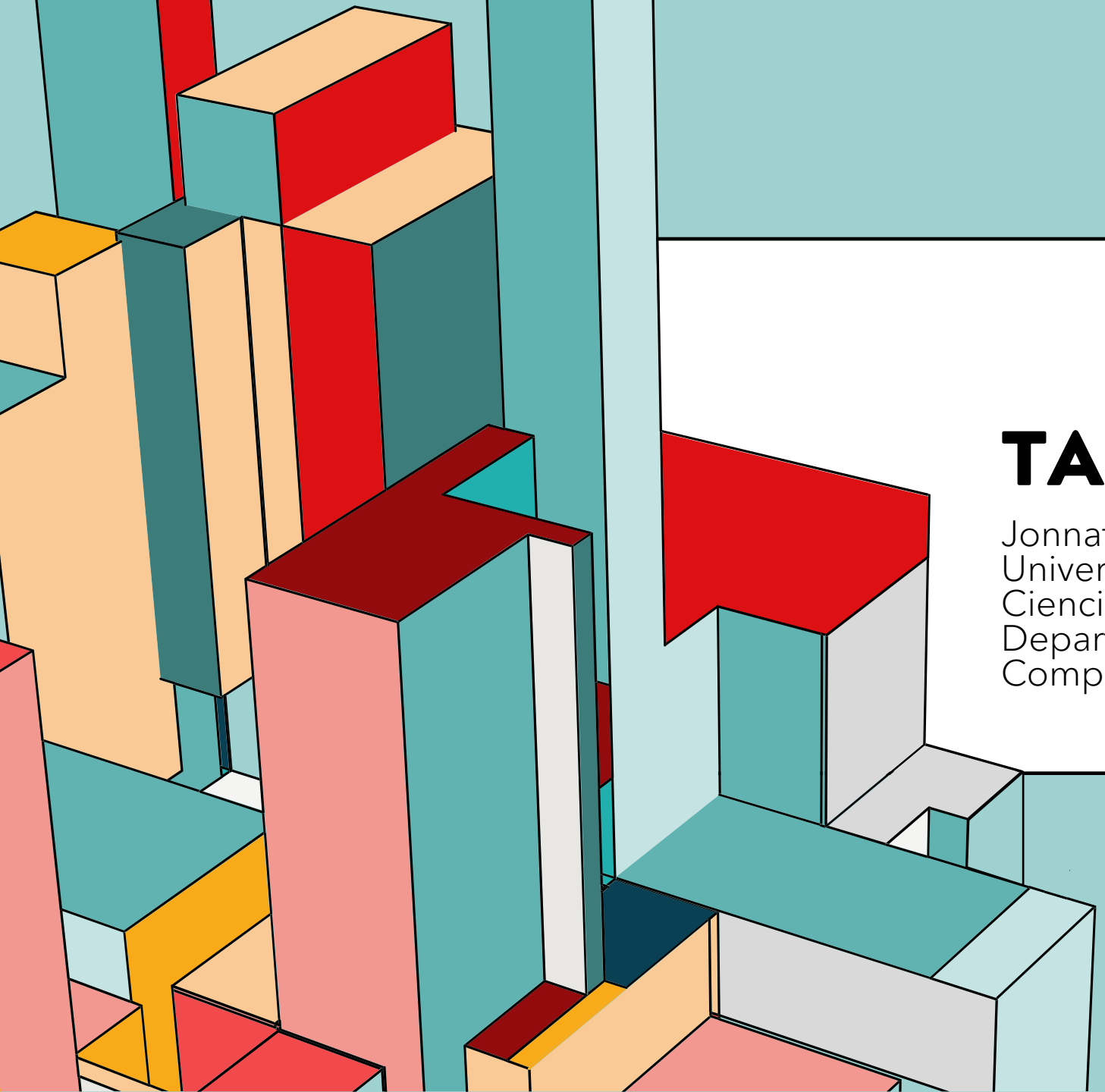


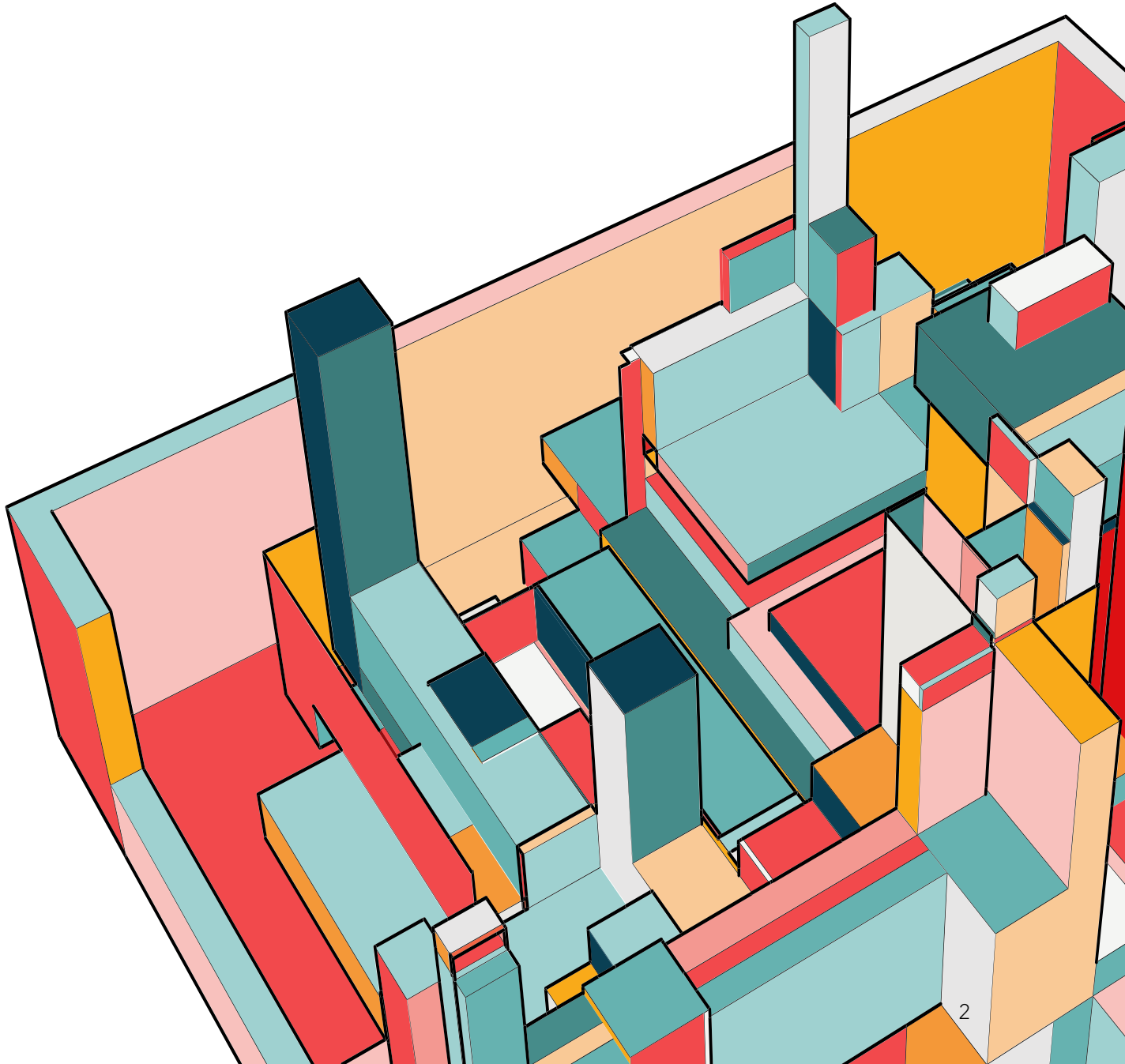
TALLER 2

Jonnatan Triana
Universidad de los Andes
Ciencia de Datos Aplicada
Departamento de Ingeniería de Sistemas y Computación

Universidad de los Andes
Colombia

¿QUÉ HAREMOS?

