Understanding Childhood Vulnerability in the City of Surrey

Varoon Mathur, Cody Griffith, Catherine Lin, Kevin Zhu

Overview

- Introduction
- Datasets
- *Top-Down: Understanding Trends of Neighborhoods*
- *Bottom-Up: Understanding City Program Reach*
- Web Application
- Conclusion and Future work

Introduction

Understanding the **community conditions that best support universal access and improved childhoods outcomes** allows ultimately to improve decision making in the areas of planning, and investing across the early and middle years of childhood development.

How do we measure this?

Early Development Instrument (EDI)

Ex. Questions for Preschool (Age 4-5) Teachers:

- Is a child ***too tired or sick*** to do school work?
- Would you say this child ***demonstrates respect*** *for other children?*

Source: Vulnerability of the EDI, The Human Early Learning Partnership

- **Language**
- **Social**
- **Emotional**
- **Physical**
- **Communication**

W2: 2004-2007	**W3: 2007-2009**	**W4: 2009-2011**	**W5: 2011-2013**	**W6: 2013-2016**

Two Approaches: Top-Down and Bottom-Up

Top-Down: Holistic Measures of Neighborhood Success in Childhood Development

- **Motivated** to understand factors that might correlate with EDI Scores across neighborhoods (and therefore childhood vulnerability)
- **Do neighborhoods that have similar EDI Scores across years (waves) behave the same?**

Bottom-Up: Granular analysis of City-wide Program Usage and Registration Data

- **Motivated** to utilize city-wide data that might better represent lived-experiences of children living in Surrey
- **Can program/resource utilization trends by families be used as an indicator for childhood vulnerability?**

Datasets used

Open Source Datasets

- Early Development Instrument (EDI) provided by UBC's Human Early Learning Partnership (HELP) for the City of Surrey
- Statistics Canada 2016 Census Data (retrieved through cansim R Package)

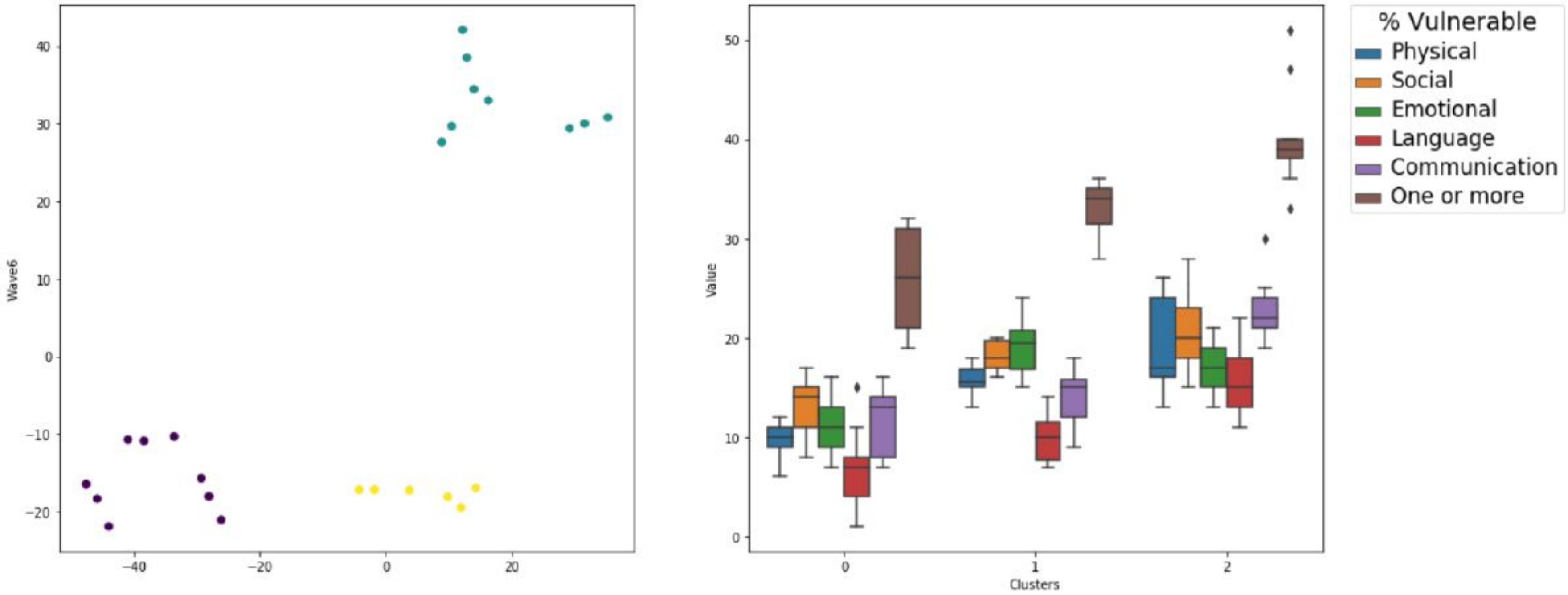
Private Dataset from Surrey

- CLASS Dataset (160Gb)

  Private Dataset - Provided by City of Surrey's Community and Recreation Services (CRS) division

Clustering Neighborhoods based on EDI Scores

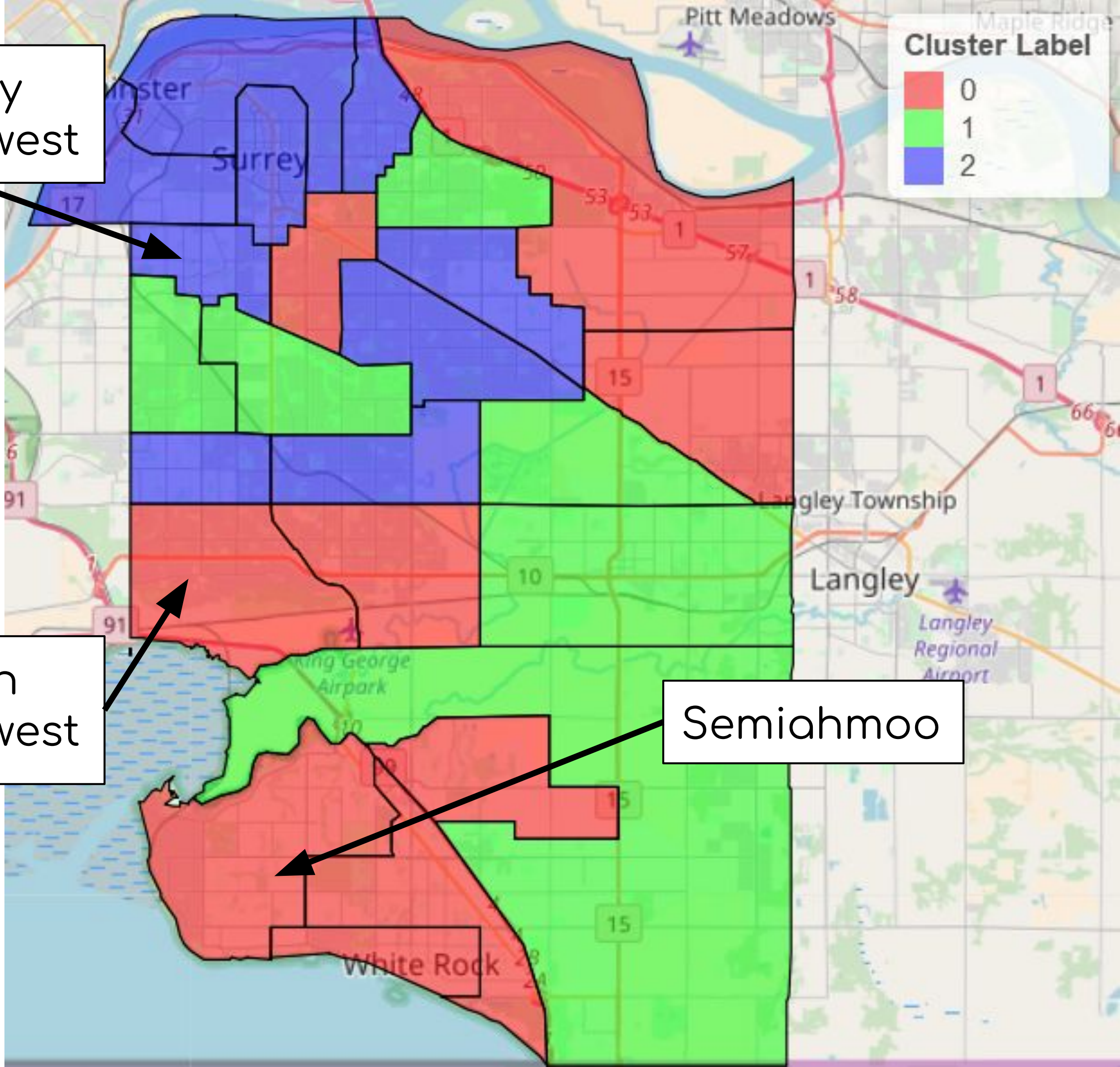
Single Wave Clusters (t-SNE) for Wave 6



Key Takeaway: t-SNE Approach shows good separation amongst all three clusters for every scale of EDI

