Predicting the best location for a new bilingual childcare in Chicago, IL.

IBM Data Science Professional Certificate: Final Capstone Project

Introduction

- Children under the age of 7 possess an innate ease of learning a secondary language.

- Early secondary language acquisition has proven to provide children with many cognitive and social benefits that will last a lifetime.

- Almost 29% of Chicago’s population is “Hispanic or Latino”

- Company ABC wants to target the Chicago market by starting up a new English-Spanish bilingual childcare in the Greater Chicago area as it seeks to expand its business into the city.

- Objective is to help the company’s business expansion team better understand the different residential zip codes within the city of Chicago, and ultimately select the best possible location for their new bilingual child daycare.

Data acquisition, processing and transformation

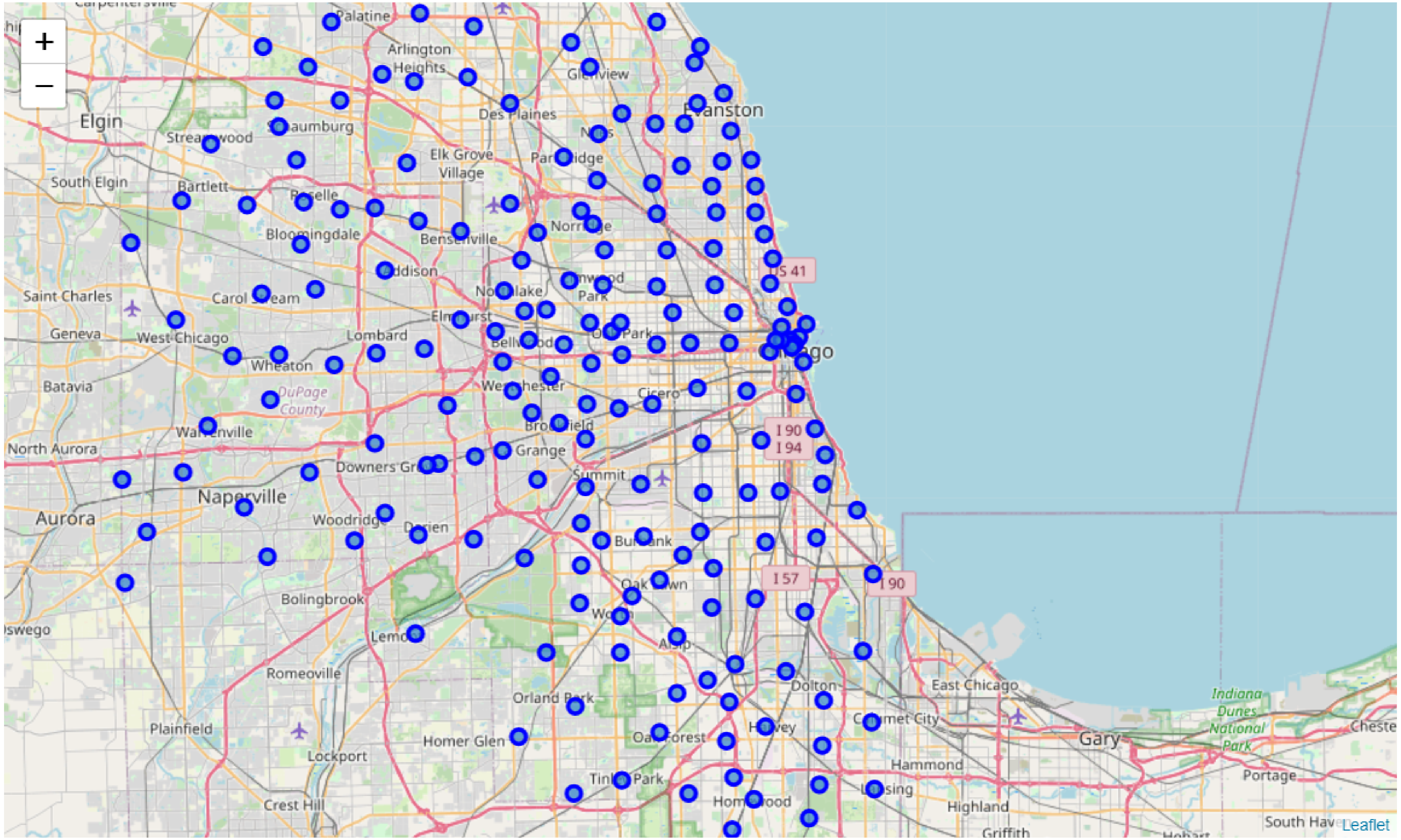
All datasets were obtained from the internet and read into pandas dataframes that were then filtered, processed and transformed. These were the 5 resulting dataframes:

