ML4VA Final Report

Matthew Whelan
University of Virginia
School of Engineering and Applied Sciences
Charlottesville, VA
mw3shc@virginia.edu

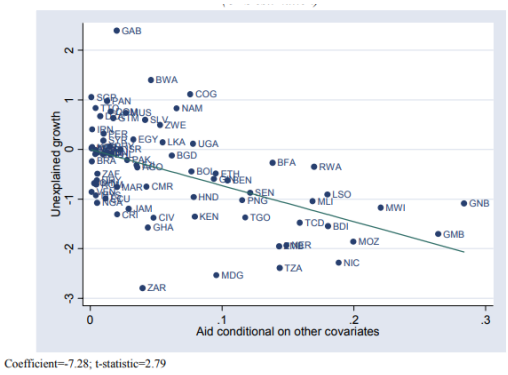
Sidhardh Burre
University of Virginia
School of Engineering and Applied Sciences
Charlottesville, VA
ssb3vk@virginia.edu

Avaneen Pinninti
University of Virginia
College of Arts and & Sciences
Charlottesville, VA
ap4xz@virginia.edu

***Abstract*—Roughly 10% of all Virginians live in poverty with millions more struggling to make ends meet. While there are many government programs set to support these families and individuals, it is quite frequent that people do not receive enough aid and assistance. In 2020, COVID-19 shook the global economy at an unprecedented scale, disproportionately affecting impoverished people. The United States in particular implemented various government assistance programs such as unemployment benefits, healthcare coverage and stimulus checks, in an attempt to curb the impacts. This research aims to evaluate the effectiveness of these government programs, with the goal of identifying their efficiency for future emergency mitigation. The methods used include k-Means clustering via Virginia counties' economic indicators, and regression analysis on counties' means with the change in government assistance items and the change in poverty over the duration of COVID-19. The results indicate that Virginia should grant more structured benefits like Food Stamps instead of unstructured benefits like stimulus checks, as those proved much more efficient at mitigating poverty.**

I. Introduction

Despite many approaches suggesting that we, as society, have a duty to provide aid to those in need, we have very little data confirming that provided aid alleviates said need. The debate for "aid is increasingly polarized between those who radically denounce aid as a vehicle for development and those who optimistically seek to reform it" [1]. This is particularly evidenced when studying the impact of aid when given to developing countries. Charts such as the one seen below in Figure 1 shows an overall negative trend between growth and aid. [2].



Coefficient=-7.28; t-statistic=2.79

Fig. 1. Conditional Correlation between Growth and Total Aid, 1960-00

On one side, we have William Easterly and Dambisa Moyo who argue that aid is ineffectual with "2.3 trillion spent and little to show for it" [3] due to aid's tendency to dissuade constituents from searching for their own solutions and corrupting existing institutions. Alternatively, we have philanthropists such as Jeffrey Sachs who argue that we instead don't spend enough and only with a huge amount of aid can we kickstart poor countries out of the poverty cycle. But a more reasonable argument is that both sides are partially wrong because no one knows what does or does not work. This is primarily due to the fundamental problem of causal inference which makes it impossible to observe the effect of a policy intervention on any single unit because the counterfactual is unknown [4]. Afterall, how can you claim that government aid improved a situation when the situation could've been addressed without government intervention?

Our group was able to find a published paper that discusses how machine learning can be used to predict and classify poverty levels. Although the paper dives deep into poverty levels in areas outside of Virginia (Jordan), the techniques used provide a lot of helpful information as to how we can structure our approach to solving our problem. To begin, the paper titled "Poverty Classification Using Machine Learning: The Case of Jordan" [1], takes into account all household expenditure and income surveys from the early 2000s and uses that data to identify and measure the poverty status of Jordanian households. Logistic regression was used to classify Jordanian households into either a "non-poor" or "poor" status. While the paper is able to properly classify the poverty status of Jordanian households from government aid metrics, it fails to correlate depth of poverty with aid given, instead relying on a binary impoverished/wealthy classifier. Furthermore, instead of attempting to assess the importance of any individual factor for determining poverty status, the researchers treat their model as a black-box and fail to incorporate the explanatory power of the model into their report.

