TU Dortmund

Introductory Case Studies

Project 1: Descriptive analysis of demographic data

Lecturers:
Prof. Dr. Jörg Rahnenführer
Dr. Philipp Adämmer
Dr. Andrea Bommert
M. Sc. Hendrik Dohme

Author: Alp Yalçın

Group number: 13

Group members: Ezgi Osmanoğlu, Elif Aykut, Yazan Alhalabi, Marges Pinderi

May 14, 2021

Contents

1 Introduction

Demographic structure of the world population has been one of the most crucial information for humanity for ages. Economic and social structures of countries, political moves of governments and marketing activities and exercise are based on the demographical structure of the countries. Additionally, demographic structures can be effected from lots of external factors and extraordinary situations.

The main aim of this report is to make descriptive analysis of demographic data of 228 countries in 2020 and the make brief comparison of years 2000 and 2020. The report mainly is focused on fertility rate and life expectancy parameters for the countries and regions.

In second part of the report, data set and data quality are clarified. Also, project objectives are explained in detail. In third part, methods are used for the statistical analysis are presented and explained. In section four, statistical analysis are performed. Density histograms, scatterplots, correlation tables and boxplots are used to enrich the quality of the results and analysis. Lastly, summary and some comments of the report is given.

2 Problem statement

2.1 Data set and data quality

This report is compiled from International Programs, Population Division, U.S. Census Bureau (2020) dataset of the U.S. Census Bureau.

The Census Bureau prepares national estimates and projections for all countries using census and survey data, vital statistics, administrative statistics from those countries, and information from multinational organizations that collect and publish data for these countries (International Programs, Population Division, U.S. Census Bureau, 2020).

IDB comprise various demographic data with the various indicators such as population by age and sex and demographic characteristics from 1950 to 2100 of the countries which have population of 5000 or more. Within this time series, the Census Bureau has developed sex ratios, population, fertility, mortality and migration measures for single ages.

The data set which is used for this report is extracted from the IDB. It contains life expectancy at birth for both sexes and for males and females, and fertility rates for 228 countries. Also countries are divided to into 5 regions and 21 sub regions geographically in the data set.

The data set includes 456 observations of 10 variables. It has 5 character variables which contains Country Name(Country), FIPS (FIPS country/area Code), GENC (Geopolitical Entities, Names, and Codes), Subregion and Region and 4 numeric variables which contains, Total Fertility Rate(TFR) (2 decimal places), Life Expectancy at Birth for both sexes(LEB), males(LEM) and females(LEF) (1 decimal places). The data set also have one integer variable Year which includes 2 values; 2000 and 2020. Year of 2020 is used for making descriptive analysis and 2000 is used for making comparisons to see how variables changed over 20 years.

Overall, data quality is adequate and data set is clean. However, 7 N/A values exist in 4 numeric variables which are Total Fertility Rates and Life Expectancy at birth (both, males and females) for countries Honduras, Libya, Puerto Rico, South Sudan, Sudan, Syria and United States year of 2000. Additionally, 2 N/A values exist in GENC variable for country of Namibia years of 2000 and 2020. N/A values are excluded from the analysis of this report.

2.2 Project objectives

The main objective of the project is making descriptive analysis of demographic data. Therefore, firstly, frequency distributions of the variables are described with separating numeric variables and central tendency and spread are measured. Histograms are used to explain frequency distributions visually for each numeric variable. Also mean of difference of life expectancy at birth between males and females is considered and is used in histogram as well. Second of all, scatter plot matrix and Pearson correlation table are used for bivariate correlation between the variables. Thirdly, values are grouped region-wise and subregion-wise to comparison of the variability of the values. Box plots are used for visualizing homogeneity within subregions and heterogeneity between different subregions. Last but not least, mean of difference of total fertility rate between years 2000 and 2020 are taken to compare the change of the variables on those years and scatter plots are used to visualize the comparison of values.

3 Statistical methods

For all analyses and calculation, the software (R Development Core Team, 2020) with packages **dplyr** (Wickham et al., 2021), **formattable** (Ren and Russell, 2021), **gridExtra** (Auguie, 2017), **moments** (Komsta and Novomestky, 2015), **skimr** (Waring et al., 2021), **ggplot2** (Wickham, 2016).

