



Final Project Report: Life Expectancy

ML: Gonzalo Ducca <gonzaloducca@gmail.com>

DA: Juan P. Bertone <bertonejpb@gmail.com>

DE: Juan E. Flórez-Coronel <juan.florez@upr.edu>

DA: Valentino Caputa <caputavalentino@gmail.com>

ML: Carlos Madoery <ccmadoery@gmail.com>

Product Owner:

Melina Griffo <meligriff@gmail.com>

Henry Mentor:

Pía Ruiz <mpiaruiz@gmail.com>

HENRY Data Science Part Time 3

Revision 1.0

November 1, 2023

Table 1. Document Revision History

Revision	Date	Reason for Revision Change(s)
0.0	10-20-23	Document Creation
1.0	11-01-23	Documentation for Sprint 1

Table of Contents

1	Problem Statement	6
1.1	Objectives	6
1.2	Possible Clientele	6
1.3	Scope of the Project	6
2	Data Acquisition	7
2.1	World Bank API	7
2.2	Possible Topics	7
3	Project Development	9
3.1	Team Roles	9
3.2	Gaant Chart	9
	Conclusions and Lessons Learned	10

List of Tables

1	Document Revision History	2
2	Team Roles	9

List of Figures

1	World Bank Data Topics	7
2	World Bank Data Economies	8
3	World Bank Data Science & Technology Factors	8
4	Project Gaant Chart	9

List of Acronyms

GDP Gross Domestic Product. 6

1 Problem Statement

We consulted World Bank databases and decided to use as reference key topics that provide indicators influencing the life expectancy of a country's inhabitants. These selected topics are education, health, economy, and the development of science and technology. We plan to gather indicators from each topic, create the databases we want to work with, and then link these variables to establish relationships. We envision this project as a product that can be offered to a company looking to invest in science and development. After all the data analysis work is completed, we will create Machine Learning models by specifying certain parameters, such as the percentage of Gross Domestic Product (GDP) invested in education or science. These models will then provide us with an estimated life expectancy for a particular country.

1.1 Objectives

To analyze life expectancy in the countries of the Americas based on five factors:

- Education
- Health
- Economy
- Science and Technology
- Security

1.2 Possible Clientele

The potential clientele for this project includes both public and private entities with a vested interest in the subject matter. Public institutions, such as government agencies, can utilize the insights and models generated to inform policy decisions and allocate resources effectively in areas related to education, healthcare, and scientific research. On the other hand, private companies, especially those involved in industries related to science and technology, can benefit from the project's findings when making investment decisions and corporate social responsibility initiatives. This project caters to a broad range of organizations looking to make informed decisions that impact the well-being and development of countries.

1.3 Scope of the Project

The project scope involves a comprehensive analysis of life expectancy in 30 countries across the World over the past 35 years. This analysis will encompass five key topics within each of the five influencing factors, namely education, health, economy, and science and technology. The goal is to provide a detailed and data-driven insight into the factors that impact life expectancy in these countries over the specified timeframe.

2 Data Acquisition

The data was acquired using the World Bank API.

2.1 World Bank API

The World Bank Data has the following topics:

id	value
1	Agriculture & Rural Development
2	Aid Effectiveness
3	Economy & Growth
4	Education
5	Energy & Mining
6	Environment
7	Financial Sector
8	Health
9	Infrastructure
10	Social Protection & Labor
11	Poverty
12	Private Sector
13	Public Sector
14	Science & Technology
15	Social Development
16	Urban Development
17	Gender
18	Millenium development goals
19	Climate Change
20	External Debt
21	Trade

Figure 1. World Bank Data Topics

And the following economies:

Each topic has a few factors:

	id	value	aggregate	longitude	latitude	region	adminregion	lendingType	incomeLevel	capitalCity
0	ABW	Aruba	False	-70.0167	12.51670	LCN		LNK	HIC	Oranjestad
1	AFE	Africa Eastern and Southern	True	NaN	NaN					
2	AFG	Afghanistan	False	69.1761	34.52280	SAS	SAS	IDX	LIC	Kabul
3	AFW	Africa Western and Central	True	NaN	NaN					
4	AGO	Angola	False	13.2420	-8.81155	SSF	SSA	IBD	LMC	Luanda
...
261	XKX	Kosovo	False	20.9260	42.56500	ECS	ECA	IDX	UMC	Pristina
262	YEM	Yemen, Rep.	False	44.2075	15.35200	MEA	MNA	IDX	LIC	Sana'a
263	ZAF	South Africa	False	28.1871	-25.74600	SSF	SSA	IBD	UMC	Pretoria
264	ZMB	Zambia	False	28.2937	-15.39820	SSF	SSA	IDX	LMC	Lusaka
265	ZWE	Zimbabwe	False	31.0672	-17.83120	SSF	SSA	IDB	LMC	Harare

Figure 2. World Bank Data Economies

	id	value
0	BM.GSR.ROYL.CD	Charges for the use of intellectual property, ...
1	BX.GSR.ROYL.CD	Charges for the use of intellectual property, ...
2	TX.VAL.TECH.MF.ZS	High-technology exports (% of manufactured exp...
3	TX.VAL.TECH.CD	High-technology exports (current US\$)
4	IP.PAT.NRES	Patent applications, nonresidents
5	IP.PAT.RESD	Patent applications, residents
6	GB.XPD.RSDV.GD.ZS	Research and development expenditure (% of GDP)
7	SP.POP.SCIE.RD.P6	Researchers in R&D (per million people)
8	IP.JRN.ARTC.SC	Scientific and technical journal articles
9	SP.POP.TECH.RD.P6	Technicians in R&D (per million people)