Economic aid programs such as Social Security, food assistance, tax credits, and housing assistance can help provide opportunities for families who fall near or below the poverty line in their respective counties/states. However, many of these programs, which are meant to support families and children, are solely based on where their family income falls with respect to the poverty threshold, do not live up to their promise. There are many instances where governments are

too late or provide insufficient assistance causing aid to be disproportionately distributed. This was brought to head during the COVID pandemic when many small business funds were improperly distributed and outright abused [5]. This brings to head the following question: “If PPP loans were misattributed during the COVID pandemic, could it be possible for other funds to be misattributed as well?” Or to put it another way; ”how can government funds most aptly be allocated so as to serve the greatest need?”

A common metric to assess need is poverty, therefore understanding poverty will help us better understand need as well and guide our project. To understand the current state of poverty in Virginia across census tracts and localities, the following choropleth map outlining poverty in various Virginian counties was created.

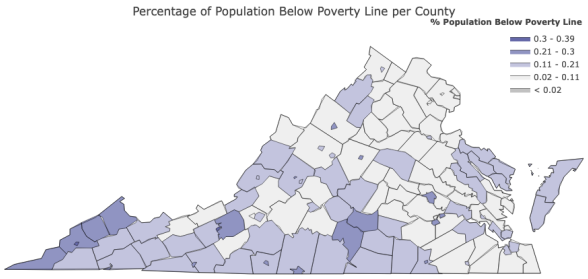


Fig. 2. A choropleth map of Virginia’s localities’ ratio of impoverished people to total population

As seen in Figure 2, poverty is clustered mostly in the more rural, Southern localities in Virginia, as well as some urban centers. This sets up our experiment well, where we will aim to investigate the root causes of this poverty, an issue that has only been worsened due to COVID-19.

A. Hypothesis

We hypothesize that the best form of government aid is cash-based public assistance income due to the fact that it can kick-start consumer demand and reel an economy out of a nose-dive. Being able to ”force” consumer demand, even artificially, makes the most conceptual sense considering that citizens with a cash-injection will be more likely to spend it and redistribute these funds across different facets of the economy, like small businesses. This discourages saving money up and thus slowing down an economy, as investing money into the economy should have a ’domino effect’ and cause businesses and others to spend more as well.

II. Data Sources

A. Approach

In this study, it is essential to be able to enable the model to delineate discrepancies in the data with explainability that is able to correlate demographic changes with aid-based changes. The COVID pandemic was able to produce a great variety of effects related to income distributions and poverty demographics. Therefore, by observing the change in these demographic metrics and attempting to correlate them with government fast-response items, we will be able to determine whether or not there is a correlation between government action or any improvement in relevant demographic metrics. The government action items will be referred to as “deltas” and measured as the change in government-controlled systems. These deltas are listed below:

- Increase in Social Security Spending
- Increase in Supplemental Security Income
- Increase in cash public assistance income
- Increase in Food Stamp/SNAP benefits
- Increase in public insurance coverage
- Change in work commute, drove alone
- Change in work commute, carpooled
- Change in work commute, Public Transportation
- Change in work commute, Walked
- Change in work commute, Other
- Change in work commute, Worked from home

The work-commute deltas are incorporated to observe if government spending to promote/improve forms of public transportation significantly improved the status of counties. We initially thought that governments might increase forms of public transportation or work from home initiatives, and this could reduce poverty overall.

