Using web data to study the informal childcare sector in Canada

Abstract submitted to PAA 2024

Jose Morales Vidales Monica Alexander

add abstract here

1 Introduction and motivation

Demand for non-parental childcare has increased substantially in recent decades, largely driven by an increase in labour force participation of women. In Canada, the use of non-parental childcare covered more than half of families in 2022 (REF). As well as centre-based and licensed or regulated childcare providers, an important part of the childcare landscape in Canada is unlicensed home-based childcare, provided by baby-sitters and nannies. Demand for this informal home-based childcare has also recently increased, potentially because of the relative flexibility of hours compared to childcare centres, and in response to health concerns surrounding the Covid-19 pandemic (REF). Compared to employees in regulated childcare options, nannies may experience less job security, more uncertainty, and poorer working conditions (REF).

Despite the increasing importance of workers in the informal childcare sector in countries such as Canada, little is known about the characteristics, working conditions, and employment outcomes of this sub-population. Data collected in censuses and surveys is limited; informal childcare workers are often hard to distinguish between care-workers in general, and information about working conditions and place of work is also limited. In Canada, in 2022 there

was a survey on the provision of childcare services, but the focus of this study was largely on formal childcare services, in response to national-level changes in childcare subsidies (REF).

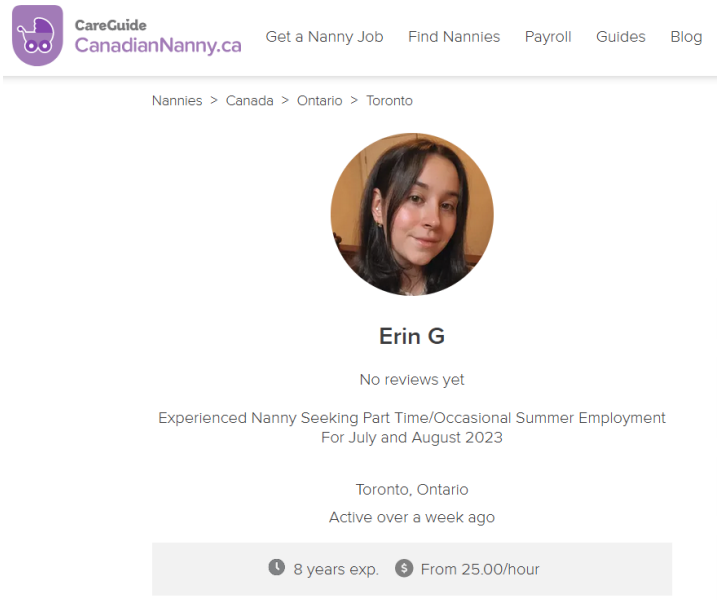
As such, in this project we utilize a large, rich, dataset constructed from public web profiles on a nanny advertising website to better understand the characteristics and systematic inequalities in the informal childcare sector in Canada. We are particularly interested in studying differences in work expectations and outcomes by country of birth and migrant status. We extracted almost 10,000 unique online profiles over three months from the website canadiannanny.ca. These profiles contain a range of different types of information, including advertised pay rate, qualifications, experience, as well as a free-text biography. We then used a variety of text-based methods to extract information on key demographic characteristics, including country of origin, age, and citizenship status.

In the remainder of this abstract, we describe the data extraction process, methods used to retrieve information on demographic characteristics, initial results and future work.

2 Data

We obtained information on characteristics of nannies seeking work in Canada from the website canadiannanny.ca. This website, which is used by over 200,000 potential nannies across Canada, contains incredibly rich data on nannies' characteristics and experiences on their profiles (example below XX ADD SCREENSHOT OF PROFILE).

We used webscraping tools in R (primarily relying on the `rvest` package) to extract information on nannies from publicly available profiles. The profiles were scraped using an R script at multiple points of July, August and September 2023. They were retrieved in order of appearance on the site which is sorted by the last time the user was active on the platform. So far more than 16,000 entries have been scraped from the site, of those around 9,700 are distinct



users.

Demographic variables are usually reported by the user in the text-based sections of their profiles, which include the short blurb right under the profile picture, a "Reasons to Hire Me" section, and a longer description in the "About Me" section. We also retrieved the users' name, url, location, reported years of experience, hourly rate, last time active on the site, number of reviews, star rating out of five, bullet points under the "I can work:" subsection (part-time, full-time, summer, etc.), children ages the user has experience with (infant, toddler, newborn, etc.), number of children they can look after, experience with children with medical conditions (diabetes, disability, epilepsy, severe allergies, etc.), transportation requirements (close to transit, has driver's license, etc.), qualifications (first aid, CPR, languages, etc.), and services they can provide (housekeeping, cooking, groceries, swim supervision, etc.).

