Certificado en Big Data

Materia de profundización - Obligatorio

Guillermo Reboledo - 204671

Docente: Eduardo García Parra

Índice

1. Letra Obligatorio

Para el obligatorio deberán utilizar las herramientas utilizadas en el curso. Deberá seleccionar un conjunto de datos tabulares con más de 4 tablas que el docente propondrá y deberá seleccionar 8 preguntas relativas a los datos para contestarlas.

Tomar los datos que fueron seleccionados junto al docente. Luego se deberá realizar un análisis exploratorio de los datos vía pandas, identificando el tipo de datos que hay en cada columna y que significado tienen dentro del dominio de los datos.

- Dentro de un Jupyter notebook se mostrará, una vista previa de las primeras filas, cantidad de columnas de cada tabla, nombre de cada columna, descripción de los datos de cada tabla, cómo está compuesto el esquema de los datos, revisar valores nulos o faltantes y limpiarlos si es necesario. Revisar registros duplicados. Claves primarias únicas.

- Los archivos resultantes se deberán almacenar en otra carpeta.

- A partir de estos nuevos archivos, se deben crear visualizaciones dentro de otro notebook con las herramientas dadas en clase u otras de elección del equipo, que ayuden a responder las preguntas seleccionadas.

El mismo análisis realizado en la parte 1 realizarlo vía spark, ya sea dentro de la máquina virtual si se tienen créditos si no dentro de Google Collab.

Se pide desarrollar un dashboard que responda algunas de las preguntas planteadas, implementado en Tableau Public o superset.

Una vez que termine con la exploración y limpieza de datos, deberá elegir una forma de modelarlos, esta puede ser, Normalizada, Diagrama Estrella, Data Vault, o OBT.

- Describir en Hive, como lo modelaría, que tablas crearía y de que tipo (externas, internas).

La entrega final consiste en un informe donde se detalle todo el proceso realizado y todo lo aprendido durante la realización del obligatorio. Se deben entregar los notebooks de análisis vía pandas, de preguntas y respuestas con las visualizaciones correspondientes, así como el notebook de análisis utilizando spark. Para el caso del dashboard se pide entregar el link a Tableau Public o Superset, o alguna captura que demuestre su funcionamiento.

2. Set de datos

El set de datos seleccionado para este análisis fue el de “Estudios de Países”. Este conjunto de datos está compuesto por una serie de archivos CSV (valores separados por coma), cada uno de los cuales contiene información relevante para analizar diferentes aspectos socioeconómicos y demográficos de diversos países en 2020 principalmente. Cada archivo contiene índices o métricas que reflejan características específicas de los países, tales como el costo de vida, estructura de edad de la población, índices de criminalidad, calidad de los servicios de salud, precios de propiedades, densidad poblacional y calidad de vida.

A continuación, se enumeran los distintos archivos del data set “Estudios de Países”:

1. Cost_of_living_index_by_country_2020.csv
2. Countries_age_structure.csv
3. Crime_index_by_countries_2020.csv
4. Health_care_index_by_countries_2020.csv
5. Properties_price_index_by_countries_2020.csv
6. Pupulation_density_by_countries.csv
7. Quality_of_life_index_by_countries_2020.csv

3. Resolución

Parte 1- Análisis exploratorio y preguntas a constestar

Parte 1a-Análisis exploratorio de los datos vía Pandas

El análisis se realizó por jupyter notebook versión 7.0.8 y se encuentra en el notebook llamado *Parte_1a_analisisExploratorioDatosPaises.ipynb* donde además podrá encontrarse todo el código y comentarios realizados para esta primera parte. Los archivos utilizados como insumo para los análisis se encuentran en la carpeta RawData, mientras que aquellos generados luego del análisis y refinado están ubicados en la carpeta RefinedData, con el prefijo “RefData_” en el nombre de cada archivo para identificarles mejor.

En las secciones siguientes se realizará un análisis exploratorio de los distintos archivos en formato CSV. Para ello, se utilizará Jupyter Notebook con Python y diversas bibliotecas como pandas, geopandas, numpy y matplotlib para análisis, y folium, seaborn o bokeh para visualizaciones. Así mismo, se identificarán campos a ser usados como primary key, tipo de datos en cada archivo, significado de los mismos y presencia o no de nulos en los distintos archivos.

Generalidades

A continuación, en la tabla 1, se muestra la dimensión de los archivos antes y después de las transformaciones realizadas luego del análisis exploratorio. Los motivos para dichas transformaciones, así como información específica obtenida a partir del análisis, se discutirá en la sección correspondiente a cada archivo.

Tabla 1. Tabla conteniendo el nombre de los archivos del set de datos y la cantidad de filas y columnas antes y post refinado.

Archivos	Raw		Refinado	
	Filas	Columnas	Filas	Columnas
Cost_of_living_index_by_country_2020.csv	132	7	131	7
Countries_age_structure.csv	191	4	186	4
Crime_index_by_countries_2020.csv	129	3	128	3
Health_care_index_by_countries_2020.csv	93	3	93	3
Properties_price_index_by_countries_2020.csv	104	8	103	8
Population_density_by_countries.csv	251	9	201	12
Quality_of_life_index_by_countries_2020.csv	80	10	80	10

Se comprobó por distintas vías que el campo Country esté presente en todos los archivos (*df.head(n)*, donde n es el número de entradas comenzando desde arriba del df a visualizar) y que además presentara valores únicos, no tiene duplicados (función utilizada *df ['Country'].duplicated().any()*). Esto junto con su relevancia para darle sentido a los datos, hacen que dicho campo sea la primary key a utilizar en las consultas que se vayan a realizar para responder las preguntas del proyecto.

En algunas tablas se vio que el campo Country no solo incluía a países, sino que también podrían incluir el territorio dependiente o una especificación de si se encuentran en disputa. Esto es un inconveniente al momento de analizar los datos ya que no es lo mismo un país que un territorio dependiente o en disputa, lo que podría alterar los estudios realizados y por ende los enunciados que podrían obtenerse de estos. Dado que estos datos se encuentran en varios de los df donde no siempre se especifica que pueden estar presente, se decidió eliminar las entradas que figuren como territorio dependiente o en disputa. Para ello se observó que todas aquellas entradas identificadas como tales traían consigo entre paréntesis el país del cual dependen o la aclaración de si se encuentran en disputa o no. Esto permitió idear una estrategia para eliminar las entradas la cual consistió en filtrar aquellas que posean un paréntesis en dato para campo Country. A su vez, para normalizar los datos y que puedan ser usados en las consultas donde se utilizan joins, luego de eliminar las entradas anteriores se pasó un script que eliminara los espacios al inicio y al final del string presente en el campo Country como dato. Estos cambios y transformaciones se realizaron en todas los df logrados a partir de los archivos para este set de datos.

Además de las peculiaridades mencionadas antes para tratar el campo Country, se aplicó una serie de funciones para completar el análisis exploratorio de los datos. Las funciones usadas fueron; *df.shape()* para cantidad de columnas y filas, *df.info()* para información sobre tipo de datos y cantidad de entrada no nulas por campo, *df.isnull().sum()* para hallar la cantidad de nulos que tiene cada campo, y *df.describe()* para descripción del df en cuanto métricas varias.

En las secciones siguientes se proseguirá al análisis exploratorio y la transformación de los datos en los distintos archivos que posee este data set y que ya fueron mencionados en la sección, *2. Set de datos*.

Cost_of_living_index_by_country_2020.csv

Análisis de los datos

Este archivo proporciona valores comparativos entre países en términos de costo de vida. Cada fila representa un país, y las columnas contienen índices numéricos que permiten evaluar aspectos específicos de los costos, desde alimentos hasta el alquiler, así como poder comparar el poder adquisitivo en diferentes países.

Al realizar el análisis y pasar los datos a un dataframe (df_Cost_of_living_index_by_country_2020) se encuentra que el archivo está compuesto por 132 entradas y siete columnas. Los tipos de datos presentes en este df son object para Country y float64 para los restantes. Ninguna de las columnas presentó datos nulos. A partir de la descripción para este df se obtienen los datos de la figura 1.

	Cost of Living Index	Rent Index	Cost of Living Plus Rent Index	Groceries Index	Restaurant Price Index	Local Purchasing Power Index
count	132.000000	132.000000	132.000000	132.000000	132.000000	132.000000
mean	49.214697	18.089470	34.327273	42.583030	42.271439	50.324697
std	18.404922	12.808608	14.989052	17.302168	22.423585	27.357111
min	21.980000	4.030000	13.660000	17.700000	14.950000	2.180000
25%	35.295000	9.720000	23.117500	29.705000	25.117500	30.040000
50%	44.170000	13.610000	30.205000	37.405000	34.890000	42.555000
75%	60.717500	23.637500	43.590000	51.620000	53.140000	67.557500
max	122.400000	79.570000	87.890000	120.270000	123.010000	119.530000

Figura 1. Información obtenida de aplicar la función *describe()* al dataframe df_Cost_of_living_index_by_country_2020. **count**: Número total de valores no nulos en la columna. **mean**: Promedio (media aritmética) de los valores. **std** (desviación estándar): Mide cuánto se desvían los valores respecto a la media. **min**: Valor mínimo de la columna. **25%** (primer cuartil): Valor que separa el 25% más bajo de los datos. Es el percentil 25. **50%** (mediana o segundo cuartil): Valor que divide los datos en dos mitades iguales (percentil 50). **75%** (tercer cuartil): Valor que deja el 75% de los datos por debajo. Es el percentil 75. **max:** Valor máximo de la columna.

En la tabla 2 se resumen algunos de los datos extraídos luego del análisis de este archivo.

Tabla 2. Nombre de los campos, tipo de dato y significado dentro del dominio de datos para el archivo Cost_of_living_index_by_country_2020.csv

Columna	Tipo de dato	Significado dentro del dominio de los datos
Country	object	País para el cual se están registrando los datos de costo de vida.
Cost of Living Index	float64	Índice que mide el costo general de bienes y servicios, sin incluir alquiler, en comparación con un valor base (como Nueva York=100).
Rent Index	float64	Índice que mide el costo de alquiler de propiedades residenciales en comparación con el valor base.
Cost of Living Plus Rent Index	float64	Combina el costo de vida con el costo del alquiler para una visión global.
Groceries Index	float64	Índice que mide el costo de comestibles básicos en comparación con el valor base.
Restaurant Price Index	float64	Índice que mide los precios en restaurantes en comparación con el valor base.
Local Purchasing Power Index	float64	Índice que evalúa el poder adquisitivo de los residentes locales, considerando los ingresos locales en comparación con el costo de vida.

Transformaciones

Para este df en específico solo se aplicaron las transformaciones que se especifican en la sección de Generalidades ya que no fue necesaria ninguna otra. Una vez realizadas las mismas, el df quedó con 131 entradas y siete columnas (tabla 1).

El archivo refinado obtenido a partir de estas transformaciones se encuentra en la carpeta RefinedData con el nombre *RefData_Cost_of_living_index_by_country_2020.csv*.

Countries_age_structure.csv

Análisis de los datos

El conjunto de datos presenta la estructura etaria de cada país, permitiendo ver la proporción de personas en diferentes grupos de edad. Cada fila representa un país, y las columnas numéricas contienen porcentajes de cada grupo de edad.

Al realizar el análisis y pasar los datos a un dataframe (df_Coutries_age_structure) se encuentra que el archivo está compuesto por 191 entradas y cuatro columnas. Los tipos de datos presentes en este df son object para todos los campos (tabla 3), lo cual es un inconveniente para su uso en análisis. Ninguna de las columnas presentó datos nulos.

Tabla 3. Nombre de los campos, tipo de dato y significado dentro del dominio de datos para el archivo Countries_age_structure.csv

Columna	Tipo de dato	Significado dentro del dominio de los datos
Country	object	País para el cual se registra la información de edad.
Age 0 to 14 Years	object	Porcentaje de la población entre 0 y 14 años.
Age 15 to 64 Years	object	Porcentaje de la población entre 15 y 64 años.
Age above 65 Years	object	Porcentaje de la población mayor de 65 años.

Transformaciones

Con la finalidad de poder trabajar con los datos y obtener mejores métricas, se realizaron dos transformaciones importantes.

1- El tipo de dato para los campos distintos a Country, se transformaron a float64, primero quitando el símbolo de porcentaje '%' y luego pasando atipo float usando astype(float) (tabla 4).

Tabla 4. Tipo de datos para los campos presentes en el archivo *RefData_Countries_age_structure.csv* luego de realizar el refinado de estos.

Columna	Tipo de dato
Country	object
Age 0 to 14 Years	float64
Age 15 to 64 Years	float64
Age above 65 Years	float64

2- Se eliminaron aquellas entradas que hacían referencia a países como territorio dependiente o en disputa. Esto hizo que el df pasara de tener 191 entradas a 186 entradas (tabla 1). A partir de la descripción para este df se obtienen los datos de la figura 2.

	Age 0 to 14 Years	Age 15 to 64 Years	Age above 65 Years
count	186.000000	186.000000	186.000000
mean	27.757097	63.598495	8.693548
std	10.588482	6.646668	6.151041
min	11.500000	47.200000	1.000000
25%	17.875000	59.225000	4.000000
50%	27.100000	64.850000	6.000000
75%	36.600000	67.575000	14.000000
max	50.200000	85.000000	27.000000

Figura 2. Información obtenida de aplicar la función *describe()* al dataframe *df_Coutries_age_structure*. **count**: Número total de valores no nulos en la columna. **mean**: Promedio (media aritmética) de los valores. **std** (desviación estándar): Mide cuánto se desvían los valores respecto a la media. **min**: Valor mínimo de la columna. **25%** (primer cuartil): Valor que separa el 25% más bajo de los datos. Es el percentil 25. **50%** (mediana o segundo cuartil): Valor que divide los datos en dos mitades iguales (percentil 50). **75%** (tercer cuartil): Valor que deja el 75% de los datos por debajo. Es el percentil 75. **max:** Valor máximo de la columna.

El archivo refinado obtenido a partir de estas transformaciones se encuentra en la carpeta RefinedData con el nombre *RefData_Coutries_age_structure.csv.*

Crime_index_by_countries_2020.csv

Análisis de los datos

Este archivo contiene información sobre los niveles de criminalidad y seguridad en diferentes países, lo cual ayuda a comparar cuán seguro es vivir en cada lugar.

Al realizar el análisis y pasar los datos a un dataframe (df_Crime_index_by_countries_2020) se encuentra que el archivo está compuesto por 129 entradas y tres columnas. Los tipos de datos presentes en este df son object para el campo Country y float64 para los campos restantes (tabla 5). Ninguna de las columnas presentó datos nulos. A partir de la descripción para este df se obtienen los datos de la figura 3.

Tabla 5. Nombre de los campos, tipo de dato y significado dentro del dominio de datos para el archivo Crime_index_by_countries_2020.csv

Columna	Tipo de dato	Significado dentro del dominio de los datos
Country	object	País para el cual se registran los índices de criminalidad y seguridad.
Crime Index	float64	Índice que representa el nivel de criminalidad en el país (un valor alto indica alta criminalidad).
Safety Index	float64	Índice de seguridad que representa lo seguro que es el país (un valor alto indica mayor seguridad).