Demographic variables were isolated to status indicators for a county’s overall health as well as relevant metrics that government officials may find relevant to assessing the effectiveness of a policy change. These variables include:

- Percent of population with income to poverty ratio:
  - Under 0.50
  - 0.50-0.74
  - 0.75-0.99
  - 1.00 to 1.24
  - 1.25 to 1.49
  - 1.50 to 1.74
  - 1.75 to 1.84
  - 1.85 to 1.99
  - 2.00 to 2.99
  - 3.00 to 3.99
  - 4.00 to 4.99
  - 5.00 and over
- Percent of population in industry:
  - Management, business, science, and arts occupations
  - Service occupations
  - Sales and office occupations
  - Natural resources, construction, and maintenance occupations
  - Production, transportation, and material moving occupations
  - Agriculture, forestry, fishing and hunting, and mining
  - Construction
  - Manufacturing
  - Wholesale trade
  - Retail trade
  - Transportation and warehousing, and utilities

  - Information
  - Finance and insurance, and real estate and rental and leasing
  - Professional, scientific, and management, and administrative and waste management services
  - Educational services, and health care and social assistance
  - Arts, entertainment, and recreation, and accommodation and food services
  - Other services, except public administration
  - Public administration
- Percent of population with income in band:
  - Less than $10,000
  - $10,000 to $14,999
  - $15,000 to $24,999
  - $25,000 to $34,999
  - $35,000 to $49,999
  - $50,000 to $74,999
  - $75,000 to $99,999
  - $100,000 to $149,999
  - $150,000 to $199,999
  - $200,000 or more
- Percent enrolled in higher education (undergraduate or graduate programs)
- Unemployment Rate (16 years or older)

B. Datasets

We used the following data sets: ratio of income to poverty level, selected economic characteristics, school enrollment, employment status and poverty status. We gathered all of these data sets from the US Census Bureau's American Community Survey program to maintain consistency between data. All of these data sets provide relevant information broken down by Virginia's county, within a supplied margin of error. We collected these data sets for 2019 and 2021, to represent pre-COVID-19 data and post-COVID-19 data, respectively.

C. Preprocessing

For each of these data sets, we decided to drop a significant amount of features, and only keep the features that were absolutely relevant to our project. Our final set of features included ratio of income to poverty level, means of commuting to work, type of occupation, industry worked in, household income, public benefits (like social security, food stamps, and health insurance), voluntary school enrollment rate (higher education), and unemployment rate.

To standardize all of our data, we decided it would be best to convert all numeric values into percentages. This way, different features can be easily compared to each other to make it easier for data analysis. We also created a few new features to simplify our model: one particular one was to combine enrollment in undergraduate and graduate school to represent enrollment in higher education.

The economic characteristics dataset was missing a few values for certain counties, so we used a SimpleImputer with a median strategy to replace these values. The employment status dataset was only missing one value for Fauquier County in 2019, so we found this value from a similar dataset (unemployment rate of 2.5%). The rest of the work in preprocessing was to format the labels and ensure the data types and indexes were correct for our dataframe.

To gain the final data, each metric was gathered at both 2019, a pre-COVID value, and in 2021, a post-COVID value. The difference between these corresponding values was found and used as the final datapoint fed into the model. This method was applied to both county status indicators as well as deltas to observe the change in health indicators as well as account for the base-line aid that the government provides.

III. Methods

Prior to beginning the experiments, we attempted to determine the optimal number of cluster centers to use when attempting cluster centering. To do so, the county's 2019 status variables were repeatedly clustered with various cluster numbers using a KMeans() model.

Four main experiments were conducted. The first one was to cluster the counties based on their relevant variables in 2019 and gain a cluster mean. We then found the same counties' 2021 cluster mean and established a correlation between the deltas and the vector difference between means. The second experiment was similar to the first one, but after clustering our 2019 counties, we also clustered 2021 counties to get a different set of clusters. We then found the vector difference between means and looked for a correlation between that and the deltas. Our third method was to simply compute the county's vector difference between 2019 and 2021 and look for a correlation between each county's unique vector difference. For a baseline model we also inspected the correlation between the deltas and the percent change in poverty. After performing each method, the linear regressor's coefficients were inspected as the final data points.