3.1 Statistics for samples

Measures of Central Tendency

Arithmetic Mean

The sample mean. $\tilde{x}$, is defined by

$$\tilde{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

The same formula, and function applies for a known population in computing the population mean µ.

The mean is typical for each sample and is not mandatorily included in the dataset for which it is computed. Mean is the most convenient measure for the homogeneous or symmetric data. For data which contain extreme values in one of the tails, another measures which are less sensitive of central tendency are preferred.

(Hay-Jahans, 2019, p. 73)

Median

Suppose the sample $x_1, x_2, ..., x_n$ is arranged in ascending order, the median, $\tilde{x}$, is obtained using the rule

$$\tilde{x} = \begin{cases} x_{(n+1)}/2, & \text{if n is odd} \\ [x_{n/2} + x_{n/2+1}]/2, & \text{if n is even} \end{cases}$$

This measure is irresponsive to excessive values in a sample and when the data contain excessive values in one of the tails, median is frequently chosen.

(Hay-Jahans, 2019, p. 76)

Measures of Spread

To understand how data are distributed across the probable values, frequency and grouped frequency distributions may be used. For numeric data, there are some general measures of spread. (Hay-Jahans, 2019, p. 76)

Consider a sample of numeric data, $x_1, x_2, ..., x_n$

Variance and Standard Deviation

The sample variance, s^2, can be computed using

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{(n-1)}$$

The sample standard deviation is simply $s = \sqrt{s^2}$

(Hay-Jahans, 2019, p. 76)

Measures of Position

Percentiles and Quantiles

The p^{th} percentile, for a given dataset $x_1, x_2, ..., x_n$ regulated in uprising order is the number q_p that corresponds to a percentile rank $100 \times p\%$, and is acquired with the rule

$$q_p = \begin{cases} x_k, & \text{where } k = \lceil np \rceil \text{ and } np \text{ is not an integer} \\ (x_k + x_{k+1})/2, & \text{where } k = np \text{ and } np \text{ is an integer} \end{cases}$$

where $\lceil \cdot \rceil$ denotes the $ceilingfunction$.
Other terminology relevant to R that appear in this context are: x_k is referred to as the k^{th} order statistic in the data, and q_p is referred to as the sample quantile corresponding to a given probability p. So, a quantile (another word for a percentile) is a data estimate for a given dataset that corresponds to a given percentile rank.

(Hay-Jahans, 2019, p. 79)

Function **quantile** is used to attain quantiles in R, which definition is:

quantile(x, probs = seq(0, 1, 0.25), na.rm = FALSE, names = TRUE, type = 7, ...)

Type = 7 is the default algorithm for the function quantile that computes the quantile suitable to a given percentile ranks p, through addition as follows. The percentile rank of the jth order statistic, x$_j$, in a given sample is defined by

$$p_j = \frac{j-1}{n-1}$$

Then, for a given percentile rank p, the order statistics x$_j$ and x$_{j+1}$ are defined for which

$$p_j \leq p \leq p_{j+1}$$

It end up that under this circumstance j= $\lfloor$(n $-$ 1)p + 1$\rfloor$, where $\lfloor \cdot \rfloor$ signify the $floorfunction$, and the quantile corresponding to p related with the given sample is computed using

$$q_p = x_j + \left[\frac{p - p_j}{p_{j+1} - p_j}\right](x_{j+1} - x_j)$$

(Hay-Jahans, 2019, p. 80)

3.2 Measures of Shape

Density Histograms

Histogram is used to demonstrate grouped frequency distributions graphically.

When comparing the distributions of two or more populations empirically,histograms can be operated. By comprising the argument assignment **freq = FALSE** in a hist function call, the result is a density histogram (if **freq = TRUE** which is default, then result is a frequency histogram)

(Hay-Jahans, 2019, p. 134)

Skewness

The *sampleskewness* formula used in package **moments**, for a sample of size n, is

$$a_3 = \left[\frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{n}\right] \Big/ \left[\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}\right]^{3/2}$$

where $\tilde{x}$ represents the sample mean

(Hay-Jahans, 2019, p. 83)

When skewness of the sample is positive, distribution of the underlying random variable longer right tail than left tail, a negative sample skewness shows the sample is skewed to the left. If the skewness of a sample is symmetric about the mean, then the skewness score becomes zero which means none of the tails is longer than the other one.