3 Retrieving demographic characteristics from profiles

In this project we are interested in studying differences in user information by key demographic characteristics, particularly age, migrant status (in particular, if the user requires sponsorship), and country of origin. However, these characteristics are not reported in profiles in a fixed

field, but rather the information is often contained somewhere in the free-text user-written biographies (XX TODO ADD EXAMPLE TEXT). As such, in order to retrieve the information of interest, we use several information retrieval approaches.

Free-text biography example.

Hi there, my name is Erin (she/her), (23 years old)!
I am a recent (2022) graduate from the University of Waterloo with an Honours degree in Therapeutic Recreation. I spent four years nannying for two families (that I connected with on this website), while completing my degree.
I moved to Toronto in the Fall of 2022 and reside in the West End, Roncesvalles neighbourhood. I have nannied for two families since being in Toronto, one who I am currently working with, and can connect you with both past and present families for recommendations! ...

Information Retrieval (IR) is the process of obtaining any type of media based on user information needs. The resulting IR system is often called a search engine (REF). The IR task that we consider returns a list of ordered sentences, taken from profile descriptions, based on a query. The IR architecture uses the vector representations of queries and sentences, which are then ordered based on a similarity function like cosine or dot product. We tested two main IR methods to extract demographic information of interest. Below is a brief description of both methods considered.

3.1 BM25Okapi

BM25 stands for "Best Matching 25" (i.e the 25th iteration of the function), and Okapi makes reference to the first IR system that used it. This is a bag-of-words model in which query and document vectors are based on unigram word counts, and each word is considered independently of its position. This function is similar to TF-IDF (term frequency - inverse document frequency) weighting, but it adds two parameters: k, which adjusts the balance between term

frequency and IDF, and b, which controls the importance of document length normalization (REF). The BM25 score of a query Q containing keywords $q_1, ..., q_n$ and a document D is:

$$\text{score}(D, Q) = \sum_{i=1}^{n} \text{IDF}(q_i) \times \frac{tf(q_i, D) \ \cdot \ (k+1)}{tf(q_1, D) + k \times \left(1 - b + \frac{b \ |D|}{|D|_{\text{avg}}}\right)}$$

Where $tf(q_i, D)$ is the number of times the keyword q_i appears on document D, $|D|$ is the length of the document and $|D|_{\text{avg}}$ is the average document length. The inverse document frequency $\text{IDF}(q_i)$ is often computed as:

$$\text{IDF}(q_i) = \log\left(\frac{N - n(q_i) + 0.5}{n(q_i) + 0.5} + 1\right)$$

Where N is the total number of documents and $n(q_i)$ is the number of documents containing the keyword q_i. To implement this method we are using the python package `rank_bm25` which sets $k = 1.5$ and $b = 0.75$ by default (REF). It also sets keywords with negative IDF to the average of the non-negative IDFs multiply by 0.25. In various IR tasks is common to remove "stop words" from the text, because these words carry little semantic knowledge and they often hurt the process. However, it's not strictly necessary since the IDF downweights these stop words.

A major disadvantage of the BM25 scoring function is that the query's keywords have to appear in the documents that it's searching through. So if the author used a synonym or expressed it in another way the model has no way of knowing that that is a relevant document. An additional disadvantage is that it doesn't take into account how close the keywords are together in the document. In our application we are dividing the profile descriptions by sentence, so the proximity between words shouldn't be an important issue.

3.2 Sentence Embeddings

Our second approach was to compute vector representations of sentences, or embeddings using transformer models. We used the `SentenceTransformers` package in python, which

was developed for semantic search tasks similar to our goal here (REF). We mainly used the `multi-qa-mpnet-base` bi-encoder model, which was specifically trained on 215 million question-answer pairs from sources such as Yahoo answers, StackExchange and Google and Bing search queries. Once the embeddings of both the query and sentences are computed we can use the cosine or dot product similarity score to rank them.

The main advantage of this approach is that we don't need to have the query keywords appear on the sentences; the model should retrieve sentences that don't contain any of the words but have synonyms or similar ones. This could be both an advantage and disadvantage, however, because the model might select certain sentences that although are similar, aren't really relevant for that specific query. For example, when feeding the query "Do you need visa sponsorship?", the model would selected sentences such as: "I have a driver's license".

