Project title: Cities, A tool to cluster and recommend ideal Israeli cities to live in

Team members info:

Reut Barack, **Email**: reut.barak2@mail.huji.ac.il , **CS id**: reut.barak2
Liora Itenberg, **Email**: liorali22@gmail.com, **CS id**: liorait
Gil Ben Cohen, **Email**: gil.bencohen@gmail.com, **CS id**: gilbc

Problem description:

We plan to create Cities, a tool for clustering and recommending the best Israeli cities to live in, based on personal preferences and on city attributes. For this project, the tool will also be used for Comparison to better understand similarities between the cities of Israel and between different city attributes.

We will use different clustering methods to identify cities with similar attributes. We will investigate correlations between attributes, identify interesting outliers, and try to collect as many independent variables as possible. The final tool should be such that the user inserts favorable attributes, and the algorithm accordingly locates the cities best suited for that user. If later executed as a website for general usage, we can include a recommendation system in our tool, to learn better recommendations from users ranking history. For example, if many users were interested in both "location near nature" and "number of private houses" simultaneously and a new user is interested in "location near nature", the tool would assume that the new user is also interested in "number of private houses". Another idea is to present city similarities based on different traits, and portray them upon the map of Israel.

Data:

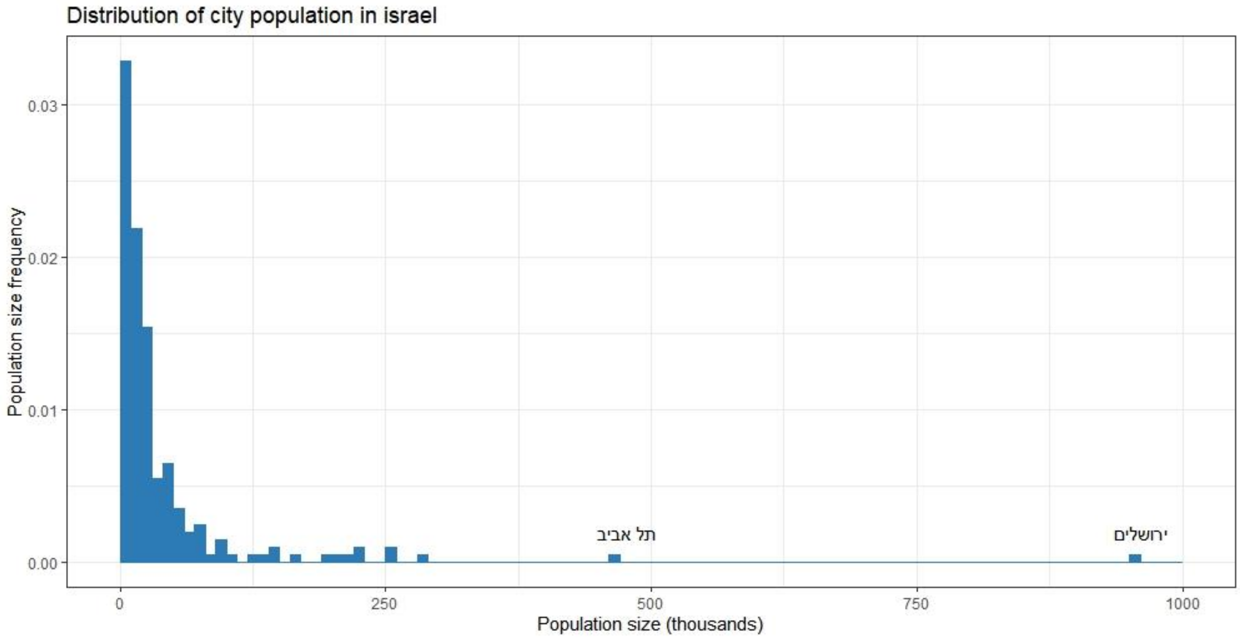
The data we collected contain numerical information about the sizes of the population and the status of education of the cities of Israel.
The data on the size of the population contains 1,216 city records for which the amount of population is given generally and by gender. The data on the status of education contain around 200 city records for which many attributes are available, e.g. data on the number of children in the educational frameworks, number of schools in the locality, percentage of those eligible for graduation, percentage of degree holders in the locality

and more. This data was extracted from The Central Bureau of Statistics (CBS).
We also collected values for employment demands based on linkedin searches for different cities, a process we hope to automate.