Single-Wave Cluster Analysis

- **Want**: Similarities between neighborhoods for each wave
- **What Worked**: t-SNE
- **What Didn't**: PCA, KPCA, Hierarchical clustering



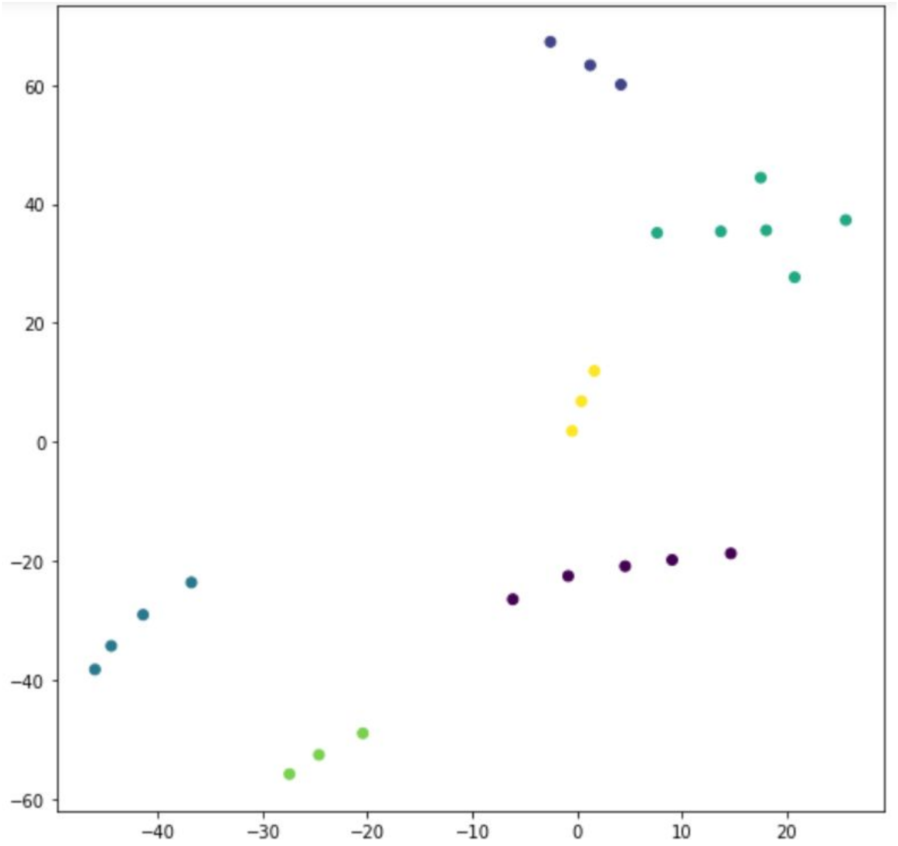
Clusters:

S0 - **low vulnerability (Good)**

S1 - **medium vulnerability (Meh)**

S2 - **high vulnerability (BAD)**

Clustering Over All Waves (t-SNE)



Key Takeaway:

t-SNE Approach incorporating all Waves of the EDI show six distinct Clusters.

All-Wave Cluster Analysis (t-SNE)

6 groups instead of 3:

A0 - Stable and low vulnerability (Good)

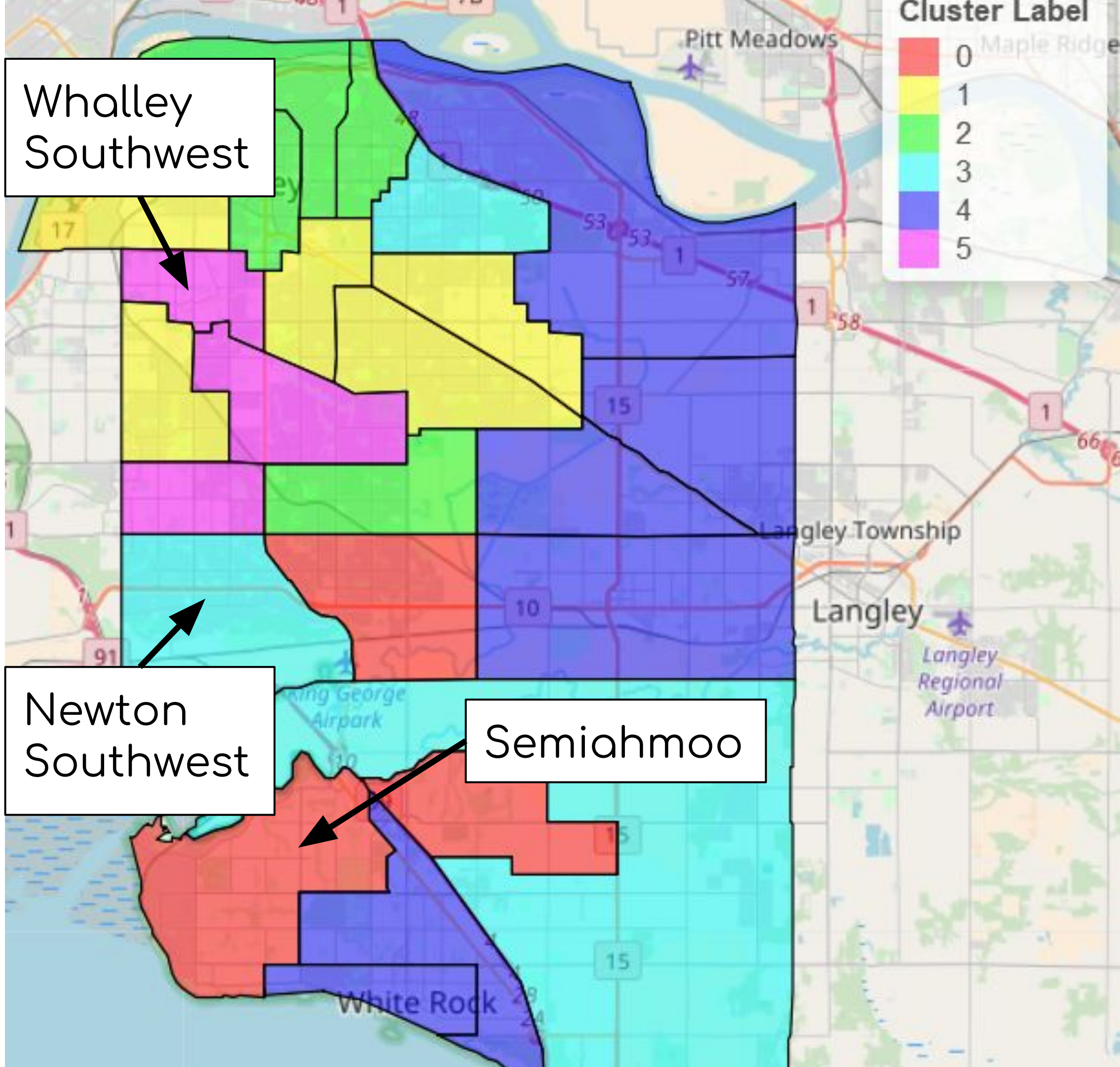
A1 - Unstable and low vulnerability (OK)

A2 - Unstable and medium vulnerability (?)

