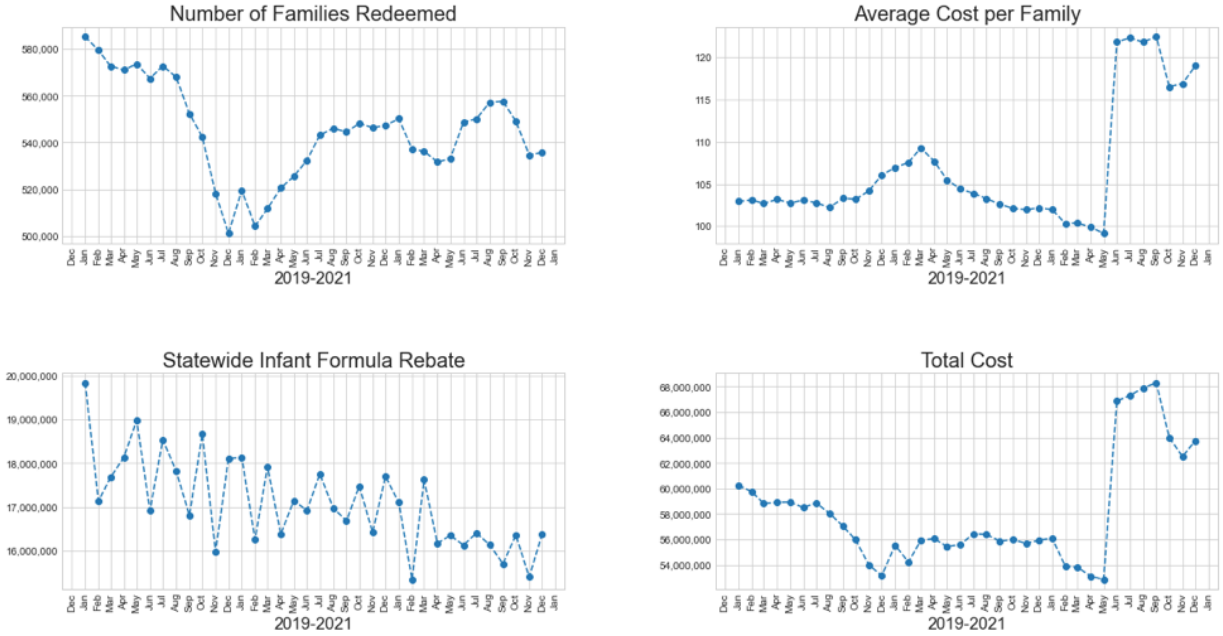
State general redemption trends from 2010 to 2021 (Figure 5)

- The Number of Families Redeemed has a steady and sharper decline in the years leading up to 2020. However, there was a sudden increase in redemption in 2021, which is different from the previous downward trend.
- Average Cost per Family has remained relatively stable with slight changes (< $50) from 2010 to 2019. However, in both 2020 and 2021, there was a sharp increase in the average cost, with a yearly speed of over 50 dollars. The increase in 2021 was particularly significant, with a speed of over 75 dollars per year.
- Statewide Infant Formula Rebate has been dropping since 2019. There is no sign of significant changes recently.

- Total Cost is also worth checking because it has been decreasing for 9 consecutive years since 2011. But after 2020, there was a significant increase of over 50 million.

Figure 6

CA WIC: Detailed redemption trends from 2019-2021



State detailed redemption trends from 2019 to 2021 (Figure 6)