3.3 Queries used and retrieval results

3.3.1 Immigration Status

All three variables were extracted with both approaches and two different queries: a keyword-based query and a question. With regards to retrieving users who needed sponsorship, Table 1 shows the queries and thresholds used, as well as, the number of users and true positives (users who actually required sponsorship as reviewed manually afterwards) retrieved. The BM25 model, with the keyword-based query, retrieved the most true positives from the texts; even though, the bi-encoder model had a better ratio of true positives, it's unclear if this holds when reducing the threshold to allow the retrieval of more users. Therefore, we consider that the BM25 model is the most appropriate for this specific task.

3.3.2 Country of origin

The two queries we tested with in order to retrieve users' country or nationality were: "I am from" and "Where are you from?". However, almost every word in these queries are included

Table 1: Users that requiere visa sponsorship retrieved by IR models.

Model	Query	Threshold	Total Users Retrieved	True Positives
BM25	Do you need sponsorship?	5.0	248	84
	immigration visa sponsor sponsorship	0.0	213	183
multi-qa-mpnet	Do you need sponsorship?	0.4	198	87
	immigration visa sponsor sponsorship	0.4	97	84

in the stop words lexicon that we are using before feeding the queries into the BM25 model, so it's not able to retrieve any valid sentence (future iterations will not remove stop words from the text so a proper comparison will be made). The transformer model is able to retrieve 243 users with valid nationalities using the keyword-based query.

3.3.3 Age

For this task we used the queries: "How old are you?" and "I am years old". There was no need to select a threshold when retrieving users' age, since out of the top sentences we just extract the numbers and performed the subsequent transformations mentioned before. The BM25 with the keyword-based query was able to retrieve 237 users with apparent valid ages and 223 with the question based one, while the bi-encoder performed better with the question query it only manage to extract 195 users with valid ages.

4 Initial results and observations

4.1 Summary characteristics by province and age

Table 2 shows summary statistics by province of specific characteristics reported by users. Ontario is by far the province with the most users with over 50% of these located there, followed by Alberta and British Columbia, which are both well over 1k users and above 15% of the total. At the national we have a mean hourly rate of $19.51 CAD with a standard deviation of $4.65, the province with highest rate in average was British Columbia with $20.98, followed

Table 2: Average value of quantitative variables by province. The statistic in parenthesis is specified in the header.

Province	N (prop.)	Rate (sd)	Years Ex. (sd)	No. Children (sd)	Age (sd)
Alberta	1674 (0.17)	19.45 (4.55)	8.38 (6.46)	3.03 (2.06)	26.93 (7.85)
British Columbia	1592 (0.16)	20.98 (4.18)	7.6 (6.18)	2.78 (1.11)	26.03 (6.76)
Manitoba	232 (0.02)	16.79 (3.53)	6.7 (5.44)	2.78 (0.94)	25 (4.55)
New Brunswick	98 (0.01)	16.44 (2.78)	8.72 (6.96)	3.31 (2.36)	52 (NA)
Newfoundland and Labrador	49 (0.01)	17.49 (2.97)	6.25 (6.74)	3.12 (0.83)	NaN (NA)
Nova Scotia	198 (0.02)	18.44 (3.49)	9.13 (7.32)	4.2 (4.45)	22.67 (4.04)
Ontario	5085 (0.52)	19.66 (4.86)	8.19 (6.89)	2.99 (2.71)	27.6 (8.72)
Prince Edward Island	15 (0)	16.01 (2.11)	8.6 (8.08)	5 (NA)	NaN (NA)
Quebec	466 (0.05)	17.93 (3.78)	7.05 (6.06)	2.7 (1.27)	25.64 (7.76)
Saskatchewan	309 (0.03)	16.32 (3.29)	7.19 (6.97)	3.83 (2.37)	22.4 (6.69)
Territories	13 (0)	22.18 (3.99)	7.25 (3.3)	3 (NA)	NaN (NA)
National	9732 (1)	19.52 (4.65)	8.02 (6.64)	2.98 (2.35)	26.85 (8.03)

by Ontario at $19.66; however, both Yukon and Northwest Territories have higher averages rates than any province. The province with the lowest average rate was Prince Edward Island with $16.01. With regards to the years of experience the national average was around eight, the province with the most experienced nannies was Nova Scotia, while the least experienced was Newfoundland.

Only 66 profiles received reviews from clients, and three of those profiles have been deleted since the time they were first retrieved. In total 70 reviews have been filled (counting profiles that received more than one), with a national average rating of 3.71. Users from Ontario are by far the most likely to receive a review and they have an average rating of 3.86, while users from BC had a very low average rating of 2.6.