Kurtosis

The sample kurtosis formula used in package **moments**, for a sample of size n, is

$$a_4 = \left[\frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{n}\right] \Big/ \left[\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}\right]^2$$

Kurtosis depends on how crested the corresponding probability distribution is close to mean; and how strong the tails are

For a symmetric sample, if the kurtosis is greater than 3 then the underlying random variable may have a leptokurtic distribution, having a higher peak and stronger tails; and if the sample kurtosis is less than 3 the underlying random variable may have a platykurtic distribution, having a lower peak and weaker tails.

(Hay-Jahans, 2019, p. 84)

3.3 Scatterplots

Scatterplots of bivariate numeric data having a *output* variable, say Y , and an *predictor* variable, say X, involve plotting ordered pairs $(xi, yi), i = 1, 2, ..., n$, as points on the 2-dimensional coordinate axes. The general use of such plots is in either finding the attitude in which the observed responses vary in relation to variations in the independent variable, or providing a graphical showing of a known relationship between two variables.

(Hay-Jahans, 2019, p. 159)

3.4 Correlation (Pearson)

The term correlation with regard to two continuous random variables X and Y means to the presence of a linear relationship between the two variables.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

(Hay-Jahans, 2019, p. 321)

Pearson correlation coefficient lies between -1.0 and +1.0, Result of 1.0 shows perfect positive correlation between two continuous random variables, whereas -1.0 explains the perfect negative correlation. Result of 0 indicates that no correlation exists between two numerical variables.

3.5 Boxplot(5 Number Summary)

Excessive miscellaneous summary function calculates and outputs prevalent statistics that contains the five-number summary for numeric variables. This summary can be used as an anterior indicator of deviation from symmetry or the probable existence of outliers in the data. If the median is majorly lower than mean, outliers can be present in the right tail, or the sample can be skewed to right; and if the median is larger than a mean, outliers may be showed in the left tail, or the sample can simply be skewed to the left.

Elementary five number summary contains the minimum value, first quartile, median, third quartile, and maximum value of a numeric univariate dataset. This conclusion is generated by the function summary when applied to numeric vectors.

Observations in a sample are described as $outliers$ if they lie $1.5 \times IQR$ or more units below the first quartile, or above the third quartile. These bypass values are called the $lower$ and $upperfences$. Observations are described as $extremeoutliers$ if they stretch out beyond $3 \times IQR$ units below the first quartile, or above the third quartile, and outliers that are not $flagged$ as being extreme are referred to as $moderateoutliers$.

Boxplots provide a graphical representation of the five-number summary.

(Hay-Jahans, 2019, p. 85)

The statistical software R (R Development Core Team, 2020), version 4.0.3 was used for analysis.

4 Statistical analysis

In this section, the statistical methods explained above are applied to the presented data set and the results are interpreted.

4.1 Frequency Distributions of the variables

Initially, structure of the dataset is analysed and 456 observations of 10 variables are obtained in the data set. There is no N/A values for the year of 2020. Measures of central tendency and spread methods are used for all observations in year of 2020. Total Fertility rate mean is 2.46 with the 1.15 standard deviation. Live expectancy at birth for males and females are calculated individually and together. The results show that the average expected life for both genders is 74.0 with the standard deviation of 7.03. While maximum life expectancy for both genders is 89.27, minimum expected life in birth is 52.84. When the genders compared, result shows that the average females expectancy life at birth is 5.13 years higher than males. When all of the observations considered, maximum live expectancy for males lies between 51.35 and 85.40, whereas females are expected live longer with the numbers of 54.41 and 93.30.

