Understanding Vulnerability of Children in Surrey

UBC Data Science for Social Good

A collaboration with Children's Partnership and Microsoft

Authors:
Catherine Lin
Cody Griffith
Kevin Zhu
Varoon Mathur

September 5, 2018

Abstract

Understanding the community conditions that best support universal access and improved childhoods outcomes allows ultimatley to improve decision making in the areas of planning, and investing across the early and middle years of childhood development. We describe a combined top-down and bottom-up approach to understanding the lived experiences of children throughout the City of Surrey. As part of the top-down approach, we find specifically that the Early Development Instrument describing childhood vulnerabilities can be used to cluster neighborhoods, and that Census variables can help explain these groupings. As part of the botton-up approach, we use program registeration data from Children's Partnership Surrey to find a crtitical age of entry and exit within childhood programs. We find that certain programs of entry can represent longer retention of children within the program. These results provide a lens to which communitiy initiatives can be strategically put forth in neighborhoods that migh experience larger vulnerabilities than others.

Contents

1 Introduction **3**
- 1.1 The City of Surrey and Children's Partnership . . . 3
- 1.2 Early Development Instrument . . . 3
- 1.3 Purpose of our Analysis . . . 3

2 Datasets used **4**
- 2.1 EDI . . . 4
- 2.2 Census . . . 4
- 2.3 CLASS . . . 4

3 Top-Down: Understanding Trends of Neighborhoods **4**
- 3.1 Approaches . . . 4
- 3.2 Validation & Results . . . 6

4 Bottom-Up: Understanding City Program Reach **9**
- 4.1 Retrieval of Data . . . 9
- 4.2 Program Activities and Grouping . . . 10
- 4.3 Summary Statistics . . . 10
- 4.4 General Activities and Retention . . . 12
- 4.5 Challenges with Predictive Modeling . . . 12

5 Discussion **14**
- 5.1 Linking Approaches . . . 14
- 5.2 Web Application Development . . . 14

6 Conclusion & Future Work **14**

A Census Variables **16**
- A.1 All Census Variables . . . 16
- A.2 Significant 2016 S-cluster Census Variables . . . 18
- A.3 Significant A-cluster Census Variables . . . 19
- A.4 Significant UA-cluster Census Variables . . . 21

B Grouping Methodology **22**
- B.1 Flow Chart for Grouping . . . 22

1 Introduction

This project builds from work done with Avenues of Change, an initiative in Guildford West, whose goals are to improve decision making for initiatives targetting children ready to enter First Grade [EPPS]. The results from last year focused largely on top-heavy metrics such as the Early Developing Instrument (EDI), as well as library registration data and crime data, to find a relationship using regression modeling. Unfortunatley, no significant results were found with this method. This years project broadens the focus of the project to all of Surrey, as well as provides granular program registration data, referred to as the CLASS Dataset, which is provided by the City of Surrey and Children's Partnership.

1.1 The City of Surrey and Children's Partnership

Through the smart cities initiative Surrey, BC has had the opportunity to grow their city using data science to create a rigorous foundation to build resources. The city has chosen that one of the best approaches is through educating the younger citizens of Surrey. It is here we find the program *Children's Partnership*, a dedicated program to expanding the network of resources children have access to throughout the city. The Children's Partnership of Surrey-White Rock aims to equip the city of Surrey with a tools and resources to support organizations and professionals working to support early and middle childhood development and positive family outcomes. The goals of this years project were to improve decision making for Children's Partnership, as well as their partner organizations. This program operates at the neighborhood level and a neighborhood is defined by UBC's *Human Early Learning Partnership* (HELP). There are a total of 24 neighborhoods that HELP considers part of Surrey we aim to understand the needs of children in these neighborhoods in a multitude of ways.

Children's Partnership has also created its own extensive database on children who have participated in their programs. These vary from aquatic programs to day care, from computer programming courses to outdoor education, from cooking to baking. The collection of these programs form the network of resources that the city of Surrey offers their children.

1.2 Early Development Instrument

A useful metric for understanding the needs of a neighborhood with children is the *Early Development Instrument* (EDI). This metric is calculated by administering a 104 question survey to kindergarten students every 3 years, we call this an *EDI wave*. This survey asks questions at scales such as physical or communicational skills to then target vulnerability, these questions were primarily motivated by the *Early Years Study* [OMM99]. Then the survey scores are aggregated at the neighborhood level to create a population level metric and a baseline of vulnerability is set with the bottom decile from the first wave. From this, we have the groundwork to identify vulnerabilities of kindergarten children at the neighborhood level for future waves.

This metric is widely used across Canada with varying degrees of historical use. For the city of Surrey (along with the entirety of British Columbia), there has been 6 waves of data collected since 2004. As per HELP's standard, the first wave was set as the baseline and there have been an effective 5 waves of usable data.

1.3 Purpose of our Analysis

We aim to provide a more data-driven approach to implementing policy development in the city through the use of modern statistical analysis and visualization, similar to that of the *Magnolia Initiative* [IB+14]. The Magnolia Community Initiative is a similar organization that aims to strengthen childhood outcomes in the Los Angeles area, and has put together a web-based dashboard to display different metrics that measure how well they are attaining their goals. Using this dashboard as a framework, we have built a pipeline of analysis to understand how neighborhoods group together in terms of their EDI scores as well as the distribution and reach of children associated with Children's Partnership. These two approaches we consider as **top-down** and **bottom-up** respectively. With both of these approaches, we have the means to address interesting anomalies present in the data and pave the way for future program development with Children's Partnership all while being able to point to what drives these decisions with statistical confidence.

2 Datasets used

2.1 EDI

HELP's EDI data is an open source dataset and could be found here. The EDI is broken into five scales: physical health, social competence, emotional maturity, language/cognitive development, and communication skills. Vulnerability is measured as percentage of children vulnerable, as well as the total count of children that are vulnerable. The percentage and count of children that are vulnerable on one or more scales, is also measured.

2.2 Census

We choose 147 Census variables from the 2016 Canada Census based on [KF10], these variables can be found in appendix A. Census does not contain data in the neighborhood levels defined by HELP, but does provide data down to the Dissemination Area (DA) level. We roughly combined DAs into neighborhoods based on whether the centroid of a DA's geometry lies within some neighborhood's boundary. The statistics of these DAs are then aggregated to the neighborhood they belong in.

2.3 CLASS

The CLASS dataset is over 160Gbs, contains information of programs, registration data, and client information that includes neighborhood of residence, age, and gender. Program and registration information primarily included the title and description of the program, as well as binary flag to describe whether a subsidy was used to pay registration fees or not. A data-sharing agreement was signed by all those involved in the project.

3 Top-Down: Understanding Trends of Neighborhoods

The biggest difference between this year's project with Children's Partnership as opposed to the first DSSG education project [EPPS] is the number of neighborhoods. Last year the education team worked with *Avenues of Change* to identify important factors contributing to EDI metrics of a specific neighborhood, Guildford West. This year our team is working with *Children's Partnership* with a larger focus of identifying the city wide span of important factors contributing to EDI metrics. As such, understanding trends at the neighborhood level helps to understand what is similar and different. In this section, we have chosen to take the aggregated EDI scores and cluster the neighborhoods that perform similarly. As we have 5 scales of EDI at our disposal (i.e Physical, Emotional, etc.), each neighborhood is a data point in $\mathbb{R}^5$. We find hidden structures of neighborhoods in this higher dimensional space using the *t-distribution Stochastic Neighbor Embedding* (t-SNE) and explain the social economical effects that could be bringing these neighborhoods together. To verify that our clusters are truly present in the data, we verify with a hypothesis test. We also compare this to another method, *Uniform Manifold Approximation and Projection* (UMAP), to make sure our results are robust and are not sensitive to the specific method and interpretation. We are especially interested in when these methods disagree and have insights into what may cause this.

3.1 Approaches

Our procedure for finding clusters of behaviors is quite simple. Since we are observing data that is outside of our ability to visualize, we choose to project down into visual space. This choice comes with a cost as we do lose information when we project and hence must be careful as to what we extract with this procedure. With this in mind, we chose to initially project our data using t-SNE and here we do not discuss much about the implementation and theory but rather refer to [**?**]. In short, t-SNE is a popular machine learning method that is engineered to work well for representing high dimensional data in a low dimensional setting for a large class of problems.