Figure 1 shows the distribution of age over all users and by province. The median over the whole population is 25, and the provinces' median ranges from 20 to 25 years old. It's clear that the distributions of the total users and for each provinces are right-skewed.

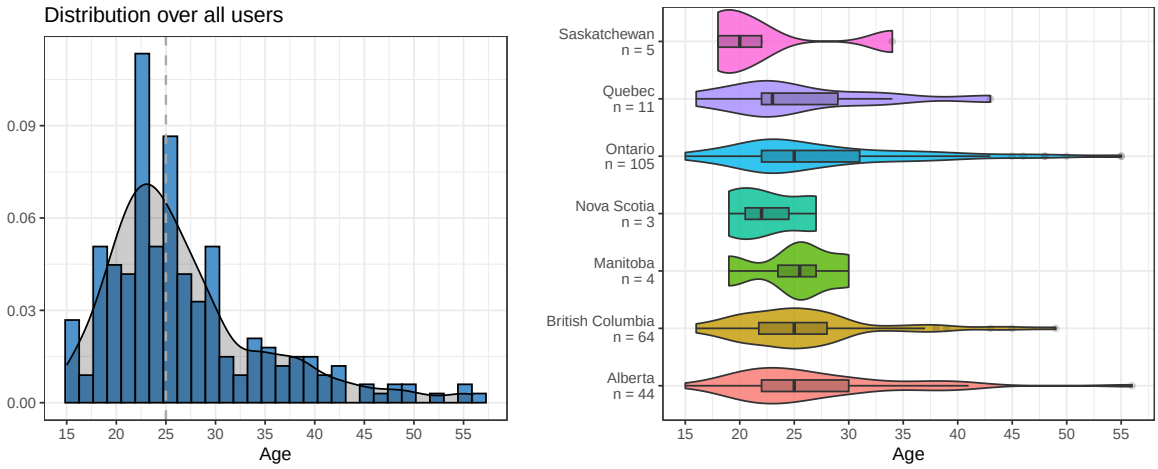


Figure 1: Age distribution of users and by province (selecting those with at least 2 users with valid ages).

Table 3: Statistics by immigrant status.

	Sponsorship		No sponsorship	
	Mean	Std dev.	Mean	Std dev.
Rate	19.474	4.254	19.517	4.659
Years Exp.	6.723	5.219	8.070	6.682
Age	28.400	6.653	26.780	8.094
No. Children	3.962	9.303	2.950	1.663

4.2 Nannies needing sponsorship

Table 3 users who need to be sponsored are willing to charge a slightly lower rate than those who doesn't. The distribution of reported years of experience of nannies who require sponsorship is much less spread out than those who doesn't. Figure XX shows the number and proportion of users requesting sponsorship by province. Unsurprisingly Ontario has the most foreign users, but provinces such as Quebec suggest higher proportions. Note that the Yukon territory has by far the highest proportion, where 2 out of 11 users where found to be looking for sponsorship, but was removed from the graph due to small counts.

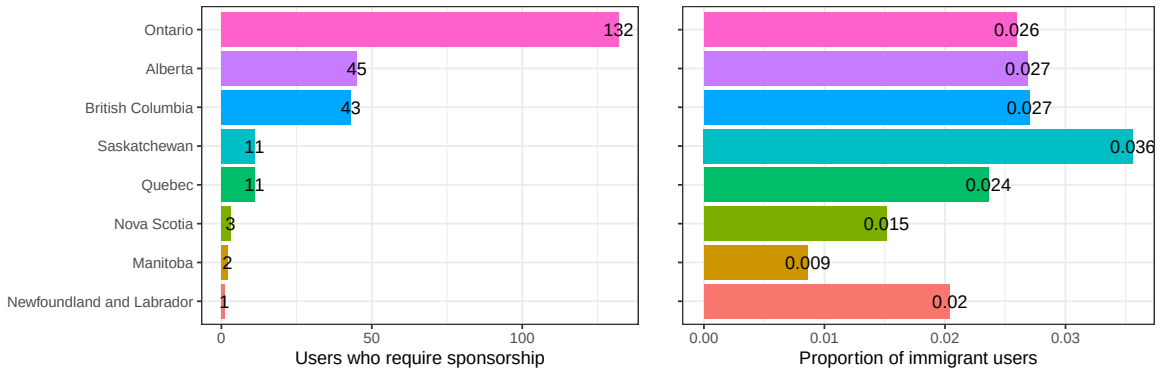


Figure 2: Number and proportion of users who require sponsorship by province or territory.