The table 1 shows the central tendency and spread

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Year	228	2,020.000	0.000	2,020	2,020	2,020	2,020
TFR	228	2.461	1.154	1.060	1.708	2.872	7.000
LEB	228	74.037	7.031	52.840	69.592	79.218	89.270
LEM	228	71.529	6.841	51.350	67.255	76.650	85.400
LEF	228	76.666	7.344	54.410	71.915	82.205	93.300

Table 1: Measures of Central Tendency and Spread

Density histograms for each numeric variable are used to visualize Frequency Distributions. Skewness is calculated to understand to degree of asymmetry observed in a probability distribution. LEM, LEF and LEB variables are skewed left, so the mean is the lower than the median. Skewness results for life expectancy at birth are between -0.68 and -0.799 which show that data is moderately symmetrical. Additionally, skewness of difference between LEM and LEF is 0.57 which is fairly symmetric. However, total fertility rate is highly skewed to right with 1.48 which shows the data for TFR is asymmetric.

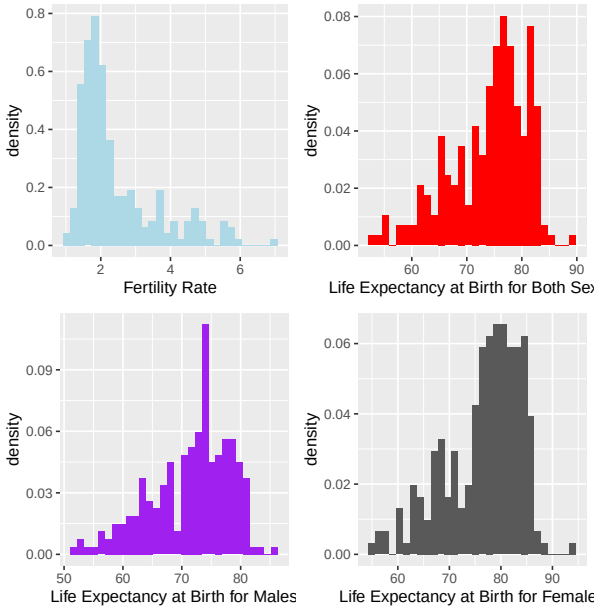

Figure 1: Density Histogram.

Figure 1 shows a density histograms of variables.

4.2 Bivariate Correlations between the variables

First of all, scatter plot matrix is used to visualize bivariate relationships between combinations of variables. Also, Pearson Correlation Table is used to test whether there is a statistically significant relationship between variables. Interestingly, there is strong negative correlation between life expectancy and total fertility rates with the score of -0.80. People who expected to live longer tends to have less fertility rate. Bivariate correlations among LEM, LEF and LEB are densely strong as expected and the scatter

	LEB	TFR	LEM	LEF
LEB	1.0000	-0.8021	0.9930	0.9934
TFR	-0.8021	1.0000	-0.7746	-0.8185
LEM	0.9930	-0.7746	1.0000	0.9729
LEF	0.9934	-0.8185	0.9729	1.0000

Table 2: Pearson Correlation

plots look perfectly correlated with the score of 0.99. Therefore, it is more logical to compare correlations between TFR and LEB.

