Clustering Toronto neighborhoods for people with children

Problem and interest

Problem:

Toronto is big city with many neighbourhoods which can offer to you different lifestyle, therefore for people with childrens is difficult to choose living area.

Interest:

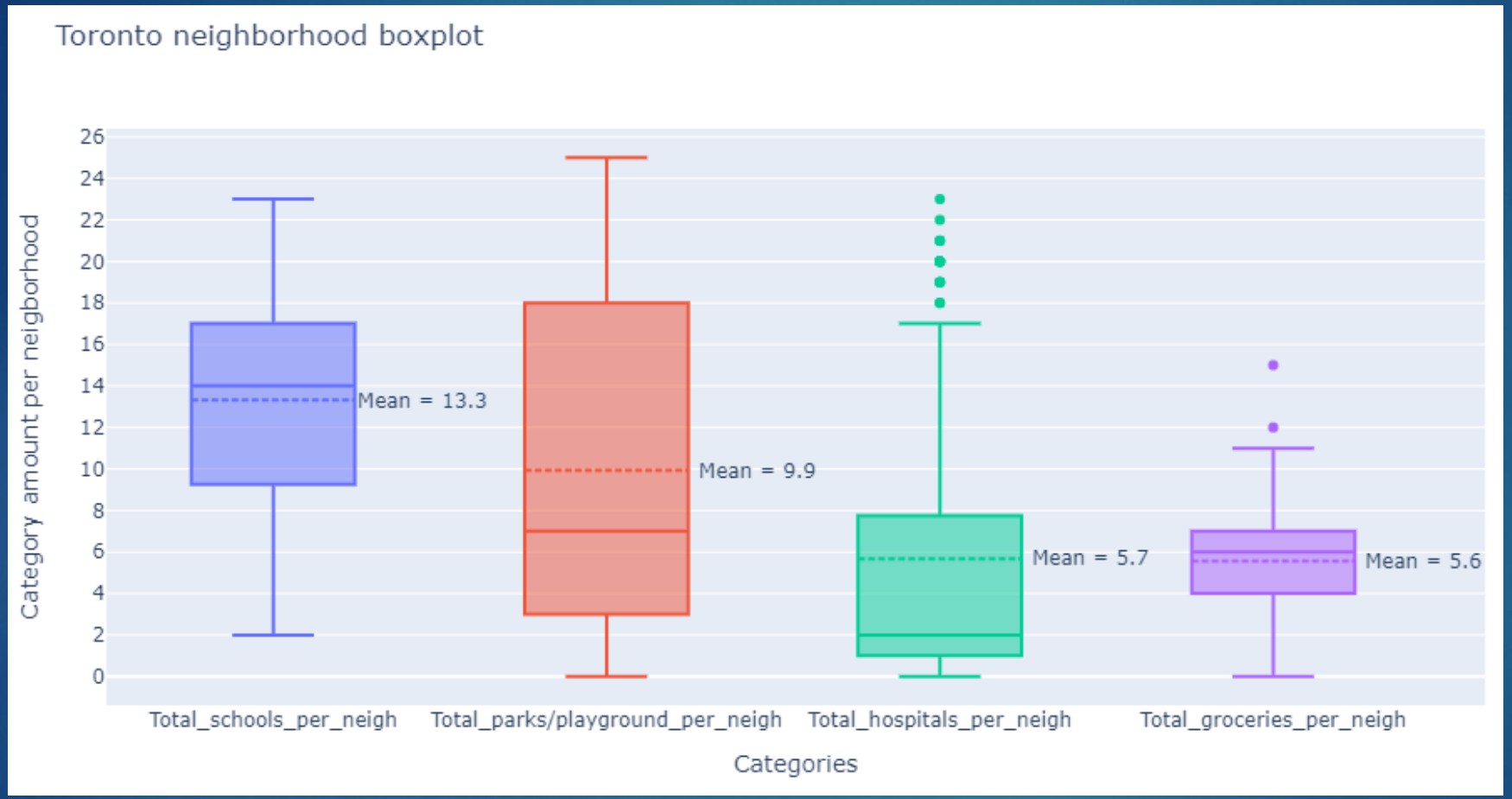
People with children who are thinking to start a life in Toronto.

People with children who wants to change Toronto neighbourhood to similar or more suitable for life with children.