	Crime Index	Safety Index
count	129.000000	129.000000
mean	44.222481	55.777519
std	15.690481	15.690481
min	11.860000	15.510000
25%	31.830000	45.210000
50%	43.710000	56.290000
75%	54.790000	68.170000
max	84.490000	88.140000

Figura 3. Información obtenida de aplicar la función *describe()* al dataframe *df_Crime_index_by_countries_2020*. **count**: Número total de valores no nulos en la columna. **mean**: Promedio (media aritmética) de los valores. **std** (desviación estándar): Mide cuánto se desvían los valores respecto a la media. **min**: Valor mínimo de la columna. **25%** (primer cuartil): Valor que separa el 25% más bajo de los datos. Es el percentil 25. **50%** (mediana o segundo cuartil): Valor que divide los datos en dos mitades iguales (percentil 50). **75%** (tercer cuartil): Valor que deja el 75% de los datos por debajo. Es el percentil 75. **max:** Valor máximo de la columna.

Transformaciones

Para este df en específico solo se aplicaron las transformaciones que se especifican en la sección de Generalidades ya que no fue necesaria ninguna otra. Una vez realizadas las mismas, el df quedó con 128 entradas y tres columnas (tabla 1).

El archivo refinado obtenido a partir de estas transformaciones se encuentra en la carpeta RefinedData con el nombre *RefData_Crime_index_by_countries_2020.csv*.

Health_care_index_by_countries_2020.csv

Análisis de los datos

El conjunto de datos en este archivo permite ver la calidad de los servicios de salud y el gasto en atención médica en cada país.

Al realizar el análisis y pasar los datos a un dataframe (df_Health_care_index_by_countries_2020) se encuentra que el archivo está compuesto por 93 entradas y tres columnas. Los tipos de datos presentes en este df son object para el campo Country y float64 para los campos restantes (tabla 6). Ninguna de las columnas presentó datos nulos. A partir de la descripción para este df se obtienen los datos de la figura 4.

Tabla 6. Nombre de los campos, tipo de dato y significado dentro del dominio de datos para el archivo Health_care_index_by_countries_2020.csv

Columna	Tipo de dato	Significado dentro del dominio de los datos
Country	object	País para el cual se registra la información sobre salud.
Health Care Index	float64	Índice que mide la calidad de la atención médica en el país.
Health Care Exp. Index	float64	Índice de gasto en atención médica, reflejando cuánto se invierte en salud en relación con el promedio global.

	Health Care Index	Health Care Exp. Index
count	93.000000	93.000000
mean	63.414194	114.197419
std	10.317796	20.350056
min	39.660000	69.140000
25%	55.730000	99.670000
50%	64.480000	116.140000
75%	71.580000	131.070000
max	86.710000	159.660000

Figura 4. Información obtenida de aplicar la función *describe()* al dataframe *df_Health_care_index_by_countries_2020*. **count**: Número total de valores no nulos en la columna. **mean**: Promedio (media aritmética) de los valores. **std** (desviación estándar): Mide cuánto se desvían los valores respecto a la media. **min**: Valor mínimo de la columna. **25%** (primer cuartil): Valor que separa el 25% más bajo de los datos. Es el percentil 25. **50%** (mediana o segundo cuartil): Valor que divide los datos en dos mitades iguales (percentil 50). **75%** (tercer cuartil): Valor que deja el 75% de los datos por debajo. Es el percentil 75. **max:** Valor máximo de la columna.

Transformaciones

Para este df en específico solo se aplicaron las transformaciones que se especifican en la sección de Generalidades ya que no fue necesaria ninguna otra. Una vez realizadas las mismas, el df quedó con las mismas dimensiones que tenía (93 entradas y tres columnas, tabla 1) por lo que se entiende que ninguna entrada en este df pertenece a un territorio en disputa o dependiente.

El archivo refinado obtenido a partir de estas transformaciones se encuentra en la carpeta RefinedData con el nombre *RefData_Health_care_index_by_countries_2020.csv*.

Pruperties_price_index_by_countries_2020.csv

Análisis de datos

El conjunto de datos en este archivo permite evaluar la accesibilidad de la vivienda en diferentes países, considerando tanto los costos de propiedad como de alquiler.

Al realizar el análisis y pasar los datos a un dataframe (df_Pruperties_price_index_by_countries_2020) se encuentra que el archivo está compuesto por 104 entradas y ocho columnas (tabla 1). Los tipos de datos presentes en este df son object para el campo Country y float64 para los restantes (tabla 7). Ninguna de las columnas presentó datos nulos. A partir de la descripción para este df se obtienen los datos de la figura 5.

Tabla 7. Nombre de los campos, tipo de dato y significado dentro del dominio de datos para el archivo Properties_price_index_by_countries_2020.csv

Columna	Tipo de dato	Significado dentro del dominio de los datos
Country	object	País para el cual se registra la información de asequibilidad de vivienda.
Price To Income Ratio	float64	Relación entre el precio de la vivienda y el ingreso promedio.
Gross Rental Yield City Centre	float64	Rentabilidad bruta de alquiler en el centro de la ciudad, calculada como el retorno anual del alquiler sobre el precio de la propiedad.
Gross Rental Yield Outside of Centre	float64	Rentabilidad bruta de alquiler fuera del centro de la ciudad.
Price To Rent Ratio City Centre	float64	Relación entre el precio de la vivienda y el costo de alquiler en el centro.
Price To Rent Ratio Outside Of City Centre	float64	Relación entre el precio de la vivienda y el costo de alquiler fuera del centro.
Mortgage As A Percentage Of Income	float64	Porcentaje del ingreso destinado a pagar una hipoteca.
Affordability Index	float64	Índice de asequibilidad que refleja cuán accesible es la compra de una vivienda en el país.

	Price To Income Ratio	Gross Rental Yield City Centre	Gross Rental Yield Outside of Centre	Price To Rent Ratio City Centre	Price To Rent Ratio Outside Of City Centre	Mortgage As A Percentage Of Income	Affordability Index
count	104.000000	104.000000	104.000000	104.000000	104.000000	104.000000	104.000000
mean	14.829712	4.841442	5.167212	24.607981	22.466731	169.583558	1.155962
std	14.561628	1.991724	1.912431	11.617980	10.065941	316.048973	0.784232
min	2.790000	1.380000	1.500000	8.800000	8.850000	20.640000	0.030000
25%	9.007500	3.402500	3.932500	16.977500	16.107500	61.495000	0.600000
50%	11.495000	4.510000	5.080000	22.175000	19.675000	100.200000	1.000000
75%	15.337500	5.892500	6.207500	29.350000	25.455000	165.465000	1.622500
max	133.290000	11.360000	11.300000	72.610000	66.600000	3025.030000	4.850000

Figura 5. Información obtenida de aplicar la función *describe()* al dataframe *df_Properties_price_index_by_countries_2020*. **count**: Número total de valores no nulos en la columna. **mean**: Promedio (media aritmética) de los valores. **std** (desviación estándar): Mide cuánto se desvían los valores respecto a la media. **min**: Valor mínimo de la columna. **25%** (primer cuartil): Valor que separa el 25% más bajo de los datos. Es el percentil 25. **50%** (mediana o segundo cuartil): Valor que divide los datos en dos mitades iguales (percentil 50). **75%** (tercer cuartil): Valor que deja el 75% de los datos por debajo. Es el percentil 75. **max:** Valor máximo de la columna.

Transformaciones

Al analizar el conjunto de datos se observó que en el campo Country existía no solo países sino también un territorio. Por lo anterior, se decidió aplicar la estrategia especificada en la sección de Generalidades del presente proyecto y eliminar dicha entrada. Luego de aplicadas las transformaciones allí mencionadas, el df pasó de tener 104 entradas a 103, manteniendo el número de columnas para ambos casos (tabla 1).

El archivo refinado obtenido a partir de estas transformaciones se encuentra en la carpeta RefinedData con el nombre *RefData_Properties_price_index_by_countries_2020.csv*.

Pupulation_density_by_countries.csv

Análisis de datos

El conjunto de datos en este archivo proporciona información sobre el área, la población y la densidad de cada país, útil para análisis demográficos y de espacio.

Al realizar el análisis y pasar los datos a un dataframe (df_Pupulation_density_by_countries) se encuentra que el archivo está compuesto por 251 entradas y nueve columnas (tabla 1). Los tipos de datos presentes en este df son object para todos los campos (tabla 8), esto al igual que lo ocurrido con el archivo Countries_age_structure.csv es un inconveniente ya que no permite el uso de los datos para análisis. De todos los campos el único que presentó datos nulos fue *Population source* con un conteo de siete campos nulos, estos se muestran en la figura 6. Además, aprovechando que estos datos poseen un campo fecha, sería interesante poder agregarle granularidad al análisis generando tres campos nuevos (día, mes y año). Dado que se decidió, como se expresa en la sección de Generalidades, utilizar los países como primary key eliminando aquellos registros que representen territorios dependientes o en disputa, se cambiará el nombre de la columna con dicho fin en esta tabla. Por lo anterior, la columna Country (or dependent territory) pasará a nombrarse Country.

Tabla 8. Nombre de los campos, tipo de dato y significado dentro del dominio de datos para el archivo Pupulation_density_by_countries.csv

Columna	Tipo de dato	Significado dentro del dominio de los datos
Rank	object	Rango o posición del país según el área o la población.
Country (or dependent territory)	object	Nombre del país o territorio dependiente.
Area km2	object	Área del país en kilómetros cuadrados.
Area mi2	object	Área del país en millas cuadradas.
Population	object	Población total del país.
Density pop./km2	object	Densidad poblacional por kilómetro cuadrado.
Density pop./mi2	object	Densidad poblacional por milla cuadrada.
Date	object	Fecha de la estimación de la población.
Population source	object	Fuente de los datos de población.

	Rank	Country (or dependent territory)	Area km2	Area mi2	Population	Density pop./km2	Density pop./mi2	Date	Population source
79	–	Sint Eustatius (Netherlands)	21	8	3,193	152	394	July 1, 2015	NaN
81	–	Saba (Netherlands)	13	5	1,947	150	388	January 1, 2016	NaN
150	–	Bonaire (Netherlands)	288	111	18,905	66	171	December 31, 2014	NaN
191	–	Abkhazia	8,660	3,344	2,43,206	28	73	April 27, 2018	NaN
207	–	Somaliland	1,76,120	68,000	35,08,180	20	52	August 17, 2017	NaN
225	–	South Ossetia	3,900	1,506	53,532	14	36	August 11, 2016	NaN
226	–	Artsakh	11,458	4,424	1,50,932	13	34	October 14, 2015	NaN

Figura 6. Se muestran las siete entradas presentes en el dataframe *df_Pupulation_density_by_countries* que poseen valor nulo para el campo Population source.

Transformaciones

Las transformaciones realizadas fueron varias para este archivo, entre ellas; se cambió el tipo de dato de object a float64 para los campos Area km2, Area mi2, Population, Density pop/km2 y Density po/.mi2. Estos cambios se realizaron de forma diferente según el campo a cambiar, ya que el formato que tenían los datos para estas columnas era distinto al habitual (usaban separación de miles con coma y cada dos dígitos). A su vez, se cambio al tipo datetime64[ns] el campo Date y a partir del mismo se agregaron tres columnas más al df desglosando la fecha en día, mes y año. Todos estos cambios pueden verse resumidos en la tabla 9.

Tabla 9. Nueva estructura y tipo de datos luego de realizar el refinado del archivo Pupulation_density_by_countries.csv.

Columna	Tipo de dato	Significado dentro del dominio de los datos
Rank	object	Rango o posición del país según el área o la población.
Country	object	Nombre del país o territorio dependiente.
Area km2	float64	Área del país en kilómetros cuadrados.
Area mi2	float64	Área del país en millas cuadradas.
Population	float64	Población total del país.
Density pop./km2	float64	Densidad poblacional por kilómetro cuadrado.
Density pop./mi2	float64	Densidad poblacional por milla cuadrada.
Date	datetime64[ns]	Fecha de la estimación de la población.
Population source	object	Fuente de los datos de población.
Day	int32	Día obtenido del campo date
Month	int32	Mes obtenido del campo date
Year	int32	Año obtenido del campo date

La próxima transformación a los datos fue cambiar el nombre de la columna Country (or dependent territory) a Country y realizar los ajustes para la misma descritos en la sección de Generalidades. Debido a que había países para los cuales el campo Population source no tenía datos, y como se considera un dato importante para la fidelidad y veracidad de los datos, se decidió eliminar aquellas entradas sin datos para este campo. Además, también se le hizo el mismo tratamiento de datos que para Country en cuanto a quitar espacios en blanco antes y después del primer y último carácter. En este momento el df, luego de realizar las transformaciones de eliminación de nulos, de países clasificados como territorios dependientes o en disputa y de eliminar aquellas entradas sin dato para el campo Population source, quedó con 201 entradas y 12 columnas (tabla 1).

Ahondando en los datos y en el análisis exploratorio se encontró que el país Uruguay en esta tabla difería de las anteriores en que se hace referencia al mismo como Uruguay[note 5]. Esto daría problemas al momento de realizar el análisis por lo que se decidió para esta entrada quitar el [note 5] y que quedara solo Uruguay.

Luego de aplicadas todas estas transformaciones se profundizó en las métricas para este df, dichas métricas se encuentran resumidas en la figura 7.

	Area km2	Area mi2	Population	Density pop./km2	Density pop./mi2	Date	Day	Month	Year
count	2.010000e+02	2.010000e+02	2.010000e+02	201.000000	201.000000	201	201.000000	201.000000	201.000000
mean	6.655587e+05	2.569737e+05	3.794880e+07	437.255721	1132.484080	2019-03-13 10:59:06.268656640	7.965174	6.139303	2018.751244
min	4.400000e-01	1.700000e-01	1.000000e+03	1.900000	4.900000	2015-01-24 00:00:00	1.000000	1.000000	2015.000000
25%	2.104000e+04	8.124000e+03	1.604528e+06	34.000000	87.000000	2018-07-01 00:00:00	1.000000	3.000000	2018.000000
50%	1.126220e+05	4.348400e+04	7.901454e+06	82.000000	212.000000	2019-07-01 00:00:00	1.000000	7.000000	2019.000000
75%	4.912100e+05	1.896570e+05	2.582307e+07	208.000000	539.000000	2019-09-30 00:00:00	18.000000	7.000000	2019.000000
max	1.712524e+07	6.612093e+06	1.401812e+09	20550.000000	53224.000000	2020-07-01 00:00:00	31.000000	12.000000	2020.000000
std	1.892733e+06	7.307883e+05	1.425811e+08	2090.089438	5413.270893	NaN	11.302822	2.956433	0.968406

Figura 7. Información obtenida de aplicar la función *describe()* al dataframe *df_Pupulation_density_by_countries_sinNan_fechas_float_country*. **count**: Número total de valores no nulos en la columna. **mean**: Promedio (media aritmética) de los valores. **std** (desviación estándar): Mide cuánto se desvían los valores respecto a la media. **min**: Valor mínimo de la columna. **25%** (primer cuartil): Valor que separa el 25% más bajo de los datos. Es el percentil 25. **50%** (mediana o segundo cuartil): Valor que divide los datos en dos mitades iguales (percentil 50). **75%** (tercer cuartil): Valor que deja el 75% de los datos por debajo. Es el percentil 75. **max:** Valor máximo de la columna.

El archivo refinado obtenido a partir de estas transformaciones se encuentra en la carpeta RefinedData con el nombre *RefData_Pupulation_density_by_countries.csv*.

Quality_of_life_index_by_countries_2020.csv

Análisis de datos

El conjunto de datos en este archivo proporciona un análisis completo de la calidad de vida en diferentes países, cubriendo aspectos de seguridad, salud, costo de vida, propiedad, tráfico, contaminación y clima.