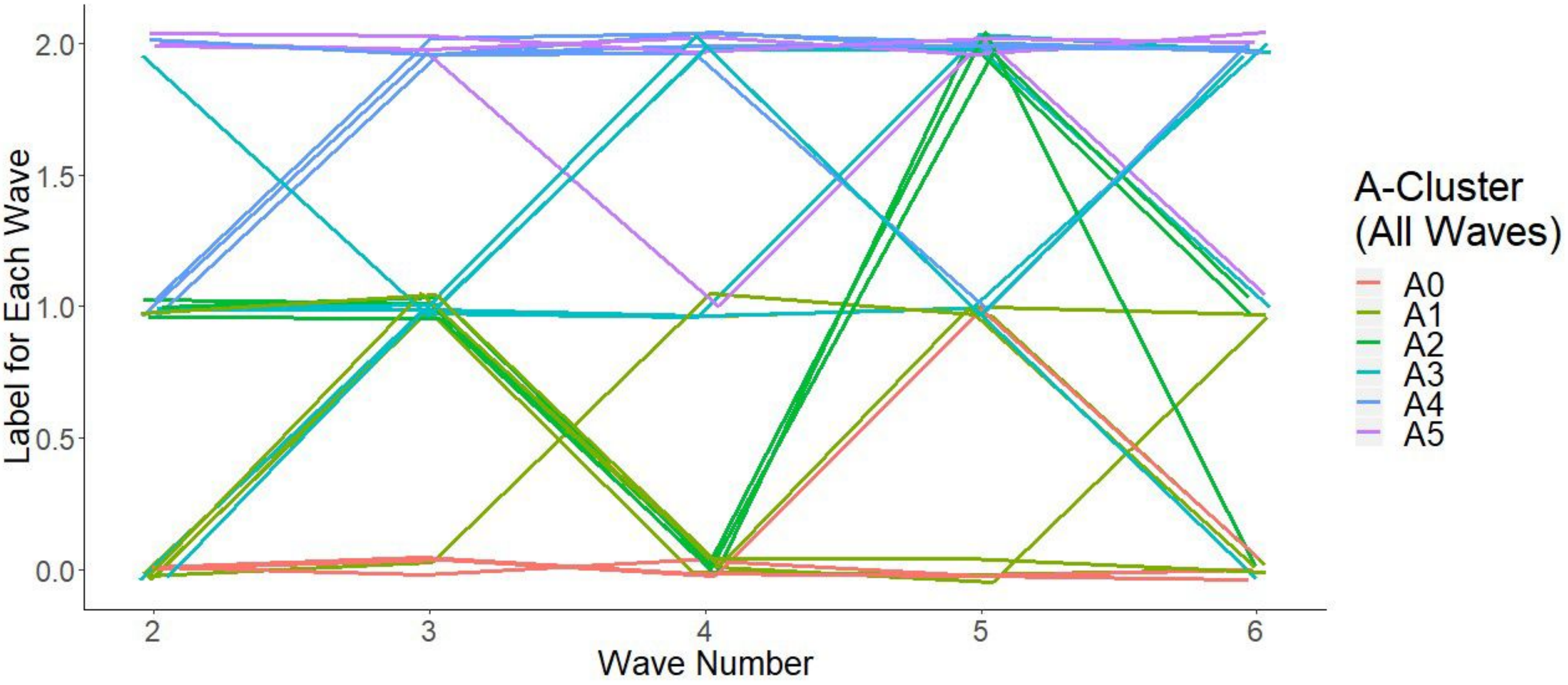
A3 - Most unstable (?)

A4 - Unstable and high vulnerability (?)

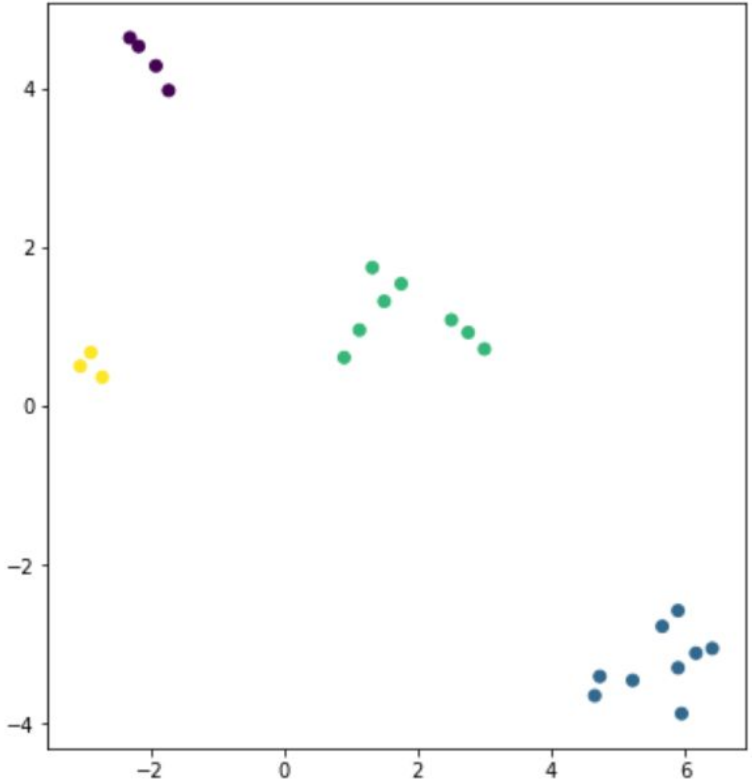
A5 - Stable and high vulnerability (BAD)



Neighborhood change each wave in relation to Single Wave Clustering



Validating Clustering results with UMAP



UMAP Clustering (Right) shows four distinct clusters on all-waves.

Hopkins Statistic (Below) to reject the null hypothesis that these clusters reasonably random.

	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.

What keeps these Clusters together? Using Census Data to describe Cluster Identity

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 (%)

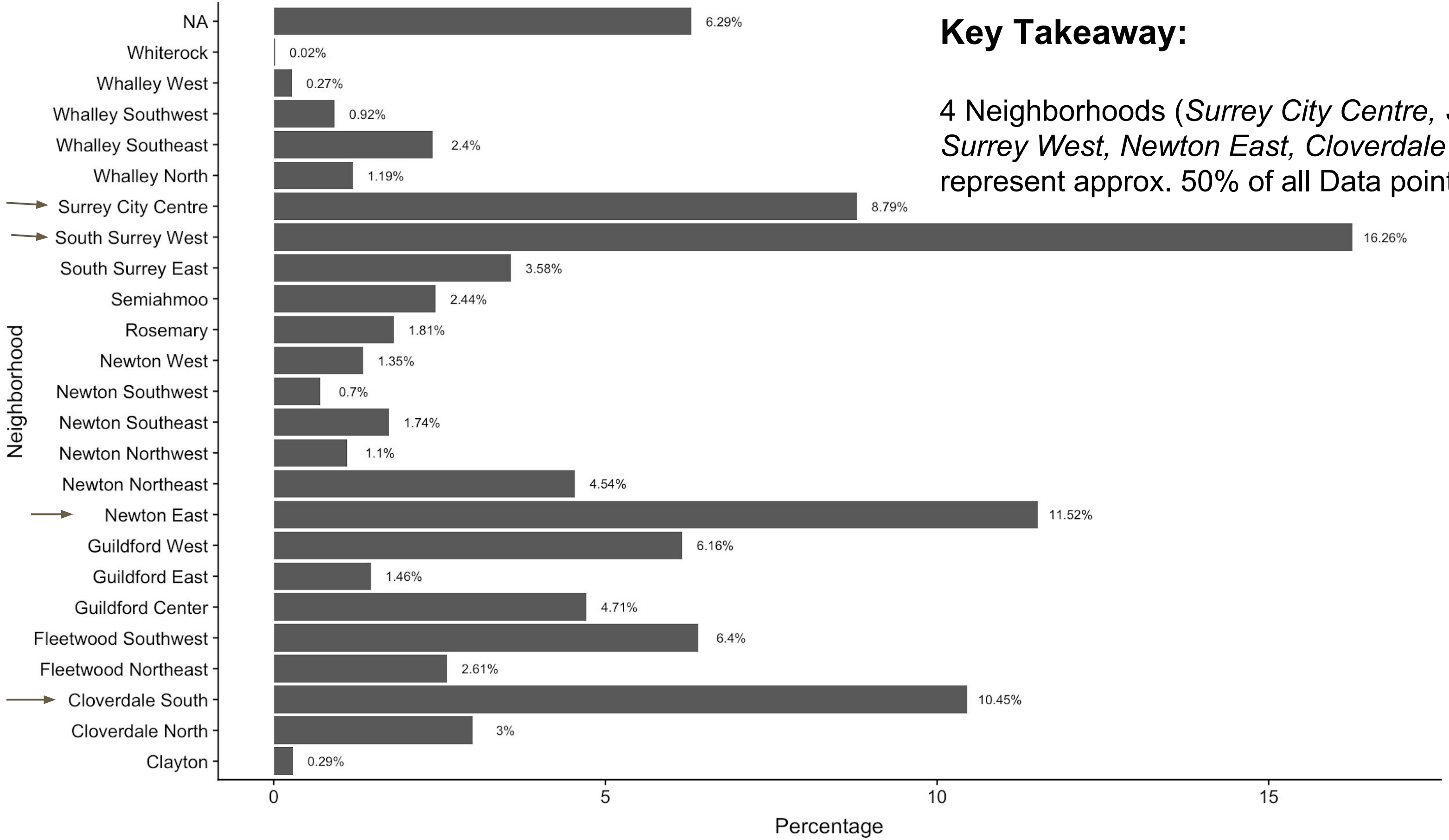
Table 5: An assortment of significant census variables for the 6 A-clusters.