We use this projection approach in two ways, first we consider all the neighborhoods at a specific time and then project our data into a two-dimensional space. We call this our **Single-Wave** approach and we aim to capture the neighborhoods that behave similarly at a particular time with this approach and thus reduce any temporal factors like population growth. See figure 1 for an example of this method. The keen observer may notice that a natural 3-cluster pattern

forms and this phenomenon occurs for each of the waves, we present this in 2. We denote these the *S-clusters* from here on due to their single-wave data. We decided to fix the number of S-clusters for this component of the analysis to be 3 for each wave due to this interesting behavior and expect this trend to continue for upcoming waves. These S-clusters are then ranked by their average *two or more vulnerabilities* EDI scale and hence consider S-cluster 0 to be the lowest in vulnerability, and S-cluster 2 to be highest. The clustering technique we impose is a simple k-means method with $k = 3$ chosen from inspection as mentioned. Even more interesting, we observe that certain neighborhoods bounce around to different clusters as we change the wave and this indicated a strong temporal influence on our data.

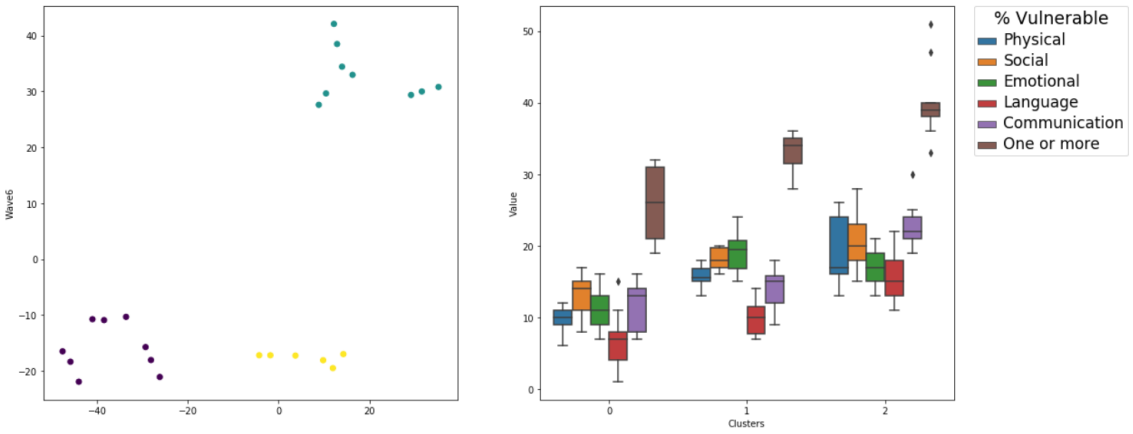

Figure 1: Single Wave Clusters (t-SNE) for Wave 6. On the left, we show the projected EDI data as well as the 3 clusters a k-means method has chosen. We choose to consider this as 3 clusters. On the right, we show the spread of the EDI scale for each S-cluster. It should be noted that we have ordered these clusters by their median *One or More* EDI score.

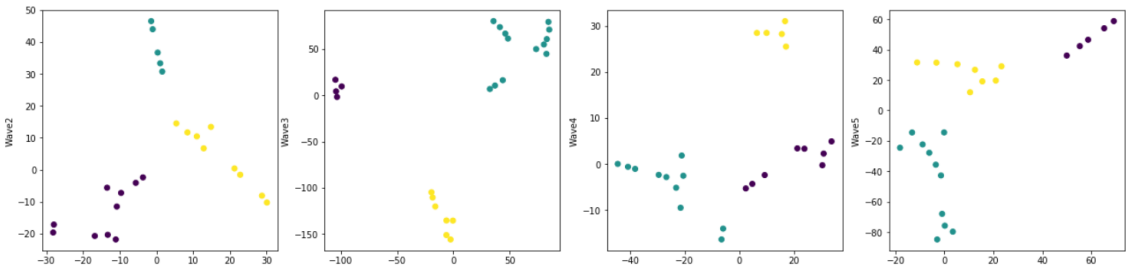

Figure 2: Single Wave Clusters (t-SNE) for Waves 2-5.

Our second approach is to project the entirety of our EDI scale data simultaneously. We call this our **All-Wave** approach and here we aim to capture the temporal effects we witness from varying the waves in the single wave approach. Before, we found a hidden 3 S-cluster structure lying within our data whereas here we find a natural 6-cluster structure, see figure 3. We denote these the *A-clusters* from here on due to all-waves are considered at the time of clustering.

These A-clusters require more than a simple increase in average vulnerability to determine their meaning. A clear pattern we observe is that neighborhoods clustered in the all-wave approach seem to have an interaction with the S-clusters over time. It is here that we noticed our all-wave cluster approach seems to be detecting a mixture of strength of vulnerability as well as the stability of the S-clusters. To best describe this pattern, we create a plot that colors the A-clusters but plots their single-wave S-cluster behavior over the waves in figure 4. From this we are able to say that A-clusters 0 and 5 are the most stable, but 0 has the lowest vulnerability. As opposed to A-clusters 2 and 3 which are the most unstable and bounce around often, the difference being whether they generally end up in a lower vulnerability S-cluster or higher. See figure 5 for an example of what we consider an unstable cluster.

Unfortunately, due to the cost of projecting this data, we are not able to directly link the reasoning for our clustering to the original data, but we do have the ability to see which neighborhood

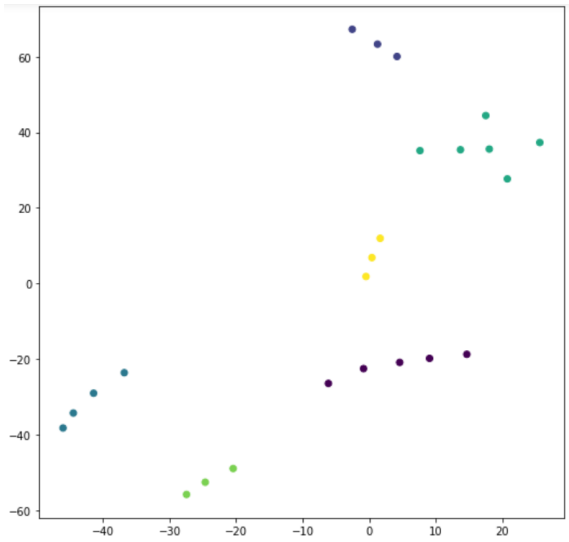

Figure 3: Clustering Over All Waves (t-SNE). Once again, the EDI data has been projected and clustered with a k-means approach to reveal 6 clusters.

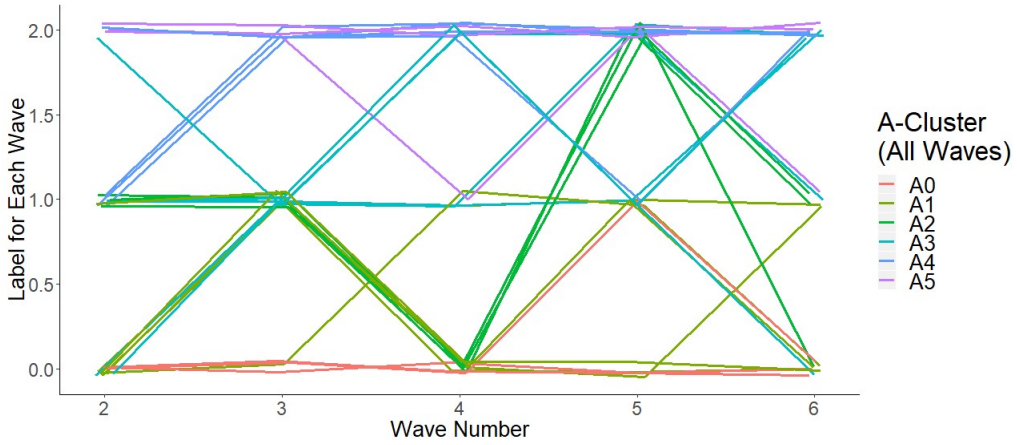

Figure 4: Neigborhoods' S-cluster over time and grouped by A-cluster. Each line represents a neighborhood,

ends in which cluster. Instead, we discuss our efforts to explain these clusters via socio-economic factors. We use [KF10] to help us determine what types of census variables in the Canadian 2016 Census could help us to discriminate between our different types of clusters. To determine if the variable is statistically significant between our clusters, we perform a one-way ANOVA test given we have sufficiently met the parametric assumptions (i.e homogeneity and normality). In the case where homogeneity of cluster variances fails, we instead perform the non-parametric Kruskal-Wallis test. We summarize a few significant variables below, but suggest for the reader to play with the data themselves within our interactive web dashboard.