4.3 Country of origin

The majority of foreign users (78) are originally from the Philippines, followed by Mexico with 21 users. Figure 3 shows the differences in rate distribution and reported years of experience by country of origin. Nannies from the Philippines are willing to take a substantially lower pay than others, with a median hourly rate of less than $16 CAD and some users even going down to $12 per hour. Users from Mexico, Japan, Colombia or Brazil have an hourly rate much closer to the global median at $20 an hour.

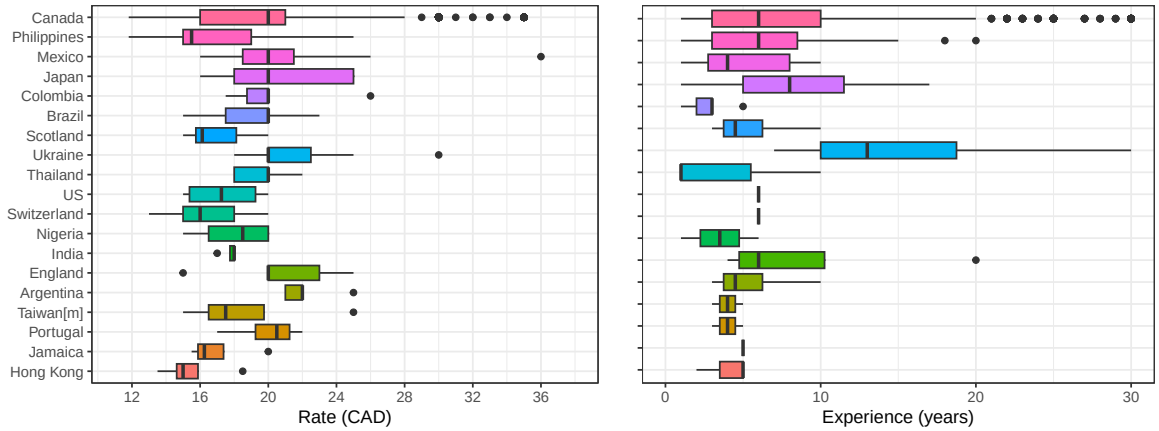


Figure 3: Distribution of rate and years of experience by country of origin.

4.4 Comparison to census data

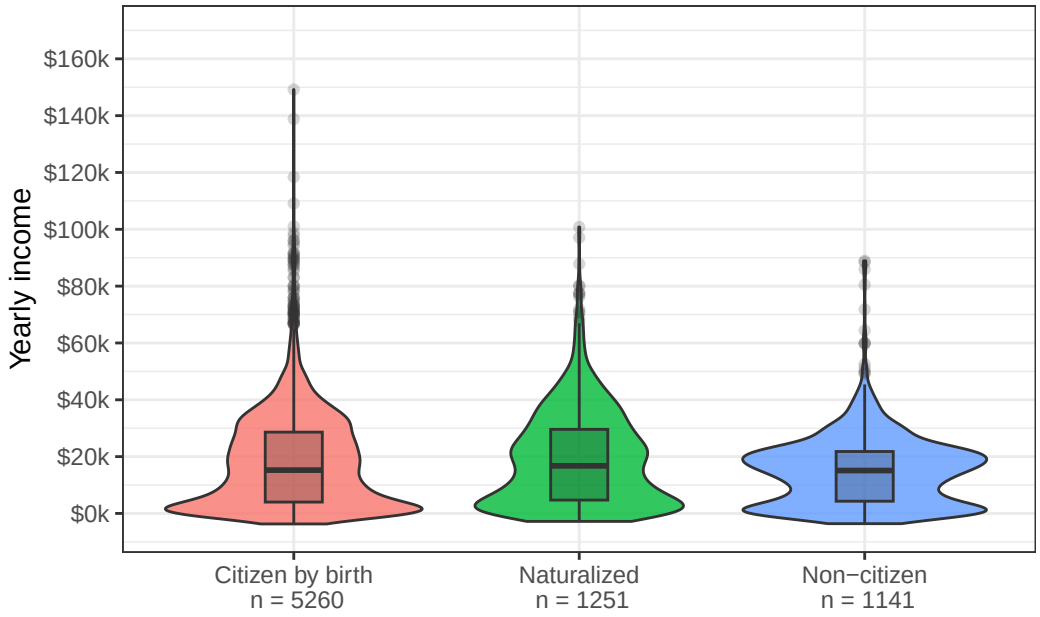
We collected data from the 2011 census from IPUMS [ref], on women working in home care and educational support occupations (code 35). Table 4 shows statistics by province. We found 7,652 women in Canada and 34% were from Ontario, significantly lower than what was found on the canadiannanny site, where 52% reported being located in this province. The discrepancy might be due to a lack of Québécois nannies on the site, only 5% reported Quebec as their location, which doesn't match true population proportions. Census data assigns 15% of people to Quebec. The reasons for this might be that, even though users can write their description in French or any language they want, the site itself is in English with no French version.

