**BDA 505 – BIG DATA MANAGEMENT**

**TERM PROJECT**

**PROPOSAL**

# TEAM

Ahmet Yetkin Eser

Berkay Soyer

Feray Ece Topcu

# DATASET

The United States Census Bureau’s International Dataset provides estimates of country populations since 1950 and projections through 2050. Specifically, the data set includes midyear population figures broken down by age and gender assignment at birth. Additionally, they provide time-series data for attributes including fertility rates, birth rates, death rates, and migration rates.

The U.S. Census Bureau provides estimates and projections for countries and areas that are recognized by the U.S. Department of State that have a population of at least 5,000.

This dataset can be available from this URL:

<https://www.kaggle.com/census/international-data/data>

Size of total dataset is 1.70 GB. It is formed by 8 different csv file that are explained below.

## age\_specific\_fertility\_rates.csv

This file includes fertility rate according to specific age categories for each country and each year from 1950 and 2050. Additionally, there is a total fertility rate info for each country for each year. Row number is 15107 and column number is 12 for this file.

Columns can be explained as below:



## birth\_death\_growth\_rates.csv

This file is formed by birth and death rates for each country for each year. Row number is 15110 and there are 8 different columns that explained as below:



## country\_names\_area.csv

This file involves country names and codes for each country. There are 229 country in this dataset.



## midyear\_population.csv

This file is comprised by midyear population information for each country and year. There are 23028 rows and 4 columns.



## midyear\_population\_5yr\_age\_sex.csv

This file includes midyear population, midyear population of male and female for different age levels (five-yeaar segments generated by dividing ages from 0 to 100 by 5). There are 330080 rows and 10 columns in this file.



## midyear\_population\_age\_country\_code.csv

This large csv file includes the midyear population for each age and year. Additionally there are some extra columns like age and total population. This data is not proper, actually. It can dived into two different tables. It has more than 1000000 rows and 107 columns.