A. Experiment 1

Counties were initially clustered via their 2019 status variables and then the cluster centers were recomputed from 2021 status variable values. Subsequently, each county's corresponding cluster center's migration was measured and assigned to each county. These cluster-center-migrations were then correlated against the delta variables. By recomputing the means of the original 2019 clusters, we can observe overall trends in clustered counties which may show high-level insights into the change in county status variables.

B. Experiment 2

Counties were initially clustered via their 2019 status variables and then re-clustered via 2021 status variables. Subsequently, each county's corresponding cluster center's migration was measured and assigned to each county. These cluster-center-migrations were then correlated against the delta variables. The purpose of this method is to take full advantage of the power of clustering and observe significant changes within a county's respective cluster. This can allow us to

observe individual counties with significant status variable changes that cause them to change clusters entirely.

C. Experiment 3

Counties 2019 status variables were subtracted from their 2021 status variables and normalized against their 2019 status variables such that a percent change in 2019 status variables could be found. These percentages were correlated against the delta variables. This would enable the direct observation of the delta variable's effect on a county's status variables.

D. Experiment 4

As a sanity check, change in a county's percentage of constituents beneath the poverty line was correlated with the delta variables. This was done to ensure that the delta variables could correspond to the change in the percent of population impoverished, a variable that was implied to have a high correlation with the health of a county. This improved the chances that the delta variables were accurately capturing some aspect of a county's health.

IV. Results

Before conducting our experiments, many of which require clustering, it was necessary to find the optimal number of clusters for the KMeans algorithm. The inertia of individual KMeans algorithms run with different cluster number configurations can be seen below in Figure 3.

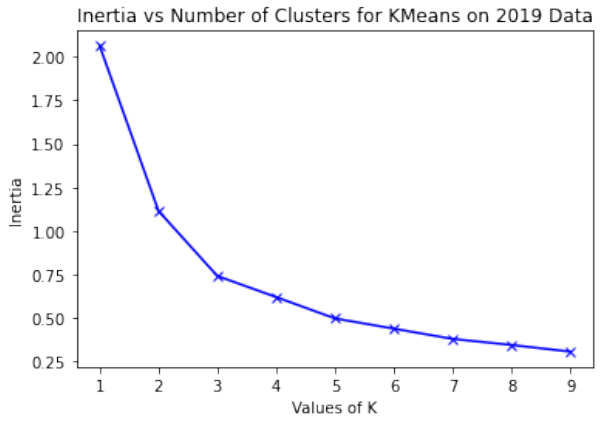


Fig. 3. Inertia vs Differing numbers of cluster centers

By observing the penultimate increase in K which still displays a significant improvement in inertia, we can find that the optimal value for k, "the elbow," is k=3. Now, we can advance with our experiments conducted knowing the optimal number of clusters.

In Figure 4, irrelevant deltas were removed from the data leaving only deltas that displayed a significant correlation with status variables. While a prescriptive method to address pandemics was not found, we did find some metrics that are reassuring for later experiments. Namely, we found a negative correlation between various income metrics and SNAP Assistance and cash public assistance income indicating

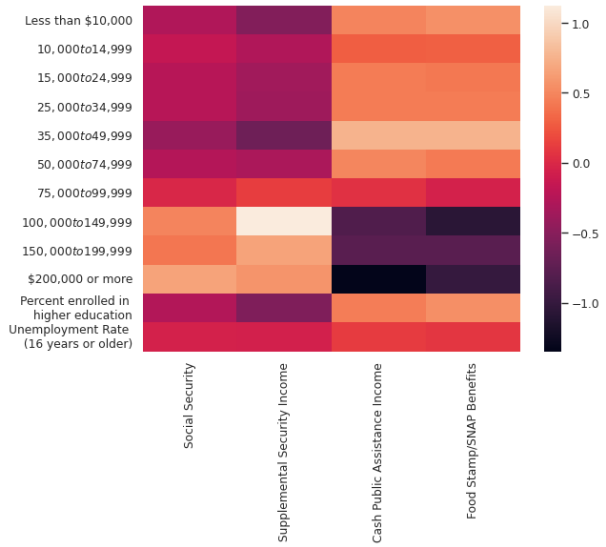


Fig. 4. Heatmap of the Coefficient Matrix within the Linear Regression used for Experiment 2