3.2 Validation & Results

To validate that the clusters we had found are truly hidden structures within our dataset, we preform an average Hopkin's hypothesis test within our clusters, see [BD04] for more theory. For this set up, the hypothesis test is as follows:

$$\begin{aligned} H_0 &: \text{The clusters are reasonably random within their clusters, } H_{av} = 0.5, \\ H_a &: \text{These clusters have further substructure, } H_{av} \neq 0.5. \end{aligned} \tag{1}$$

It is worthwhile to note that to see a value of $H_{av} > .5$ generally indicates that a sub-cluster would exist whereas $H_{av} < .5$ indicates that the cluster itself is regularly spaced (not random, nor

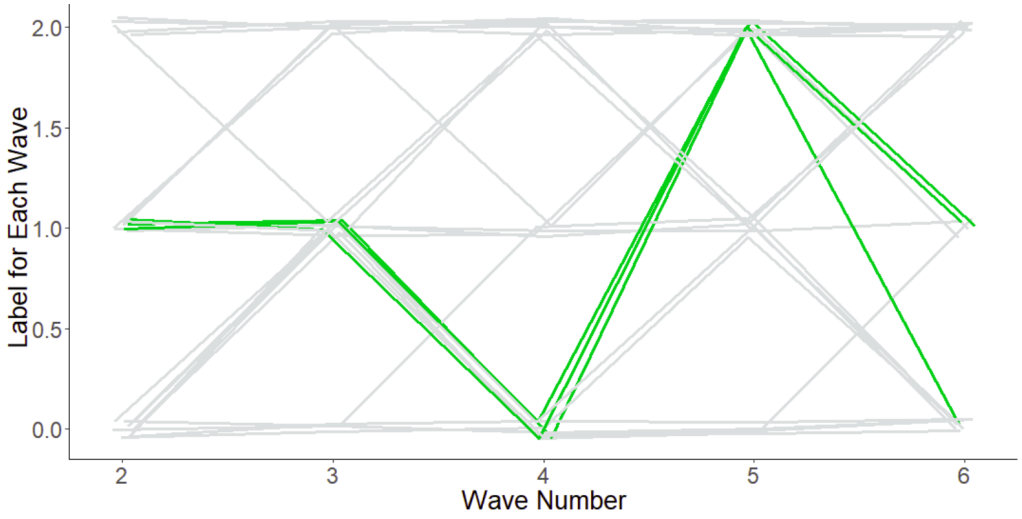

Figure 5: An example of one of the A-cluster's changes in S-Cluster over time. We consider A-cluster 2 as being the most unstable due to it's tendency to constantly change S-clusters at each wave.

sub-clustered). We conduct this test over each single-wave as well as over the all-wave approaches to verify our findings. The results can be found in table 1. Given that we stay within a reasonable range of $H_{av} = .5$ for the entirety of our clustering, we may conclude that what we have found meaningful hidden structure within the data. It may be noted that for waves 3 and 6, we are seeing some evidence of substructure, but with the minimal amount of data we do not consider further exploration here.

	S-clusters					A-clusters
Wave	2	3	4	5	6	All
H_{av}	0.4817	0.4327	0.4734	0.4759	0.5226	0.5051

Table 1: Average Hopkin's statistic over the t-SNE clusters.

As another means to verify our clusters are robust, we consider a different method entirely to project our data into visual space. We choose to use UMAP [MH18] which has the advantage of being much more mathematically rigorous as opposed to the engineered t-SNE. At a high-level, both of these methods roughly approach the problem of dimension projection in a similar way (graph-based), but UMAP has the additional benefit of preserving more global structure from our data than t-SNE does. We apply this method in a similar manor as we had with t-SNE, both across single-waves and over all-waves. For the resulting projection and clustering for the single-waves, we found no difference, see figure 6. Although to our surprise, there is quite a difference between the methods in the all-wave problem.

Recall we had understood t-SNE to capture both vulnerability as well as stability with the A-clusters, from UMAP we instead find 4 clusters, which we denote *UA-clusters*. Upon further inspection, we find that UMAP coalesces the t-SNE A-clusters 0 and 3 as well as 1 and 4 into single clusters. This is interesting for a variety of reasons, this indicates that when the t-SNE A-clusters and UMAP UA-clusters agree, this is truly hidden structure present in our data (i.e 2 and 5 with $H = 0.5706$ and 0.4311 respectively). Maybe more important, UMAP has combined A-clusters 0 and 3 which independently had $H = 0.4563$ and 0.4166 into a single UA-cluster with $H = 0.5023$. This has also occurred with A-clusters 1 and 4 which had $H = 0.5478$ and 0.6080 independently and together the UA-cluster has $H = 0.5308$. These values can be seen in tables 2 and 3. In both cases, we see the Hopkin's statistic of the combined UA-clusters being closer to $H = 0.5$ which indicates a better randomly distributed nature within the cluster. In essence, this means that UMAP has identified global behavior that we did not pick up through t-SNE and has refined our cluster analysis to find the most important groups. What remains to be understood is what exactly this global behavior is, whether that be minute differences between neighborhood EDI scales or a similarity between EDI scales themselves.

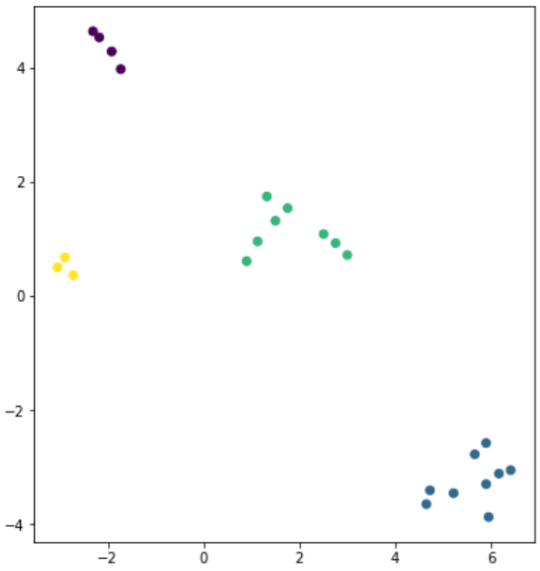

Figure 6: Clustering over all waves (UMAP). The EDI data has been projected in a different manor as to preserve global structure. This effectively combines certain A-clusters we previously had.

	t-SNE A-clusters					
Cluster	0	1	**2**	3	4	**5**
H	0.4563	0.5478	**0.5706**	0.4166	0.6080	**0.4311**

Table 2: Hopkin's statistic over the t-SNE all-wave clusters.

	UMAP UA-clusters			
Cluster	**0**	1	2	**3**
H	**0.5706**	0.5023	0.5308	**0.4311**

Table 3: Hopkin's statistic over the UMAP all-wave clusters.

From the one-way ANOVA we list off a few statistically significant census variables that offer some discrimination between the S-clusters in table 4, A-clusters in table 5, and UA-clusters in table 6. For the full set of significant census variables see appendix A. Note, we choose to run this test at the $\alpha = 0.05$ level and we also use this significance level to check the assumptions of normality and homogeneity as well, but in the interactive dashboard all of this can be adjusted to the user's input. This also is only a snapshot in time, we use only the 2016 Canadian census and thus these variables can pick up only information from 2011-2016. This means that only partially can we explain the differences between clusters for both the A-clusters and UA-clusters as these capture information from 2004-2016, but we demonstrate the idea for these clusters as well. We pick a wide spread of census variables that are indicative of how diverse and complex the regions our clusters act over. We emphasize a variable we expected to see as a discriminant: *Total Income of Households in 2015 (Median)*. There is an outstanding amount of literature to verify that income separates quality of life and we anticipate this to be a massive indicator of EDI vulnerability. Among other interesting variables found, we notice that *unemployment rate*, *use of transit*, *occupation*, and *inter-family relations* are common indicators across these clusters. We also choose to represent ethnic origins in this report but with a great sense of responsibility.

It is interesting that sometimes the percentile version of a variable appears significant while it's count does not. This is indicative that the population density of each neighborhood influences the spread of the variable. We believe this better reflects the true sizes of neighborhoods and distinguishes the larger but sparsely populated neighborhoods in southern Surrey. Furthermore, as evidence towards EDI as a useful metric and our clustering scheme capturing valuable information, we notice that no physical geography census variables were found to be significant discriminators between any cluster approach. We recap this in appendix A.

Disclaimer: The interpretation of our results can be lead awry if not handled properly. We

intend to represent ethnic origin as a discriminant of our cluster analysis to show that groups may be under-represented or not given the resources they need to prosper. This is not meant to be misconstrued to provide evidence of ethnic inequality or for victim blaming as this data cannot capture the full extent of these social issues. We urge the reader to be careful to note that this conclusion is purely to show that ethnic origin may be an indicator of EDI vulnerability. With this in mind, we can lead to growing the network of resources available to all people of Surrey.