Al realizar el análisis y pasar los datos a un dataframe (df_Quality_of_life_index_by_countries_2020) se encuentra que el archivo está compuesto por 80 entradas y diez columnas (tabla 1). Los tipos de datos presentes en este df son object para el campo Country y float64 para los restantes (tabla 10). Ninguna de las columnas presentó datos nulos. A partir de la descripción para este df se obtienen los datos de la figura 8.

Tabla 10. Nombre de los campos, tipo de dato y significado dentro del dominio de datos para el archivo Quality_of_life_index_by_countries_2020.csv

Columna	Tipo de dato	Significado dentro del dominio de los datos
Country	object	País para el cual se registra el índice de calidad de vida.
Quality of Life Index	float64	Índice general de calidad de vida en el país.
Purchasing Power Index	float64	Índice de poder adquisitivo, indicando cuánta capacidad de compra tienen los ciudadanos.
Safety Index	float64	Índice de seguridad en el país.
Health Care Index	float64	Índice de calidad en el cuidado de la salud.
Cost of Living Index	float64	Índice de costo de vida en el país.
Property Price to Income Ratio	float64	Relación entre el precio de la propiedad y el ingreso promedio.
Traffic Commute Time Index	float64	Índice que mide el tiempo promedio de traslado en el tráfico.
Pollution Index	float64	Índice de contaminación ambiental.
Climate Index	float64	Índice que mide la calidad climática del país.

	Quality of Life Index	Purchasing Power Index	Safety Index	Health Care Index	Cost of Living Index	Property Price to Income Ratio	Traffic Commute Time Index	Pollution Index	Climate Index
count	80.000000	80.000000	80.000000	80.000000	80.000000	80.000000	80.000000	80.000000	80.000000
mean	134.100375	59.751000	61.079125	64.840500	52.461250	13.332875	35.989250	53.227625	77.350250
std	33.921748	26.995155	13.808162	9.840011	19.772413	7.139386	8.236982	20.698200	16.520585
min	55.650000	13.520000	22.510000	42.800000	21.980000	2.790000	20.100000	11.550000	20.220000
25%	106.377500	36.560000	53.177500	56.357500	37.127500	8.830000	29.855000	36.157500	68.705000
50%	132.755000	55.620000	60.055000	66.000000	48.810000	11.680000	34.755000	57.125000	79.215000
75%	162.395000	83.087500	72.075000	72.565000	67.365000	14.935000	39.650000	67.435000	90.612500
max	192.670000	119.530000	88.140000	86.710000	122.400000	47.460000	61.080000	88.370000	99.790000

Figura 8. Información obtenida de aplicar la función *describe()* al dataframe *df_Quality_of_life_index_by_countries_2020*. **count**: Número total de valores no nulos en la columna. **mean**: Promedio (media aritmética) de los valores. **std** (desviación estándar): Mide cuánto se desvían los valores respecto a la media. **min**: Valor mínimo de la columna. **25%** (primer cuartil): Valor que separa el 25% más bajo de los datos. Es el percentil 25. **50%** (mediana o segundo cuartil): Valor que divide los datos en dos mitades iguales (percentil 50). **75%** (tercer cuartil): Valor que deja el 75% de los datos por debajo. Es el percentil 75. **max:** Valor máximo de la columna.

Transformaciones

Los datos contenidos en este archivo tuvieron mayoro complejidad en cuanto a tipo de datos, presencia de nulos o alguna otra particularidad. Por lo anterior, solo se aplicaron las transformaciones necesarias descritas en la sección de Generalidades. Luego de realizar dichas transformaciones la estructura y dimensión del df quedó incambiada, continuó con 80 entradas y diez columnas.

El archivo refinado obtenido a partir de estas transformaciones se encuentra en la carpeta RefinedData con el nombre *RefData_Quality_of_life_index_by_countries_2020.csv*.

Datos Refinados

Como se fue mencionando en las secciones anteriores del análisis exploratorio, para cada archivo se realizó análisis exploratorio y distintas transformaciones para limpiar los datos. El resultado fueron archivos nuevos en formato csv y con los cuales se seguirá de ahora en adelante. Estos archivos nuevos se ubican dentro de la carpeta RefinedData.

A continuación, se enumeran los distintos archivos refinados a partir del conjunto de datos "Estudios de Países" en la carpeta RefinedData:

1. RefData_Cost_of_living_index_by_country_2020.csv
2. RefData_Countries_age_structure.csv
3. RefData_Crime_index_by_countries_2020.csv
4. RefData_Health_care_index_by_countries_2020.csv
5. RefData_Properties_price_index_by_countries_2020.csv
6. RefData_Population_density_by_countries.csv
7. RefData_Quality_of_life_index_by_countries_2020.csv

Elección del modelo de datos

Previo a la implementación de las visualizaciones y el análisis de datos, se llevó a cabo una evaluación exhaustiva de los modelos de datos más adecuados para el presente proyecto.

Inicialmente, se consideró el modelo One Big Table (OBT) debido a su simplicidad de construcción y la minimización de operaciones de unión. Esta característica permite centrar los esfuerzos en la lógica de negocio, agilizando el desarrollo y reduciendo el mantenimiento del modelo. Sin embargo, se identificó que este enfoque podría presentar escalabilidad limitada y complejizar la formulación de consultas a medida que la cantidad de datos aumentara.

Tras un análisis detallado de las tablas, los datos y las preguntas de negocio específicas, se optó por un modelo de datos normalizado. Esta decisión se fundamenta en las siguientes razones:

- **Volumen de datos:** El conjunto de datos presenta un tamaño reducido, con un número limitado de tablas y registros. En este contexto, las operaciones de unión, que suelen ser un punto crítico en modelos normalizados, no representan una carga significativa.

- **Perfil de los usuarios:** El grupo de usuarios es reducido y posee un conocimiento profundo de los datos, lo que facilita la comprensión y utilización del modelo normalizado.

- **Adaptabilidad:** La estructura normalizada ofrece mayor flexibilidad para futuras ampliaciones del modelo, garantizando la integridad de los datos y facilitando la incorporación de nuevas dimensiones de análisis.

En conclusión, considerando las características específicas del conjunto de datos y los requerimientos del proyecto, se determinó que un modelo de datos normalizado es la opción más adecuada. Esta decisión permite aprovechar las ventajas de este enfoque, como la integridad de los datos y la facilidad de mantenimiento, sin comprometer el rendimiento o la escalabilidad del sistema.

Parte 1b-Análisis de las preguntas planteadas

Pregunta 1: ¿Existe una relación entre el costo de vida y el poder adquisitivo local en los países más y menos densamente poblados?

El objetivo de esta pregunta es analizar si los países más densamente poblados tienen un mayor costo de vida en relación con su poder adquisitivo, comparado con los menos densos. Esto permitiría comprender cómo la densidad poblacional afecta el equilibrio económico puede informar políticas de desarrollo.

Para responder esta incógnita se utilizarán las siguientes tablas y campos:

1- RefData_Cost_of_living_index_by_country_2020.csv (campo: Cost of Living Index).

2- RefData_Population_density_by_countries.csv (campo: Density pop./km2).

3- RefData_Quality_of_life_index_by_countries_2020.csv (campo: Purchasing Power Index).

Al momento de realizar el merge entre las tablas se considera que para este juego de datos es preferible dejar los valores nulos ya que, si un país no tiene datos en cualquiera de los índices, excluirlo es preferible, ya que rellenar con 0 o eliminar podría distorsionar la correlación. Otra información para destacar es que hay dos tablas (la 1 y la 3 en el punteo del párrafo anterior) que poseen un campo con el mismo nombre, *Cost of Living Index*. Por tal motivo, al realizar el merge, automáticamente se renombraron como *Cost of Living Index_x* y *Cost of Living Index_y* (correspondientes a las tablas 1 y 3 del punteo del párrafo anterior respectivamente). Dado que el primero era el que poseía menos datos nulos, y que corresponde al archivo específico de costo de vida, se decidió realizar la gráfica con el campo *Cost of Living Index_x*.

A continuación, se presenta una gráfica de dispersión mostrando la relación entre costo de vida y el poder de compra (figura 9).

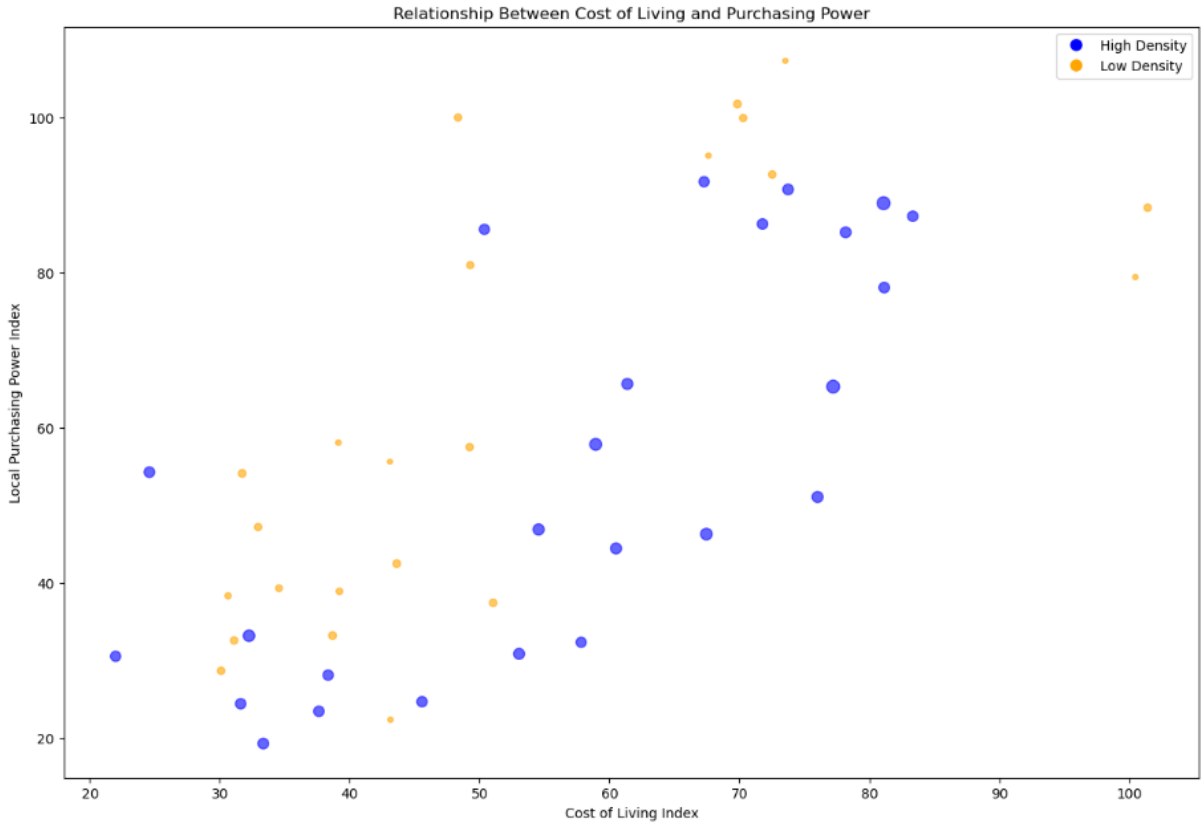

Figura 9. Gráfico de dispersión mostrando la relación entre el costo de vida y el poder de compra registrado en los distintos países en base a la densidad poblacional. En el eje de las X se muestra el índice de costo de vida, mientras en el eje Y índice de poder adquisitivo. La leyenda de la gráfica corresponde a la densidad poblacional, siendo azul para aquellos países de alta densidad y amarillo para aquellos de baja densidad.

Para chequear la incidencia en el estudio de utilizar una tabla donde se hayan excluido los valores nulos, se obtuvo una tabla nueva con todos los países incluyendo aquellos que poseían índices en nulo y se volvió a realizar el análisis. El resultado fue muy parecido al que se visualiza en la figura 9, manteniéndose el aspecto y características generales de la misma. El código utilizado para esta prueba quedará comentado en el notebook y sección correspondientes.

Análisis y conclusión de los resultados

Como tendencia general, puede decirse que existe una relación positiva entre el índice de costo de vida y el índice de poder adquisitivo. Esto significa que, en general, los países con un alto costo de vida también tienden a tener un mayor poder adquisitivo. Esto puede deberse a que economías más desarrolladas tienen tanto costos como ingresos altos.

Los países con alta densidad poblacional parecen estar concentrados en la parte superior derecha de la gráfica (alto costo de vida y alto poder adquisitivo). Esto podría indicar que las áreas densamente pobladas tienden a ser económicamente activas y con ingresos elevados, pero también más caras para vivir.

Los países con baja densidad están distribuidos en diferentes regiones de la gráfica. Algunos tienen un costo de vida bajo y un poder adquisitivo bajo, mientras que otros tienen un poder adquisitivo más alto en relación con su costo de vida.

Los países de alta densidad tienden a estar mejor representados en la esquina superior derecha de la gráfica (alto costo y poder adquisitivo), mientras que los países de baja densidad parecen tener una mayor dispersión, lo que sugiere que la relación costo-poder adquisitivo es más variada en estos países.

Aunque la densidad no parece tener un impacto directo sobre la relación entre costo de vida y poder adquisitivo, es notable que los países más densamente poblados tienden a concentrarse en la parte más alta de la gráfica en ambos índices.

La gráfica respalda la hipótesis de que, en general, los países con un alto costo de vida también tienen un alto poder adquisitivo. Sin embargo, la densidad poblacional introduce variaciones interesantes, ya que los países de baja densidad presentan más diversidad en la relación entre estos dos índices. Para profundizar en este resultado, podría realizarse un análisis de correlaciones o tendencias separadas para los grupos de alta y baja densidad, o analizar los países atípicos (puntos alejados de la tendencia principal).

Pregunta 2: ¿Cómo varía la distribución de edades en países con diferentes niveles de calidad de vida?

Con esta pregunta se pretende analizar si los países con mayor proporción de población joven (0-14 años) o adulta (15-64 años) tienen un mayor índice de calidad de vida. Esta pregunta es interesante ya que la demografía de un país podría influir en su calidad de vida debido a las políticas sociales y económicas.

Para responder esta incógnita se utilizarán las siguientes tablas y campos:

1- RefData_Countries_age_structure.csv (*Age 0 to 14 Yeaers, Age 15 to 64 Years, Age above 65 Years*)

2- RefData_Quality_of_life_index_by_countries_2020.csv (*Quality of Life Index*)

Para responder esta pregunta se decidió separar los datos por rango de calidad de vida. Para ello inicialmente se establecieron rangos para definir los límites o bordes de los intervalos en los qu se dividirán los datos. Los rangos propuestos fueron de 0 a 50, de 50 a 75 y de 75 a 100 (denominados como Low, Medium y High respectivamente). Sin embargo, no se registraban conteos de países para el rango Low. Por este motivo se decidió ahondar más en los datos y analizar la tabla obtenida a partir del merge generado para esta pregunta. Utilizando la función describe con el df merged_NAN_df, se logró evidenciar que el valor mínimo para el campo *Quality of Life Index* en dicho df es 55.65 mientras el máximo es de 192.67, lo cual explica por qué el rango Low no captura ningún dato. Este dato también indica que deben ajustarse los rangos para reflejar mejor la distribución de los datos. Con la información actualizada se propone redefinir los rangos de la siguiente manera:

- **Low:** 55-100 (valores más bajos dentro del conjunto de datos)
- **Medium:** 100-150
- **High:** 150-200 (aproximadamente, considerando el valor máximo)

Una vez realizado este cambio en el código se verifican los nuevos rangos mediante la función value_counts() y se vuelve a graficar el df obteniéndose como resultado la figura 10.