Data acquisition and cleaning

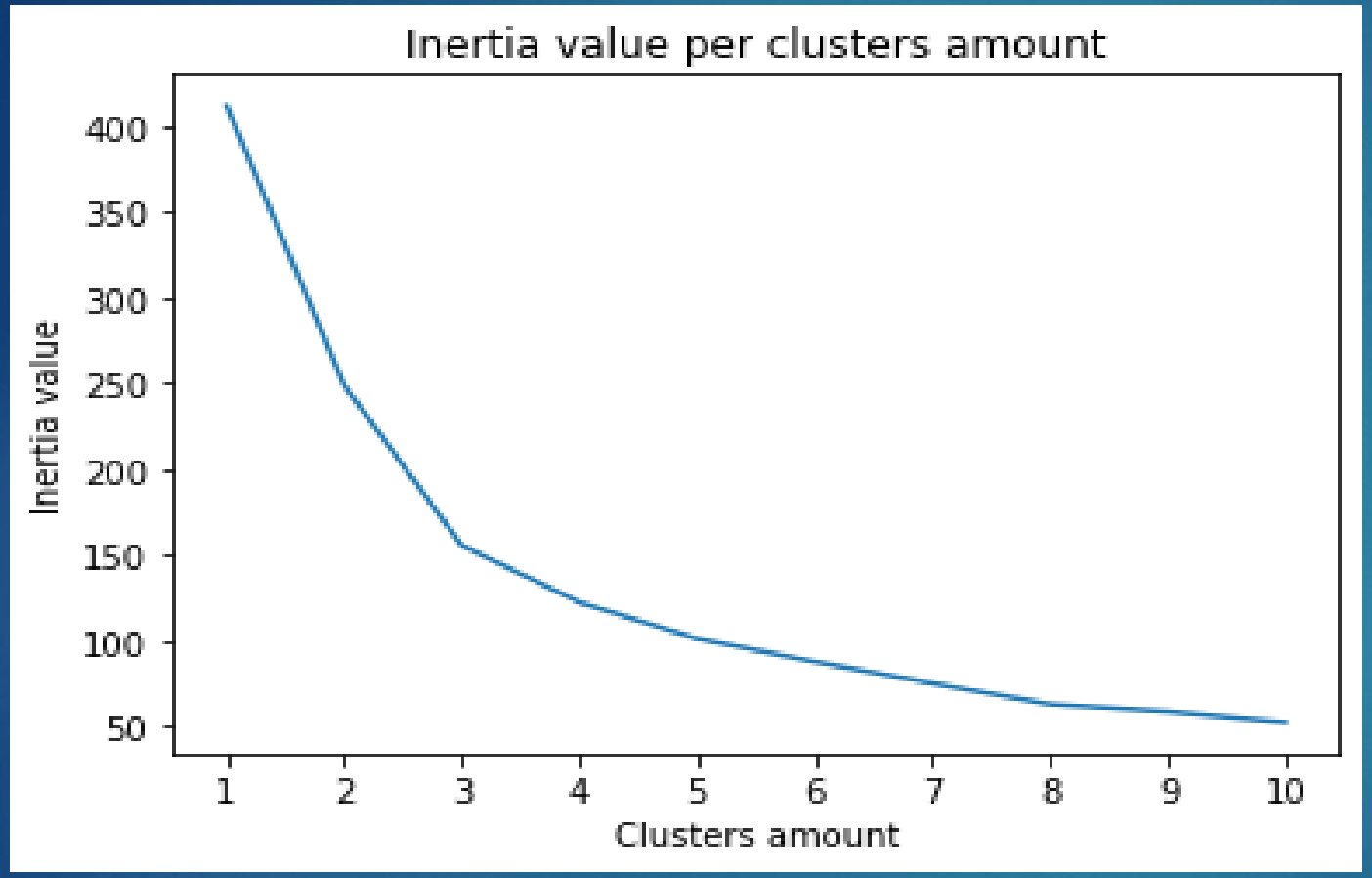
- Toronto neighbourhoods was scraped from https://en.wikipedia.org/ wiki/List_of_postal_codes_of_Canada: ;
- Data for studies was used from Foursquare API: it was search for schools, parks, playgrounds, hospitals and groceries stores.
- Data was cleaned from not relevant information and if Toronto neighbourhood does not have feature it was given 0 value.

Exploratory Data Analysis



- Some Toronto neighbourhoods do not have around them parks/playgrounds, hospitals or grocery stores.
- Hospitals data per neighbourhoods has biggest difference between mean and median and has more outliers than other features.

Predictive model



Higher clusters amount has lower Inertia value: The highest Inertia value changes are from cluster amount 1 to 3.
From cluster amount 4 Inertia value starts to decrease slower.