- **zip_df**: Zip codes dataset obtained by joining two zip code lists: https://www.zipcodestogo.com/Cook/IL/ and https://www.zipcodestogo.com/Dupage/IL/ .
  - Resulted in 276 zip codes in Cook and DuPage counties.
- **IRS_df** : IRS dataset obtained from https://www.irs.gov/pub/irs-soi/18zpallnoagi.csv.
  - Resulted in 197 zip codes with reported household incomes in Cook and DuPage counties.
  - Included in dataframe: Number of households, annual income per household and dependants per household for each zip code.

Data acquisition, processing and transformation – Cont’d

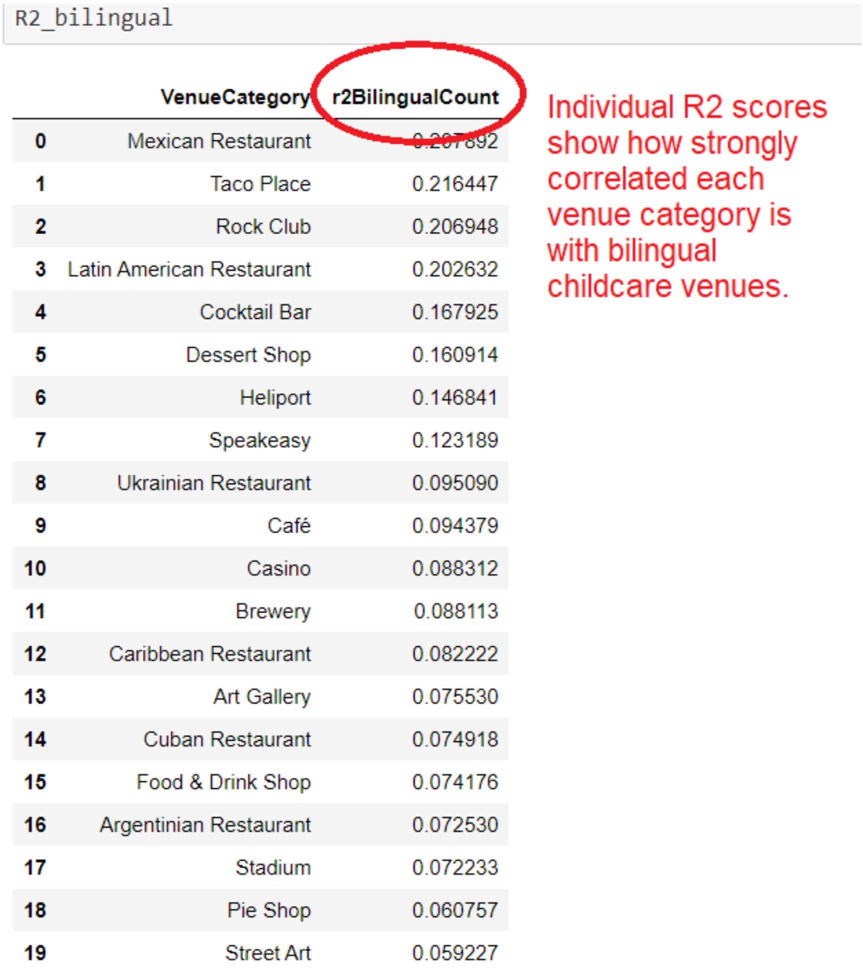
- **Chicago_df**: Latitude and longitude coordinates obtained from https://www2.census.gov/geo/docs/maps-data/data/gazetteer/2019_Gazetteer/2019_Gaz_zcta_national.zip
  - Contains all information in IRS_df + latitude and longitude coordinates + zip code areas in sqmi.
- **Chicago_venues**: Obtained from Foursquare API through a loop in python.
  - Contains 15,965 venues with 444 categories in 197 zip codes.
- **daycare_df**: Information about childcare venues obtained from https://sunshine.dcfs.illinois.gov/Content/Licensing/Daycare/ProviderLookup.aspx
  - Returned childcare venues in 186 zip codes out of the 197 zip codes.
  - Total of 1619 active childcare venues, 94 of which were English-Spanish bilingual.