Table 4: Census statistics by province.

Province	N (prop.)	Non-citizens (%)	Employed (%)	Mean Age (sd)
Alberta	917 (0.12)	182 (20%)	788 (86%)	41.166 (13.41)
British Columbia	1086 (0.142)	200 (18%)	918 (85%)	43.022 (14.14)
Manitoba	379 (0.05)	29 (8%)	330 (87%)	40.343 (13.79)
New Brunswick	258 (0.034)	5 (2%)	220 (85%)	44.736 (13.51)
Newfoundland	230 (0.03)	0 (0%)	175 (76%)	46.43 (13.16)
Nova Scotia	255 (0.033)	7 (3%)	211 (83%)	42.506 (13.37)
Ontario	3006 (0.393)	593 (20%)	2515 (84%)	42.17 (13.88)
Prince Edward Island	48 (0.006)	2 (4%)	39 (81%)	41.938 (13.85)
Quebec	1207 (0.158)	108 (9%)	1032 (86%)	44.535 (13.7)
Saskatchewan	237 (0.031)	12 (5%)	205 (86%)	40.958 (14.11)
Yukon	29 (0.004)	3 (10%)	23 (79%)	43.069 (13.41)
National	7652 (1.00)	1141 (15%)	6456 (84%)	42.643 (13.85)

Average age range from 40 to 46 years old with a standard deviation of around 13-14 years across all provinces and at the national level. These numbers are much higher that what was observed from the users on the site, where the mean age was 26.85 at the national level and ranging from 22 to 27 across all provinces. This might indicated that a higher proportion of

young people are using the site, and they're using it to find part-time or temporary jobs.



Around 6,511 (85%) people were Canadian citizen, out of these 1,251 (16% out of the total) were naturalized immigrants, and 1141 didn't weren't citizens. Table 5 shows census statistics by citizenship status. We can see that naturalized citizens earn a higher yearly income on average than non-citizens and citizens by birth, they also have a higher proportion of people earning more than $50k a year (4.1%) than the other 2 groups, only 0.9% of non-citizens earned more than $50k in 2011. The income distributions of the both groups of citizens were significantly right-skewed, while the income distribution for non-citizens was slightly skewed to the left. Citizens by birth were more likely to earn negative income, but the proportion of people who are in this situation is quite low in all three groups. Non-citizen had the lowest unemployment rate, with only 2.8%; they were also the youngest group, with a mean age of 39 years, while naturalized citizens were the oldest group at an average of 47 years, makes sense given the time it takes to became a citizen in Canada.

Table 5: Census statistics by citizenship status.

	Citizen by birth n = 5260	Naturalized n = 1251	Non-citizen n = 1141
Income			
Mean (k)	18.131	19.081	14.866
Median (k)	15.200	16.700	15.100
Prop. >50k (%)	3.479	4.077	0.964
Prop. <0k (%)	0.475	0.320	0.351
Unemployment			
Number	246.000	52.000	32.000
Prop. (%)	4.677	4.157	2.805
Age			
Mean (years)	42.158	47.908	39.107
Std. dev. (years)	14.289	12.760	11.052

5 Summary and future work

In this project we utilize the large amounts of data available online to glean insights into the informal childcare sector in Canada. Our work to date has involved extracting almost 10,000 unique nanny profiles and using information retrieval techniques to extract information on age, country of origin, and immigration status. Initial observations suggest some evidence for systematic differences in advertised rates and other characteristics by immigration status and country of origin, with nannies requiring sponsorship willing to charge a lower rate than those who don't on average.

Future work will focus on several aspects. Firstly, we are working on improving the information retrieval algorithm to extract more information on potential migrants. Secondly, we will investigate image processing techniques to extract information about age and possibly gender from user profile pictures. Finally, we plan to collect multiple waves of profiles over time, in order to get estimates for the implied turnover rates of nannies seeking employment.

6 References