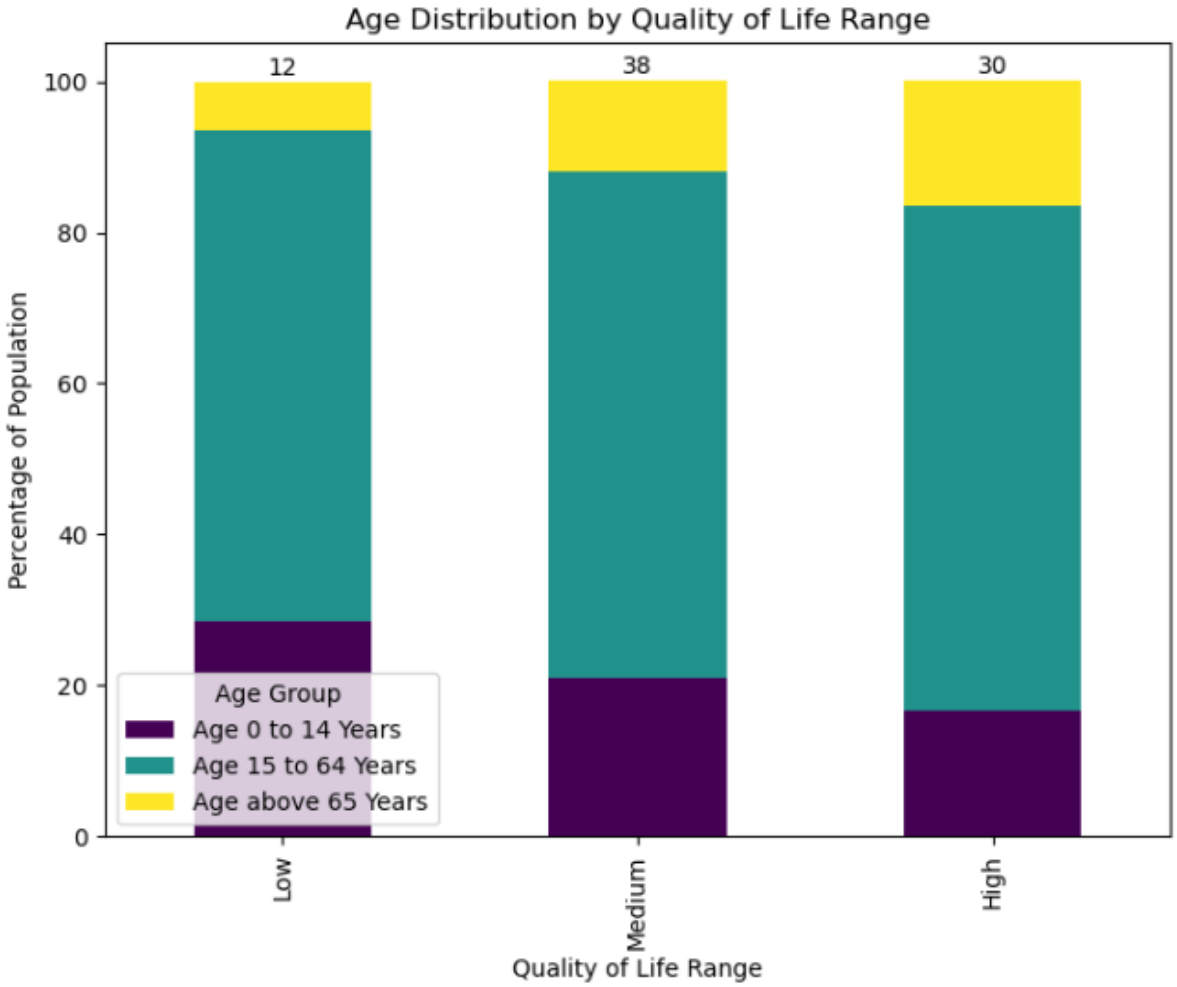

Figura 10. Gráfica de barras apiladas mostrando la distribución promedio de la población por grupo etario (Age 0 to 14 Years, Age 15 to 64 Years, Age above 65 Years) en función de los rangos de calidad de vida (Quality Life Range): Low, Medium y High. Los números en la parte superior de cada barra indican el número de países en cada rango.

Análisis y conclusión de los resultados

Al observar la figura 10 vemos que, en la distribución de la población, el grupo predominante en todos los rangos son las edades entre los 15 y 64 años (verde), lo cual tiene sentido ya que corresponde a la población económicamente activa. Por otro lado, los grupos de 0 a 14 años (morado) y mayores de 65 (amarillo) representan proporciones menores, siendo los mayores de 65 años el segmento más reducido.

Al pasar a la diferencia entre rangos, vemos que a medida que aumenta la calidad de vida la proporción de personas mayores de 65 tiende a crecer. Esto es consistente con la expectativa de que una mayor calidad de vida se asocia con una mayor esperanza de vida. A su vez, la proporción de personas jóvenes (0 a 14 años) disminuye, lo que puede estar relacionado con menores tasas de natalidad en países con alta calidad de vida. Por último, la población entre 15 y 64 años se mantiene estable, aunque con ligeras variaciones.

A modo de conclusión, podría decirse que el gráfico ilustra cómo la composición etaria varía con los rangos de calidad de vida. Países con mejor calidad de vida tienen una población más envejecida, mientras que aquellos con menor calidad de vida tienen más personas jóvenes y menos personas mayores. Esto refleja dinámicas demográficas y económicas relacionadas con el desarrollo social y económico.

Pregunta 3: ¿Qué países ofrecen el mejor balance entre costo de vida y calidad de vida?

Esta pregunta nos permitirá conocer aquellos países donde el costo de vida es razonable en comparación con un alto índice de calidad de vida. Esta información es importante para las personas que buscan optimizar su calidad de vida sin incurrir en altos costos.

Para responder esta incógnita se utilizarán las siguientes tablas y campos:

1- RefData_Cost_of_living_index_by_country_2020.csv (*Cost of Living Index*)

2- RefData_Quality_of_life_index_by_countries_2020.csv (*Quality of Life Index*)

Para responder esta pregunta se propone un mapa geográfico. Sin embargo, el inconveniente es que no se poseen las latitudes ni longitudes para los distintos países de la tabla conseguida. Para solucionarlo se utiliza la librería geopandas la cual permite a partir del nombre de un país obtener la latitud y longitud de este. Dicho acercamiento solo funciona si el nombre del país en el campo Country de la tabla mergeada coincide con nombre del país del archivo shapefile de la librería geopandas. Como solución debe descargarse el archivo shapefile desde la página Natural Earth Data (https://www.naturalearthdata.com/downloads/110m-cultural-vectors/) y con este archivo realizar una comparación con los países en la tabla obtenida a partir del merge y obtener aquellos países que no coincidan. Dicho archivo se encuentra en la carpeta AuxiliarData.

A partir de las consideraciones anteriores, se realizó el análisis para conocer el balance entre costo de vida y calidad de vida teniendo como resultado la figura 11. La visualización generada utiliza representación del balance mediante heatmap donde los colores claros (verde/amarillos) sugieren países que ofrecen un buen nivel de calidad de vida en relación con su costo de vida. Por otro lado, los colores oscuros (azules/violetas) reflejan un mayor costo relativo para una menor calidad de vida.

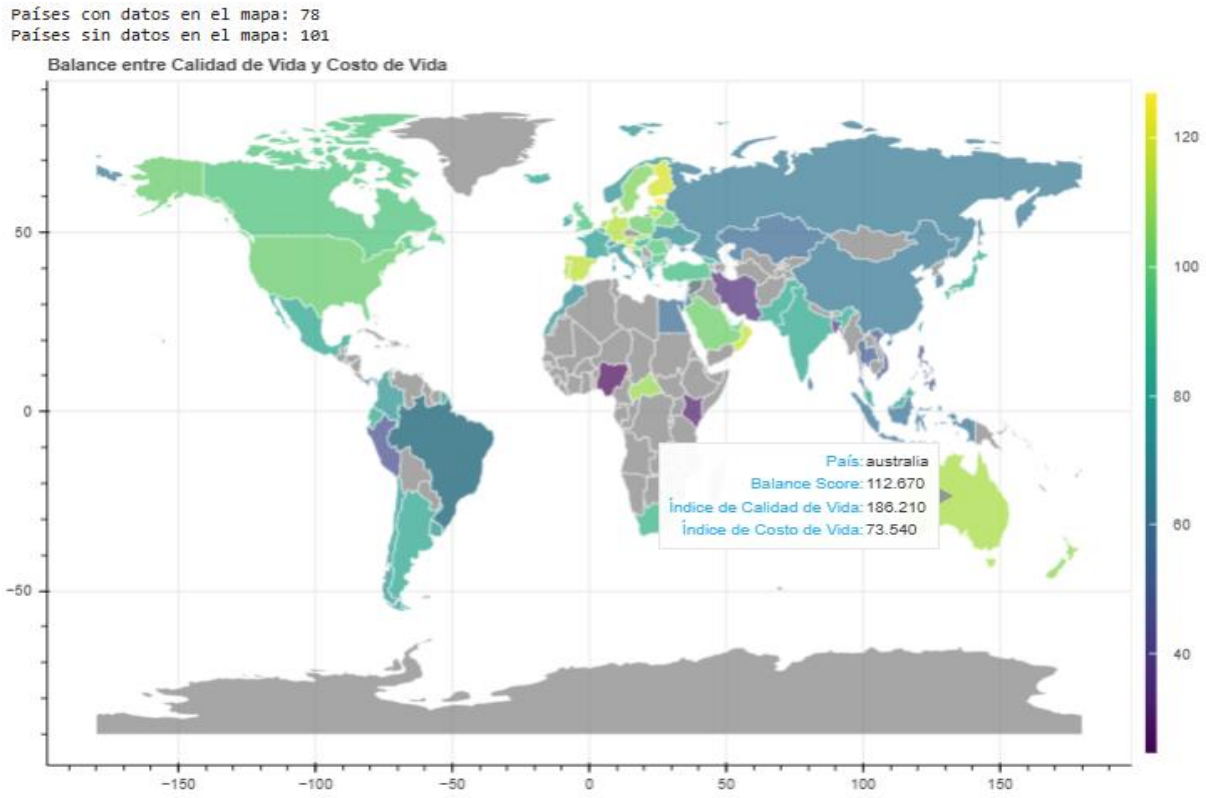

Figura 11. Visualización por mapa geográfico del análisis del balance entre calidad de vida y costo de vida por países. Los colores claros (verde/amarillo) indican un balance más alto entre calidad de vida y costo de vida. Los colores oscuros (azul/violeta) indican un balance más bajo entre calidad de vida y costo de vida. El gris significa que no hay datos disponibles. En su versión html, la visualización presenta funcionalidad de obtención de los datos al cliquear arriba de un país. Se especifica la cantidad de países con datos y sin datos en la parte superior de la visualización.

Análisis y conclusión de los resultados

A partir de los datos obtenidos podemos observar que Europa occidental y Escandinavia parecen tener un balance relativamente alto según la escala y colores del heatmap. Latinoamérica y África muestran un balance más variado, con algunos países en azul y otros con datos faltantes. Por último, Asia y Oceanía tienen un rango diverso de balances, pero algunos países destacan con valores altos.

El mapa es útil para comparar rápidamente las condiciones de vida relativas entre los países, pero tiene ciertas limitaciones. Una de ellas es la limitación en los datos, los paíes sin datos deben ser considerados para interpretar el mapa en su contexto global. Por otro lado, tampoco se especifican causas, por lo que no se poseen los factores contribuyentes al balance en cada país (como políticas económicas, salud, educación, etc.)

Pregunta 4: ¿Qué impacto tienen los índices de seguridad y crimen en el costo de vida?

La finalidad de esta pregunta es evaluar si los países más seguros tienen un costo de vida significativamente mayor o menor. Pareció una pregunta importante ya que la seguridad es un factor clave para muchas personas al elegir un lugar para vivir.

Para responder esta incógnita se utilizarán las siguientes tablas y campos:

1- RefData_Cost_of_living_index_by_country_2020.csv (*Cost of Living Index*)

2- RefData_Crime_index_by_countries_2020.csv (*Crime Index, Safety Index*)

Para contestar esta pregunta se eligió visualizar los datos a partir de un gráfico de dispersión de dos ejes y escala de color en formato heatmap para un tercer dato. Dado que el índice de seguridad y criminalidad están relacionados, si se usaban ambos para los ejes daba como resultado un gráfico esperable y poco informativo (el código para este gráfico se encuentra comentado en el notebook *Parte_1b_preguntasVisualizaciones.ipynb* dentro de la sección para esta pregunta). La visualización más informativa resultó ser aquella que tenía en los ejes del gráfico de dispersión el índice de criminalidad y el índice de costo de vida, usando el índice de seguridad para el heatmap (figura 12).

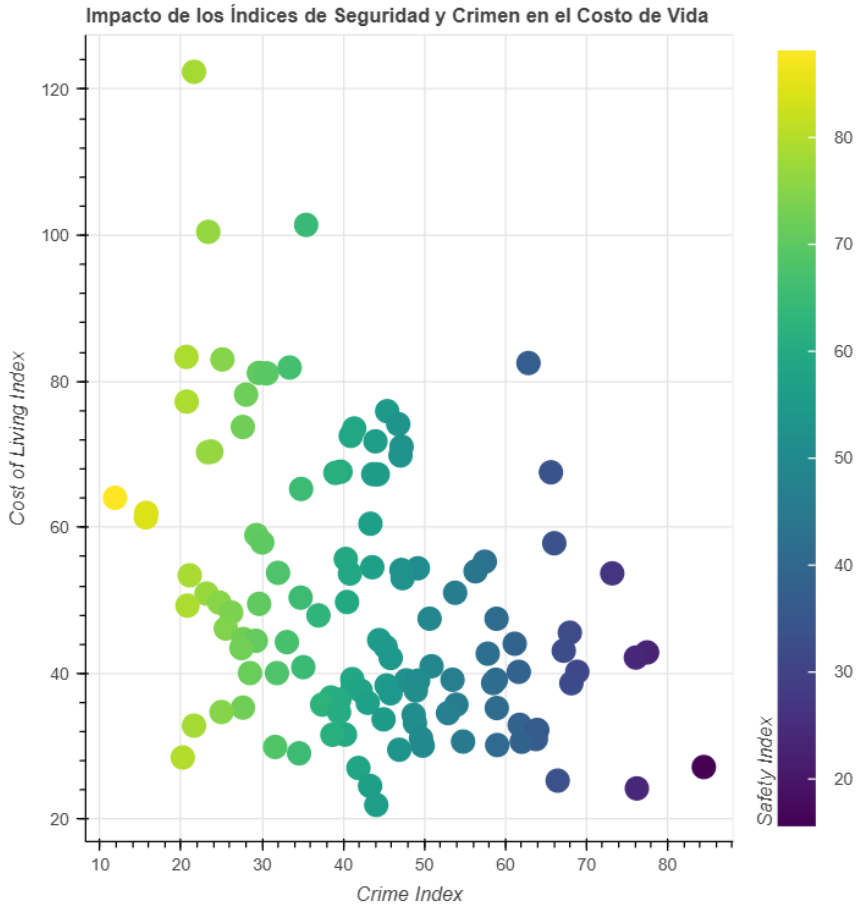

Figura 12. Gráfico de dispersión que muestra la relación entre el índice de criminalidad (Crime Index), el costo de vida (Cost of Living Index), y el índice de seguridad (Safety Index). Cada punto representa un país, donde el color

indica el nivel de seguridad (de morado a amarillo, siendo amarillo más seguro).