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.

Analysis of the CLASS Dataset (Program registration for the City of Surrey)

Representation of Neighborhoods in CLASS Dataset



Key Takeaway:

4 Neighborhoods (*Surrey City Centre, South Surrey West, Newton East, Cloverdale South)* represent approx. 50% of all Data points.

Extracting Child Registration Data from CLASS

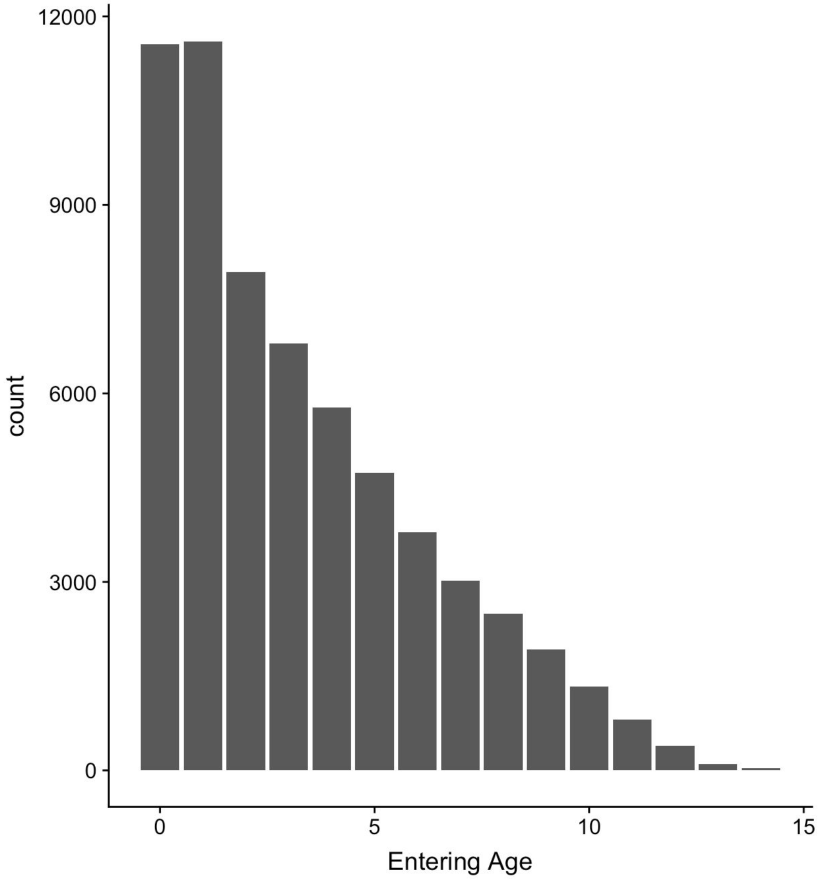
- **PostgreSQL Search Terms:**
  - Accounts with registered Birth Dates greater or equal to 01/01/2000
  - Course with a Max Registration count >= 1
  - Course must have been completed (no Withdrawals)

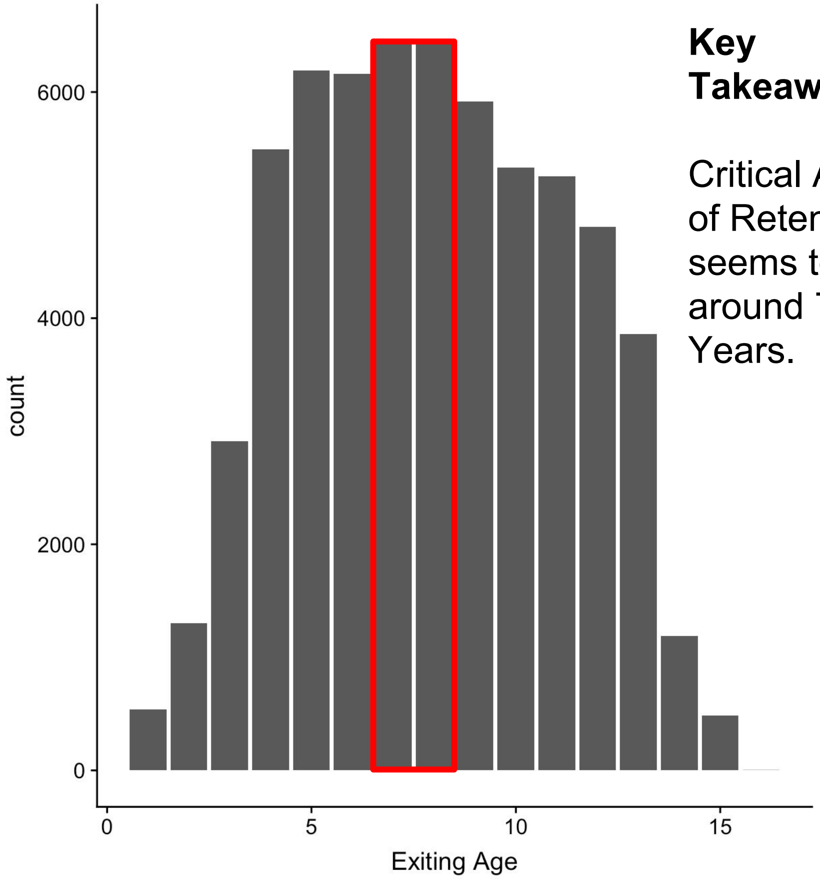
- **High-Level Classification of Courses offered and visible in CLASS:**
  - Aquatics
  - Arena and Skating
  - Arts and Crafts
  - Day Camps
  - General Activities
  - Music, Dance and Theatre
  - Parent Participation and Family
  - Sports, Fitness and Wellness

General Activities: *(e.g Arts and General - Children Computer, Arts and General - Children Personal Development, Youth Outdoor Recreation, Youth Personal Development)*

Parent Participation and Family: *(e.g Arts and General - Parent Participation Performing Arts-Arts Centre, Family Environment and Parks)*