Table 2 shows Pearson correlation between variables

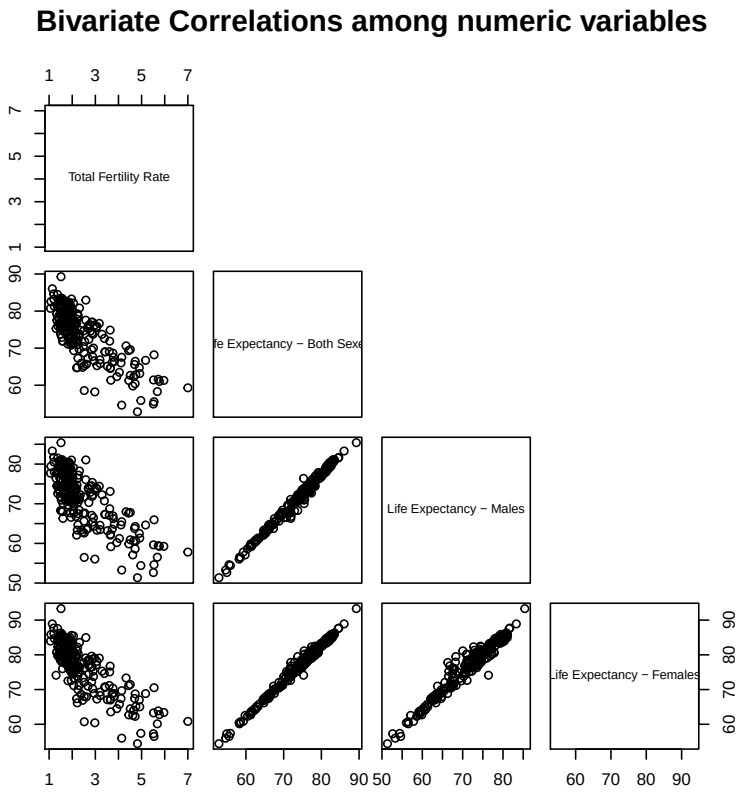

Figure 2: Bivariate Correlation

Figure 2 shows Bivariate correlation of variables.

4.3 Homogeneity and Heterogeneity on the basis of subregions

4.3.1 Life Expectancy for Regions and Subregions

When regions are considered, the highest live expectancy is in Europe with the mean of 79.59. Then Americas, Oceania and Asia follows the Europe with the means of 76.9, 75.05 and 74.72. Africa has the lowest life expectancy at birth with the mean of 65.59. Furthermore, the standard deviation for the life expectancy in Africa is the highest one among regions, shows the greatest difference between subregions.

When life expectancy of Europe's subregions are examined, the most homogeneous part is Western Europe with the standard deviation of 2.48 and the highest mean life with 82.79. Northern Europe part follows in the second place with 2.59 standard deviation and 80.60. The homogeneity and life expectancies decreases for the other subregions of the Europe.

Central America subregion is the homogeneous part for the America region with the standard deviation of 2.41 with the mean life expectancy of 75.80. Northern America subregion has the highest life expectancy with the mean of 79.94.

Australia/New Zealand subregion of the Oceania region is the most homogeneous subregion of the data set with standard deviation of 0.39 and the mean life expectancy of 82.41.

Asia region has the second and third most heterogeneous subregions of the dataset which are South-Eastern Asia with the standard deviation of 6.01 and South-Central Asia with the standard deviation of 5.90. The other subregions of Asia are also can considered as heterogeneous.

Northern Africa is the most heterogeneous subregion of the dataset with the standard deviation of 7.68. Also the other subregions of the Africa can easily considered as heterogeneous and have the lowest mean life expectancy among regions. Especially, Middle Africa and Southern Africa regions are in critical point with the mean life expectancies of 62.2 and 62.3

Figure 3 visualize the life expectancies of subregions with boxplot

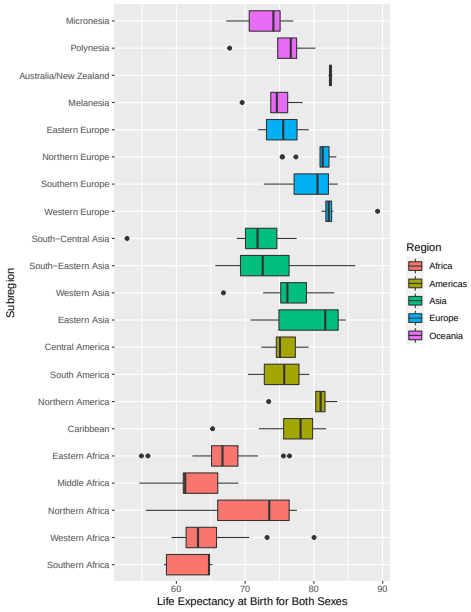

Figure 3: Total Life Expectancy

4.3.2 Total Fertility Rates for Regions and Subregions

Africa region holds the highest mean total fertility rate with 3.94 and the highest standard deviation of fertility rates with 1.23. Europe region, which has the highest rate of life expectancy rate at birth, has the lowest fertility rate with the mean of 1.62 and also is the most homogeneous region with the 0.21 standard deviation. Eastern Europe is the most homogeneous subregion of the dataset with the standard deviation of 0.07. The other parts of Europe can also be counted as homogeneous. Also Southern Europe has the lowest fertility rate of the Europe with the mean of 1.54. The second most homogeneous subregion of the data set for fertility rate is Australia/New Zealand with the standard deviation of 0.096.