Análisis y conclusión de los resultados

Al analizar el gráfico, a primera vista no parece haber una relación fuerte y directa entre el índice de criminalidad y el costo de vida. Dicho lo anterior, sí es verdad que algunos países con índices de criminalidad bajos (menores a 20) tienden a tener costos de vdia altos, hay una gran dispersión en los valores para índices de criminalidad entre 20 y 60. Además, países con índices de criminalidad altos (superiores a 60) no necesariamente tienen costo de vdia bajos, ya que algunos se distribuyen en diferentes rangos de costo.

Cuando se observa el impacto del índice de seguridad se aprecia que un índice bajo (azul a morado) tiende a estar asociado con países donde el costo de vida es más bajo, mientras que países con índices de seguridad altos (amarillo a verde claro) tienden a ubicarse más hacia la mitad o el tercio superior del costo de vida. Esto sugiere que la seguridad puede tener un leve impacto positivo en el costo de vida, ya que países más seguros suelen tener mayores costos de vida.

La mayoría de los países se concentran en un rango medio de criminalidad (entre 20 y 60) y costos de vida moderados (entre 30 y 80). Existen algunos valores atípicos donde países con índices de criminalidad bajos presentan costos de vida extremadamente altos, lo que podría corresponder a economías desarrolladas con altos niveles de seguridad (por ejemplo, países escandinavos o similares)

A partir de las observaciones anteriores, podemos inferir que la relación entre el *Cost of Living Index* y el *Crime Index* no es lineal ni claramente fuerte. Esto implica que otros factores externos, como el desarrollo económico, la infraestructura, o las políticas gubernamentales, podrían tener un mayor impacto en el costo de vida. Por otro lado, el *Safety Index* parece tener una correlación más notable, donde países más seguros tienden a tener costos de vida más altos, aunque esto no es una regla universal. Al analizar los patrones de coloración parece ser que países con índices de criminalidad extremadamente bajos tienen una mayor probabilidad de tener costos de vida altos, mientras que los países con costos de vida muy bajos suelen presentar índices de seguridad bajos. Tratando de fundamentar lo anterior, podría pensarse que los países más desarrollados económicamente suelen tener mayores costos de vida y mejores índices de seguridad, ya que invierten más en sistemas de justicia, educación, y bienestar social. Por el contrario, en países menos desarrollados, un costo de vida bajo puede estar asociado con mayores niveles de criminalidad y menor seguridad.

Pregunta 5: ¿Cómo varía la calidad del sistema de salud según la densidad poblacional?

Otro punto importante cuando se evalúan opciones de vivienda por ejemplo, es tener a disposición un buen sistema de salud. Este puede verse afectado por varios motivos, entre ellos la densidad de población dado que puede provocar una presión fuerte sobre el sistema de salud afectando la calidad del servicio brindado, la disponibilidad de medicamentos, la atención, entre otras. Por tal motivo se analizará si los países más densamente poblados tienen sistemas de salud más eficientes o menos.

Para responder esta incógnita se utilizarán las siguientes tablas y campos:

1- RefData_Health_care_index_by_countries_2020.csv (*Health Care Index*)

2- RefData_ Pupulation_density_by_countries.csv (*Density pop./km2*)

Para la visualización de los datos de esta pregunta se pensó utilizar box plot dado que nos permitirá analizar la distribución del Health Care Index agrupando los países en rangos de densidad poblacional (ejemplo: baja, media, alta). El box plot es útil para visualizar cómo varía la calidad del sistema de salud en diferentes niveles de densidad poblacional y también resalta posibles valores atípicos. En este análisis se descartarán las entradas con valores nulos luego de realizar el merge entre las tablas involucradas. Cada diagrama de caja representa una categoría de densidad poblacional (baja, media, alta y muy alta), y la posición y forma de cada caja nos proporciona información valiosa sobre la calidad del sistema de salud en cada categoría. El resultado se muestra en la figura 13.

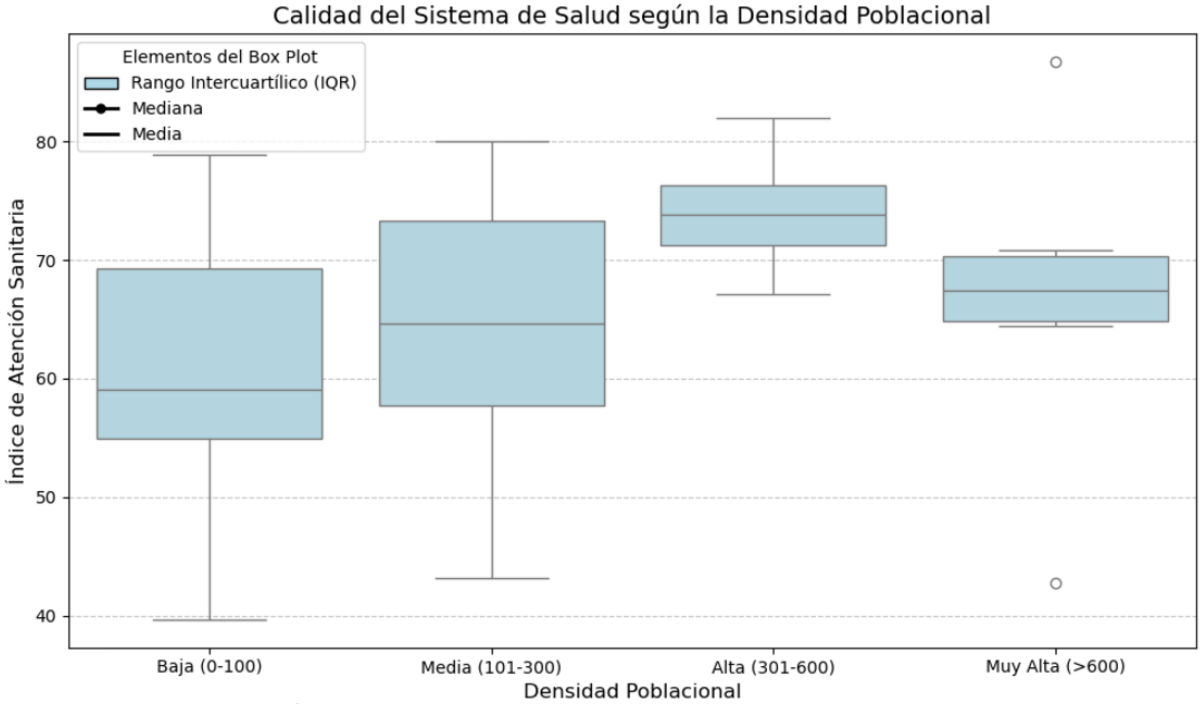

Figura 13. Variación del Índice de Atención Sanitaria según la Densidad Poblacional. Los diagramas de caja muestran la distribución del índice de atención sanitaria (eje Y) en diferentes categorías de densidad poblacional (eje X). La línea dentro de la caja representa la mediana, el rectángulo representa el rango intercuartílico (IQR) y el punto negro indica la media.

Análisis y conclusión de los resultados

De la figura resultante se desprende que a medida que aumenta la densidad poblacional, el índice de atención sanitaria tiende a ser mayor. Esto sugiere que, en promedio, las regiones con mayor densidad poblacional tienen sistemas de salud de mejor calidad.

A pesar de la tendencia general, existe una considerable variabilidad en la calidad de la atención sanitaria dentro de cada categoría de densidad poblacional. Esto nos indica que otros factores además de la densidad poblacional influyen en la calidad del sistema de salud.

En cuanto a los valores atípicos que se visualizan como puntos blancos en el gráfico, su presencia nos sugiere que hay algunos países que no siguen el patrón general. Por ejemplo, en la categoría de "Muy Alta (>600)" densidad poblacional, hay una región con un índice de atención sanitaria relativamente bajo y otro relativamente alto. Esto podría deberse a factores específicos de esa región, como problemas de infraestructura, falta de recursos o políticas de salud inadecuadas.

En cuanto a la pregunta original planteada de “Cómo varía la calidad del sistema de salud según la densidad poblacional” el gráfico sugiere que la densidad poblacional es un factor importante que influye en la calidad del sistema de salud. Sin embargo, no es el único factor. Otros factores como el nivel de desarrollo económico, la inversión en salud, la cobertura universal de salud, entre otros, también pueden jugar un papel importante.

Pregunta 6: ¿Cuáles son los países más accesibles para comprar vivienda considerando los índices económicos?

La accesibilidad a la vivienda es un tema central en la calidad de vida de las personas, ya que representa uno de los mayores desafíos económicos que enfrentan las familias en todo el mundo. Poder determinar qué países ofrecen mejores condiciones para adquirir una vivienda es clave no solo para quienes buscan establecerse en un lugar más accesible económicamente, sino también para los gobiernos y desarrolladores inmobiliarios que buscan evaluar las necesidades habitacionales de sus poblaciones.

Entender los factores que hacen que un país sea más asequible para la compra de vivienda permite identificar tendencias globales relacionadas con el costo de vida, los ingresos locales, y las dinámicas del mercado inmobiliario. Además, proporciona una herramienta valiosa para quienes buscan maximizar su calidad de vida a través de decisiones estratégicas sobre dónde vivir o invertir.

Esta pregunta no solo aborda la relación entre el costo de la vivienda y los ingresos promedio, sino también otros aspectos económicos clave, como el poder adquisitivo y los índices de rentabilidad.

Para responder esta incógnita se utilizarán las siguientes tablas y campos:

1- RefData_Properties_price_index_by_countries_2020.csv (*Affordability Index, Price To Income Ratio, Gross Rental Yield City Centre*)

2- RefData_Cost_of_living_index_by_country_2020.csv (*Local Purchasing Power Index*)

A partir del análisis de los datos se obtuvo una tabla de los 10 países más accesibles para comprar vivienda considerando los índices económicos (tabla 10). Estos datos fueron luego usados para generar una visualización que permitiera la interpretación del análisis de una forma más directa y sencilla. La estrategia de visualización elegida es un gráfico de burbujas de los 10 países más accesibles para comprar vivienda (figura 14). Esta gráfica permite comparar múltiples dimensiones de los datos en una sola representación visual. Se podrá ver

cómo el poder adquisitivo y la relación precio-ingreso afectan la accesibilidad de la vivienda en diferentes países entre otros datos.

Tabla 10. La tabla clasifica a los países según el Índice de Asequibilidad, que mide la accesibilidad de la vivienda considerando factores como ingresos locales y precios. Evaluaremos cómo interactúan las variables

Country	Price To Income Ratio	Affordability Index	Gross Rental Yield City Centre	Local Purchasing Power Index
Saudi Arabia	2.79	4.85	7.22	100.00
United States	3.52	3.79	10.36	109.52
Puerto Rico	3.53	3.67	9.34	79.38
United Arab Emirates	5.29	2.55	9.05	91.58
Belgium	6.91	2.42	4.92	86.28
Denmark	7.45	2.24	4.04	100.88
Qatar	5.82	2.21	7.19	111.69
Palestine	4.38	2.15	7.16	46.91
Netherlands	7.51	2.12	5.28	90.73
South Africa	3.93	2.11	9.85	73.61

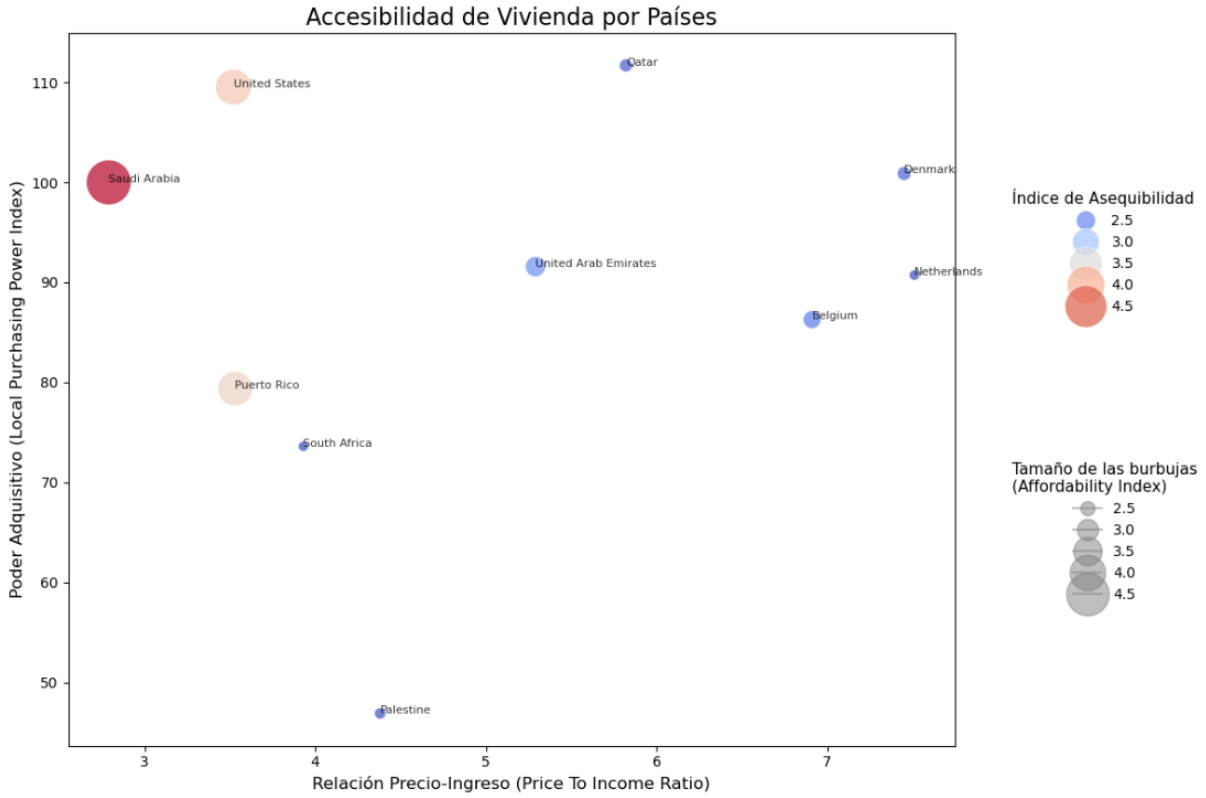

Figura 14. Gráfico de dispersión en formato burbuja que evalúa la accesibilidad de vivienda por países en función de tres variables clave: Eje X: La relación Precio-Ingreso (*Price To Income Ratio*), que representa el costo promedio de una propiedad residencial en relación con el ingreso anual promedio. Valores más altos indican que las viviendas son menos accesibles en términos relativos. Eje Y: El índice de poder adquisitivo (*Purchasing Power Index*), que mide la capacidad de los ciudadanos para adquirir bienes y servicios con los ingresos locales. Valores más altos indican mayor poder adquisitivo. Tamaño de las burbujas: El índice de asequibilidad (*Affordability Index*), que sintetiza cuán accesible es adquirir una vivienda considerando diversos factores económicos. Burbujas más grandes indican mejor asequibilidad. Color de las burbujas: Representa también el índice de asequibilidad (*Affordability Index*) mediante una escala de colores desde azul (menos

asequible) hasta rojo (más asequible), lo que facilita la identificación visual de patrones. Cada burbuja corresponde a un país, etiquetado por su nombre, permitiendo observar patrones en los datos a nivel global. Las leyendas proporcionan información adicional sobre la escala de colores y tamaños.