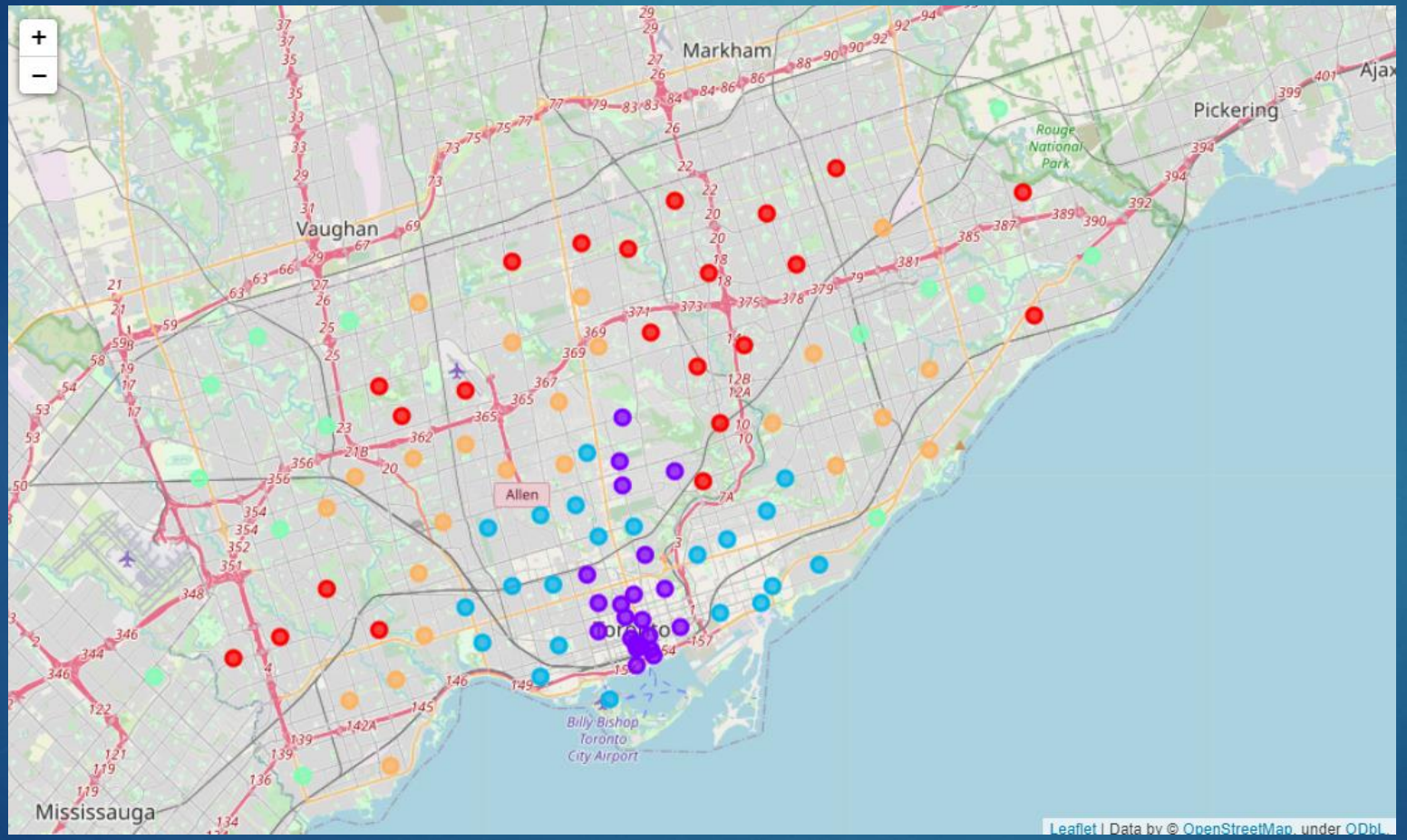
Even model with higher amount of clusters have lower Inertia value in this study is not practical to used high amount of cluster because for people will be harder to interpret results.

Models with 4,5 and 6 clusters

In study was more analysed models with 4, 5 and 6 clusters:

- A model with 4 clusters has high feature variation per cluster, for example, in cluster label 0 is included Toronto neighborhoods which have 2 schools and Toronto neighborhoods which have 20 schools.
- Even model with 6 clusters has lower Inertia value than models with 4 and 5 clusters, it is difficult to interpret clusters results therefore model with 5 cluster was chosen to cluster Toronto data.

Clustered Toronto neighbourhoods into 5 clusters



Red – Cluster label 0: S↑, P↓, H↓, G↓;
Purple – Cluster label 1: S↓, P↑, H↑, G↑;
Blue – Cluster label 2: S↑, P↑, H↓, G↑;
Green – Cluster label 3: S↓, P↓, H↓, G↓;
Orange – Cluster label 4: S↑, P↓, H↓, G↑.

↑ - High amount
↓ - Low amount
S – Schools
P – Parks/ Playground
H – Hospitals
G – Grocery stores

Conclusion and Discussion

In this study a k-mean clustering model was built for Toronto neighborhoods data.

Further improvements:

- To use more features which are important to people with children.
- To get information which Toronto neighbourhood are still in development process and they surroundings will change in the future.