BDA 505 – BIG DATA MANAGEMENT

TERM PROJECT

PROPOSAL

TEAM

Ahmet Yetkin Eser

Berkay Soyer

Feray Ece Topcu

1. 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.

2.1. 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:

Column Name	Data Type	Explanation
country_code	STRING	Federal Information Processing Standard (FIPS) country/area code
country_name	STRING	Country or area name
year	INTEGER	Year
fertility_rate_15_19	FLOAT	Age specific fertility rate for age 15-19 (births per 1,000 population)
fertility_rate_20_24	FLOAT	Age specific fertility rate for age 20-24 (births per 1,000 population)
fertility_rate_25_29	FLOAT	Age specific fertility rate for age 25-29 (births per 1,000 population)
fertility_rate_30_34	FLOAT	Age specific fertility rate for age 30-34 (births per 1,000 population)
fertility_rate_35_39	FLOAT	Age specific fertility rate for age 35-39 (births per 1,000 population)
fertility_rate_40_44	FLOAT	Age specific fertility rate for age 40-44 (births per 1,000 population)
fertility_rate_45_49	FLOAT	Age specific fertility rate for age 45-49 (births per 1,000 population)
total_fertility_rate	FLOAT	Total fertility rate (lifetime births per woman)
gross_reproduction_rate	FLOAT	Gross reproduction rate (lifetime female births per woman)
sex_ratio_at_birth	FLOAT	Sex ratio at birth (male births per female birth)

2.2. 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:

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
crude_birth_rate	FLOAT	Crude birth rate (births per 1,000 population)
crude_death_rate	FLOAT	Crude death rate (deaths per 1,000 population)
net_migration	FLOAT	Net migration rate (net number of migrants per 1,000 population)
rate_natural_increase	FLOAT	Rate of natural increase (percent)
growth_rate	FLOAT	Growth rate (percent)

2.3. country_names_area.csv

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

Columns	Data Type	Explanation
country_code	STRING	Federal Information Processing Standard (FIPS) country/area code
country_name	STRING	Country or area name
country_area	FLOAT	Area in square kilometers

2.4. midyear_population.csv

This file is comprised by midyear population information for each country and year. There are 23028 rows and 4 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
midyear_population	INTEGER	Both sexes midyear population

2.5. 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.

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
total_flag	STRING	Total flag: "*"=Total, all ages; "A"=Individual age group
starting_age	INTEGER	Starting age (0 to 100)
age_group_indicator	STRING	Age group indicator: "-"=5-year age group; "+"=open-ended age group
ending_age	INTEGER	Ending age (4 to 99; set to 0 if G="+")
midyear_population	INTEGER	Both sexes midyear population in the age group
midyear_population_male	INTEGER	Male midyear population in the age group
midyear_population_female	INTEGER	Female midyear population in the age group

2.6. 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.

2.7. 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.

2.8. mortality_life_expectancy.csv

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

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
infant_mortality	FLOAT	Both sexes infant mortality rate (infant deaths per 1,000 population)
infant_mortality_male	FLOAT	Male infant mortality rate (infant deaths per 1,000 population)
infant_mortality_female	FLOAT	Female infant mortality rate (infant deaths per 1,000 population)
life_expectancy	FLOAT	Both sexes life expectancy at birth (years)
life_expectancy_male	FLOAT	Male life expectancy at birth (years)
life_expectancy_female	FLOAT	Female life expectancy at birth (years)
mortality_rate_under5	FLOAT	Both sexes under-5 mortality rate (probability of dying between ages 0 and 5)
mortality_rate_under5_male	FLOAT	Male sexes under-5 mortality rate (probability of dying between ages 0 and 5)
mortality_rate_under5_female	FLOAT	Female sexes under-5 mortality rate (probability of dying between ages 0 and 5)
mortality_rate_1to4	FLOAT	Both sexes child mortality rate (probability of dying between ages 1 and 4)
mortality_rate_1to4_male	FLOAT	Male sexes child mortality rate (probability of dying between ages 1 and 4)
mortality_rate_1to4_female	FLOAT	Female sexes child mortality rate (probability of dying between ages 1 and 4)

3. 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.