Análisis y conclusión de los resultados

Al estudiar la relación entre el índice de asequibilidad y el *Price to income ratio* a partir del procesamiento de los datos, vemos que Arabia Saudita posee el primer lugar en asequibilidad con un valor de 4.85. Además, analizando el resto de los números podemos decir que Arabia Saudita lidera en accesibilidad debido a una combinación de precios bajos y un poder adquisitivo adecuado (*Price to Income Ratio* de 2.79 y Poder Adquisitivo Local de 100).

Como tendencia general los países con un menor *Price to income ratio* (2.79-3.93) suelen tener un índice de asequibilidad alto, mayor a tres. Ejemplo, Arabia Saudita, Estados Unidos, y Puerto Rico. Los países como Dinamarca (7.45) y los Países bajos (7.51), con ratios más altos, presentan menores índices de asequibilidad (<2.3). Esto podría indicar que una mayor proporción de ingresos destinados a la compra de vivienda disminuye la asequibilidad.

Países con bajo *Price to Income Ratio* y un poder adquisitivo alto, como Arabia Saudita, tienen el mejor desempeño en asequibilidad. En cuanto al impacto de la estabilidad económica, países con altos precios relativos y bajos rendimientos de alquiler (Dinamarca y Países Bajos) son menos asequibles, incluso con un poder adquisitivo elevado. En cuanto a los países con bajo poder adquisitivo, ejemplo Palestina, enfrentan ciertos desafíos significativos de accesibilidad debido a ingresos insuficientes, a pesar de tener precios no tan elevados.

Pregunta 7: ¿Cómo afecta la contaminación al índice de calidad de vida y salud?

La contaminación es un factor crítico que puede impactar negativamente la calidad de vida y la salud en los países. Lo anterior, puede tiene incidencia en indicadores clave como la esperanza de vida, el poder adquisitivo (relacionado con la capacidad de acceder a servicios de salud) y la asequibilidad de la vivienda en ciudades con alta contaminación. Se espera que un mayor *Pollution Index* (más contaminación) se asocie con una reducción en el *Quality of Life Index* debido al deterioro ambiental y sus efectos en el bienestar. Un mejor *Health Care Index* podría compensar en cierta medida los efectos negativos de la contaminación, contribuyendo a un mejor *Quality of Life Index*.

Para responder esta incógnita se utilizarán las siguientes tablas y campos:

1- RefData_Health_care_index_by_countries_2020.csv (*Health Care Index*)

2- RefData_Quality_of_life_index_by_countries_2020.csv (*Quality of Life Index, Pollution Index*)

Para esta pregunta se eligió realizar la visualización por heatmap (figura 15) dado que es compacto y fácil de interpretar y resalta las relaciones clave en un solo gráfico. Permite identificar rápidamente patrones entre la contaminación, la salud y la calidad de vida. Como resultados esperados, se tendría una correlación negativa entre el índice de contaminación y los otros dos índices. Por otro lado es esperable encontrar una relación positiva entre el índice de

calidad de vida y el índice de salud, mostrando cómo un mejor sistema de salud puede mejorar la calidad de vida.

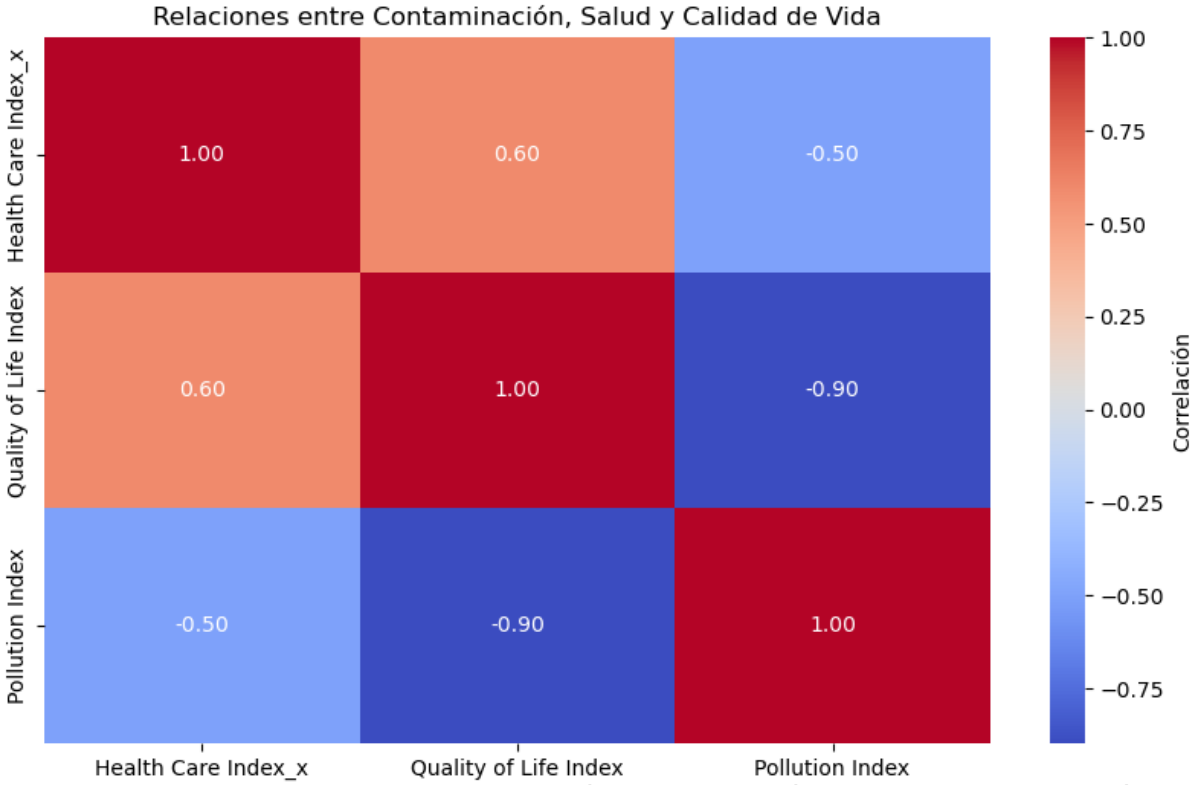

Figura 15. El gráfico muestra la correlación entre el Índice de Salud, el Índice de Calidad de Vida y el Índice de Contaminación (Health Care Index_x, Quality of Life Index y Pollution Index respectivamente). Los valores dentro de las celdas representan el coeficiente de correlación de Pearson. Una correlación positiva fuerte (cercana a 1) indica que ambas variables tienden a aumentar juntas, mientras que una correlación negativa fuerte (cercana a -1) indica que una variable aumenta mientras la otra disminuye. Los colores oscilan entre azul (correlaciones negativas) y rojo (correlaciones positivas), con el blanco indicando ausencia de correlación.

Análisis y conclusión de los resultados

El heatmap refleja las correlaciones entre tres índices principales: Índice de Contaminación, Índice de Salud e Índice de Calidad de Vida. A partir de la matriz de correlación, se pueden identificar patrones importantes en los datos:

Índice de Contaminación vs. Índice de Calidad de Vida:

- Correlación: -0.90 (negativa fuerte).

- Interpretación: A medida que aumenta la contaminación en un país, el índice de calidad de vida disminuye significativamente. Esto sugiere que la contaminación tiene un impacto directo y muy negativo en la calidad de vida.

Índice de Contaminación vs. Índice de Salud:

- Correlación: -0.50 (negativa moderada).

- Interpretación: Una mayor contaminación también afecta negativamente al sistema de salud. Esto puede estar relacionado con un aumento en las enfermedades respiratorias, cardiovasculares y otros problemas asociados a la contaminación.

Índice de Salud vs. Índice de Calidad de Vida:

- Correlación: +0.60 (positiva moderada).

- Interpretación: Un sistema de salud robusto contribuye de manera importante a mejorar la calidad de vida. Este resultado confirma que una mejor atención médica y acceso a servicios de salud elevan los estándares de vida en general.

La contaminación no solo disminuye la calidad de vida de los ciudadanos, sino que también tiene un impacto en los sistemas de salud, posiblemente al incrementar la carga sobre ellos. Países con bajos índices de contaminación suelen tener sistemas de salud más eficientes y una mejor calidad de vida, como se evidencia en la tabla (ej., Dinamarca, Finlandia, Suiza). Por otro lado, países con índices de contaminación elevados, como Egipto, Nigeria o Bangladesh, tienden a tener menores índices de calidad de vida y sistemas de salud más débiles.

A modo de conclusión, a partir de los datos recolectados por país se puede afirmar que la contaminación es un factor crítico que afecta negativamente tanto la salud pública como la calidad de vida. Las políticas para reducir la contaminación no solo tienen beneficios ambientales, sino que también mejoran significativamente las condiciones de vida y alivian la presión sobre los sistemas de salud. Países que logran mantener bajos niveles de contaminación pueden mejorar ambos índices de manera simultánea, como se observa en los datos de países como Dinamarca y Finlandia.

Pregunta 8: ¿Qué países destacan por su equilibrio entre calidad de vida, salud, seguridad y vivienda?

Es importante identificar los países que logran un equilibrio entre calidad de vida, salud, seguridad y vivienda, ya que estos aspectos son indicadores clave del bienestar social y económico. Conocer estos países puede ser útil para decisiones relacionadas con políticas públicas, migración, estudios de sostenibilidad urbana, o incluso para personas que buscan un lugar para vivir con alta calidad de vida. Este tipo de análisis permite destacar naciones que han implementado estrategias exitosas para optimizar estos factores esenciales.

Para responder esta incógnita se utilizarán las siguientes tablas y campos:

1- RefData_Health_care_index_by_countries_2020.csv (*Health Care Index*)

2- RefData_Crime_index_by_countries_2020.csv (*Safety Index*)

3- RefData_Pruperties_price_index_by_countries_2020.csv (*Affordability Index*)

4- RefData_Quality_of_life_index_by_countries_2020.csv (*Quality of Life Index*)

Para esta pregunta, dado que se trata de evaluar varios índices simultáneamente, una visualización del tipo radar chart (gráfico de telaraña) me pareció buena idea. Este gráfico permite comparar múltiples dimensiones para cada país, mostrando de forma intuitiva qué países logran un equilibrio en los factores seleccionados. Sin embargo, por el tipo de gráfico no pueden verse muchos países a la vez ya que complejizaría la imagen. Por esto, luego de generar la tabla con los índices que se usarán, se establecerán criterios de equilibrio para filtrar

la lista y así quedarse con un grupo reducido de países para graficar (figura 16). Para obtener la tabla principal con los distintos índices que se usarán, se hará un merge utilizando el df llamado *health_care* como base (ya que de todas las tablas es la que menos países tiene), de esta forma nos aseguramos de que cada entrada posea valores para los índices. Luego de realizado este merge se filtrarán los cuatro campos que se usarán, cada campo un índice. Esta tabla final será utilizada para establecer un criterio de equilibrio por el cual filtrar los países que destaquen por el mismo. Finalmente, la lista de países así obtenidos será la que se use para el gráfico de telaraña.

Para definir los criterios de equilibrio y seleccionar los países destacados, se considerará que los valores de cada índice deben ser superiores o iguales a la mediana de su respectiva métrica, ya que la mediana permitirá identificar países con indicadores relativamente buenos en cada categoría.

A partir de la tabla con la totalidad de los países y los índices se calcularon sus medianas (50%) siendo éstas:

- Quality of Life Index: 133.07
- Health Care Index: 66.08
- Safety Index: 60.33
- Affordability Index: 1.02

Dadas las métricas calculadas se fijaron los criterios de equilibrio:

- Un país debe tener un Quality of Life Index ≥ 133.07
- Un país debe tener un Health Care Index ≥ 66.08
- Un país debe tener un Safety Index ≥ 60.33
- Un país debe tener un Affordability Index ≥ 1.02

Estos criterios serán los usados para filtrar a los países que tienen un desempeño balanceado en calidad de vida, salud, seguridad y asequibilidad.

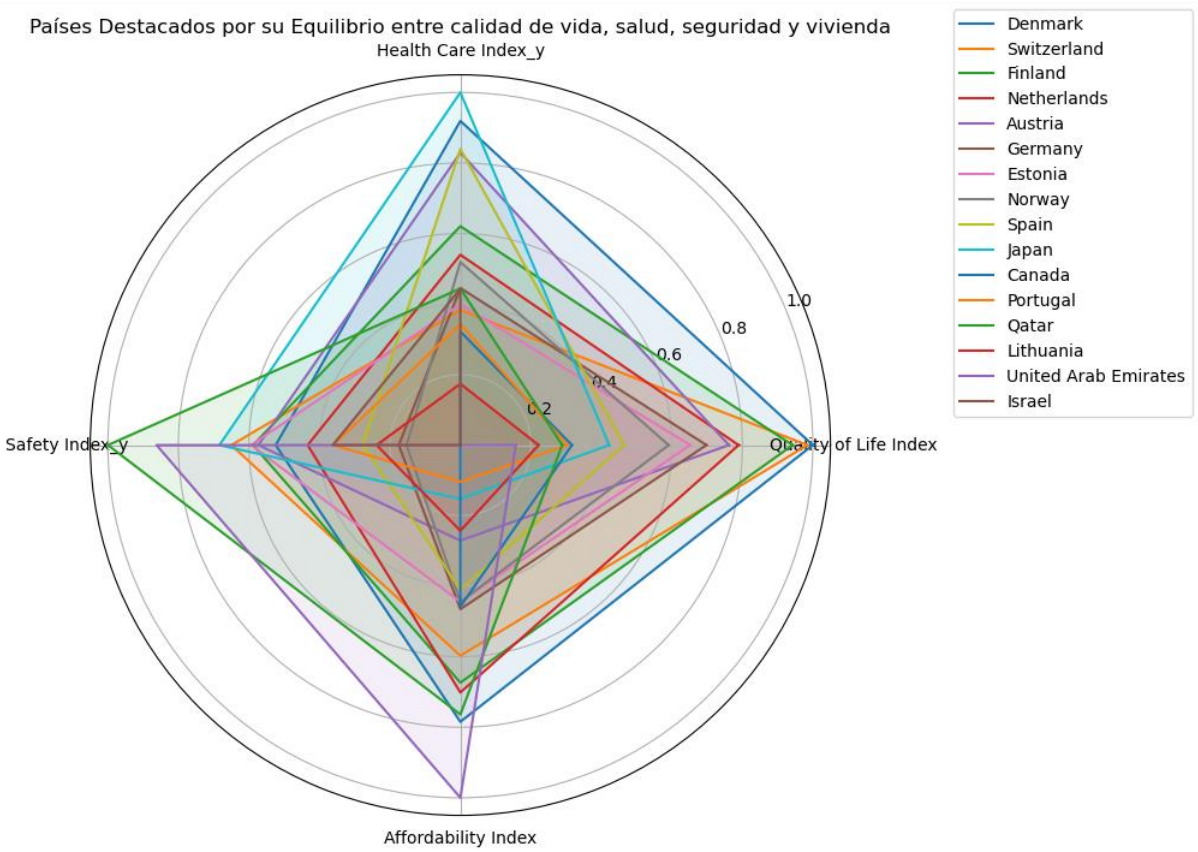

Figura 16. Gráfico de radar que compara países destacados en cuatro índices clave: Calidad de Vida, Índice de Salud, Seguridad y Asequibilidad. Los valores están normalizados (escala de 0 a 1) para permitir comparaciones directas. La figura destaca cómo los países equilibran diferentes aspectos, reflejando similitudes y diferencias en el desempeño global en estas dimensiones.