that aid was going to the right places, people that were impoverished. This is because a negative correlation between SNAP Assistance and cash public assistance income and high income band earners (with an income 5 times that of the poverty line) indicates that states with ballooning high-income earners lose SNAP Assistance and cash public assistance income. Furthermore, because our data primarily analyzes the change in variables, we can assess that governments were highly responsive to changes in demographics as well. Reallocating and redistributing resources to people in need when necessary. This clearly demonstrates how governments adapt their funding distributions to demographic changes in income distribution.

Another aspect to note is the high correlation between high-income band earners and SS/SSI. Note that SSI is "designed to help aged, blind, and disabled people, who have little or no income." This will be explored further in Experiment 2.

The second experiment differed from the first due to the reclustering of counties by 2021 data. The result of this method was that some counties effectively had 0-vectors when it came to their cluster-migration vector. This is due to the fact that some counties, despite migrating slightly, had no significant change in their overall cluster center position and therefore, indicated no change in the county overall. In effect, this method isolated county's with significant changes that either migrated clusters, or caused new clusters to form.

By isolating only county's with significant cluster changes, much of the data was "normalized" leading to cleaner gradients displayed in Figure 5. Our deltas are clearly classified into two separate classes, the first class, consisting of SS and SSI, conditional government aid and the second class being cash public assistance income and Food Stamps/SNAP benefits. The first class increases as income increases and the second class decreases as income increases. This is in-line with the expected

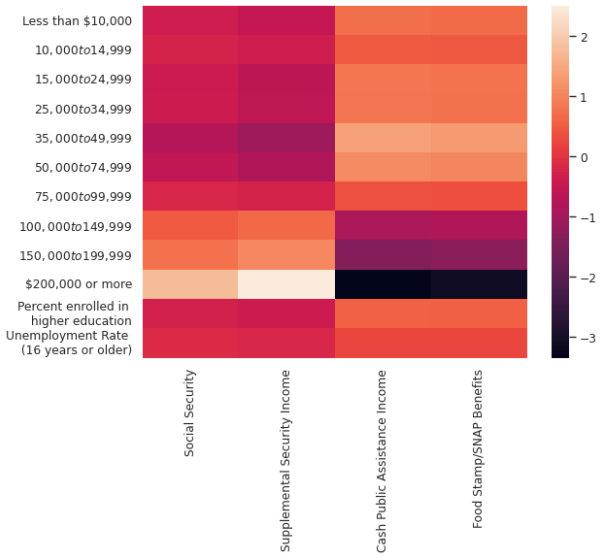


Fig. 5. Heatmap of the Coefficient Matrix within the Linear Regression used for Experiment 2

functions of both classes of aid.

Most notable about the first class of government aid is that it is targeted mainly at the older population and it is known that COVID-19 had a devastating impact on the older population. So overall, it can be expected that SS and SSI decrease but a positive correlation with higher-income band earners indicates that the proportion of the population that was in high income bands decreased over the course of the pandemic.