197 Zip Codes in Cook and DuPage Counties:



Simple linear regression models

- Defined simple linear regression models through a loop (one for each venue category) to find correlations between individual venue categories and bilingual childcare venues in each zip code.
  - Resulting dataframe with R2 scores for 20 venue categories:



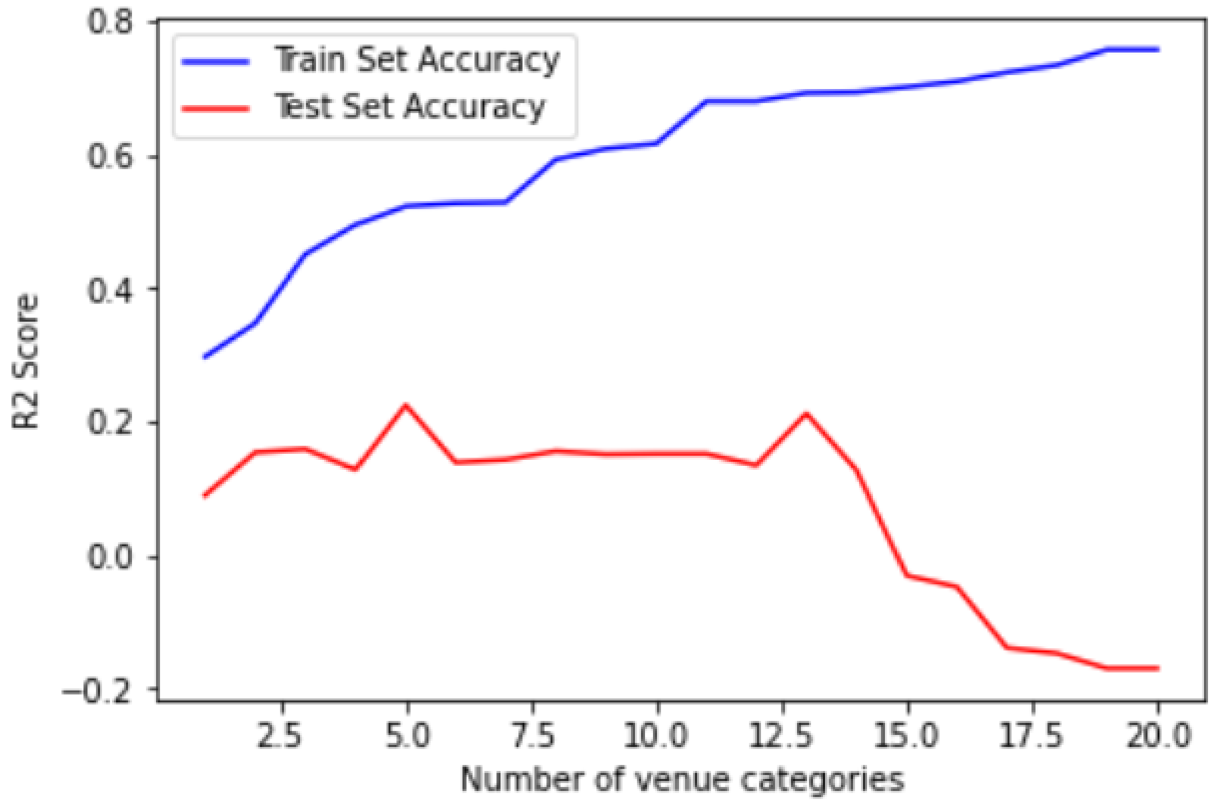
```
R2_bilingual
```

	VenueCategory	r2BilingualCount
0	Mexican Restaurant	0.207892
1	Taco Place	0.216447
2	Rock Club	0.206948
3	Latin American Restaurant	0.202632
4	Cocktail Bar	0.167925
5	Dessert Shop	0.160914
6	Heliport	0.146841
7	Speakeasy	0.123189
8	Ukrainian Restaurant	0.095090
9	Café	0.094379
10	Casino	0.088312
11	Brewery	0.088113
12	Caribbean Restaurant	0.082222
13	Art Gallery	0.075530
14	Cuban Restaurant	0.074918
15	Food & Drink Shop	0.074176
16	Argentinian Restaurant	0.072530
17	Stadium	0.072233
18	Pie Shop	0.060757
19	Street Art	0.059227

Individual R2 scores show how strongly correlated each venue category is with bilingual childcare venues.

