# Data Extraction and Factorize

The database containing 82 columns and 56,000 rows (each row presenting a jobseeker have joined the ES program and each column presenting a jobseeker characteristic like family status, religion, age, number of children, education, etc.). The data was collected for the period from 2016-2019. The data is was analyzed to recognize empty cells or unsuitable cells for each characteristic, 14% of empty cells, and 86% complete data.

1860 unique activities categorized to 30 unique. Age column is categorized to groups of 18-29, 30-39, 40-49, 50-54, 55+. Education' column was divided into categories: elementary, high school, degree, professional certificate, and no education. Religion column also divided into Jewish, Christian, Druze, Muslim, or other. The country of the birth column has grouped to Judea and Samaria, Golan Heights, Gaza, and Sinai. Disability Percentage column was divided into categories: [0,1), [1,20), [20, 40), [40,60), [60,100]. The language column has many options because each language has its level, therefore, all extensions could be subtracted and left with only the name of the language. Licenses column has been converted from subdivision of categories into categories themselves. Family Status was controversial (values such as vacancy, polygamy, alliance, prescription) so they were converted to main: single, married, unknown including the regular of a widow and divorced. The number of children up to the age of 18 have been cataloged by 0-7, 8+.

Changing the "Activities that went through the program" column was a must because when employees of the IES open a new activity, they can give a free text to the program name and not choose one from a predefined list. Therefore, 1860 unique values were created and had to be filtered, categorized, and creating a predefined list of smaller and pre-agreed activities. After manipulating the data, only 30 unique activities were left. After these actions, all activities had to be deployed to indicate whether the jobseeker was in the same activity, if the jobseeker was in the activity receiving the value 1 and if not receiving value 0.

'Depth of unemployment in months' column values are from 0 to 100 were divided by 10 with the result rounded to get a categorical column. The age column is categorized according to what is used in the IES, groups of 18-29, 30-39, 40-49, 50-54, 55+.  
The 'Education' column also has a lot of values that can be entered in free text, so it was divided into categories: elementary, high school, degree, professional certificate, and no education. The "Religion" column is also divided into the major religions: Jewish, Christian, Druze, Muslim or other. Of course, each one contains several types and so it was decided to group them. Edge cases will not affect the data so be grouped to 'Other'.  
The 'country of birth' column has strange values that belong to Israel but appear as a different value from it: Judea and Samaria, Golan Heights, Gaza, and Sinai. By logical thinking, I decided to associate you with the value of 'Israel'. The 'Disability Percentage' column was divided into categories: [0,1), [1,20), [20, 40), [40,60), [60,100]. The 'Language' column has many options because each language has its level. Therefore, it was decided that all extensions could be subtracted and left with only the name of the language. The 'Licenses' column has been converted from subdivision of categories into categories themselves. Values in the 'family status' column were controversial (values such as vacancy, polygamy, alliance, prescription) so they were converted to main: single, married, unknown. Of course, there are also: widows, divorced. 'Number of children' up to the age of 18 have been cataloged by 0-7, 8+.

***Table 1****First part of Socio-demographics features*

|  |  |  |  |
| --- | --- | --- | --- |
| **Socio-demographics features** | | **Number of Job Seekers** | **Weight (%)** |
| Gender | Male | 22,765 | 41% |
| Female | 33,224 | 59% |
| Age | 18-29 | 9,192 | 16% |
| 30-39 | 15,213 | 27% |
| 40-49 | 19,304 | 34% |
| 50-54 | 8,386 | 15% |
| 55+ | 3,894 | 7% |
| Religion | Jews | 23,887 | 43% |
| Muslims | 25,876 | 46% |
| Druze | 1,761 | 3% |
| Christianity | 1,361 | 2% |
| Other | 3,102 | 6% |
| No Religion | 2 | 0% |
| Singal Parent | Yes | 5,880 | 11% |
| No | 50,109 | 89% |
| Education | Primary School | 11,140 | 20% |
| Part Secondary school | 10,650 | 19% |
| Secondary school | 25,467 | 45% |
| Bagrut certificate | 2,436 | 4% |
| First Degree | 1,943 | 3% |
| Second Degree | 732 | 1% |
| Third Degree | 68 | 0% |
| Associate degree | 1,372 | 2% |
| License | 10 | 0% |
| No Education | 2,171 | 4% |

***Table 2****Second part of Socio-demographics features*

|  |  |  |  |
| --- | --- | --- | --- |
| **Socio-demographics features** | | **Number of Job Seekers** | **Weight (%)** |
| Disability | None | 50,825 | 91% |
| 1% to 19% | 657 | 1% |
| 20% to 39% | 1,966 | 4% |
| 40% to 59% | 1,372 | 2% |
| 60% to 100% | 1,169 | 2% |
| Family Status | Single | 17,205 | 31% |
| Married | 25,117 | 45% |
| Divorced | 11,840 | 21% |
| Widow | 1,511 | 3% |
| Unknown | 316 | 1% |
| Number of kids under 18 | None | 22,400 | 40% |
| 1 to 3 | 23,428 | 42% |
| 4 to 6 | 8,367 | 15% |
| 7+ | 1,794 | 3% |
| Military Service / Sherut Leumi | None | 45,752 | 82% |
| Sherut Leumi | 941 | 2% |
| Military Service | 9,296 | 17% |
| Unemployment duration | <1 year | 4,519 | 8% |
| 2 - 3 years | 5,166 | 9% |
| 3> years | 46,100 | 82% |
| None | 204 | 0% |