Distribution of Children's Age at time of First and Last Registration

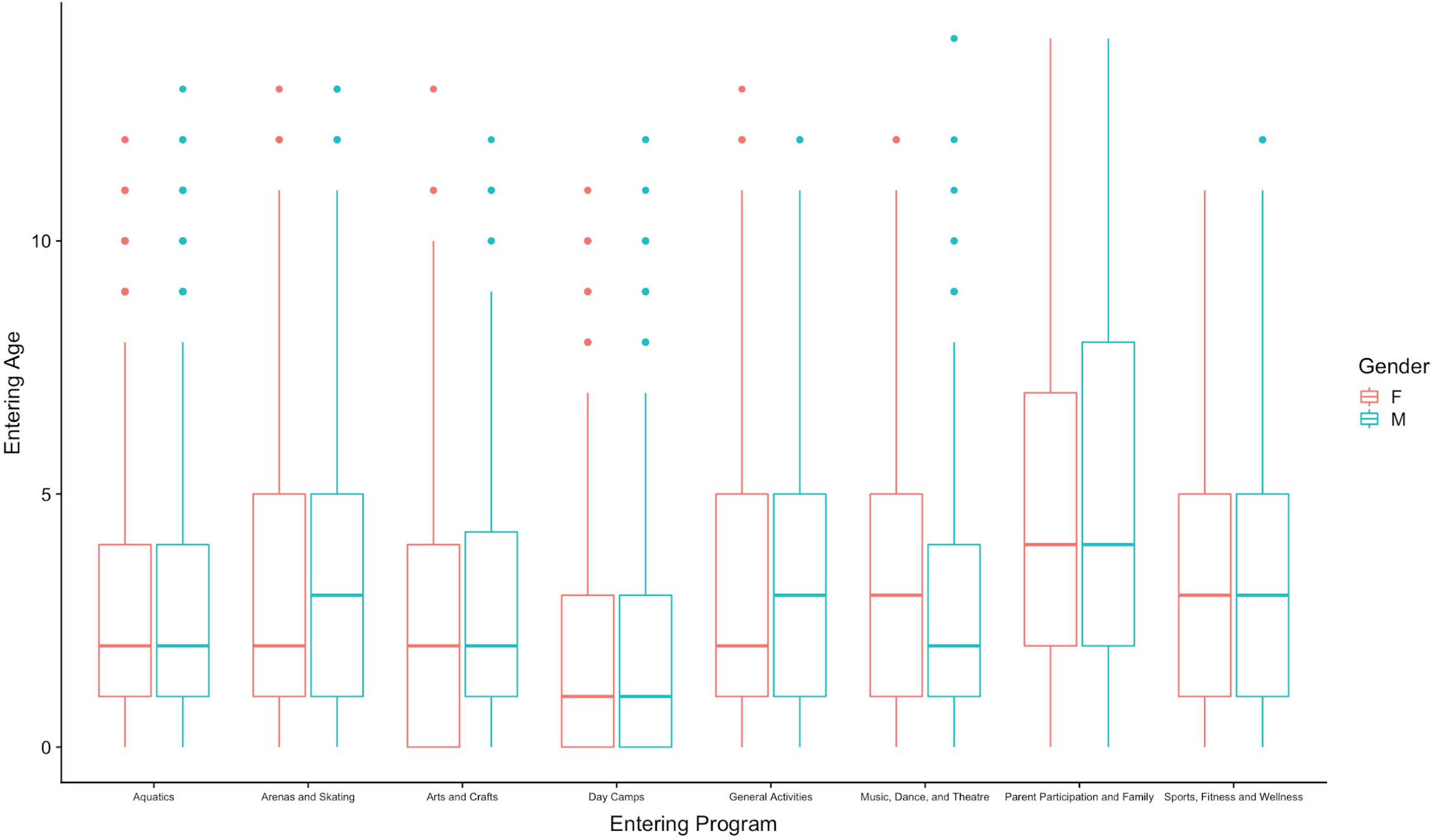




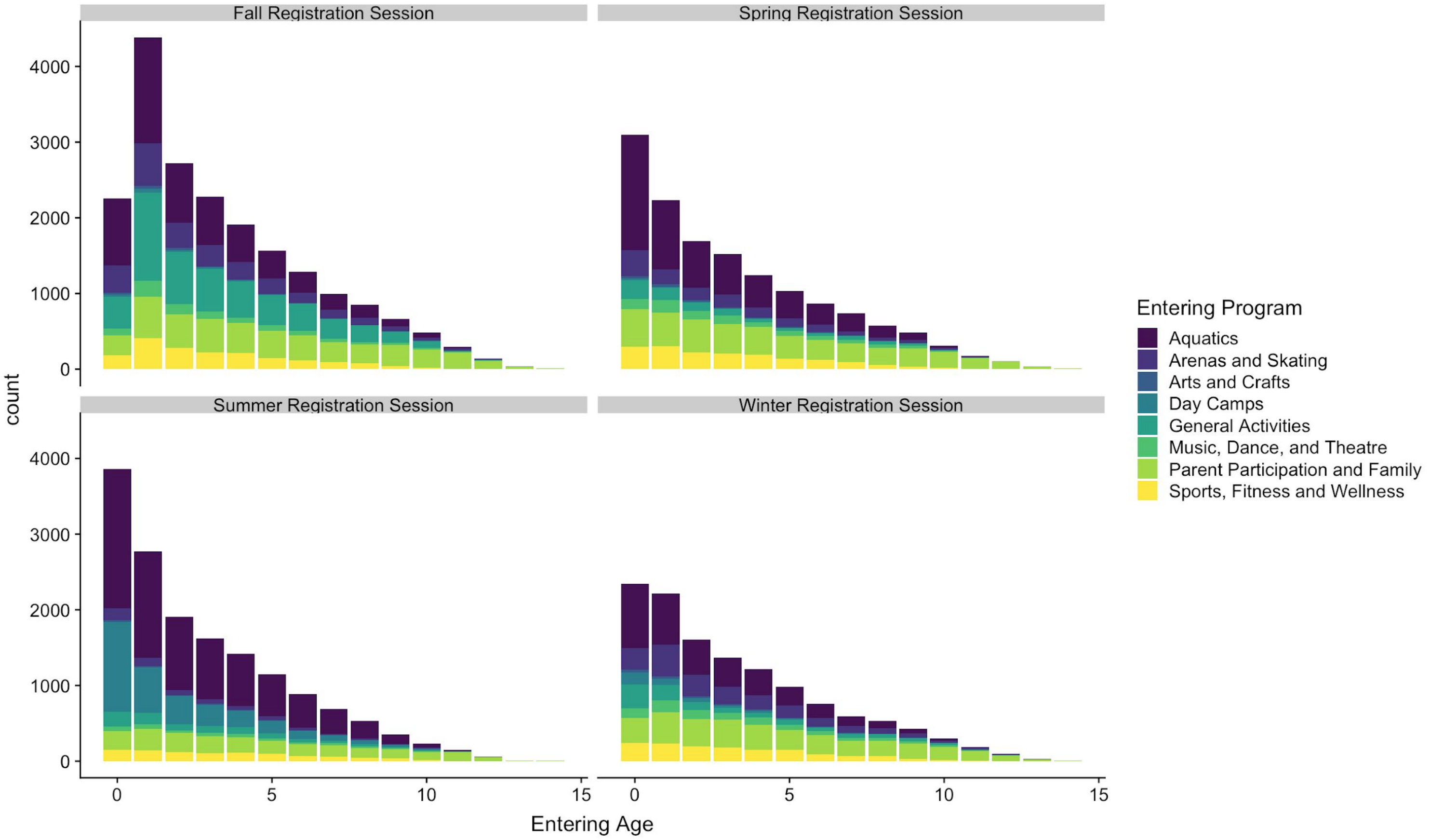
Key Takeaway:

Critical Age of Retention seems to be around 7-8 Years.