S-cluster significant census variables	
Total Income of Households in 2015 (Median)	Unemployment Rate
Renters	Non-Permanent Residents
Native Tongue – English and Non-Official Language	Management Occupations
People of African Origins	People of West and Central Asian and Middle Eastern Origins

Table 4: An assortment of significant census variables for the 3 S-clusters in 2016.

A-cluster significant census variables	
Total Income of Households in 2015 (Median)	Male Unemployment Rate
Employed that use Transit	Production Occupations
Native Tongue – Hindi	Immigrants from Oceania and Other
People of European Origins	Lone Parent (%)

Table 5: An assortment of significant census variables for the 6 A-clusters.

UA-cluster significant census variables	
Total Income of Households in 2015 (Median)	Female Unemployment Rate
Employed that Commutes for over 60 Minutes	Art/Sport Occupations
Native Tongue – Punjabi	Immigrants
People of South Asian Origins	Married (%)

Table 6: An assortment of significant census variables for the 4 UA-clusters.

4 Bottom-Up: Understanding City Program Reach

The availability of CLASS presented an opportunity to build predictive modeling around children registering for programs. A question that lingered from the previous year was whether CLASS could help predict when a child might not return to re-register for any affiliated programs [EPPS]. A critical age for children being introduced to programs run by the City of Surrey and Children's Partnership was defined between the ages of 8-10. Retention of children was then defined as how long could a program have children continue to register for programs every year throughout early and middle childhood development. Retaining children during these critical development periods could then help inform the impact organizations might have on EDI as well as the Middle Development Instrument (MDI).

While summary statistics and visualizing the data provided unique insights, we describe here the challenges of trying to use CLASS for predictive modeling or other machine learning techniques. We also describe the efforts we undertook to link our top-down approach to the analysis done in CLASS.

4.1 Retrieval of Data

CLASS Data was retrieved using a PostgreSQL database. Only clients whose accounts were created after the 01/01/2000, and whose birth years were later than 01/01/2000 were chosen for analysis.

This was done due to large inconsistencies in data records before the year 2000. Programs that were also selected must have had a maximum number of registrants greater than 1, in order to avoid selecting programs that were not community inclusive. All registration records that were selected were of records that indicated a child had succesfully completed the course, and was not withdrawn before hand. A data table was created in which each client ID was associated with a registration ID, as well as a course ID. In this manner, we could analyze the first and last program each child would have registered in, as well as the length of time they were retained by the Children's Partnership network.

4.2 Program Activities and Grouping

Each program or course is associated with both a title and a subtitle to describe the program in-depth. Due to this, there are 237 unique program titles and descriptors. We grouped together activities, in order to pool granular data to ensure that meaningful associations could be drawn more conclusively after analysis. Figure in B details the decision making process to arrive at the 8 total groups of activities used in analysis.

General Activities mainly consistes of activities describes in the database as general interest, computer literacy, personal development, and social recreation.

4.3 Summary Statistics

In total, 62313 unique registrants were identified, with nearly 2000 more males included than females. A distribution of neighborhood representation in this dataset is exhibited in Figure 7 below. Four neighborhoods (South Surrey West, Newton East, Cloverdale South, and Surrey City Centre) account for just under 50% of the data.

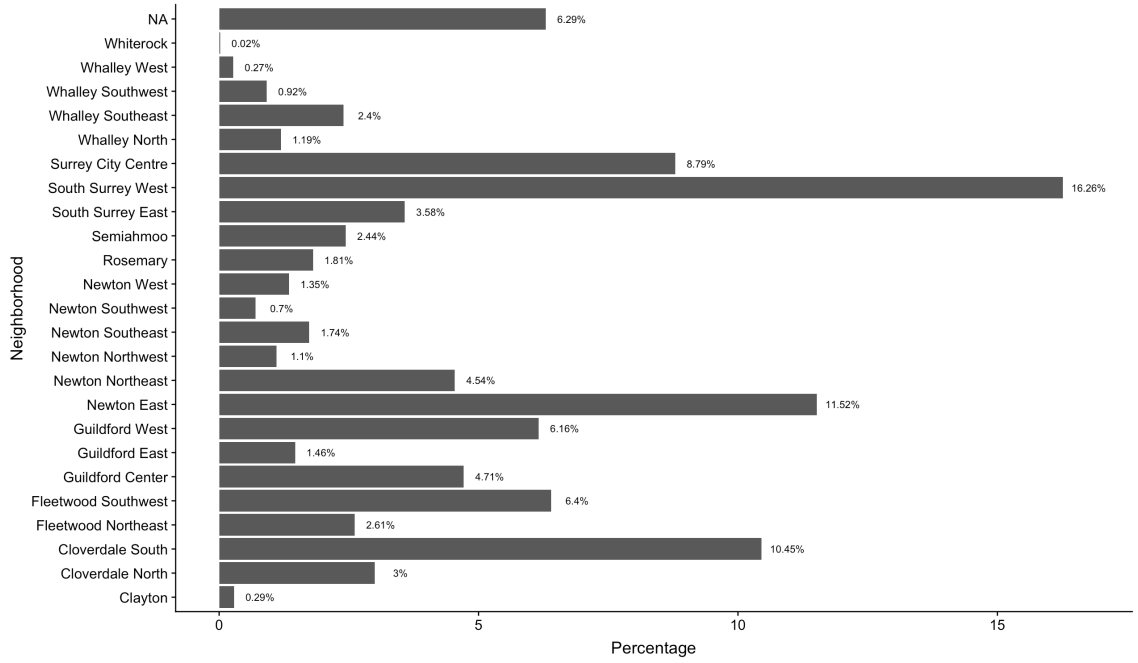

Figure 7: Proportion of neighborhoods represented in children born and registering in programs after 01/01/2000. A neighborhood of NA indicates no area was designated to the child within the dataset or none was specified.

Figure 8 displays the distribution of entry and exit ages for children within the dataset. The exponential decay like distrbution for the ages of entry suggests that the majority of children first introduced to Children's Partnership and other organizations are in the early phases of child development.

However, the bell-like distribution of the exit ages suggests a critical age of retention between the ages of 7-9. Here, a majority of children seem to leave Children's Partnership just after Kindergartern through to second grade.

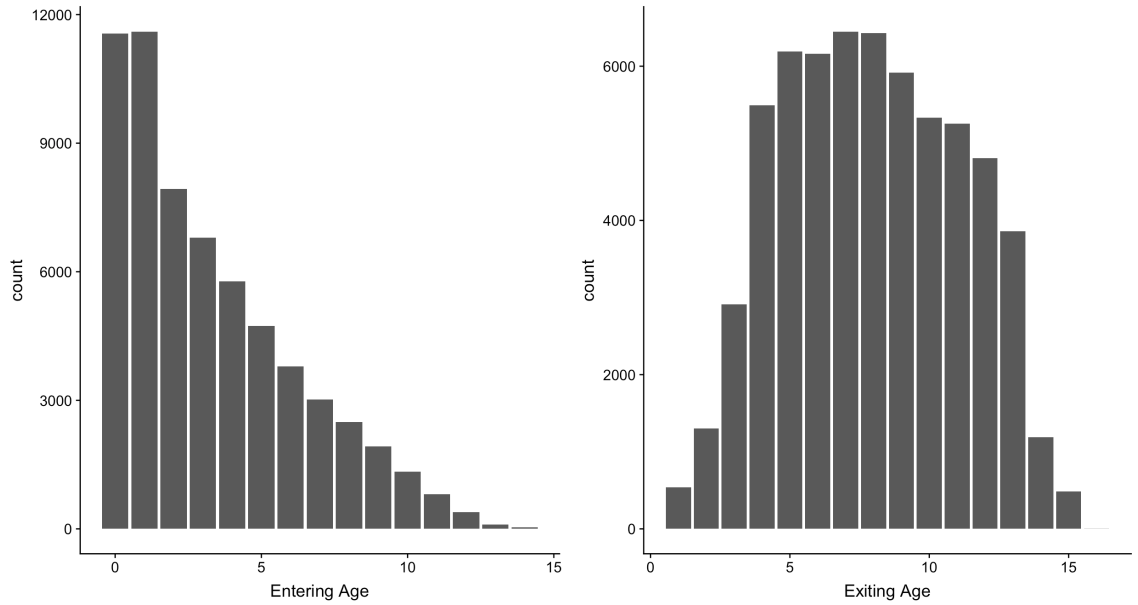

Figure 8: Distribution of a child's age at the time of registering for their first program (left). Distribution of a child's age of last registration (right). Count denotes total number of children.