Análisis y conclusión de los resultados

Muchas líneas de los países están bastante cercanas entre sí, indicando que los valores de estos índices no son significativamente diferentes entre los países representados. Esto puede reflejar que los países seleccionados pertenecen a un grupo homogéneo (países con un alto desarrollo humano o economías avanzadas).

Algunos países, como Qatar y Emiratos Árabes Unidos, tienen valores más altos en el índice de asequibilidad ("Affordability Index") pero caen en otros índices como la Seguridad o la Calidad de Vida. Países como Japón y Dinamarca tienen valores equilibrados en la mayoría de los índices, con diferencias mínimas entre sus dimensiones.

Quality of Life Index: Tiende a ser alto y consistente entre la mayoría de los países.

Safety Index: Hay mayor variabilidad; algunos países, como Japón, destacan con valores relativamente altos, mientras que otros, como Canadá, están en la parte baja.

Health Care Index_y: Es uno de los índices más consistentes, con pocos países destacándose claramente.
Affordability Index: Tiene la mayor variabilidad, lo que podría reflejar diferencias significativas en el costo de vida o el acceso a bienes y servicios básicos.

Las disparidades en "*Affordability Index*" podrían estar influenciadas por factores económicos locales, como subsidios, costos de vida o ingresos promedio en cada país.

Países con un bajo índice de "*Safety Index*" podrían priorizar políticas de seguridad.

Países con bajo "*Affordability Index*" deben analizar cómo mejorar el acceso económico a recursos básicos para su población.

A modo de conclusión se podría decir que la mayoría de los países seleccionados muestran un buen balance en los índices considerados, lo que refleja que estos países ofrecen altos estándares de vida en general cumpliendo así con la premisa de la pregunta.

Parte 2- Análisis exploratorio de los datos vía Spark

Esta parte fue realizada utilizando Google Colab. Para ello se generó la estructura de carpetas que se muestra en la figura 17. Dicha estructura presenta en el primer nivel una carpeta RawData, una carpeta SparkRefinedData y el archivo "*Parte_2_análisisSpark.ipynb*" donde se encuentra el código utilizado durante el análisis exploratorio. Este archivo además de realizar el análisis exploratorio y obtener las tablas refinadas se encarga de la persistencia de las mismas guardándolas en carpetas dentro del archivo SparkRefinedData.

Figura 17. Estructura de carpetas del proyecto, mostrando datos sin procesar (RawData) y datos refinados mediante Spark (SparkRefinedData), organizados por índices y características específicas por país.

A continuación, se mostrarán los primeros bloques del archivo de Google Colab "*Parte_2_análisisSpark.ipynb*" donde aparece la configuración del entorno y carga de datos en Spark utilizando PySpark (figura 18).

Preparación del ambiente

```
# Instalación de Java
!apt-get install openjdk-8-jdk-headless -qq > /dev/null

# Descargar e instalar Spark
!wget -q https://archive.apache.org/dist/spark/spark-3.1.2/spark-3.1.2-bin-hadoop2.7.tgz
!tar xf spark-3.1.2-bin-hadoop2.7.tgz

# Instalación de findspark
!pip install -q findspark

# Configuración de las variables de entorno
import os
os.environ["JAVA_HOME"] = "/usr/lib/jvm/java-8-openjdk-amd64"
os.environ["SPARK_HOME"] = "/content/spark-3.1.2-bin-hadoop2.7"

# Inicialización de findspark
import findspark
findspark.init()
```

```
from pyspark.sql.functions import *
from google.colab import drive
drive.mount('/content/drive')

# Crear una sesión de Spark
from pyspark.sql import SparkSession
spark = SparkSession.builder.master("local[*]").getOrCreate()
```

```
Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount("/content/drive", force_remount=True).
```

```
# Rutas en Drive
RawInput = '/content/drive/MyDrive/BigData/RawData'
RefinedOutput = '/content/drive/MyDrive/BigData/SparkRefinedData'
```

Carga de tablas

```
# Ruta a tablas
costLivingIndex_access = RawInput + '/Cost of living index by country 2020.csv'
countriesAgeStructure_access = RawInput + '/Coutries age structure.csv'
crimeIndex_access = RawInput + '/Crime index by countries 2020.csv'
healthCareIndex_access = RawInput + '/Health care index by countries 2020.csv'
propertiesPriceIndex_access= RawInput + '/Properties price index by countries 2020.csv'
populationDensityIndex_access= RawInput + '/Pupulation density by countries.csv'
qualityLifeIndex_access = RawInput + '/Quality of life index by countries 2020.csv'



# Cargar la tabla de datos
df_Cost_of_living_index_by_country_2020 = spark.read.csv(costLivingIndex_access, header=True, inferSchema=True)
df_Coutries_age_structure= spark.read.csv(countriesAgeStructure_access, header=True, inferSchema=True)
df_Crime_index_by_countries_2020=spark.read.csv(crimeIndex_access, header=True, inferSchema=True)
df_Health_care_index_by_countries_2020=spark.read.csv(healthCareIndex_access, header=True, inferSchema=True)
df_Properties_price_index_by_countries_2020=spark.read.csv(propertiesPriceIndex_access, header=True, inferSchema=True)
df_Pupulation_density_by_countries=spark.read.csv(populationDensityIndex_access, header=True, inferSchema=True)
df_Quality_of_life_index_by_countries_2020=spark.read.csv(qualityLifeIndex_access, header=True, inferSchema=True)
```

Figura 18. Configuración del entorno y carga de datos en Spark utilizando PySpark en Google Colab. La imagen muestra pasos como la instalación de Java, configuración de variables de entorno, inicialización de Spark y carga de archivos CSV desde Google Drive.

Para el código, resto del contenido y análisis realizado, referirse al archivo *Parte_2_análisisSpark.ipynb* que se encuentra en el archivo comprimido que contiene la entrega de este obligatorio.

Parte 3- Dashboard

Justificación de Tableau como herramienta de visualización

Como herramienta de visualización se utilizó Tableau (*Tableau Public 2024.3.0*) ya que posee ciertas características ideales para este proyecto. Es capaz de conectarse a una amplia variedad de fuentes de datos. Brinda una gran cantidad de conectores nativos que permiten a quien los use conectarse directamente a bases de datos, archivos y servicios en la nube sin necesidad de configuraciones complicadas. Además, permite la creación de gráficos básicos de manera intuitiva y en caso de ser necesario también tiene otros complejos capaces de proporcionar insights más profundos. Otra característica importante es que Tableau permite combinar múltiples visualizaciones, provenientes de una misma base de datos o de diferentes, en un solo dashboard, proporcionando una visión más completa y cohesiva de los datos. Por último, los datos provenientes o no de diferentes fuentes, pueden vincularse para dar interactividad a la visualización. Lo anterior hace las visualizaciones en Tableau no sean estáticas lo que significa que se puede interactuar con los datos para explorar y descubrir insights relevantes.

Preguntas seleccionadas

Para explorar Tableau generando las visualizaciones se seleccionaron cuatro de las ocho preguntas tratadas en el presente informe (sección 3, *Parte 1b- Análisis de las preguntas planteadas*). El criterio para elegir las preguntas consistió en seleccionar aquellas que se considera bridan más información y que por lo tanto tuvieran un mayor beneficio o provecho analizar su visualización por Tableau. Las preguntas así seleccionadas fueron las preguntas:

- **Pregunta 3:** ¿Qué países ofrecen el mejor balance entre costo de vida y calidad de vida?
- **Pregunta 4:** ¿Qué impacto tienen los índices de seguridad y crimen en el costo de vida?
- **Pregunta 5:** ¿Cómo varía la calidad del sistema de salud según la densidad poblacional?
- **Pregunta 6:** ¿Cuáles son los países más accesibles para comprar vivienda considerando los índices económicos?

Carga de tablas y configuración de la relación entre estas en Tableau

Para cargar los datos se descargaron las tablas en formato csv obtenidas a partir del análisis de cada pregunta, parte 1b. Estas tablas contienen solamente los campos utilizados para las visualizaciones realizadas por jupyter notebook de dicha parte. Dado lo anterior, estas tablas son solo una parte de las tablas refinadas obtenidas a partir de las tablas crudas iniciales, contienen los campos necesarios para las visualizaciones. Para cargar las tablas se subieron estos archivos como texto para que reconociera el formato de los mismos.

Una vez cargadas las tablas en Tableau, se prosiguió a configurar la relación entre éstas para permitir la interactividad entre las distintas visualizaciones. El campo por el cual se realizó la vinculación entre tablas fue Country. Esto fue así dado que Country es la clave primaria en

todas las tablas lo que permite trabajar con dicho campo generando filtros que actúen sobre las diferentes visualizaciones a la vez, mejorando así la interactividad.

Visualizaciones y filtros

Luego de la etapa de cargado de tablas y vinculación por campo entre estas, se prosiguió a replicar las visualizaciones obtenidas en la parte 1b. Se mantuvieron los features logrados en dicha sección y se agregaron nuevos, como ser leyendas y etiquetas pop-up que aparecen al cliquear sobre alguno de los países representados en dichas visualizaciones, así como un filtro de valores múltiples desplegable por país. Posteriormente se generó un dashboard con estos gráficos (figura 19). Este dashboard puede encontrarse en el siguiente link: https://public.tableau.com/views/Parte3_tableau/Dashboard1?:language=es-ES&publish=yes&:sid=&:redirect=auth&:display_count=n&:origin=viz_share_link

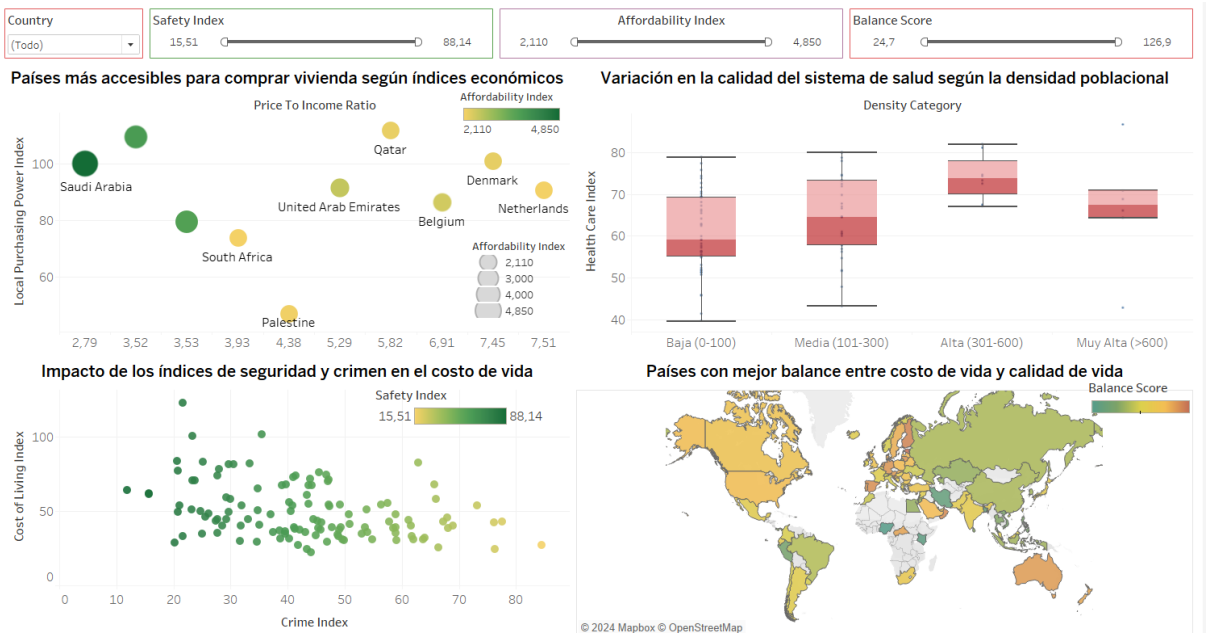

Figura 19. Dashboard interactivo en Tableau que visualiza indicadores clave de accesibilidad económica, calidad de vida y seguridad a nivel global**. Gráfico superior izquierdo:** Evalúa la accesibilidad a la vivienda considerando la relación Precio-Ingreso (eje X), el índice de poder adquisitivo (eje Y) y el índice de asequibilidad (tamaño y color de las burbujas). **Gráfico superior derecho:** Boxplot que compara la calidad del sistema de salud según categorías de densidad poblacional. **Gráfico inferior izquierdo:** Dispersión que relaciona el impacto del índice de seguridad y crimen (eje X) con el costo de vida (eje Y), resaltando patrones regionales mediante el índice de seguridad**. Mapa inferior derecho:** Representación geográfica del balance entre costo de vida y calidad de vida, utilizando un índice de balance como escala de color, aquellos países para los cuales no hay datos se visualizan en gris. El filtro de país (en la parte superior) permite ajustar dinámicamente la visualización a una o varias naciones seleccionadas. Los filtros específicos -Safety Index, Affordability Index y Balance Score- permiten ajustar individualmente los datos que impactan, respectivamente, el mapa geográfico, el gráfico de burbujas y el gráfico de dispersión.

Las gráficas del dashboard presentan interactividad con el usuario como también vinculación entre ellas a partir de filtros. Estos filtros pueden aplicarse tanto a partir de una lista de valores múltiples desplegables de países como también al cliquear sobre alguno de los puntos en las gráficas. El filtro de países en la lista de valores múltiples desplegable aplica a todas las gráficas del dashboard.

De todas las figuras en el dashboard, la única que no se configuró como filtro (al cliquear en uno de sus datos no filtra el resto de las gráficas) es la del histograma ya que dado

la información que muestra no tiene mucha relevancia usarla como filtro. Por otro lado, los tres gráficos restantes fueron configurados para funcionar como filtro sobre todas las visualizaciones en el dashboard. El filtro puede realizarse tanto por cliquear en uno múltiples datos de una visualización como por selección de una sección del gráfico. Un ejemplo de esto último se observa en la figura 20 donde al seleccionar una sección del gráfico de dispersión abajo a la izquierda, filtra el resto de las figuras en base a dichos datos seleccionados. Esto es útil ya que, de encontrarse patrones o agrupación de países aislados, seleccionado dicho grupo automáticamente se modificaría el resto de los gráficos para corresponder con el grupo de países seleccionados. Lo anterior, permitirá concentrar el análisis y visualización de datos en dicho grupo, lo que podría resultar en mejores o nuevos insights.