Figure 3. World Bank Data Science & Technology Factors

2.2 Possible Topics

Counting missing values:

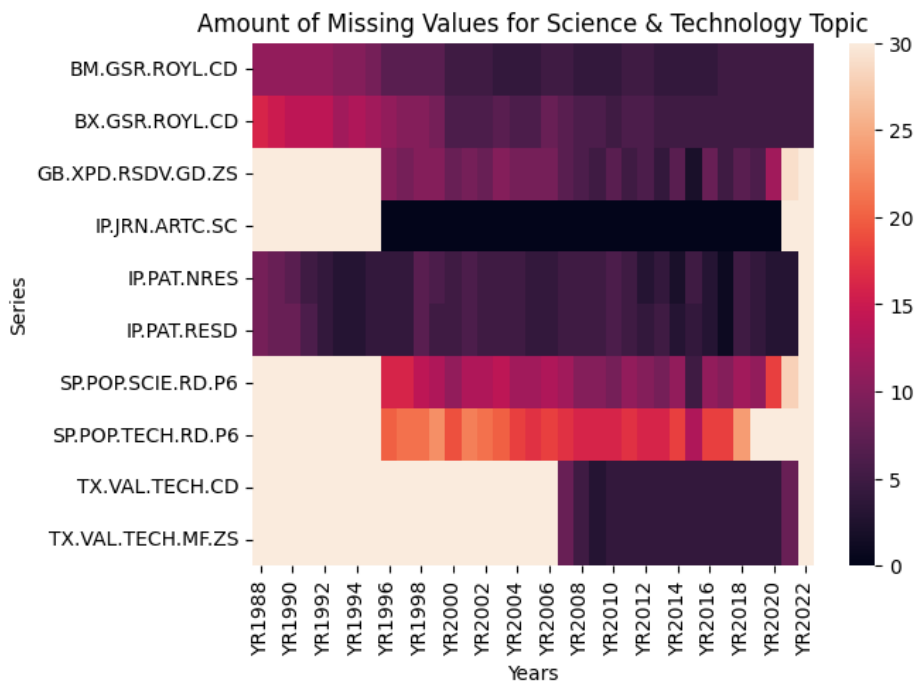


Figure 4. Nan Analysis

3 Project Development

3.1 Team Roles

Table 2. Team Roles

Role	Name
Machine Learning	Gonzalo Ducca
Machine Learning	Carlos Madoery
Data Analytics	Valentino Caputa
Data Analytics	Juan P. Bertone
Data Engineer	Juan E. Flórez-Coronel

3.2 Gaant Chart

Esperanza de vida

Sprint 1

Nombre	Responsable	Fecha	Estado	Cronograma - Start	Cronograma - End
KPI	caputavalentino@gmail.com	2023-11-01	Listo	2023-10-23	2023-11-01
Stack tecnológico	Juan Pablo Bertone	2023-10-27	Listo	2023-10-23	2023-11-01
Documentación	Gonzalo Ducca, Carlos Madoery	2023-10-28	En curso	2023-10-23	2023-11-01
EDA preliminar	Juan E Florez-Coronel	2023-10-31	Listo	2023-10-23	2023-11-01
Github	Juan E Florez-Coronel	2023-10-27	Listo	2023-10-23	2023-11-01
Desde el 2023-10-27 hasta el 2023-11-01				2023-10-23	2023-11-01

Sprint 2

Nombre	Responsable	Fecha	Estado	Cronograma - Start	Cronograma - End
DW automatizado	Juan E Florez-Coronel	2023-11-07	No iniciado	2023-11-01	2023-11-08
Tablas	Juan E Florez-Coronel, Gonzalo Ducca		No iniciado	2023-11-01	2023-11-08
Carga incremental	Carlos Madoery	2023-11-10	No iniciado	2023-11-08	2023-11-15
Big Data y servicios cloud.	Juan Pablo Bertone	2023-11-13	No iniciado	2023-11-08	2023-11-15
Desde el 2023-11-07 hasta el 2023-11-13				2023-11-01	2023-11-15

Sprint 3

Nombre	Responsable	Fecha	Estado	Cronograma - Start	Cronograma - End
Storytelling	caputavalentino@gmail.com, Juan Pablo Bertone	2023-11-28	No iniciado	2023-11-15	2023-11-29
Dashboard y reportes	valentino@gmail.com, Juan E Florez-Coronel, Juan Pablo Bertone	2023-11-25	No iniciado	2023-11-15	2023-11-29
Recomendaciones	Carlos Madoery	2023-11-29	No iniciado	2023-11-15	2023-11-29
KPIs y datos hallados	Carlos Madoery	2023-11-27	No iniciado	2023-11-15	2023-11-29
Retomar hitos	Gonzalo Ducca	2023-11-25	No iniciado	2023-11-15	2023-11-29
Modelos ML	Gonzalo Ducca, Carlos Madoery	2023-11-21	No iniciado	2023-11-15	2023-11-29
Reporte Vis. geografica	Juan Pablo Bertone	2023-11-23	No iniciado	2023-11-15	2023-11-29
Desde el 2023-11-21 hasta el 2023-11-29				2023-11-15	2023-11-29

Figure 5. Project Gaant Chart

Conclusions and Lessons Learned

bla