Identifying potential differences among male and females in regards to entering ages, as well as the their entering program (the program type they first register in) was an important next step in analyzing the data. Figure 9 shows that no significant differences exist between male and female registrants depending on the program type they are first introduced too, suggesting that male and female children entering the Children's Partnership network have relatively similar experiences at the start of their journeys through programs offered by the organization.

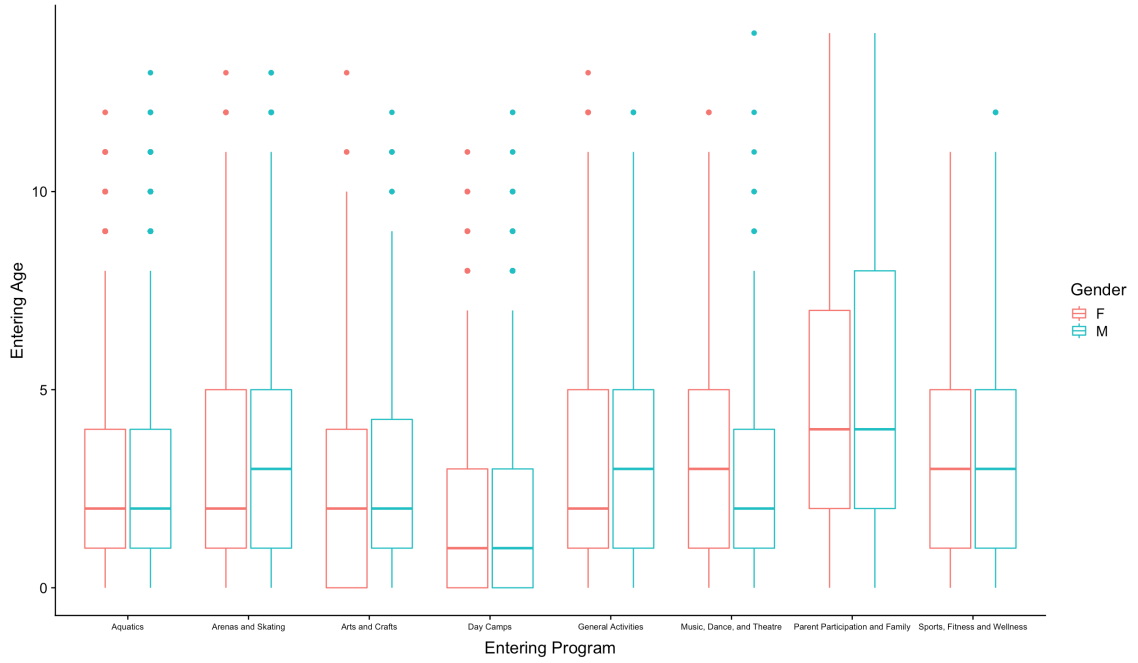

Figure 9: Boxplot of entering ages for male and female children for each program type from 2000-2017.

While male and female program entry might be similar, season effects could be of value to understanding entry age by program type. Within the CLASS Dataset, four period of seasonal registration periods are associated with each course offering (Winter, Spring, Summer and Fall).

Figure 10 shows that General Activities as an entry program increases for all ages during the Fall Registration period. While Day Camp registration increases for all ages during the Summer period, this is largely intuitive given the academic calendar for all schools in Surrey.

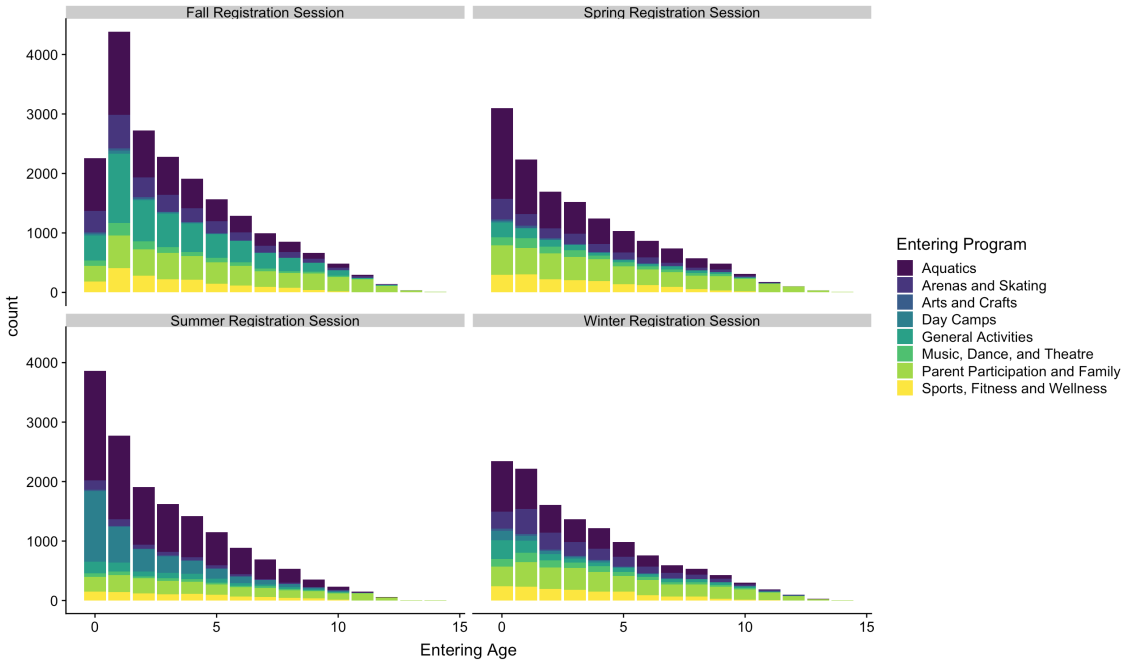

Figure 10: Registration season by entering age and stratified by program type. We see Day Camps increase signficiantly during Summer registration periods, while General Activities increase during the Fall. Count denotes total number of children within the data table.

4.4 General Activities and Retention

Distribution of exiting ages based on program types could help point to potential indicators of child retention within the network. Figure 11 shows interesting results when it comes to comparing the program type of General Activities to other programs. While most programs show a bell-like distribution curve for age of exit, General Activity programs display a more bimodel distribution, showing a majority of children exiting after the age of 10. Male and Female proportions within this program type are also fairly equal. Parent Participation programs are the other anomaly within this figure, with virtually no child exiting after the age of 5. However, this can more than likely be explained by the fact that most of these programs are geared towards early childhood development only. Finally Music, Dance, and Theatre program types are the only sub-category of programs that have different age of exit distributions for male and female children. Female children seem to exit primarily before the age of 7, while male children are more normally distributed (though they are far fewer in number).

To investigate further whether the length of retention might be longer for children exiting from General Activities than from other program types, we visualized years of stay within the program as proportions of total children exiting by program type (Figure 12). We see here that General Activities enjoys the largest proportion of children exiting after 7 or more years of registering within programs associated with the City of Surrey and Children's Partnership.

4.5 Challenges with Predictive Modeling

While the aim of collecting the data table initially was in the hopes of building regression models and other machine learning models to predict when a child might leave the network, fundamentally a child's last program registration does not conclusively mean that the child would not return to the network. The data queried from the CLASS dataset encompassed children registration information from 2000-2017, and while most children whose last program may have been in 2010 would more likely never return, the cutoff would be difficult to delineate for children whose last program came