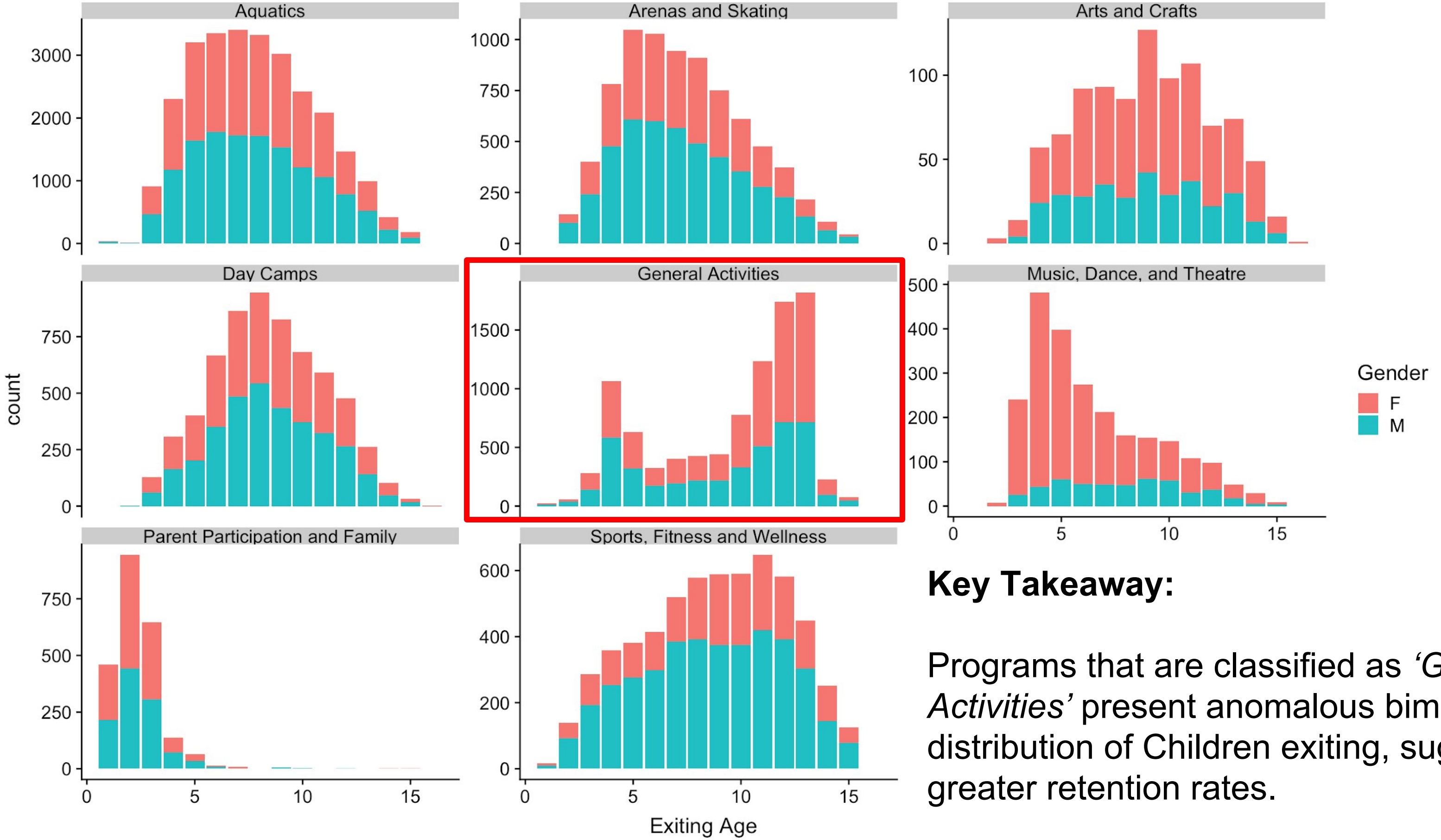
Age of First Registration for Male and Female Children



Number of Children Registering for Programs by Season



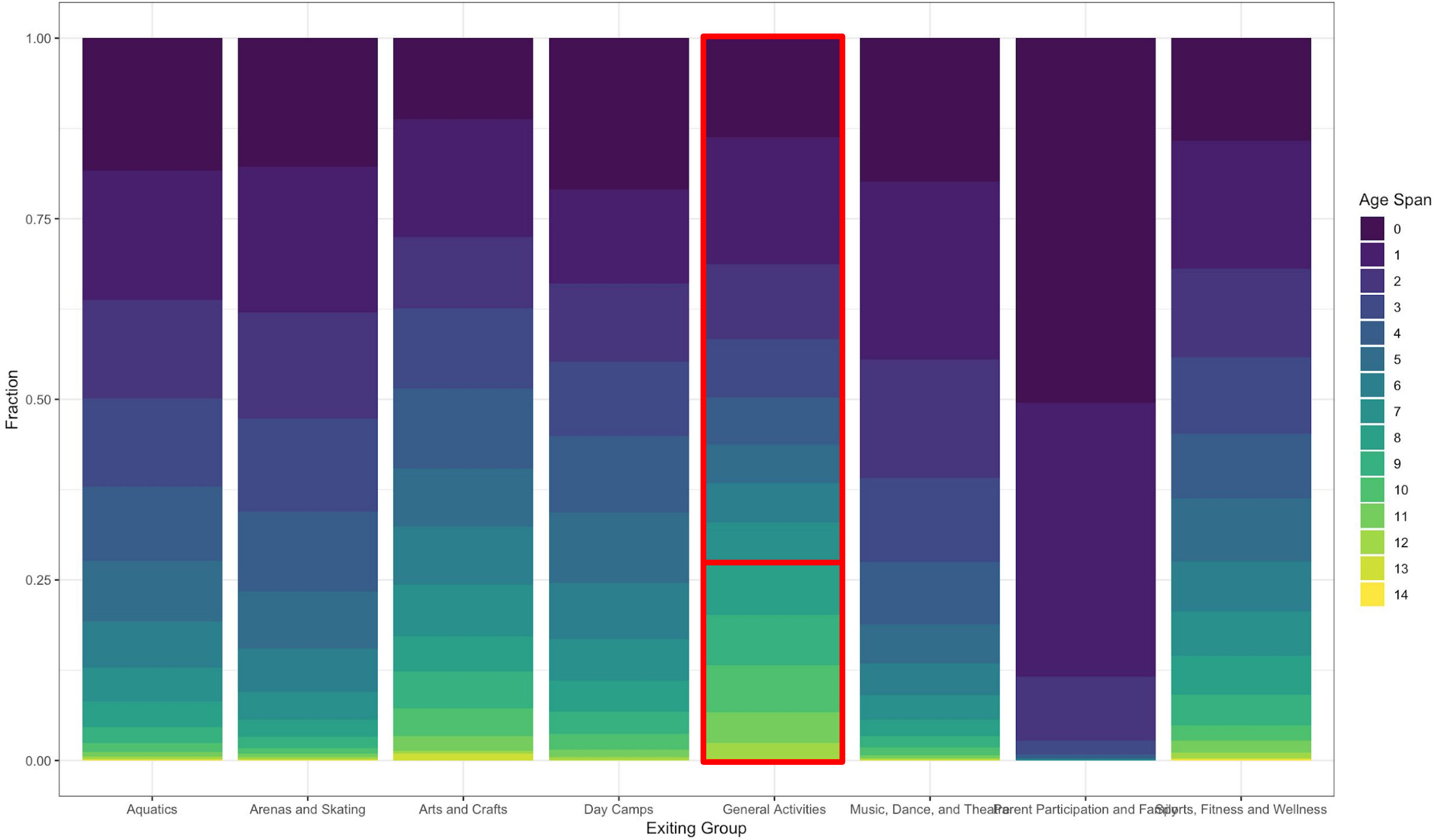
Distribution of Total Number of Children per Exit Age



Key Takeaway:

Programs that are classified as *'General Activities'* present anomalous bimodal distribution of Children exiting, suggesting greater retention rates.