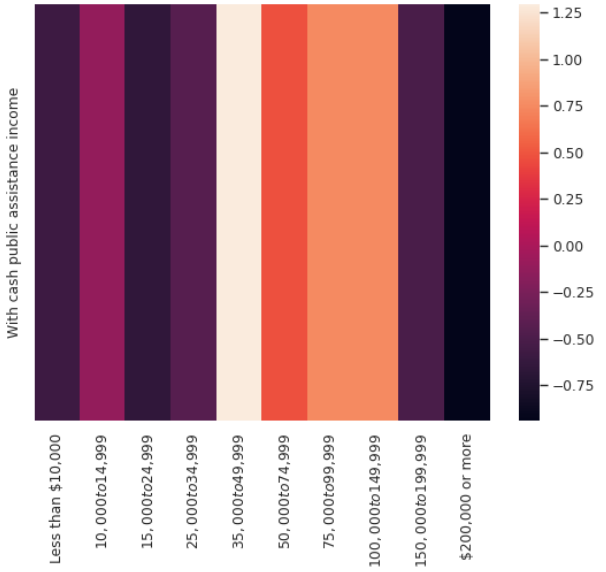


Fig. 6. Heatmap of the Cash Public Assistance Feature used for Experiment 3

The third experiment conducted yielded maybe the most important results of all experiments so far. From the results, income band migration was observed. Figure 6 showed that families and individuals in poorer income bands actually migrated to higher income bands. This was specifically observed in the cases where people made less than $35,000 and changed to making between $35,000 and $50,000. This is also true for the above $150,000 brackets, as they also increased in income over COVID-19.

We thought that the clustering would help characterize Virginia's counties in specific groups in Experiments 1 and 2, but it did not seem to help in analysis. This is because Experiment 3 showed that lower income groups benefited from cash assistance, which was not the case in the other two experiments surprisingly.

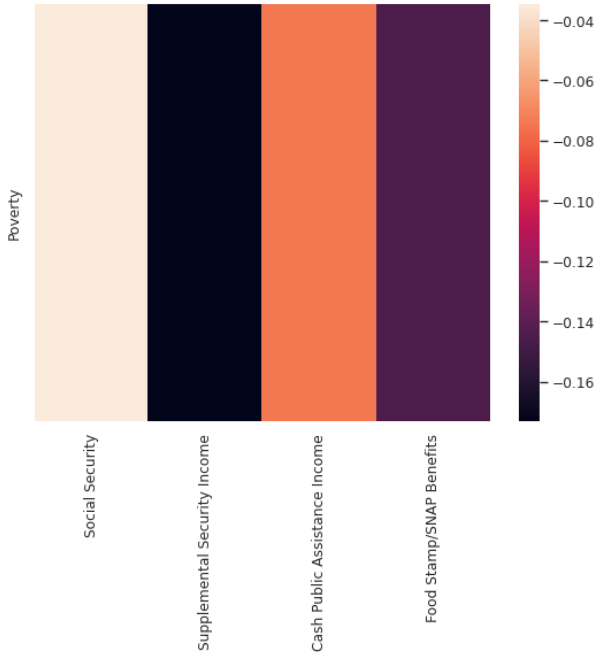


Fig. 7. Heatmap of the Change in Poverty used for Experiment 4

Experiment 4 showed us that the most effective measures of decreasing poverty level were SSI (correlation coefficient of -0.173), then Food Stamps/SNAP benefits (-0.145), then Cash Public Assistance (-0.074) and finally SS (-0.035). These results are depicted in Figure 7. This means that SSI had nearly 3 times the impact of Cash Assistance had, and Food Stamps had over twice the impact. Interestingly, increasing funding for healthcare led to an increase of people eligible for public healthcare, which explains the positive correlation.

Experiment Number	RMSE Score
1	0.022
2	0.045
3	0.015
4	0.020

TABLE I
RMSE SCORES FOR EACH EXPERIMENT

Table 1 shows the RMSE scores from the linear regression ran for each experiment. We can deduce that experiment 3 gave the lowest RMSE, which means the the deltas had the best correlation to the vector difference in the health of counties.

Experiment 4 gave the second-lowest score, which indicates that deltas correlated well to the change in poverty over the same time period. Experiment 1 gave the second-highest score, which means that the deltas were correlated well to the cluster migration for 2021 data. Experiment 2 had the highest RMSE, which shows that re-clustering the 2021 data gave a lower correlation than just migrating the cluster over for the 2021 data.

Overall, we found that the first four deltas had the greatest impact on reducing poverty. These were the labels on each of the graphs: SS, SSI, Cash Public Assistance and Food Stamps. We found that most of the other deltas, namely the work-commute deltas, had little impact on the change in poverty over the timespan of COVID-19. In our hypothesis, we thought that governments could increase forms of public transportation or sponsoring initiatives to encourage working from home like covering internet costs. However, we found this to be false due to negligible (close to 0) correlation for all of these attributes.