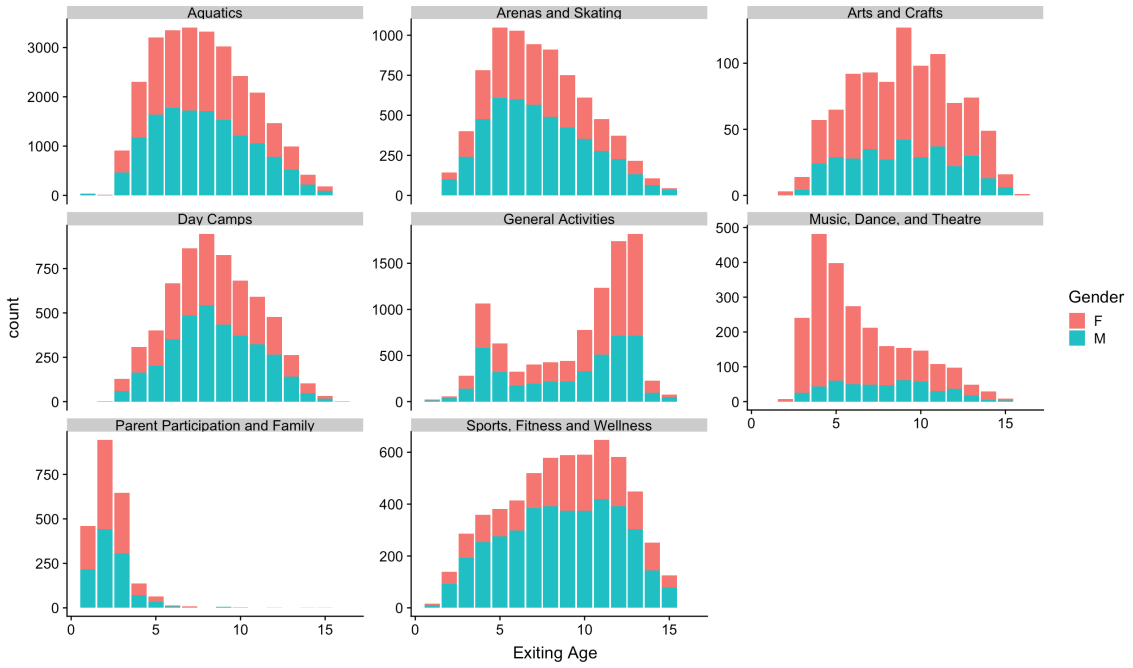

Figure 11: Exiting age by program type, stratified by gender. Count denotes total number of children within the data table. General Activities presents a much more bimodal distribution than the rest.

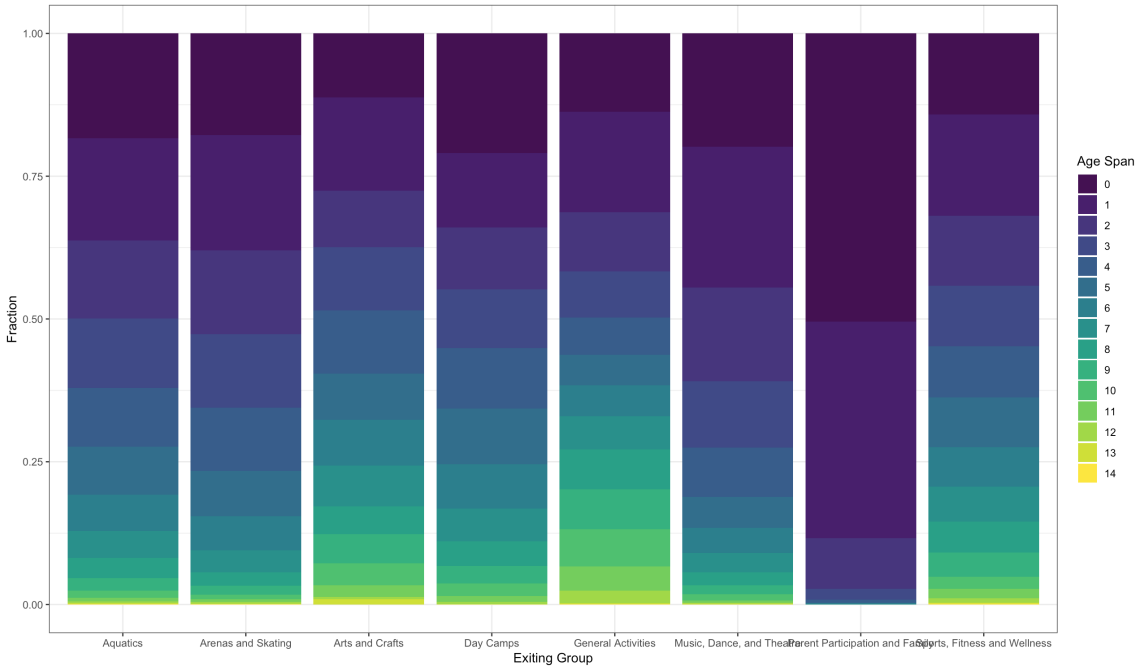

Figure 12: Proportion of age spans for each exiting program type. Age spans denote length of years between first program registration and last program registration for any given child within the data table.

between 2015-2017. Due to this irregularity in the modeling hypothesis and the true representation of the data, no machine learning models were used within CLASS.

5 Discussion

Through the use of a two-pronged approach, we attempted to understand the lived experiences of children in the city of Surrey through several lenses. Through the use of clustering around the EDI, we have good evidence to suggest that neighborhoods that enjoy lower or higher vulnerabilities on average may have similar traits that might be elucidated through Census variables.

In the case of understanding children and their relationship to organizations such as Children's Partnership, our surface-level understanding of the CLASS dataset suggests that programs that are more likely to involve Parents, creativity, and enhanced social engagement outside of sports or a competitive environment might be better positive indicators for a child being retained past the critical age within the program network. This is important due to the fact that organizations can then prioritive these programs further in neighborhoods that are experiencing higher rates of childhood vulnerabilities across a number of different scales.

5.1 Linking Approaches

While the city of Surrey represents a complex adaptive system, in which many different factors probably contribute to the rise of childhood vulnerability, we investigates whether any linear relationship might exist among neighborhoods that have higher rates of children enrolled in General Activities programs and EDI. Neither clustered neighborhoods, nor EDI scores correlated strongly with General Activities enrollment. the Middle Development Index (MDI) was also used to try and resolve clusters as well as their General Activities program enrollment. However, this method did not yield any statisticially significant results.

5.2 Web Application Development

As for one of the foundations on which this project may continue, we have designed a deployable web application. This application has many useful features:

- EDI Tab: Visualize the EDI data across the city of Surrey and see simple trends down to neighborhood level. The user may change wave as well as what scale they would like to visualize.
- Cluster Tab: Perform the entire cluster analysis and visualize this across the city of Surrey. The choice in using t-SNE or UMAP can be made and the user then has the ability to choose census variables to see the spread across each cluster. The application will also recommend which variables may be interesting to look at and then provide the results of a ANOVA test.
- CLASS Tab: Sensitive data can be uploaded according to a specific data format to visualize the flow of children through the city of Surrey.

All together, this application allows for the user to conduct their own exploratory analysis with smart suggestions. We believe the value in allowing for a dynamic analysis is more facilitative towards understanding the data the city has. Furthermore, this application has been built in mind that the EDI datasets and CLASS datasets may be updated over time to re-conduct the analysis we have preformed throughout this paper. We claim that this allows further research to be done on top of our current platform and stronger results can be drawn with more data. You may find our application currently deployed on Cody's Shiny server.

6 Conclusion & Future Work

There do exist alternatives to our projection methods, for example hierarchical clustering could be done on the original data. Although we did consider and initially implement this approach, see Figure 13, we are not skilled in distinguishing between clusters and do not have the expertise to fine tune this algorithm. Further study into the alternatives could warrant an entire paper on their own.

We see value in pursuing the hierarchical clustering approach as this method avoids any error induced from projecting data into a visual spectrum. A decision boundary that indicates what defines a cluster also may be learned as to find the regions of EDI that define a cluster. This is

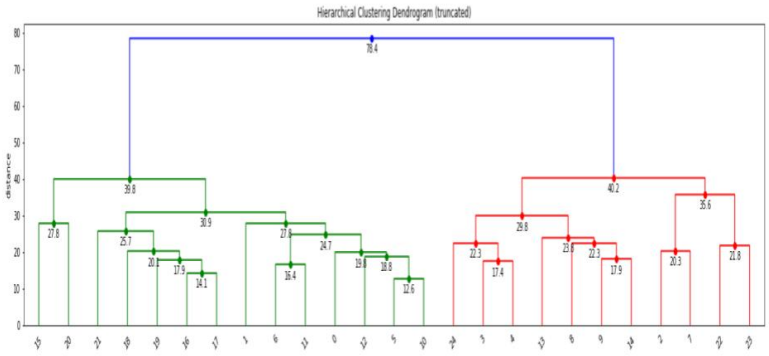

Figure 13: An example of a choice of hierarchical clustering. The y-axis represents the distances between points and the x-axis is just a numeric representation of the neighborhoods in Surrey. Depending on your interpretation, we could see as few as two clusters (colored green and red) or as many as 12.

incredibly useful for finding which scale of the EDI is most influential for each cluster and can directly suggest what programs Children's Partnership should improve upon.

We also had the opportunity to briefly look at the sub-scale data that define the EDI scales. We believe our approach in considering single-wave and all-waves to find structure could still be useful here to draw conclusions on what programs Children's Partnership. The difficulty in this direction is that we do not currently know what weights each sub-scale has towards their respective scale. We advise caution in naively treating them as equal as this may lead to false clusters forming where these are constructs rather than true hidden structure. With a proper approach, these conclusions can add much value to the results already presented in this paper.