Eastern Asia is the only homogeneous subregion of the Asia region with standard deviation of 0.35 and also it holds the lowest fertility rate of the dataset with the mean of 1.42. The other subregions of the Asia can be categorized as heterogeneous.

Americas is the second most homogeneous region for the fertility rate of the dataset with standard deviation of 0.21. All of the subregions of Americas are homogeneous and the most homogeneous subregion is South America with the standard deviation of 0.22

Region	mean fertility	mean LE	median Fertility	median LE	sd fertility	sd LE
Africa	3.944430	65.59482	3.87500	64.965	1.2342933	5.988592
Americas	1.928998	76.90320	1.90340	77.515	0.3070334	3.500851
Asia	2.191604	74.72212	1.97285	75.235	0.7962473	5.730182
Europe	1.627174	79.59776	1.57070	81.080	0.2108705	3.634900
Oceania	2.379833	75.05286	2.35000	74.790	0.4523014	4.337030

Table 3: Region-wise Comparison

Africa is the most heterogeneous region of the dataset and also it includes the most heterogeneous subregion which is Western Africa. All subregions of Africa can be called as heterogeneous except Southern Africa.

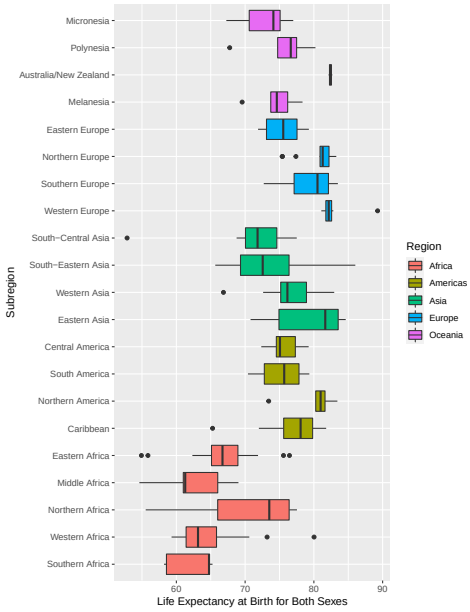

Figure 4: Total Fertility Rate

Figure 4 visualize the total fertility rate of subregions with boxplot

Table 3 shows Region-wise comparison of the variables

4.4 Variables comparison between 2000 and 2020

In this last subsection of the statistical analysis, variables of years 2000 and 2020 are compared and Scatter plots are used for visualization of the comparison of values. First of all, mean differences of life expectancy at birth between these years is 5.93 which

skim variable	Region	.mean	.sd	.p0	.p25	.p50	.p75	.p100
TFR	Africa	5.07	1.50	1.53	4.20	5.30	6.04	8.09
TFR	Americas	2.48	0.77	1.34	1.92	2.36	2.80	4.53
TFR	Asia	3.05	1.54	0.89	2.01	2.63	3.79	8.00
TFR	Europe	1.57	0.39	1.10	1.32	1.46	1.74	2.96
TFR	Oceania	3.35	1.02	1.72	2.57	3.54	4.33	4.70

Table 4: Total Fertility Rate of 2000

skim variable	Region	mean	sd	p.0	p.25	p.50	p.75	p.100
LEB	Africa	55.37	8.21	39.84	50.18	53.60	59.06	76.70
LEB	Americas	72.68	4.41	56.45	70.17	73.51	75.93	79.85
LEB	Asia	69.17	7.20	45.49	64.34	70.59	73.81	82.97
LEB	Europe	75.57	4.37	66.02	72.55	76.71	78.55	87.65
LEB	Oceania	70.39	5.68	60.37	67.85	71.27	73.91	79.36

Table 5: Life Expectancy of 2000

means the expected life is increased by approximately 6 years after 20 years. Whereas, the mean of total fertility rate is decreased by -0.67 in 20 years.

There are 7 N/A values in numeric variables in 2000 which is omitted from the dataset.