|  |  |  |
| --- | --- | --- |
| **Columns** | **Data Type** | **Explanation** |
| country\_code | STRING | Federal Information Processing Standard (FIPS) country/area code |
| country\_name | STRING | Country or area name |
| year | INTEGER | Year |
| sex | STRING | Gender |
| max\_age | INTEGER | The last age in the distribution with a value greater than zero |
| population\_age\_0 | INTEGER | Population at Age 0 |
| population\_age\_1 | INTEGER | Population at Age 1 |
| population\_age\_2 | INTEGER | Population at Age 2 |
| population\_age\_3 | INTEGER | Population at Age 3 |
| population\_age\_4 | INTEGER | Population at Age 4 |
| population\_age\_5 | INTEGER | Population at Age 5 |
| population\_age\_6 | INTEGER | Population at Age 6 |
| population\_age\_7 | INTEGER | Population at Age 7 |
| population\_age\_8 | INTEGER | Population at Age 8 |
| population\_age\_9 | INTEGER | Population at Age 9 |
| population\_age\_10 | INTEGER | Population at Age 10 |
| population\_age\_11 | INTEGER | Population at Age 11 |
| population\_age\_12 | INTEGER | Population at Age 12 |
| population\_age\_13 | INTEGER | Population at Age 13 |
| population\_age\_14 | INTEGER | Population at Age 14 |
| population\_age\_15 | INTEGER | Population at Age 15 |
| population\_age\_16 | INTEGER | Population at Age 16 |
| population\_age\_17 | INTEGER | Population at Age 17 |
| population\_age\_18 | INTEGER | Population at Age 18 |
| population\_age\_19 | INTEGER | Population at Age 19 |
| population\_age\_20 | INTEGER | Population at Age 20 |
| population\_age\_21 | INTEGER | Population at Age 21 |
| population\_age\_22 | INTEGER | Population at Age 22 |
| population\_age\_23 | INTEGER | Population at Age 23 |
| population\_age\_24 | INTEGER | Population at Age 24 |
| population\_age\_25 | INTEGER | Population at Age 25 |
| population\_age\_26 | INTEGER | Population at Age 26 |
| population\_age\_27 | INTEGER | Population at Age 27 |
| population\_age\_28 | INTEGER | Population at Age 28 |
| population\_age\_29 | INTEGER | Population at Age 29 |
| population\_age\_30 | INTEGER | Population at Age 30 |
| population\_age\_31 | INTEGER | Population at Age 31 |
| population\_age\_32 | INTEGER | Population at Age 32 |
| population\_age\_33 | INTEGER | Population at Age 33 |
| population\_age\_34 | INTEGER | Population at Age 34 |
| population\_age\_35 | INTEGER | Population at Age 35 |
| population\_age\_36 | INTEGER | Population at Age 36 |
| population\_age\_37 | INTEGER | Population at Age 37 |
| population\_age\_38 | INTEGER | Population at Age 38 |
| population\_age\_39 | INTEGER | Population at Age 39 |
| population\_age\_40 | INTEGER | Population at Age 40 |
| population\_age\_41 | INTEGER | Population at Age 41 |
| population\_age\_42 | INTEGER | Population at Age 42 |
| population\_age\_43 | INTEGER | Population at Age 43 |
| population\_age\_44 | INTEGER | Population at Age 44 |
| population\_age\_45 | INTEGER | Population at Age 45 |
| population\_age\_46 | INTEGER | Population at Age 46 |
| population\_age\_47 | INTEGER | Population at Age 47 |
| population\_age\_48 | INTEGER | Population at Age 48 |
| population\_age\_49 | INTEGER | Population at Age 49 |
| population\_age\_50 | INTEGER | Population at Age 50 |
| population\_age\_51 | INTEGER | Population at Age 51 |
| population\_age\_52 | INTEGER | Population at Age 52 |
| population\_age\_53 | INTEGER | Population at Age 53 |
| population\_age\_54 | INTEGER | Population at Age 54 |
| population\_age\_55 | INTEGER | Population at Age 55 |
| population\_age\_56 | INTEGER | Population at Age 56 |
| population\_age\_57 | INTEGER | Population at Age 57 |
| population\_age\_58 | INTEGER | Population at Age 58 |
| population\_age\_59 | INTEGER | Population at Age 59 |
| population\_age\_60 | INTEGER | Population at Age 60 |
| population\_age\_61 | INTEGER | Population at Age 61 |
| population\_age\_62 | INTEGER | Population at Age 62 |
| population\_age\_63 | INTEGER | Population at Age 63 |
| population\_age\_64 | INTEGER | Population at Age 64 |
| population\_age\_65 | INTEGER | Population at Age 65 |
| population\_age\_66 | INTEGER | Population at Age 66 |
| population\_age\_67 | INTEGER | Population at Age 67 |
| population\_age\_68 | INTEGER | Population at Age 68 |
| population\_age\_69 | INTEGER | Population at Age 69 |
| population\_age\_70 | INTEGER | Population at Age 70 |
| population\_age\_71 | INTEGER | Population at Age 71 |
| population\_age\_72 | INTEGER | Population at Age 72 |
| population\_age\_73 | INTEGER | Population at Age 73 |
| population\_age\_74 | INTEGER | Population at Age 74 |
| population\_age\_75 | INTEGER | Population at Age 75 |
| population\_age\_76 | INTEGER | Population at Age 76 |
| population\_age\_77 | INTEGER | Population at Age 77 |
| population\_age\_78 | INTEGER | Population at Age 78 |
| population\_age\_79 | INTEGER | Population at Age 79 |
| population\_age\_80 | INTEGER | Population at Age 80 |
| population\_age\_81 | INTEGER | Population at Age 81 |
| population\_age\_82 | INTEGER | Population at Age 82 |
| population\_age\_83 | INTEGER | Population at Age 83 |
| population\_age\_84 | INTEGER | Population at Age 84 |
| population\_age\_85 | INTEGER | Population at Age 85 |
| population\_age\_86 | INTEGER | Population at Age 86 |
| population\_age\_87 | INTEGER | Population at Age 87 |
| population\_age\_88 | INTEGER | Population at Age 88 |
| population\_age\_89 | INTEGER | Population at Age 89 |
| population\_age\_90 | INTEGER | Population at Age 90 |
| population\_age\_91 | INTEGER | Population at Age 91 |
| population\_age\_92 | INTEGER | Population at Age 92 |
| population\_age\_93 | INTEGER | Population at Age 93 |
| population\_age\_94 | INTEGER | Population at Age 94 |
| population\_age\_95 | INTEGER | Population at Age 95 |
| population\_age\_96 | INTEGER | Population at Age 96 |
| population\_age\_97 | INTEGER | Population at Age 97 |
| population\_age\_98 | INTEGER | Population at Age 98 |
| population\_age\_99 | INTEGER | Population at Age 99 |
| population\_age\_100 | INTEGER | Population at Age 100 |
| Age | INTEGER | Age |
| Permutation | STRING | All rows are tagged as ‘child’ |
| Population | INTEGER | Total population for the country for the year. |

## midyear\_population\_age\_sex.csv

This file includes clean data of midyear\_population\_age\_country\_code.csv file. The dataset is formed by midyear population information for each age group and for each year according to genders. Interesting part of this file, dataset of it not vertical, it is horizontal like above file.

The columns are same with midyear\_population\_age\_country\_code.csv except the last three.

## mortality\_life\_expectancy.csv

This file involves the dataset of mortality rate information for different parameters like sex, age ranges and etc.



# Objective & Methodology

Our aim is focusing on Turkey data. The main objective is understanding Turkey population distribution based on the different parameters like age, sex, year and etc. Exploring some information about population growing rate according to gender may be additional part. Furthermore, we want to visualize this numbers with graphs to make this dataset more understandable.

Firstly, we want to import these files into PostgreSQL database system by creating one table for each csv file. The primary key of each table will be Country Code and Year combination, so the tables can be combined with these two columns as foreign key.

Secondly, we want to use R language for making data analysis. It will be some descriptive analysis about Turkey data according to the main objective of project. This part will be evaluated with R-markdown format, so we planned to generate an html report.

To sum up, the main aim is understanding the distribution of Turkey population and exploring some interesting information by using PostgreSQL and R Language.