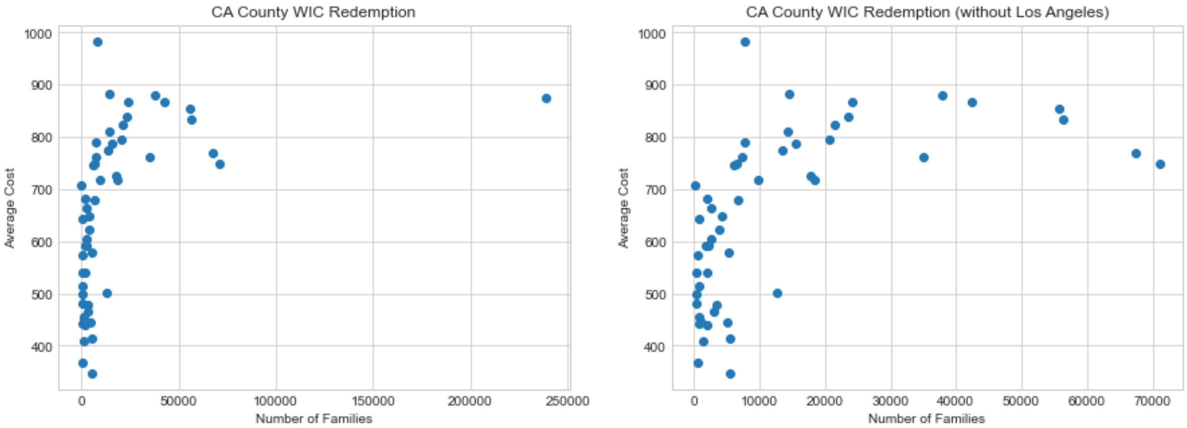
- The Number of Families Redeemed sharply turned from a decreasing to an increasing trend at the beginning of 2020, but started to decline again towards the end of 2021. However, despite the decline, the number of families redeemed at the end of 2021 was still higher than at the end of 2019, which explains the slight increase in the yearly plot from 2020 to 2021.
- The sudden increase in both Average Cost per Family and Total Cost by over 20% in May 2021, which remained stable for four months before slightly decreasing, is also reflected in

the yearly plots. It is notable that the increase in these two variables occurred at the same time, indicating a strong correlation between them.

- In contrast, the Statewide Infant Formula Rebate program has continued to decline throughout this period. It appears that the pandemic has had no significant impact on this program.

Figure 7

CA COUNTY WIC: Number of Families VS Average Cost (2021)



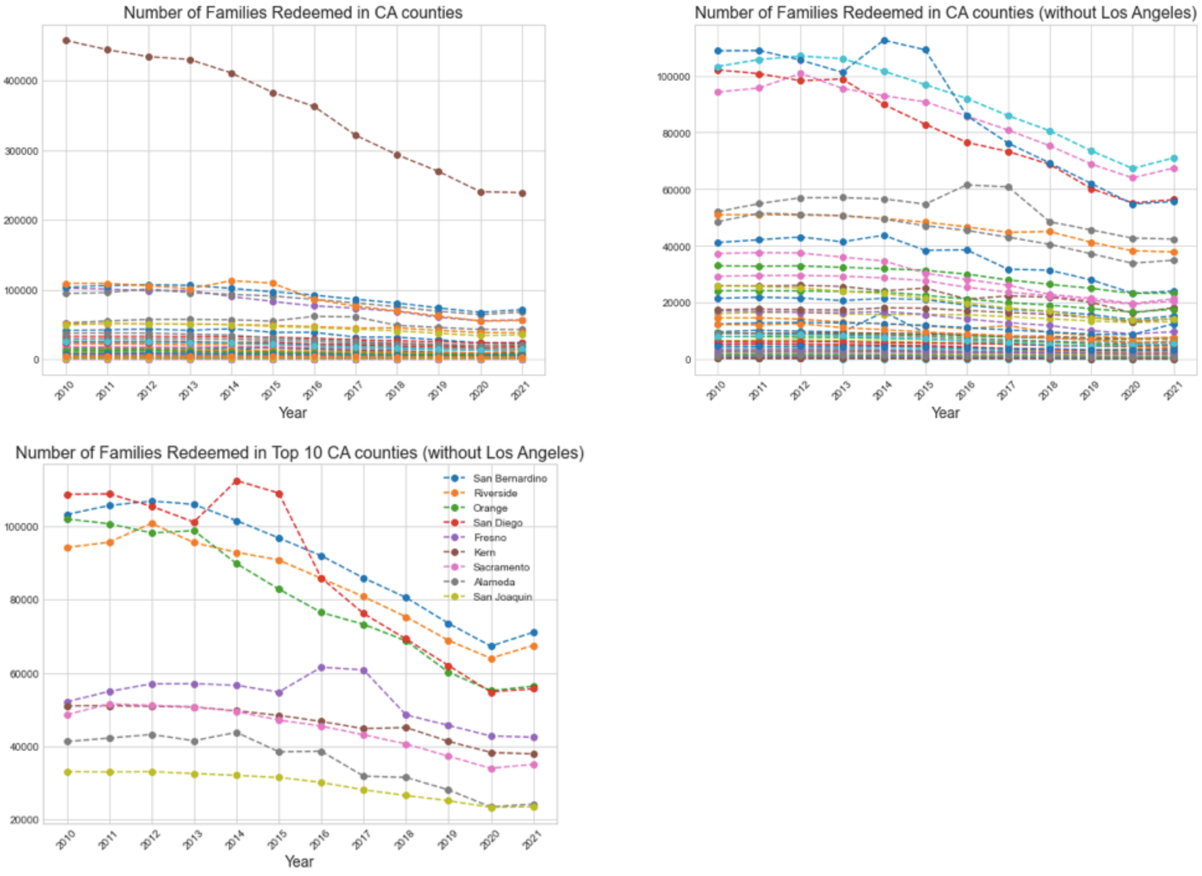
County Number of WIC Families versus Average Cost (2021) (Figure 7)