When numeric variables are grouped by regions, it shows that the highest life expectancy at birth is Europe region with the mean of 75.6 and also the most homogeneous region at this variable. In 2020, the result is quite similar for Europe with the highest mean of the regions and the second most homogeneous region of the dataset after Americas. Live expectancy in Africa has increased considerably 55.4 to 65.59 and also heterogeneity of the region has decreased considerably in 20 years.

While life expectancy and homogeneity has increased in every region in between 2000 and 2020, Total fertility rate tends to decrease for each region except Europe. Especially for Africa in 2000, total fertility rate is significantly high with the mean of 5.07. In 20 years, it is changed drastically to the mean of 3.94. Also, standard deviations are decreased for every region which shows that homogeneity is increased for also total fertility rate in 2020

The table 4 is Total Fertility Rate of 2000
The table 5 is Life Expectancy of 2000

Figure 5 visualize the differences between the year of 2000 and 2020

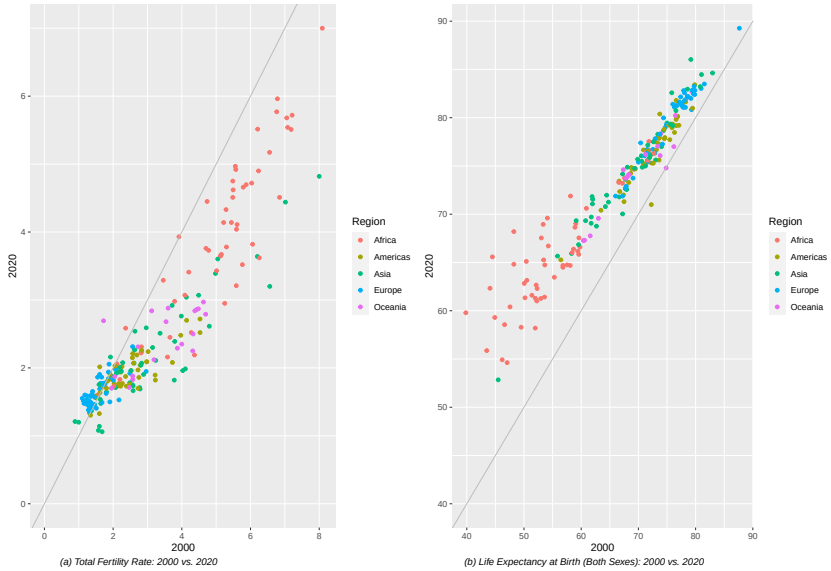

Figure 5: Total Fertility Rate

5 Summary

The data set analysed in this report was compiled by the instructors of the course Case Studies I at TU Dortmund University in the summer term 2020/21 and originates from International Programs, Population Division, U.S. Census Bureau (2020). It includes life expectancy and fertility rates for 228 countries from 2000 and 2020.The countries are divided geographically into 5 regions and 21 subregions.

Firstly, frequency distributions of the variables were described for both genders individually and together. Result showed that averaged life expectancy of females 5 years more than males. Also, density histograms indicated that distribution of population for averaged life expectancy is fairly symmetric whereas the total fertility rate is asymmetric. Second of all, bivariate correlations between the variables were examined. Interestingly, strong negative correlation showed up between total fertility rate and averaged life expectancy variables. It can be said that, people who live in a region where life expectancy rate is high such as Europe, tend not to give birth for some reasons. Also, almost perfect correlation emerged between life expectancies between genders as expected. Thirdly, homogeneity and heterogeneity on the basis of subregions were examined. The most homogeneous subregion was Australia/New Zealand for the life expectancy and the the most heterogeneous region was Africa. Interestingly, for fertility rate variable, Asia was found the second heterogeneous region, however Eastern Asia was counted as a homoge-

neous subregion. China's one child policy may be the one of the reasons of that result. Lastly, variables comparison made between years of 2000 and 2020. It is found that life expectancy for all regions has been increased in 20 years, whereas fertility rates has been decreased drastically.