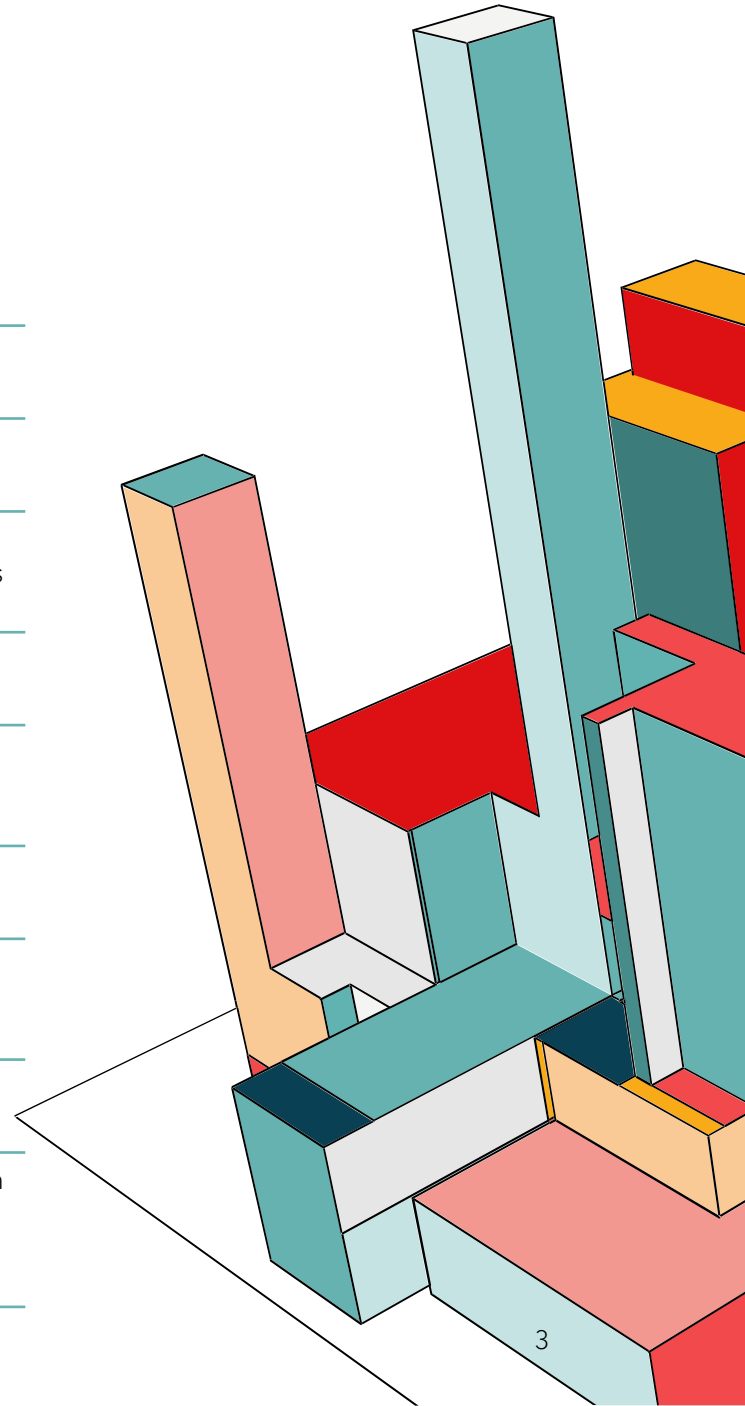
Con base en datos del Banco Mundial recomendaremos un set de políticas que puedan ser efectivas basadas en el análisis exploratorio de datos y la aplicación de un algoritmo de Machine Learning que encuentre relaciones entre diferentes variables de comportamiento para 166 países.



TIPOLOGÍA DE LOS DATOS

Se cuenta con información de diferentes años para 166 países acerca de indicadores macroeconómicos globales.