The scatter plot on the left shows a wide range of average redemption costs across different counties, but a large number of WIC participants in Los Angeles County appears as an outlier and makes it difficult to see the patterns among other counties. Therefore, we created a second scatter plot on the right that excludes Los Angeles County. From this plot, we can observe that most counties with over 10,000 participating families have an average redemption cost between 700 and 900 dollars. In contrast, for counties with less than 10,000 participating families, the majority of the average costs are below 700 dollars.

As a result, we observe a low correlation (0.44) between the number of participating families and the average redemption cost in the entire state. However, when we divide counties into two groups based on the number of participant families, i.e., counties with over 10,000 participant families and counties with less than 10,000 families, we notice a difference. The correlation in each group is 0.27 and 0.54, respectively. Therefore, in counties with less than 10,000 WIC families, we observe a moderate correlation between the number of families and the average redemption cost. In contrast, in counties with a larger number of WIC families, the correlation is even lower.

Figure 8

CA COUNTY WIC: Number of Families VS Time (2010-2021)



County Number of WIC Families versus Time (2010-2021) (Figure 8)

- The first graph indicates that LA County is an outlier, which skews the rest of the data and makes it difficult to observe patterns. However, the second graph provides a comprehensive view of how the number of participants has changed in each county, excluding LA.
- The second graph shows that although the number of participants is not decreasing every year, there is a general trend of declining numbers from 2010 to 2020 in most counties. However, this trend changed in 2021, with most counties experiencing either a flat slope or even a positive slope with an increase in the number of participants.
- The third graph highlights that in the top 2-10 counties with the most participants, eight out of nine counties experienced an increase in the number of WIC families in 2021, even San Diego County, which had the most rapid drop in numbers in the past. It is noteworthy that although the increases in numbers were not as significant as the drops in previous years, which explains the statewide data, the increase in 2021 was almost shown in a flat slope.

County Average Cost versus Time (2010-2021) (Figure 9)

- The left graph displays the changes in the average cost of WIC redemption across all counties in California over time. From 2010 to 2019, the majority of lines fall within the range of 400 to 800, indicating a relatively stable average cost. However, in 2021, many counties have seen a significant increase in average cost.
- The right graph shows the changes in average cost for the top 10 counties with the highest number of WIC participants. It shows that 9 out of 10 counties experienced a rapid growth in average cost in both 2020 and 2021. This finding is consistent with our observation of statewide growth in average cost.

Figure 9

CA COUNTY WIC: Average Cost VS Time (2010-2021)