References

[BD04] Amit Banerjee and Rajesh N Dave. Validating clusters using the hopkins statistic. In *Fuzzy systems, 2004. Proceedings. 2004 IEEE international conference on*, volume 1, pages 149–153. IEEE, 2004.

[EPPS] E.Gomez, P.Angkiriwang, P.Laflamme, and S.Pan. A data-driven approach to early childhood intitiatives. not submitted.

[IB+14] Moira Inkelas, Patricia Bowie, et al. The magnolia community initiative: The importance of measurement in improving community well-being. *Community Investments*, (01):18–24, 2014.

[KF10] Paul Kershaw and Barry Forer. Selection of area-level variables from administrative data: an intersectional approach to the study of place and child development. *Health & place*, 16(3):500–511, 2010.

[MH18] Leland McInnes and John Healy. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[OMM99] Early Years Study Ontario, Margaret Norrie McCain, and James Fraser Mustard. *Reversing the real brain drain: Early years study*. Canadian Institute for Advanced Research= Institut canadien de recherches avancées, 1999.

A Census Variables

A.1 All Census Variables

There are a total of 147 census variables chosen with [KF10] in mind. These variables range in terms of overall categories they belong to and we have tried to choose variables that would be most meaningful and discriminate the most between our clusters. The variables are counts unless specified. The percentages (%) are among the population unless otherwise specified.

- **Geography**

1. Shape Area
2. Dwellings
3. Households
4. Population
5. Area (sq km)

- **Ethnic Origins**

1. People of Aboriginal Origins
2. People of European Origins
3. People of West and Central Asian and Middle Eastern Origins
4. People of South Asian Origins
5. People of East and Southeast Asian Origins
6. People of Latin Central and South American Origins
7. People of African Origins
8. Total Population with Ethnic Origin Data for Private Households
9. People of Aboriginal Origins (%)
10. People of European Origins (%)
11. People of West and Central Asian Origins (%)
12. People of South Asian Origins (%)
13. People of East Southeast Asian Origins (%)
14. People of Latin Central and South American Origins (%)
15. People of African Origins (%)
16. Other Origins (%)

- **Language and Immigration**

1. Native Tongue – English
2. Native Tongue – Aboriginal Languages
3. Native Tongue – Chinese Languages
4. Native Tongue – Punjabi
5. Native Tongue – Hindi
6. Native Tongue – Tagalog
7. Native Tongue – English and Non-Official Language
8. Non-Immigrants
9. Non-Permanent Residents
10. Immigrants
11. Immigrants from the Americas
12. Immigrants from Europe
13. Immigrants from Africa
14. Immigrants from Asia
15. Immigrants from Oceania and Other
16. Native Tongue – English (%)
17. Native Tongue – Aboriginal Languages (%)
18. Native Tongue – Chinese Languages (%)
19. Native Tongue – Punjabi (%)
20. Native Tongue – Hindi (%)
21. Native Tongue – Tagalog (%)
22. Native Tongue – English and Non-Official Language (%)
23. Immigrants (%)
24. Non-Immigrants (%)
25. Non-Permanent Residents (%)
26. Immigrants from the Americas (% of immigrants)
27. Immigrants from Europe (% of immigrants)
28. Immigrants from Africa (% of immigrants)
29. Immigrants from Asia (% of immigrants)
30. Immigrants from Oceania and Other (% of immigrants)

- **Income**

1. Total Income of Households in 2015 (Median)
2. Government Transfers Recipients in Private Households
3. Amount of Government Transfers Recipients in Private Households (Median)
4. Income Recipients in Private Households
5. Male Income Recipients in Private Households
6. Female Income Recipients in Private Households
7. Income Amoung Recipients (Median)
8. Income Amoung Male Recipients (Median)
9. Income Amoung Female Recipients (Median)
10. Composition of Income from Government Transfers (% of income)
11. Income of Couple Economic Families with Children (Median)
12. Income of Couple Economic Families without Children (Median)
13. Income of Lone Parent Economic Families (Median)
14. Economic Families' Income in the Bottom Decile
15. Income Recipiants in Private Housholds (%)
16. Government Transfers Recipiants in Private Housholds (%)
17. Income Recipiant Male/Female Ratio
18. Economic Families' Income in the Bottom Decile (% of economic families)

- **Cost of Living**

1. Rooms per Dwelling (Mean)
2. Owner Households Spending 30% or more of Income on Shelter Costs
3. Owner Households Spending 30% or more of Income on Shelter Costs (% of owners)
4. Monthly Shelter Costs for Owned Dwellings (Median)
5. Monthly Shelter Costs for Owned Dwellings (Mean)
6. Value of Dwellings (Median)
7. Value of Dwellings (Mean)
8. Tenant Households Spending 30% or more of Income on Shelter Costs (% of tenants)
9. Monthly Shelter Costs for Rented Dwellings (Median)
10. Monthly Shelter Costs for Rented Dwellings (Mean)

- **Employment**

1. Labour Force
2. Male Labour Force
3. Female Labour Force
4. Employment Rate
5. Male Employment Rate
6. Female Male Employment Rate
7. Unemployment Rate
8. Male Unemployment Rate
9. Female Unemployment Rate
10. Not in the Labour Force
11. Males not in the Labour Force
12. Females not in the Labour Force
13. Commute Duration
14. Employed that Commutes for over 60 Minutes
15. Employed that use Transit
16. Labour Force (%)
17. Labour Force Male/Female Ratio
18. Not in the Labour Force Male/Female Ratio
19. Employment Rate Male/Female Ratio
20. Unemployment Rate Male/Female Ratio
21. Employed that Use Transit (% of employed population)

- **Occupation**

1. Management Occupations
2. Finance Occupations
3. Science Occupations
4. Health Occupations
5. Liberal Arts Occupations
6. Art/Sport Occupations
7. Sales and Service Occupations
8. Transit/Industrial Occupations
9. Production Occupations
10. Manufacturing Occupations
11. All Occupations
12. Management Occupations (% of all occupations)
13. Finance Occupations (% of all occupations)
14. Science Occupations (% of all occupations)
15. Health Occupations (% of all occupations)
16. Liberal Arts Occupations (% of all occupations)
17. Art/Sport Occupations (% of all occupations)
18. Sales and Service Occupations (% of all occupations)
19. Transit/Industrial Occupations (% of all occupations)
20. Production Occupations (% of all occupations)
21. Manufacturing Occupations (% of all occupations)

- **Population**

1. Private Dwellings
2. Private Dwellings Occupied by Usual Residents
3. Population Density (per sqr km)
4. Population in 2011
5. Private Households from Tenure Data
6. Owners
7. Renters
8. Total Number of Census Families in Private Households
9. Total Couple Families
10. Total Lone Parent Families by Sex of Parent
11. Female Parent
12. Male Parent
13. Couples without Children
14. Couples with Children
15. Average Size of Census Families
16. Married
17. Economic Families
18. Private Dwellings Occupied by Usual Residents (% of private dwellings)
19. Renter/Owner Ratio
20. Couple with Children (% of census families)
21. Male Lone Parent (% of census families)
22. Female Lone Parent (% of census families)
23. Lone Parent (% of census families)
24. Married (% of census families)
25. Couple With/Without Child Ratio
26. Lone Parent Male/Female Ratio

A.2 Significant 2016 S-cluster Census Variables

These are the 41 significant census variables for the wave 6 S-clusters. We note that none are from the geography category.

- **Ethnic Origins**

1. People of West and Central Asian and Middle Eastern Origins
2. People of African Origins

- **Language and Immigration**

1. Native Tongue – Hindi
2. Native Tongue – Tagalog
3. Native Tongue – English and Non-Official Language
4. Non-Permanent Residents
5. Immigrants

6. Immigrants from Asia
7. Immigrants from Oceania and Other
8. Non-Permanent Residents (%)

- **Income**

1. Total Income of Households in 2015 (Median)
2. Income Amoung Recipients (Median)
3. Income Amoung Male Recipients (Median)
4. Income Amoung Female Recipients (Median)
5. Composition of Income from Government Transfers (%)
6. Income of Couple Economic Families with Children (Median)
7. Income of Couple Economic Families without Children (Median)
8. Economic Families' Income in the Bottom Decile
9. Income Recipiant Male/Female Ratio
10. Economic Families' Income in the Bottom Decile (%)

- **Cost of Living**

1. Owner Households Spending 30% or more of Income on Shelter Costs (%)

- **Employment**

1. Unemployment Rate
2. Male Unemployment Rate
3. Employed that Commutes for over Minutes
4. Employed that use Transit
5. Employed that Use Transit (%)
6. Not in the Labour Force Male/Female Ratio

- **Occupation**

1. Management Occupations
2. Art/Sport Occupations
3. Sales and Service Occupations
4. Transit/Industrial Occupations
5. Manufacturing Occupations
6. Management Occupations (%)
7. Art/Sport Occupations (%)
8. Sales and Service Occupations (%)
9. Transit/Industrial Occupations (%)
10. Manufacturing Occupations (%)

- **Population**

1. Renters
2. Female Lone Parent (%)
3. Lone Parent (%)
4. Renter/Owner Ratio

A.3 Significant A-cluster Census Variables

These are the 58 significant census variables for the A-clusters. Once again, note that geography is not present either as a significant factor in cluster separation.