Proportion of Age Groups vs. Last Program Type

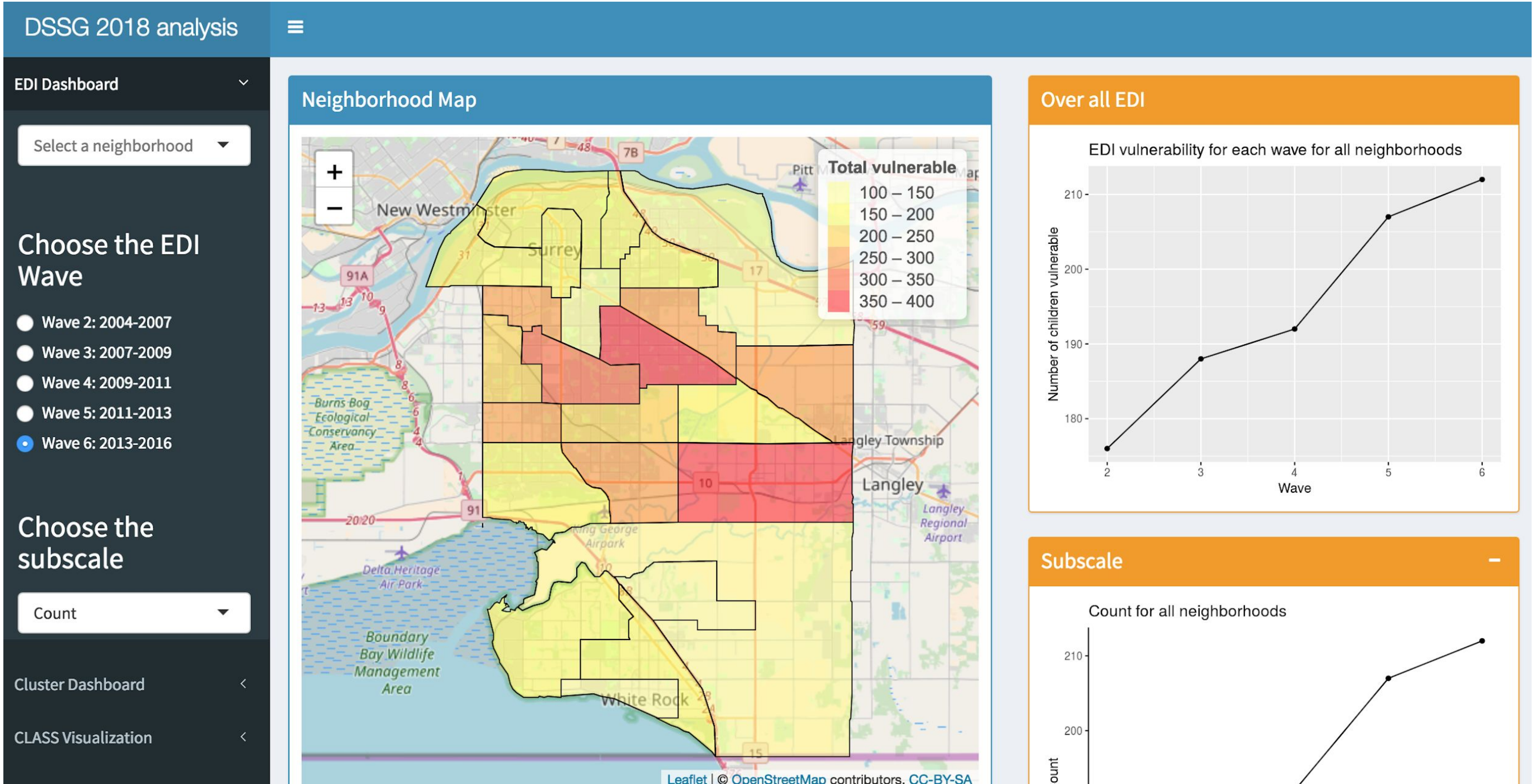


Key Takeaway:

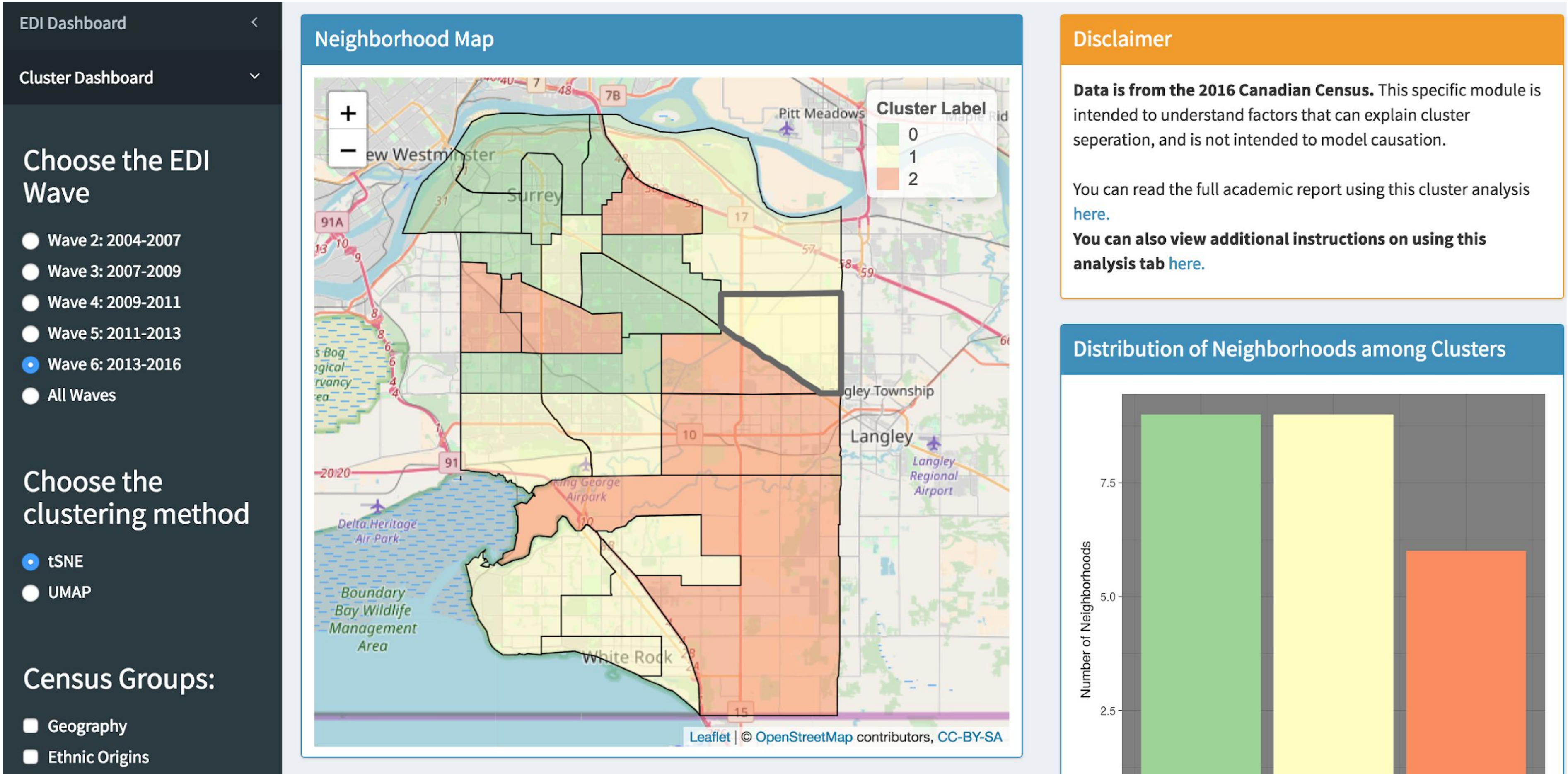
Programs that are classified as *‘General Activities’* present the largest proportion of Children having spent 8 or more years within the Program Pipeline when they leave.