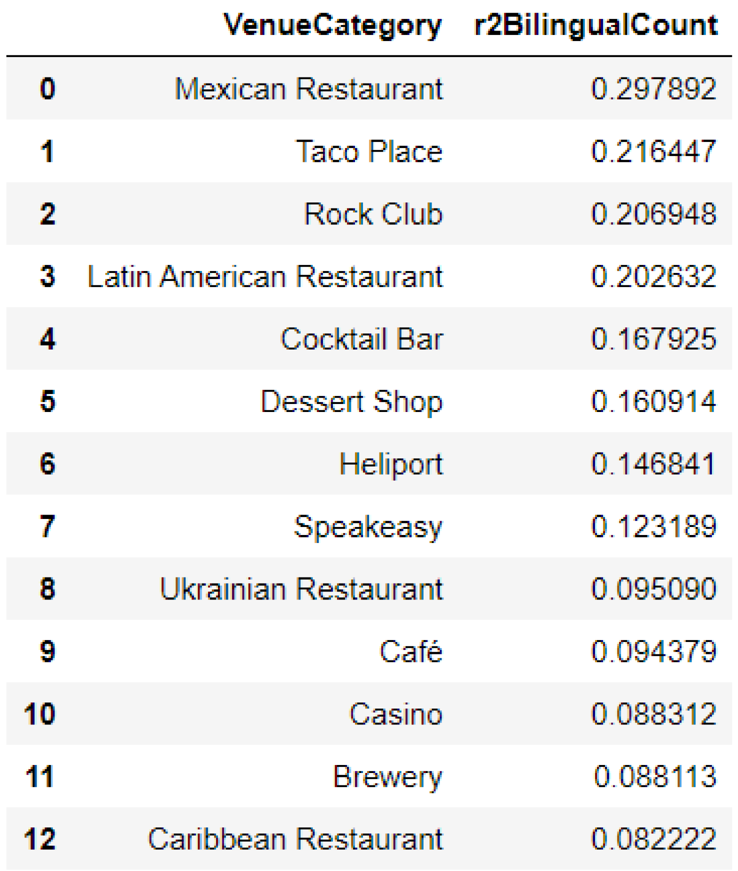
Multiple linear regression model

- Determined the optimal number of venue categories to fit in the model as independent variables.
  - Designed a loop to find the optimal model accuracy by creating this plot:

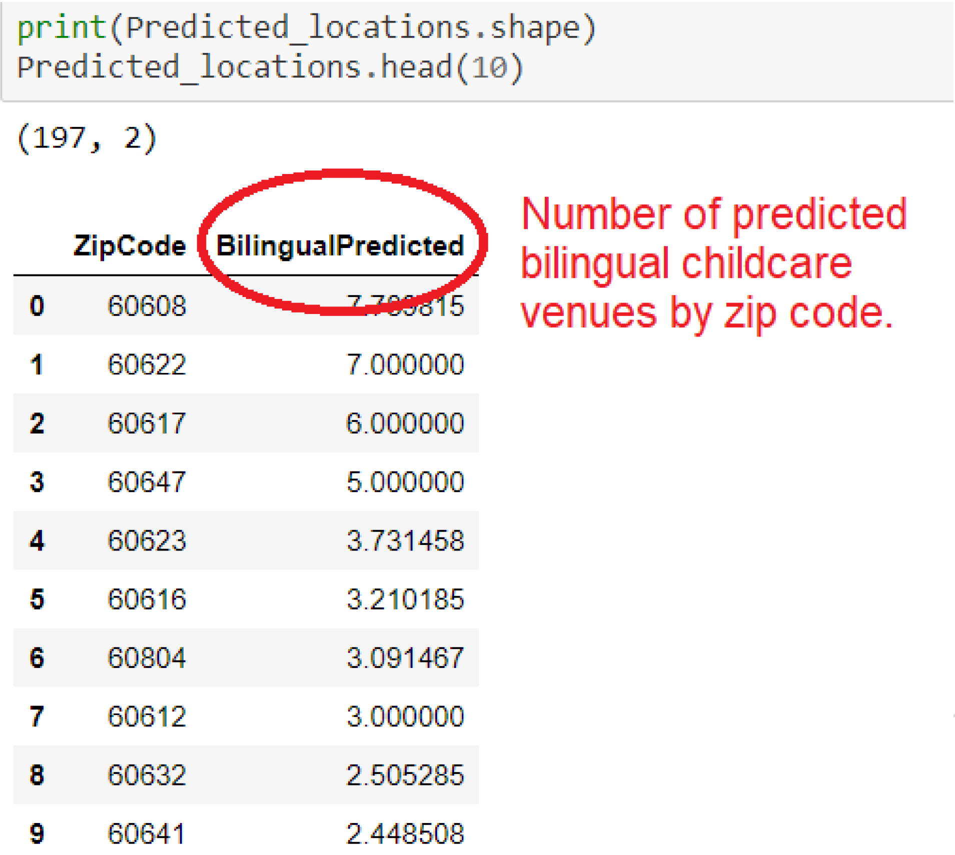


Multiple linear regression model – Cont'd

- Selected the top 13 venue categories and fit them into the model.
  - Resulted in in-sample R2 accuracy of 0.6931 and out-of-sample accuracy of 0.2126.
- Used model to predict bilingual daycare venues in all 197 zip codes.



	VenueCategory	r2BilingualCount
0	Mexican Restaurant	0.297892
1	Taco Place	0.216447
2	Rock Club	0.206948
3	Latin American Restaurant	0.202632
4	Cocktail Bar	0.167925
5	Dessert Shop	0.160914
6	Heliport	0.146841
7	Speakeasy	0.123189
8	Ukrainian Restaurant	0.095090
9	Café	0.094379
10	Casino	0.088312
11	Brewery	0.088113
12	Caribbean Restaurant	0.082222



```
print(Predicted_locations.shape)
Predicted_locations.head(10)
```

```
(197, 2)
```

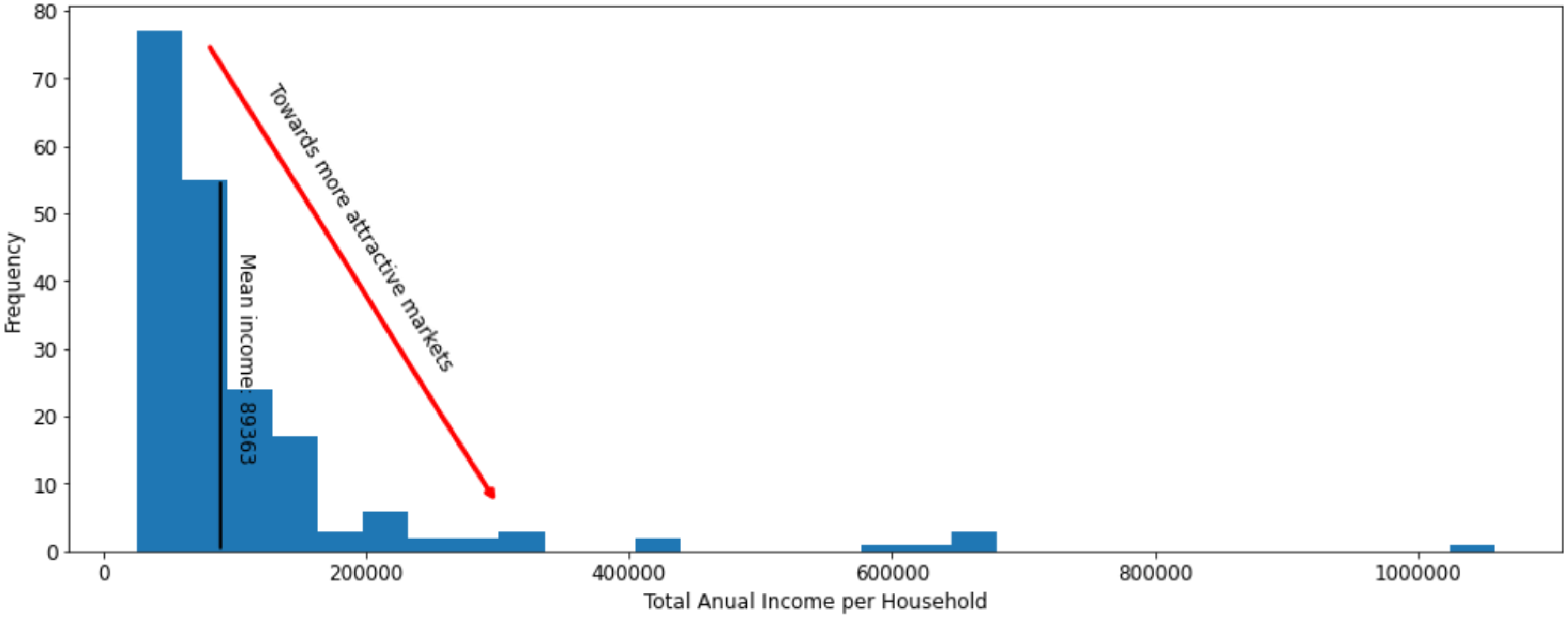
	ZipCode	BilingualPredicted
0	60608	7.788815
1	60622	7.000000
2	60617	6.000000
3	60647	5.000000
4	60623	3.731458
5	60616	3.210185
6	60804	3.091467
7	60612	3.000000
8	60632	2.505285
9	60641	2.448508

Number of predicted bilingual childcare venues by zip code.