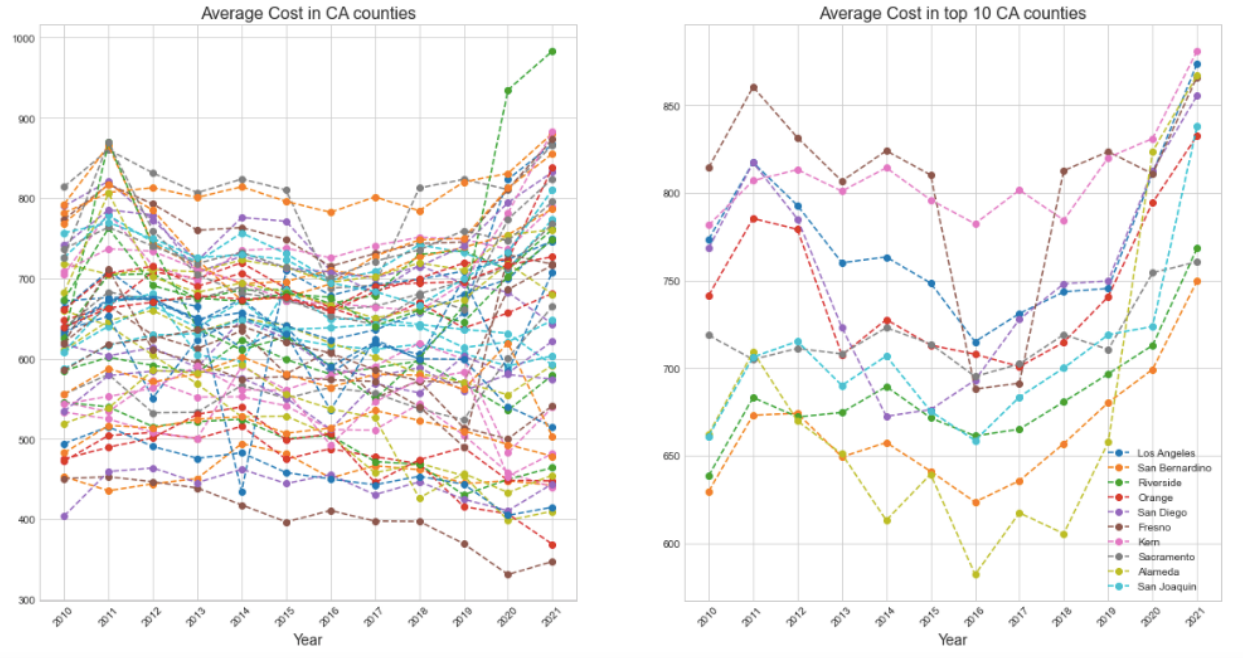
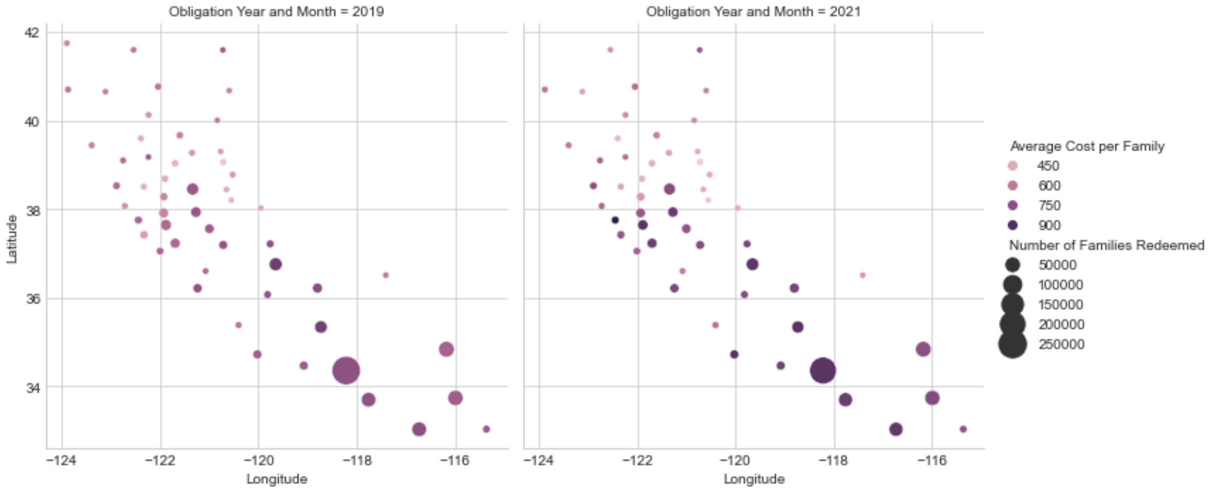


Figure 10

CA COUNTY WIC: Redemption VS Location (2019 vs 2021)



County Redemption versus Location (2019 vs 2021) (Figure 10)