V. Conclusion

The main takeaways from this project and the four experiments conducted show that the state of Virginia indeed did supply adequate monetary aid and benefits to those who needed it the most during the COVID-19 pandemic. The first experiment showed how the government is responsive to changes in demographics, as the highest income earners saw their aid decrease over the pandemic. Experiment 2 showed that the amount of aid aimed at the older population (SS and SSI) decreased, likely due to the unfortunate deaths that COVID-19 caused. The results of the third experiment showed that families and individuals who were making less than $35,000 moved up to making between $35,000 and $50,000. This confirms our hypothesis that Virginia supplied money to the people who needed it most, and it clearly benefited them in the long term. However, despite this good news, our results in the fourth experiment show that it would be much more effective to distribute aid to Virginians through structured programs such as SS, SNAP benefits and other structured cash benefits. This is not to say that the stimulus checks provided by the US were not helpful - rather the trillions of dollars spent could have been allocated more effectively. This is a conversation that should happen in Washington D.C. as there are many other factors involved in a large decision like this (public opinion, how quickly money needs to be injected into the economy, etc).

Future work would include attempting to generalize the methods/conclusions here to other states and countries. Further, more advanced models and methods would be used to assess the same metrics. One prominent model would be the use of some form of Euler's method on each of the status variables and attempting to fit Euler's method with a variety of delta variables to the month-by-month change of the status variables.

References

[1] Alsharkawi A, Al-Fetyani M, Dawas M, Saadeh H, Alyaman M. Poverty Classification Using Machine Learning: The Case of Jordan. Sustainability. 2021; 13(3):1412. https://doi.org/10.3390/su13031412

[2] "The Great Aid Debate," CIHA Blog, Aug. 06, 2010. http://www.cihablog.com/the-great-aid-debate/ (accessed Dec. 07, 2022).

[3] R. G. Rajan and A. Subramanian, "Aid and Growth: What Does the Cross-Country Evidence Really Show?," p. 49.

[4] S. Engel, "The not-so-great aid debate," Third World Quarterly, vol. 35, no. 8, pp. 1374–1389, 2014.

[5] "Causal Inference - an overview — ScienceDirect Topics." https://www.sciencedirect.com/topics/social-sciences/causal-inference (accessed Dec. 07, 2022).

[6] "'Biggest fraud in a generation': The looting of the Covid relief program known as PPP." https://www.nbcnews.com/politics/justice-department/biggest-fraud-generation-looting-covid-relief-program-known-ppp-n1279664 (accessed Dec. 07, 2022).

[7] https://data.census.gov/table?t=Poverty&g=0400000US51$0500000&tid=ACSDT1Y2019.B17002

[8] https://data.census.gov/table?t=Poverty&g=0400000US51$0500000&tid=ACSDP1Y2019.DP03

[9] https://data.census.gov/table?t=Poverty&g=0400000US51$0500000&tid=ACSDT1Y2021.B14

[10] https://data.census.gov/table?q=employment&g=0400000US51$0500000&tid=ACSST1Y

[11] https://data.census.gov/table?q=B17020:+POVERTY+STATUS+IN+THE+PAST+12+MONTHS+BY+AGE&t=Poverty&g=0400000US51$0500000&tid=ACSDT1Y2021.B17020

VI. Contribution

Matthew Whelan

- Research the datasets relevant to the project and metadata
- Perform data extraction and preprocessing operations
  - Dropping irrelevant features
  - Combining features for simplified data
  - Filling in missing values with SimpleImputer
  - Cleaned data in one dataframe with consistent formatting of labels

Sidhardh Burre

- Data Modeling
  - Training models/performing regression
  - Experimenting with alternative setups
- Data Visualization
  - Generate KMeans graphs
  - Heatmaps for each experiment
- Dataset Discovery

Avaneen Pinninti

- Final project video editing
- Researching relevant experiments and studies for the work related section