- **Ethnic Origins**

1. People of European Origins
2. People of South Asian Origins
3. People of Aboriginal Origins (%)
4. People of Euroupean Origins (%)
5. People of South Asian Origins (%)

- **Language and Immigration**

1. Native Tongue – English
2. Native Tongue – Punjabi
3. Native Tongue – Hindi
4. Native Tongue – English and Non-Official Language
5. Non-Permanent Residents
6. Immigrants
7. Immigrants from Asia
8. Immigrants from Oceania and Other
9. Native Tongue – English (%)
10. Native Tongue – Punjabi (%)
11. Native Tongue – Hindi (%)
12. Native Tongue – English and Non-Official Language (%)
13. Immigrants (%)
14. Non-Immigrants (%)
15. Non-Permanent Residents (%)
16. Immigrants from the Americas (%)
17. Immigrants from Europe (%)
18. Immigrants from Africa (%)
19. Immigrants from Asia (%)
20. Immigrants from Oceania and Other (%)

- **Income**

1. Total Income of Households in 2015 (Median)
2. Income Amoung Recipients (Median)
3. Income Amoung Male Recipients (Median)
4. Income Amoung Female Recipients (Median)
5. Composition of Income from Government Transfers (%)
6. Income of Couple Economic Families with Children (Median)
7. Income of Couple Economic Families without Children (Median)
8. Government Transfers Recipiants in Private Housholds (%)
9. Income Recipiant Male/Female Ratio
10. Economic Families' Income in the Bottom Decile (%)

- **Cost of Living**

1. Owner Households Spending 30% or more of Income on Shelter Costs (%)

- **Employment**

1. Unemployment Rate
2. Male Unemployment Rate
3. Female Unemployment Rate
4. Employed that use Transit
5. Labour Force Male/Female Ratio
6. Not in the Labour Force Male/Female Ratio
7. Employed that Use Transit (%)

- **Occupation**

1. Management Occupations
2. Art/Sport Occupations
3. Sales and Service Occupations
4. Transit/Industrial Occupations
5. Production Occupations
6. Manufacturing Occupations
7. Management Occupations (%)
8. Art/Sport Occupations (%)
9. Sales and Service Occupations (%)
10. Transit/Industrial Occupations (%)
11. Production Occupations (%)
12. Manufacturing Occupations (%)

- **Population**

1. Renter/Owner Ratio
2. Female Lone Parent (%)
3. Lone Parent (%)

A.4 Significant UA-cluster Census Variables

These are the 64 significant census variables for the UA-clusters. We note once more that geography is not present as a significant factor in cluster separation.

- **Ethnic Origins**

1. People of European Origins
2. People of West and Central Asian and Middle Eastern Origins
3. People of South Asian Origins
4. People of Euroupean Origins (%)
5. People of West and Central Asian Origins (%)
6. People of South Asian Origins (%)
7. People of African Origins (%)

- **Language and Immigration**

1. Native Tongue – English
2. Native Tongue – Punjabi
3. Native Tongue – Hindi
4. Native Tongue – English and Non-Official Language
5. Non-Permanent Residents
6. Immigrants
7. Immigrants from Asia
8. Immigrants from Oceania and Other
9. Native Tongue – English (%)
10. Native Tongue – Punjabi (%)
11. Native Tongue – Hindi (%)
12. Native Tongue – Tagalog (%)
13. Native Tongue – English and Non-Official Language (%)
14. Immigrants (%)
15. Non-Immigrants (%)
16. Non-Permanent Residents (%)
17. Immigrants from the Americas (%)
18. Immigrants from Europe (%)
19. Immigrants from Africa (%)
20. Immigrants from Asia (%)

- **Income**

1. Total Income of Households in 2015 (Median)
2. Income Amoung Recipients (Median)
3. Income Amoung Male Recipients (Median)
4. Income Amoung Female Recipients (Median)
5. Composition of Income from Government Transfers (%)
6. Income of Couple Economic Families with Children (Median)
7. Income of Couple Economic Families without Children (Median)
8. Economic Families' Income in the Bottom Decile
9. Government Transfers Recipiants in Private Housholds (%)
10. Income Recipiant Male/Female Ratio
11. Economic Families' Income in the Bottom Decile (%)

- **Cost of Living**

1. Owner Households Spending 30% or more of Income on Shelter Costs (%)
2. Renter/Owner Ratio
3. Labour Force Male/Female Ratio
4. Not in the Labour Force Male/Female Ratio
5. Employed that Use Transit (%)

- **Employment**

1. Unemployment Rate
2. Male Unemployment Rate
3. Female Unemployment Rate
4. Employed that Commutes for over 60 Minutes
5. Employed that use Transit

- **Occupation**

1. Management Occupations
2. Art/Sport Occupations
3. Sales and Service Occupations
4. Transit/Industrial Occupations
5. Production Occupations
6. Manufacturing Occupations
7. Management Occupations (%)
8. Art/Sport Occupations (%)
9. Sales and Service Occupations (%)
10. Transit/Industrial Occupations (%)
11. Production Occupations (%)
12. Manufacturing Occupations (%)

- **Population**

1. Renters
2. Female Lone Parent (%)
3. Lone Parent (%)
4. Married (%)

B Grouping Methodology

B.1 Flow Chart for Grouping

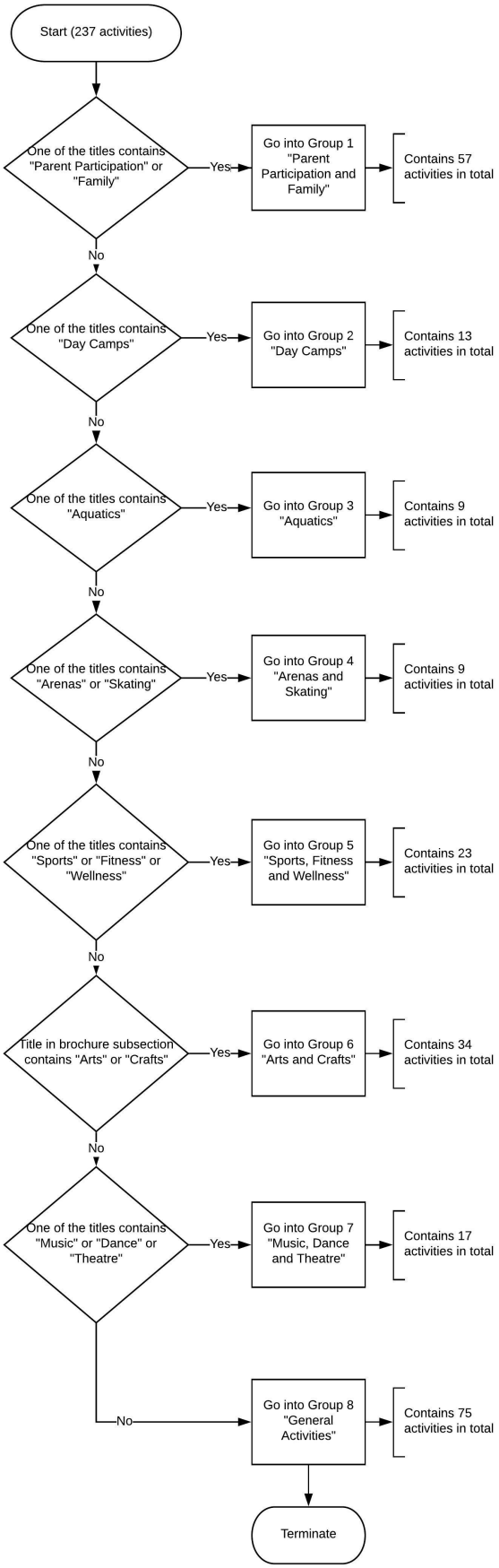

Figure 14: A Flow Chart describing the grouping methodology.