Putting it all Together:
A Web Dashboard Application

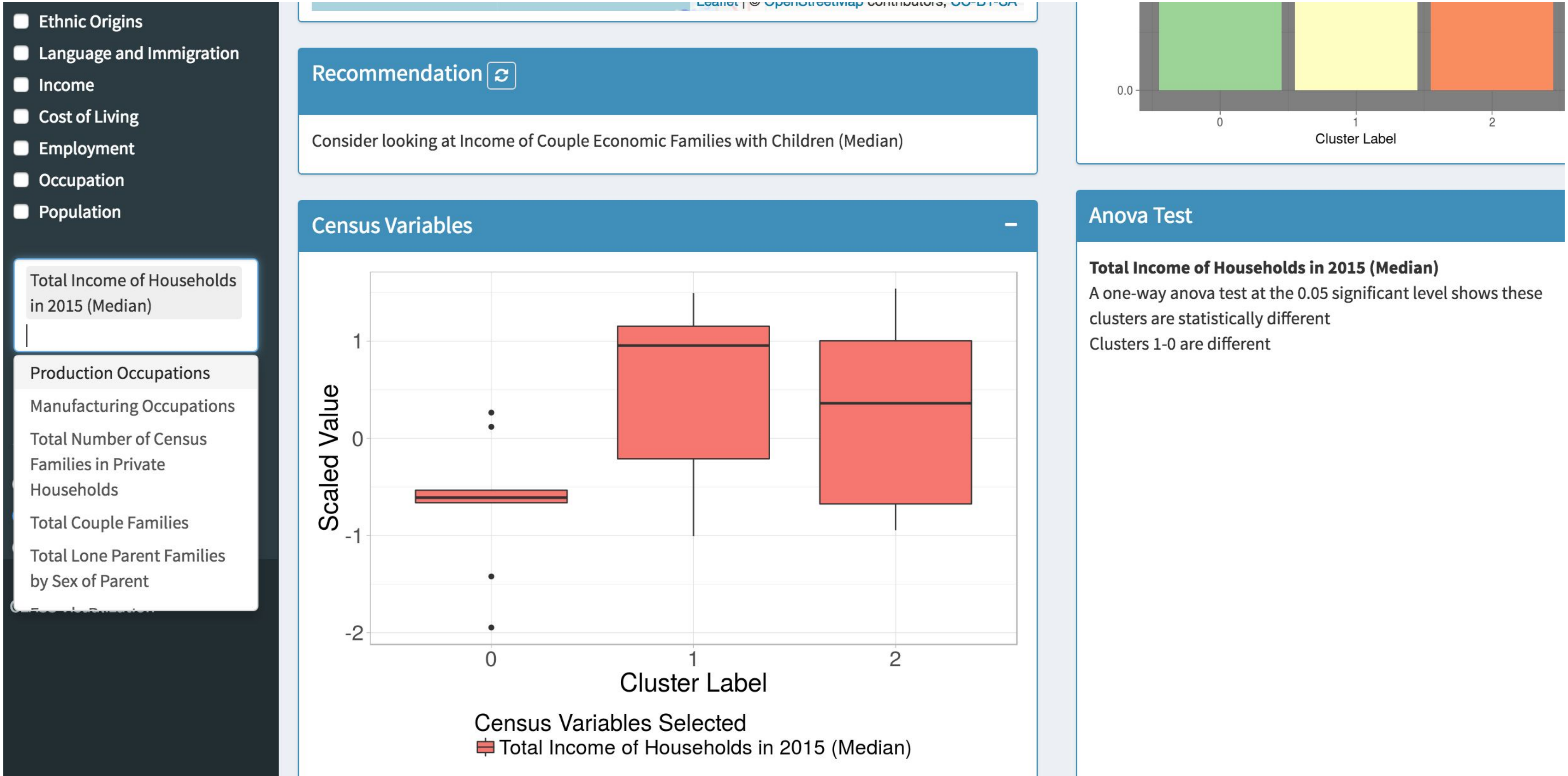
Visualizing EDI Scores by Neighborhood



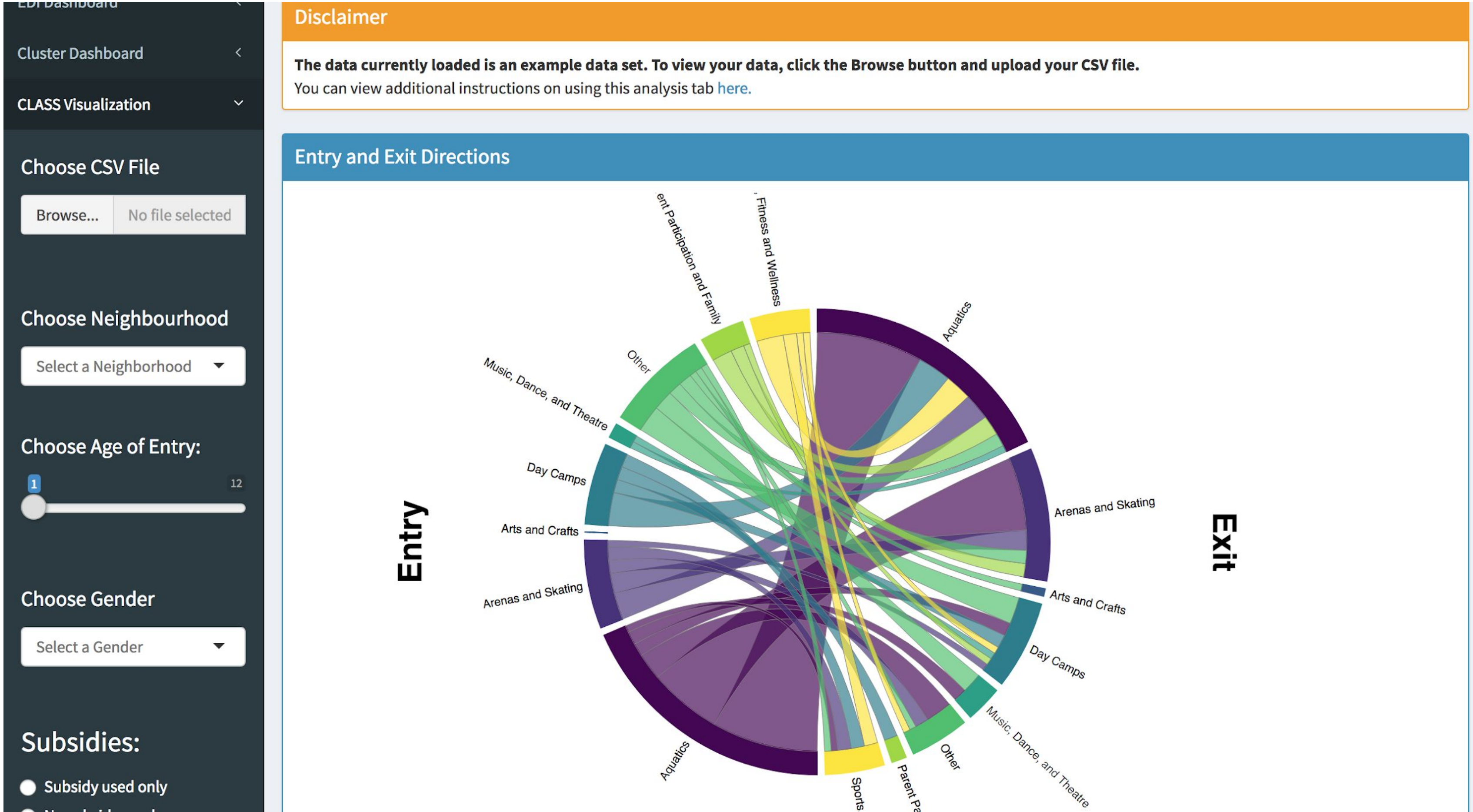
Visualizing Cluster Analysis Results



Using Census Data to describe Cluster Variation



Visualizing a Child’s First and Last Registered Program



Conclusions

Results from Clustering with t-SNE and UMAP suggests that **Clusters are real**, and may provide useful in understanding underlying factors that drive Childhood Vulnerability rates (i.e EDI Scores)

Ethnicity and SES Census variables emerging as significant discriminants between clusters suggests **different groups access programs differently**

CLASS Analysis suggests that **certain Programs and their enrollment can influence retention of Children**, allowing for greater engagement of Children within the community and City

Challenges and Future Work

When is Machine Learning “appropriate”

- In the case of CLASS Dataset, modeling “Exit-Age” to build a predictor makes little sense since the data does not accurately reflect this
- Combining the Top-Down and Bottom-Up approaches in a unifying model led to no statistically significant results (**Connecting EDI to CLASS**).

Future Work can include

- Analyzing Sub-Scale Data for EDI, utilization of MDI as well as future Census Data, and City of Surrey COSMOS Data (e.g Greenspace)

Acknowledgements

Stacey Rennie (City of Surrey) - Project Contact and Lead

Raymond Ng and Kevin Lin (UBC DSI)

Biljana Stojkova, Joe Watson, Carolyn Taylor (UBC ASDA)

Pippa Rowcliffe, Barry Forer (UBC HELP)

Sarah D'Ettore (Microsoft Vancouver)

Patrick Laflamme (UBC Psychology, DSSG '17)

Thank you so much for your mentorship and guidance!