Analysis of markets and demographics

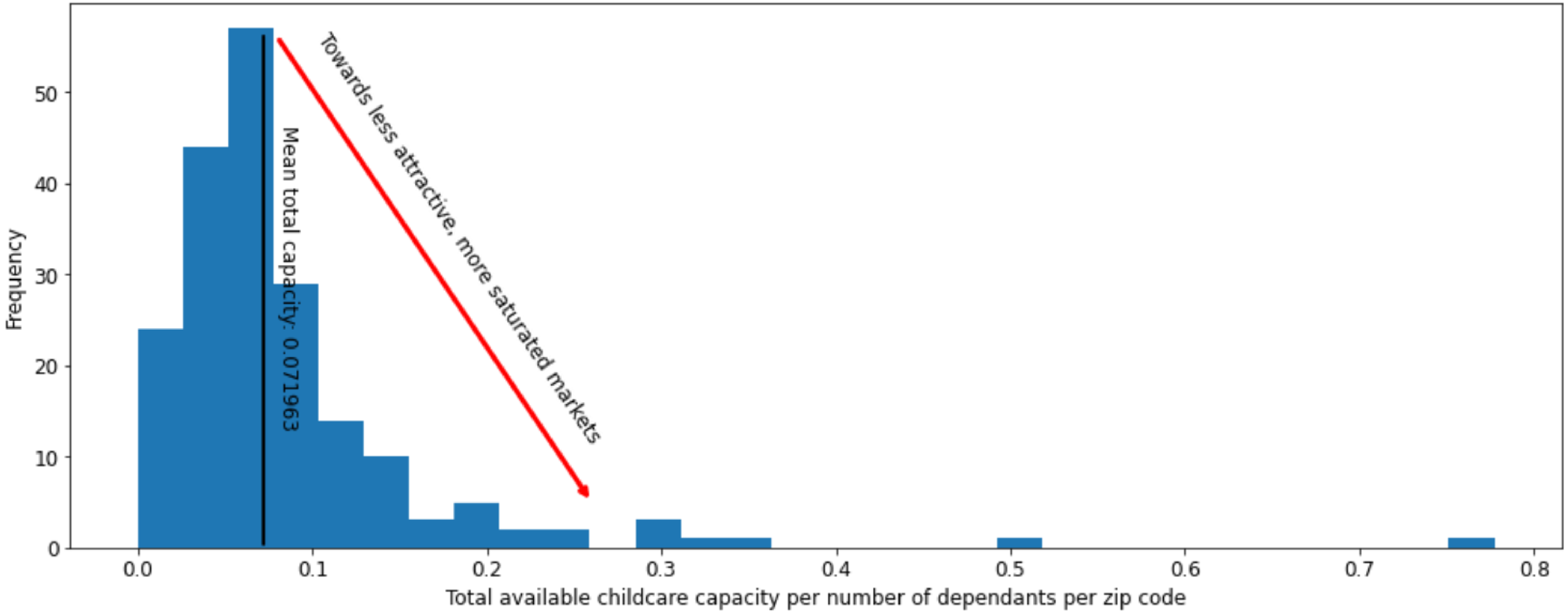
The following factors were analyzed:

- Annual income per household: Measures the household affordability by zip code.