- Each dot in the plot represents the center of a county in California. The hue of the dots represents the average cost, with darker colors indicating higher costs. From the 2021 plot, we can observe that many of the dots have become darker than their corresponding dots in 2019, particularly in the central and southern parts of California. This finding corresponds to our earlier conclusion that the average cost is growing significantly at both county and state levels. In contrast, the majority of the dots in northern California remained in light colors. Additionally, we found that if a county had a high average cost in 2019, it was more likely for the cost to grow after 2021, while counties with low costs tended to remain low.
- The size of the dots in the plot represents the number of participants in each county. We can see that there is not much difference in dot sizes between the 2019 and 2021 plots. As we noted earlier, the number of WIC families dropped in 2020 and only increased slightly in 2021. Therefore, it is not surprising to see that the sizes of the dots in the two plots are comparable.

Conclusion

In summary, this project provides insight into the changes and trends of the WIC program in California over the past decade, with a focus on the impact of the COVID-19 outbreak. The findings suggest that while the WIC program was facing a decrease in participants and total cost before the pandemic, the situation has changed in both 2020 and 2021 with a slight increase in the number of participants and a significant rise in average cost. But the cost of the Statewide Infant Formula Rebate is still dropping no matter of time. The analysis also highlights the differences in redemption patterns between counties, with counties with larger numbers of participants showing higher redemption costs, especially in southern CA. After the outbreak of COVID-19, the geography of WIC redemption changed. The differences in average costs between northern and southern CA became more visible. Counties with a larger number of participants and higher redemption costs are mainly located in southern CA, and their data shows greater changes after 2019.

Big data can be used to further improve this project in several ways. Firstly, big data provides a wider scope for data collection. We can utilize a greater variety of data sources, including transactional data, and use innovative data collection methods such as Internet of Things (IoT) sensors. Secondly, big data analytics can help us identify patterns and trends in the WIC program data more timely and accurately. For example, some data analytics can be applied to the data to detect correlations between various factors, such as statewide distribution of population, income levels, employment rates, and WIC participation rates. What is more, big data can be used to predict future trends in WIC participation and costs. By analyzing historical data and identifying patterns, machine learning methods can be trained to predict how the WIC program may change in the coming years. This can help policymakers and organizations to generate insights, and make informed decisions accordingly. It also enables the identification of service delivery gaps, monitoring of progress and outcomes, and ultimately leads to the development of more effective social welfare programs and services.

Datasets

California Women, Infants and Children Program Redemption by County - California Open Data. (2021). Ca.gov. https://data.ca.gov/dataset/california-women-infants-and-children-program-redemption-by-county

Women, Infants and Children (WIC) Authorized Vendors - California Open Data. (2023). Ca.gov. https://data.ca.gov/dataset/women-infants-and-children-wic-authorized-vendors

California Counties - California Open Data. (2023). Ca.gov. https://data.ca.gov/dataset/california-counties

References

Al-Barashdi, & Al-Karousi, R. (2019). Big Data in academic libraries: literature review and future research directions. Journal of Information Studies & Technology (JIS&T), 2018(2). https://doi.org/10.5339/jist.2018.13

How WIC Helps - California Women, Infants & Children Program. (2022). Ca.gov. https://www.myfamily.wic.ca.gov/Home/HowWICHelps#interestedTitle

UN WOMEN. (2020, September 16). COVID-19 and its economic toll on women: The story behind the numbers. UN Women. https://www.unwomen.org/en/news/stories/2020/9/feature-covid-19-economic-impacts-on-women

UN WOMEN. (2020). From Insights to Action: Gender Equality in the Wake of COVID-19. UN Women. https://www.unwomen.org/en/digital-library/publications/2020/09/gender-equality-in-the-wake-of-covid-19

ECONOMIC IMPACT OF COVID-19 ON SINGLE MOTHERS. (n.d.). Women's Council Wisconsin. https://womenscouncil.wi.gov/Documents/SingleMothersCOVIDInfoSheet_F_Corr.pdf

McHenry, P. (2011). The Relationship between Location Choice and Earnings Inequality. RePEc: Research Papers in Economics. http://economics.wm.edu/wp/cwm_wp112.pdf

Women, Infants & Children Program. (n.d.). California Department of Public Health. https://www.cdph.ca.gov/Programs/CFH/DWICSN/pages/program-landing1.aspx