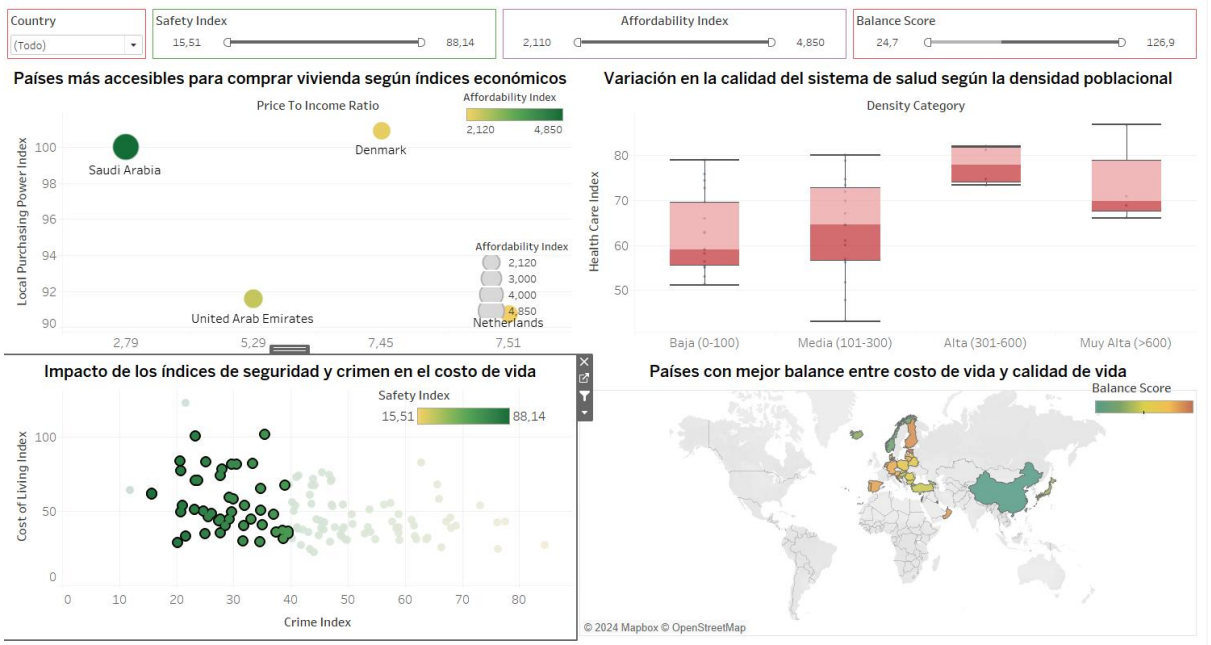

Figura 20. Dashboard interactivo en Tableau que visualiza indicadores clave de accesibilidad económica, calidad de vida y seguridad a nivel global luego de aplicar filtros. En esta figura se muestra la vinculación entre gráficos del dashboard, específicamente cómo la selección de un grupo de puntos en el gráfico de dispersión (gráfico inferior izquierdo) cambia la visualización gráfica en las tres gráficas restantes.

Por otro lado, el gráfico de países más accesibles para comprar vivienda según índices económicos (Figura 19, gráfico superior izquierdo) es el resultado de graficar los 10 países más accesibles para dicho enunciado. Por lo que al cliquear en alguno de estos países podremos observar en el resto de las visualizaciones los índices para el o los países seleccionados, esto permitirá tener más información sobre estos (Figura 21).

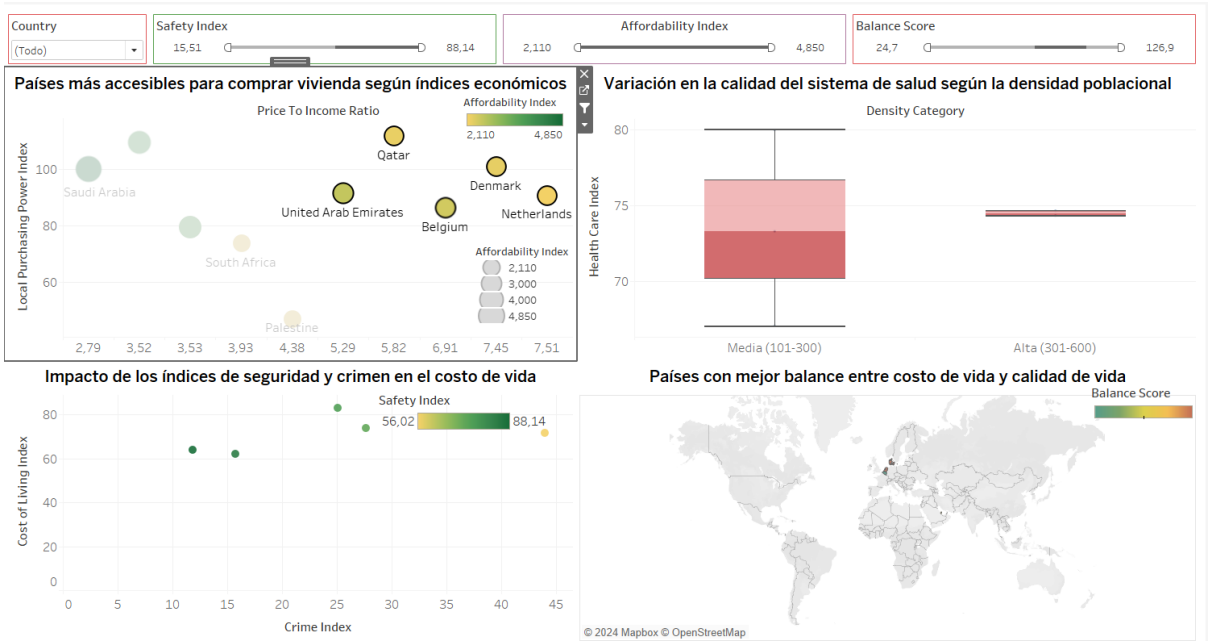

Figura 21. Dashboard interactivo en Tableau que visualiza indicadores clave de accesibilidad económica, calidad de vida y seguridad a nivel global luego de aplicar filtros. En esta figura se muestra la vinculación entre gráficos del dashboard, específicamente cómo la selección de un grupo de algunas de las burbujas en el gráfico superior izquierdo condiciona el resto de las visualizaciones.

Por último, lo mismo es aplicable si seleccionamos países en el mapa geográfico. El resto de las visualizaciones se modificarán acorde a los países seleccionados. (Figura 22).

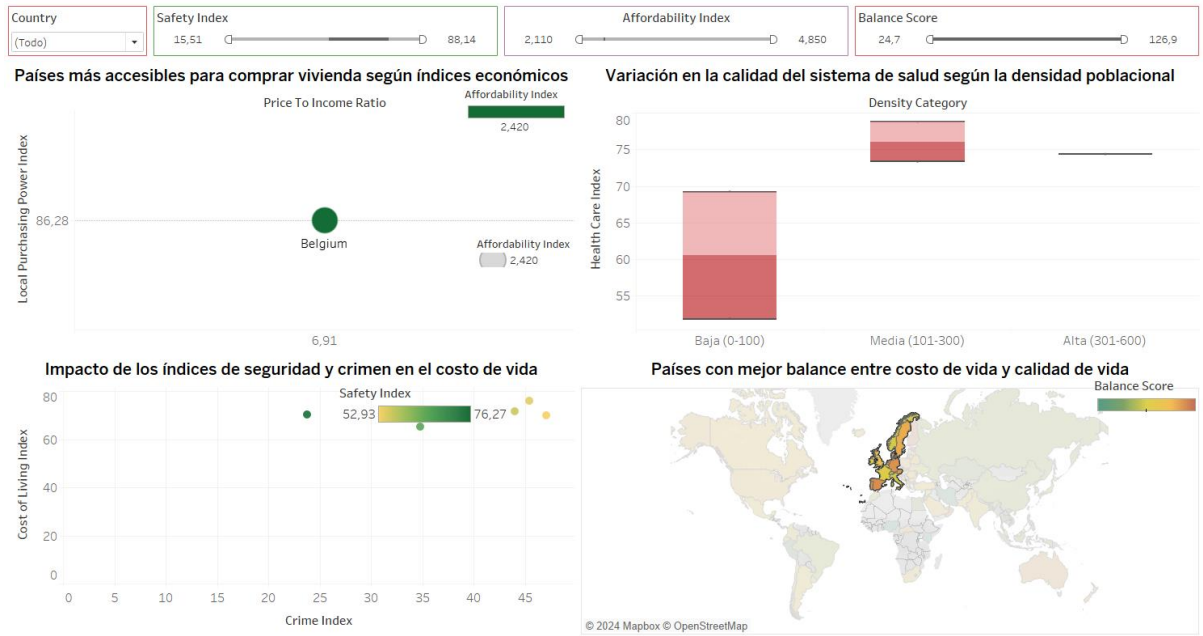

Figura 22. Dashboard interactivo en Tableau que visualiza indicadores clave de accesibilidad económica, calidad de vida y seguridad a nivel global luego de aplicar filtros. En esta figura se muestra la vinculación entre gráficos del dashboard, específicamente cómo la selección de países en el mapa geográfica en la figura inferior derecha condiciona el resto de las visualizaciones.

El dashboard cuenta con filtros generales e individuales, estratégicamente ubicados en la parte superior para facilitar su acceso y uso (Figura 23). Estos filtros cumplen una función esencial: permiten segmentar y enfocar los datos, ofreciendo la posibilidad de profundizar en ellos mediante interacciones que afectan el resto de las visualizaciones.

Figura 23. La imagen muestra los filtros utilizados en el dashboard para segmentar y analizar los datos presentados. El primer filtro a la izquierda es el filtro general Country, que afecta a todas las visualizaciones del dashboard. Los filtros específicos -Safety Index, Affordability Index y Balance Score- permiten ajustar individualmente los datos que impactan, respectivamente, el mapa geográfico, el gráfico de burbujas y el gráfico de dispersión. Estos filtros están diseñados para facilitar el análisis interactivo, permitiendo seleccionar rangos de índices y observar cómo estas selecciones influyen en el resto de las visualizaciones relacionadas.

El filtro general, que abarca todas las figuras del dashboard, corresponde a Country. Por otro lado, los filtros individuales -Safety Index, Affordability Index y Balance Score- están diseñados para actuar exclusivamente sobre el mapa geográfico, el gráfico de burbujas y el gráfico de dispersión, respectivamente.

El uso de estos filtros es intuitivo: primero, se aplica el filtro individual para acotar los datos de interés en la gráfica específica, seleccionando un rango de índices según sea necesario. Posteriormente, al interactuar con los datos filtrados en esa gráfica (por ejemplo, seleccionando un punto o área de interés), los efectos de la selección se reflejan automáticamente en las visualizaciones relacionadas, facilitando un análisis más profundo y detallado de las conexiones entre los diferentes indicadores.

Parte 4- Modelado de tablas en Hive

El modelo de datos elegido es el normalizado, la fundamentación de la decisión se encuentra en la sección "Elección del modelo de datos" dentro del apartado "*Parte 1a-Análisis exploratorio de los datos vía Pandas*" del presente informe.

En cuanto a la decisión sobre el tipo de tablas a utilizar, internas o externas, se optó por emplear tablas externas. Esta elección se fundamenta en las ventajas que ofrecen en términos de seguridad e integridad de los datos, dado que, al eliminar una tabla de este tipo, únicamente se elimina la estructura de la tabla, mientras que los datos asociados permanecen intactos. Además, considerando que estamos trabajando con entradas que representan países (una entrada por país), y que estas columnas actúan como claves primarias (sin valores duplicados en este campo), no se prevén problemas relacionados con la cantidad de datos o la escalabilidad. Esto se debe a que los campos de las tablas ya están definidos y no existe más de un registro por país, ya que no se ha planificado un seguimiento temporal que pudiera incrementar significativamente el volumen de las tablas actuales, la mayoría de las tablas presentan los datos para un año específico, ejemplo *Cost_of_living_index_by_country_2020.csv*.

Se decidió generar ocho tablas externas, las cuales corresponden a réplicas de las tablas utilizadas para las visualizaciones y el análisis de la sección Parte 1b - Análisis de las preguntas planteadas de este informe. Estas tablas representan los datos fundamentales que sustentan las respuestas y gráficos desarrollados.

La elección de tablas externas se fundamenta en su capacidad para mantener la integridad y seguridad de los datos originales. A diferencia de las tablas internas, las tablas externas permiten que los datos persistan en el sistema de archivos incluso si la tabla en Hive es eliminada, lo que resulta especialmente útil para preservar la información fuente.

En total, las tablas que se generarán son las siguientes:

costoVidaPoderAdquisitvoDensidadPoblacional: Contiene índices como costo de vida, renta, y poder adquisitivo local.

distribucionEdadesCalidadVida: Incluye rangos de calidad de vida y datos demográficos por edad.

paisesMejorBalanceCostoCalidadVida: Combina índices de calidad de vida, costo de vida, y coordenadas geográficas.

impactoSeguridadCrimenCostoVida: Relaciona índices de seguridad, crimen y costo de vida.

sistemaSaludDensidadPoblacion: Explora densidad poblacional y calidad del sistema de salud.

paisesAccesiblesCompraVivienda: Analiza la accesibilidad económica mediante índices de asequibilidad y renta.

incidenciaContaminacionSalud: Incluye índices de calidad de vida, salud y contaminación.

paisesDestacadosEqulibrioEntreIndices: Relaciona índices de calidad de vida, salud, seguridad y asequibilidad.

Para implementar estas tablas, es necesario crear previamente un directorio en HDFS donde se almacenará la información. Este directorio, que llamaremos HiveP4, puede generarse con el siguiente comando:

HDFS DFS -MKDIR /HiveP4

El código específico para crear las tablas externas en Hive, apuntando al directorio previamente mencionado, está documentado en el archivo *Parte 4_Hive.ipynb*. Este archivo contiene todos los scripts necesarios para configurar y gestionar las tablas en el entorno Hive.

4. Conclusiones Generales

El desarrollo del presente trabajo ha permitido explorar y comprender de manera integral diversas técnicas y herramientas de análisis de datos, aplicadas a un contexto de Big Data. A continuación, se destacan las principales lecciones aprendidas en los distintos temas tratados:

Análisis Exploratorio de Datos

Se reforzó la importancia de entender las características de los datos antes de utilizarlos, incluyendo la identificación de valores nulos, duplicados, y transformaciones necesarias para preparar los conjuntos de datos. Esto subraya el papel crucial de la limpieza y normalización de los datos para asegurar resultados fiables.

Herramientas para Big Data

La comparación entre pandas y Spark permitió apreciar las ventajas y limitaciones de cada herramienta. Mientras pandas es ágil para pequeños volúmenes de datos, Spark demostró ser más eficiente en la manipulación de grandes volúmenes, mostrando la relevancia de seleccionar las herramientas adecuadas según el contexto.

Visualización de Datos

Tableau y otras herramientas de visualización ofrecieron un enfoque claro para comunicar hallazgos complejos de manera intuitiva. Esto resaltó cómo las visualizaciones no solo facilitan la interpretación de los datos, sino también la toma de decisiones basadas en evidencia.

Modelado de Datos

La decisión de utilizar un modelo normalizado mostró ser efectiva para manejar el tamaño reducido del conjunto de datos, garantizando integridad y flexibilidad en el análisis. Este proceso subrayó la importancia de seleccionar un esquema de datos adecuado para responder preguntas específicas del negocio.

Preguntas y Respuestas Analíticas

Cada una de las preguntas planteadas permitió entender cómo diversos factores interactúan en contextos socioeconómicos y demográficos. Desde la relación entre densidad poblacional y calidad del sistema de salud hasta el impacto de la contaminación en la calidad de vida, estos análisis ejemplificaron cómo Big Data puede desentrañar patrones relevantes para la toma de decisiones.

Impacto Global del Big Data

Finalmente, este proyecto demostró cómo el manejo adecuado de Big Data puede proporcionar perspectivas valiosas para abordar desafíos globales, como mejorar la accesibilidad a la vivienda, optimizar la calidad de vida o diseñar políticas públicas más efectivas.

Dicho lo anterior, y a modo de resumir el trabajo realizado, puede afirmarse que este no solo permitió aplicar las herramientas aprendidas en el curso, sino que también resaltó la importancia de un enfoque estructurado y metódico en el manejo y análisis de datos para extraer valor de grandes volúmenes de información.