To sum up, this descriptive analysis and results show that life expectancy is expected to increase and fertility rates seems to decrease in future. However, the most crucial parts of this analysis are the heterogeneity and significant differences of these variables between regions and subregions. When the results are compared, people who live in Africa are expected to live approximately 15 years less than who live in Europe. Hopefully, descriptive analyses of demographic data will show different results for humanity in future

Bibliography

Baptiste Auguie. *gridExtra: Miscellaneous Functions for "Grid" Graphics*, 2017. URL https://CRAN.R-project.org/package=gridExtra. R package version 2.3.

Christopher Hay-Jahans. *R Companion to Elementary Applied Statistics*. CRC Press, 2019.

International Programs, Population Division, U.S. Census Bureau. International data base: Population estimates and projections methodology. 2020. URL https://www2.census.gov/programs-surveys/international-programs/technical-documenta

Lukasz Komsta and Frederick Novomestky. *moments: Moments, cumulants, skewness, kurtosis and related tests*, 2015. URL https://CRAN.R-project.org/package=moments. R package version 0.14.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020.

Kun Ren and Kenton Russell. *formattable: Create 'Formattable' Data Structures*, 2021. URL https://CRAN.R-project.org/package=formattable. R package version 0.2.1.

Elin Waring, Michael Quinn, Amelia McNamara, Eduardo Arino de la Rubia, Hao Zhu, and Shannon Ellis. *skimr: Compact and Flexible Summaries of Data*, 2021. URL https://CRAN.R-project.org/package=skimr. R package version 2.1.3.

Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN 978-3-319-24277-4. URL https://ggplot2.tidyverse.org.

Hadley Wickham, Romain François, Lionel Henry, and Kirill Müller. *dplyr: A Grammar of Data Manipulation*, 2021. URL https://CRAN.R-project.org/package=dplyr. R package version 1.0.5.

Appendix

A Additional figures

Figure 6: Difference of life expectancy between genders.

Figure 6 life expectancy mean difference between genders which is explained in section 4.1

B Additional tables

Table 6 shows Subregion-wise comparison of the variables which is explained in section 4.3

Subregion	mean fertility	mean LE	median Fertility	median LE	sd fertility	sd LE
Australia/New Zealand	1.804850	82.41000	1.80485	82.410	0.0968029	0.3959798
Caribbean	1.822676	77.45640	1.80560	78.070	0.2834580	3.5197548
Central America	2.214175	75.80875	2.13975	75.070	0.3369493	2.4167596
Eastern Africa	3.838088	66.68941	3.78000	66.710	1.1818877	5.5966949
Eastern Asia	1.423563	79.25125	1.28910	81.655	0.3592071	5.6881919
Eastern Europe	1.485440	75.35800	1.48145	75.545	0.0764605	2.6924743
Melanesia	2.558000	74.49000	2.79000	74.610	0.4539493	3.2629741
Micronesia	2.473429	72.82286	2.68000	74.110	0.4197010	3.9539504
Middle Africa	4.666867	62.24667	4.45000	61.290	0.9254633	4.3966095
Northern Africa	3.416012	70.53125	3.23000	73.515	1.2011863	7.6869842
Northern America	1.767060	79.94400	1.84360	80.980	0.1827736	3.8243732
Northern Europe	1.798743	80.60143	1.80900	81.285	0.2068461	2.5899149
Polynesia	2.323257	75.58286	2.35000	76.640	0.4649966	3.9377181
South America	2.027858	75.21333	2.06000	75.685	0.2297943	3.1055942
South-Central Asia	2.375929	71.14357	2.07500	71.810	0.8511281	5.9045531
South-Eastern Asia	2.253391	73.13818	2.07000	72.560	0.8777031	6.0115403
Southern Africa	2.648000	62.31800	2.52000	64.750	0.3630014	3.6024464
Southern Europe	1.545294	79.57188	1.49555	80.525	0.1858119	3.1247639
Western Africa	4.298277	64.91353	4.51000	63.150	1.2822667	5.4391267
Western Asia	2.343400	76.36895	1.98570	76.150	0.6926465	3.4234987
Western Europe	1.663333	82.79333	1.62400	82.160	0.1856223	2.4879459

Table 6: Subregion-wise Comparison