Future plans are to extract data from various sites that monitor other characteristics and may interest the user, such as air pollution indices from the Ministry of Environmental Protection's site, areas of nature from the Nature and Parks Authority, or from RATAG, the number of jobs from the AllJobs site, covid19 vaccination percentages from the health ministry site. We also hope to find data regarding proximity to nature, number of private houses compared to buildings, and to come up with other relevant attributes.

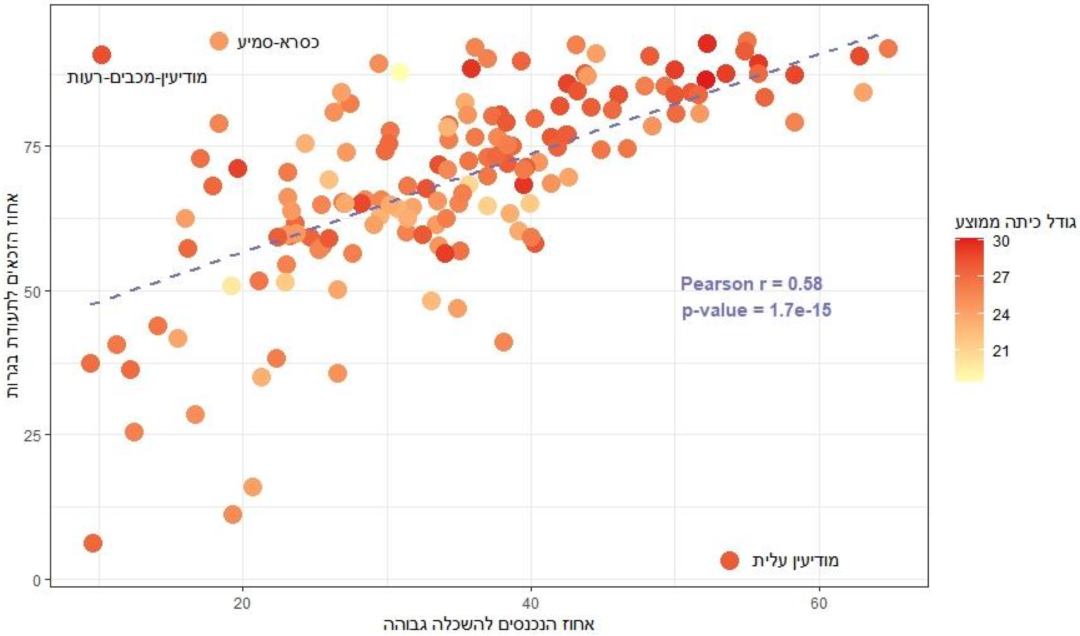
Visualizations:

1. With regards to **city population size**, we assumed that the frequency of cities will drop exponentially as the population size gets larger, because larger cities are much rarer. We also expected to see Jerusalem and Tel Aviv as the largest outliers. All population size values should be smaller than Tel Aviv and Jerusalem, and no cities should have negative values.



We have created a histogram for city population size, and indeed the frequency drops as population size gets larger. This also fits similar distributions we saw for cities in other countries (for example: https://jmichaelbatty.wordpress.com/fractals/city-size-distributions/). Tel Aviv and Jerusalem indeed appear at the extreme, and no values are negative.

2. We assumed that some education parameters will have strong correlation with each other. For example, the number of students graduating from high school in a city should correlate with the percentage of students accepted into academy from the same city.



To examine our assumption, we have created a scatterplot of these two variables, and calculated the Pearson correlation coefficient. Indeed, there appears to be a statistically significant correlation between these variables, with r = 0.58 and a very small p-value, as shown in the plot. We colored the plot for a third educational variable- average class size, based on the assumption that size of classes may affect graduation and academy stats. Surprisingly, it seems that educationally weaker cities have medium to low class sizes, while the largest class sizes tend to be in stronger cities. We also noticed some outliers (e.g: “מודיעין עלית”), also annotated on the plot, that might have false values, or otherwise would be interesting to investigate.