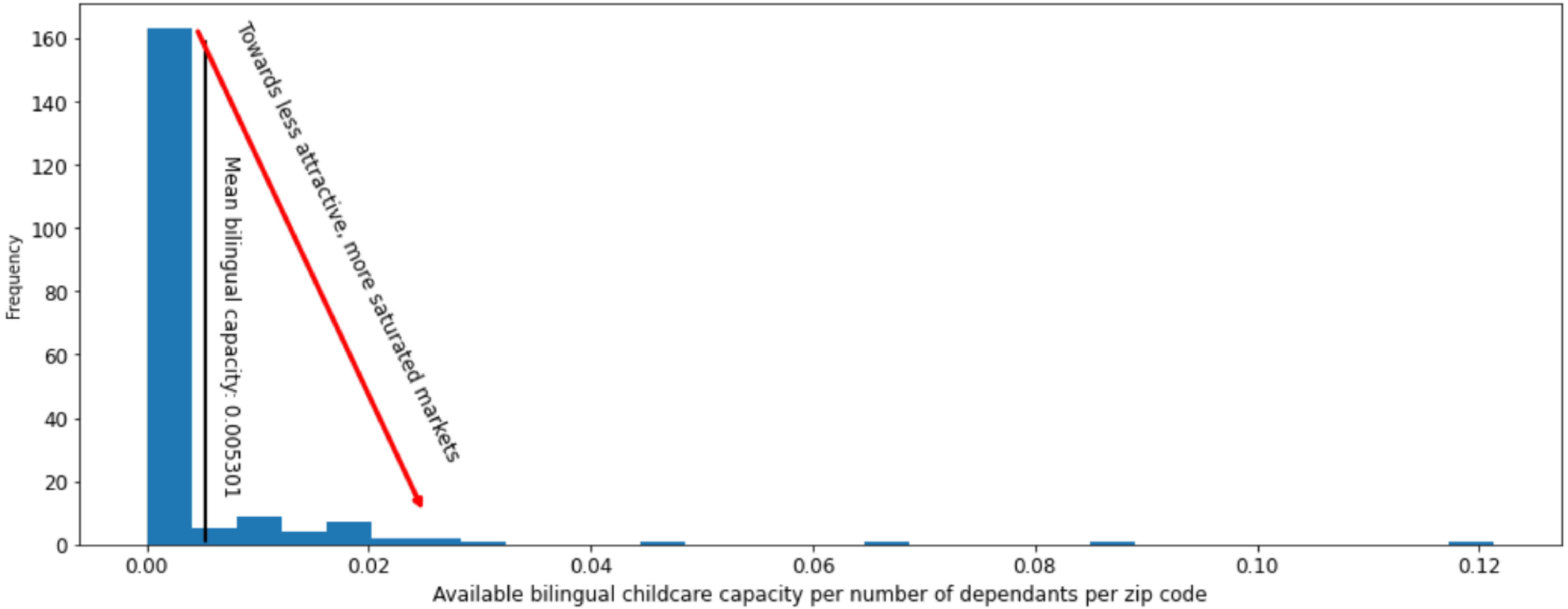
Analysis of markets and demographics – Cont'd

- Total childcare children capacity: Measures the market supply of all childcare venues.



Analysis of markets and demographics – Cont’d

- Bilingual childcare children capacity: Measures the market supply of bilingual childcare venues.



Analysis of markets and demographics – Cont’d

- Number of dependants by zip code: Measures the approximate size of each market.

Further narrowed the data from 197 to 72 zip codes by using the market criteria provided by the company. Only the following zip codes were considered:

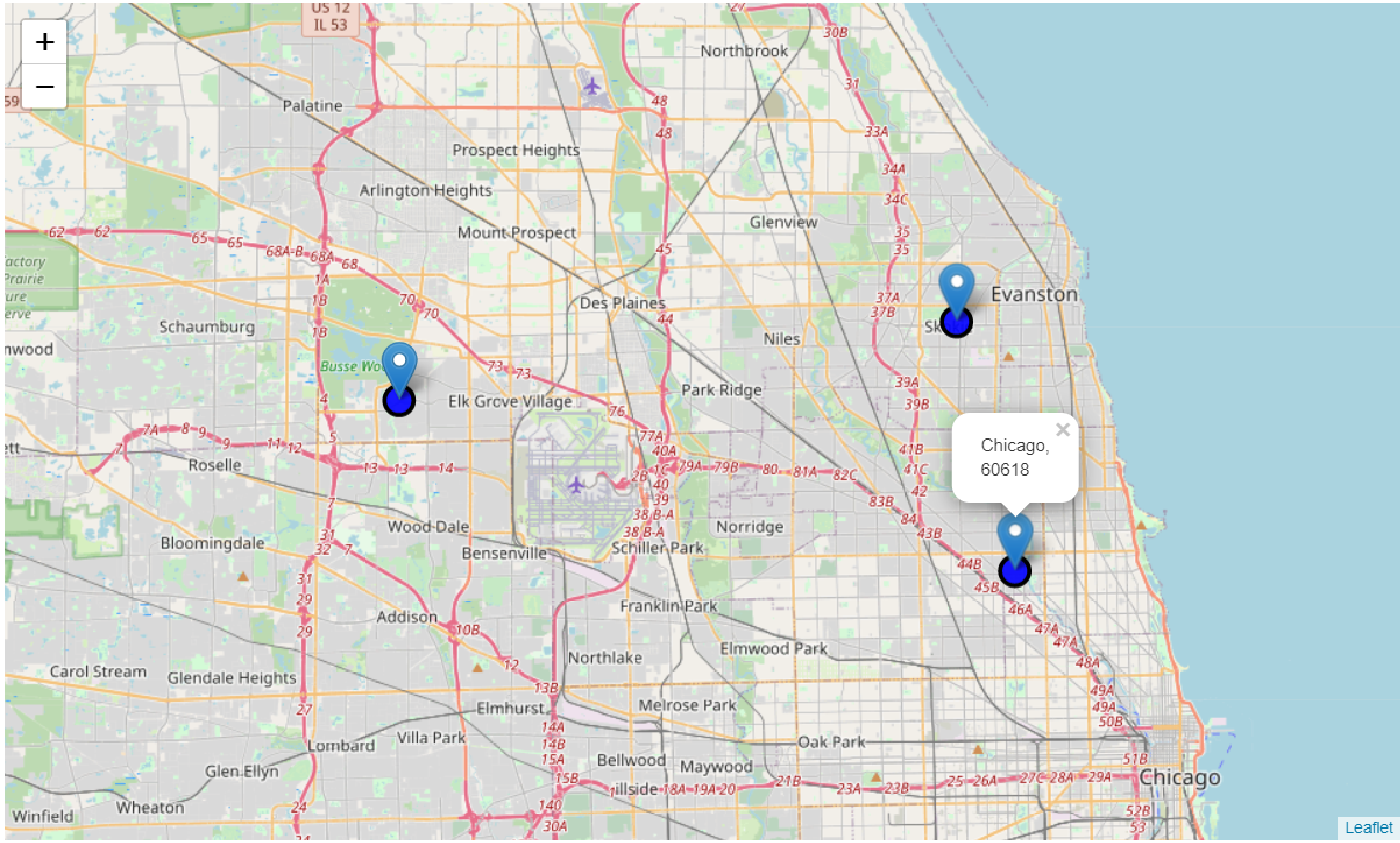
- Zip codes where annual income per household is greater than 75% of the mean annual income (at least 67k USD average household income).
- Zip codes where total available childcare capacity per dependant is less than or equal to 150% of the mean (less than 0.108 children per dependant)
- Zip codes where the available bilingual childcare capacity per dependant is less than or equal to 0.02.

The final recommendation

- 60618, 60007 and 60076 were selected as the top 3 zip codes based on the predictions of the multiple linear regression model, then analyzed separately.
  - 60618
    - 30 childcare venues, 4 of which are bilingual.
    - Highest percentage of Hispanics and Latinos, at more than 40% of its population.
    - Most populated with the largest market for childcare venues of the three.
    - Strategically located between Evanston and Chicago.
    - High affordability: Mean Household income of 87.8k USD.
  - 60007
    - Only 2 childcare venues, none of which are bilingual.
    - Only 11.8% of Hispanics and Latinos.
    - Least number of child care dependants of the three.
    - Mean Household income of 76k USD.
  - 60076
    - 12 childcare venues, none of which are bilingual.
    - Lowest percentage of Hispanics and Latinos, at just 10.6% of its population.
    - Least populated zip code.
    - Mean Household income of 84.2k USD.

The final recommendation – Cont’d

Zip code 60618 was ultimately chosen as the optimal location for the company’s new bilingual childcare venue.


+
−
Northbrook
Palatine
Prospect Heights
Arlington Heights
Mount Prospect
Glenview
Evanston
Des Plaines
Schaumburg
Niles
Busse Wo
Elk Grove Village
Park Ridge
Chicago,
60618
Roselle
Wood Dale
Norridge
Bloomingdale
Bensenville
Schiller Park
Addison
Franklin Park
Northlake
Elmwood Park
Carol Stream
Glendale Heights
Elmhurst
Melrose Park
Lombard
Villa Park
Oak Park
Glen Ellyn
Bellwood
Maywood
Winfield
Wheaton
Chicago
Leaflet