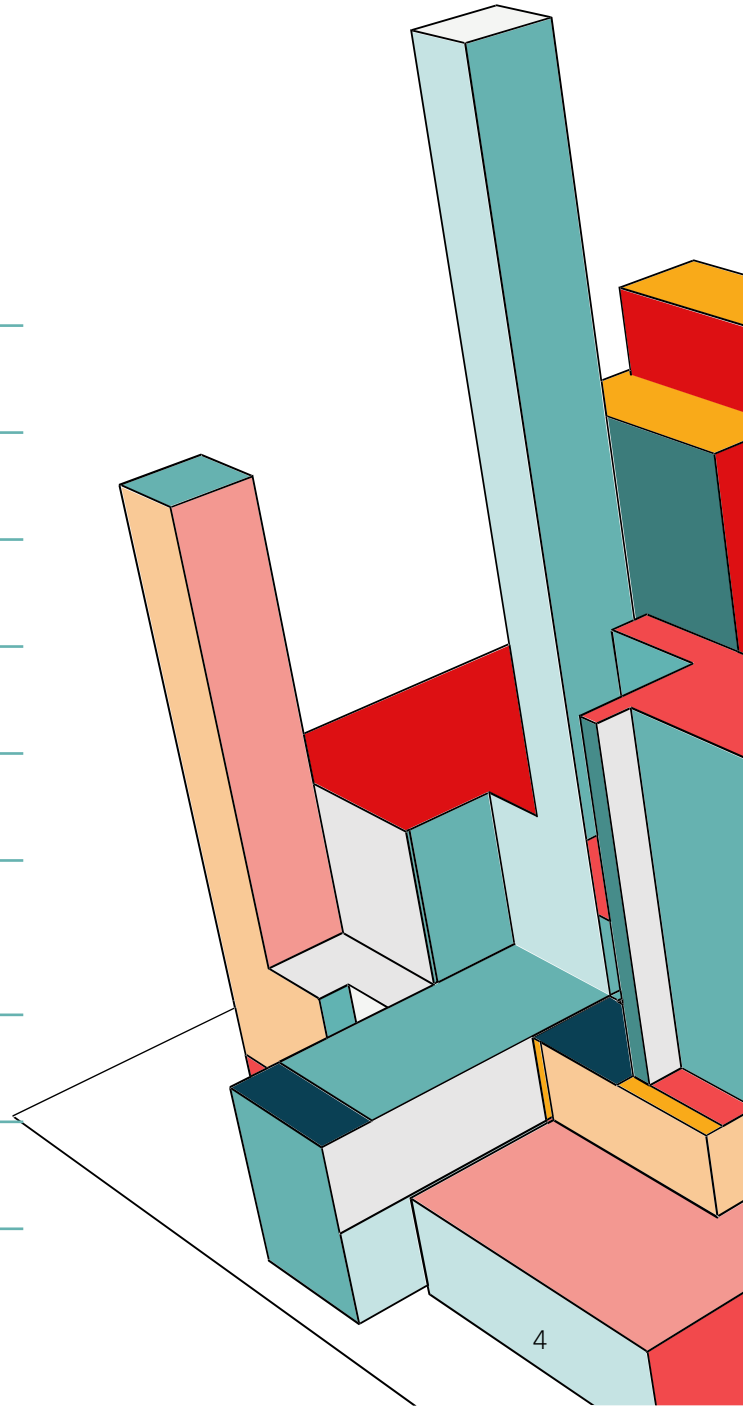
Variable	Identificador	Descripción
Ingreso per Cápita	incomeperperson	Producto Interno Bruto per cápita de 2010 en dólares constantes de 2000. Se ha tenido en cuenta la inflación, pero no las diferencias en el costo de vida entre países.
Consumo de Alcohol	alcconsumption	Consumo de alcohol per cápita en 2008 por adulto (15 años o más), en litros. Consumo promedio de alcohol registrado y estimado, consumo per cápita de adultos (15 años o más) en litros de alcohol puro.
Personal Militar	armedforcesrate	Personal de las fuerzas armadas (% de la fuerza laboral total)
Prevalencia de Cáncer de Seno	breastcancerper100TH	Nuevos casos de cáncer de mama en 2002 por 100.000 mujeres. Número de nuevos casos de cáncer de mama en 100.000 mujeres residentes durante un año determinado.
Emisiones de CO2	co2emissions	Emisión acumulada de CO2 en 2006 (millones de toneladas métricas), Cantidad total de emisiones de CO2 en millones de toneladas métricas desde 1751.
Empleo Femenino	femaleemployrate	Mujeres empleadas en 2007 de 15 años o más (% de la población). Porcentaje de la población femenina, mayor de 15 años, que ha estado empleada durante el año dado.
Tasa de Empleo	employrate	Empleados totales en 2007 de 15 años o más (% de la población). Porcentaje de la población total, mayor de 15 años, que ha estado empleada durante el año dado.
Tasa de Urbanidad	urbanrate	Población urbana (% del total) La población urbana se refiere a las personas que viven en áreas urbanas según lo definido por las oficinas nacionales de estadística (calculado utilizando estimaciones de población del Banco Mundial y tasas urbanas de las Perspectivas de urbanización mundial de las Naciones Unidas).

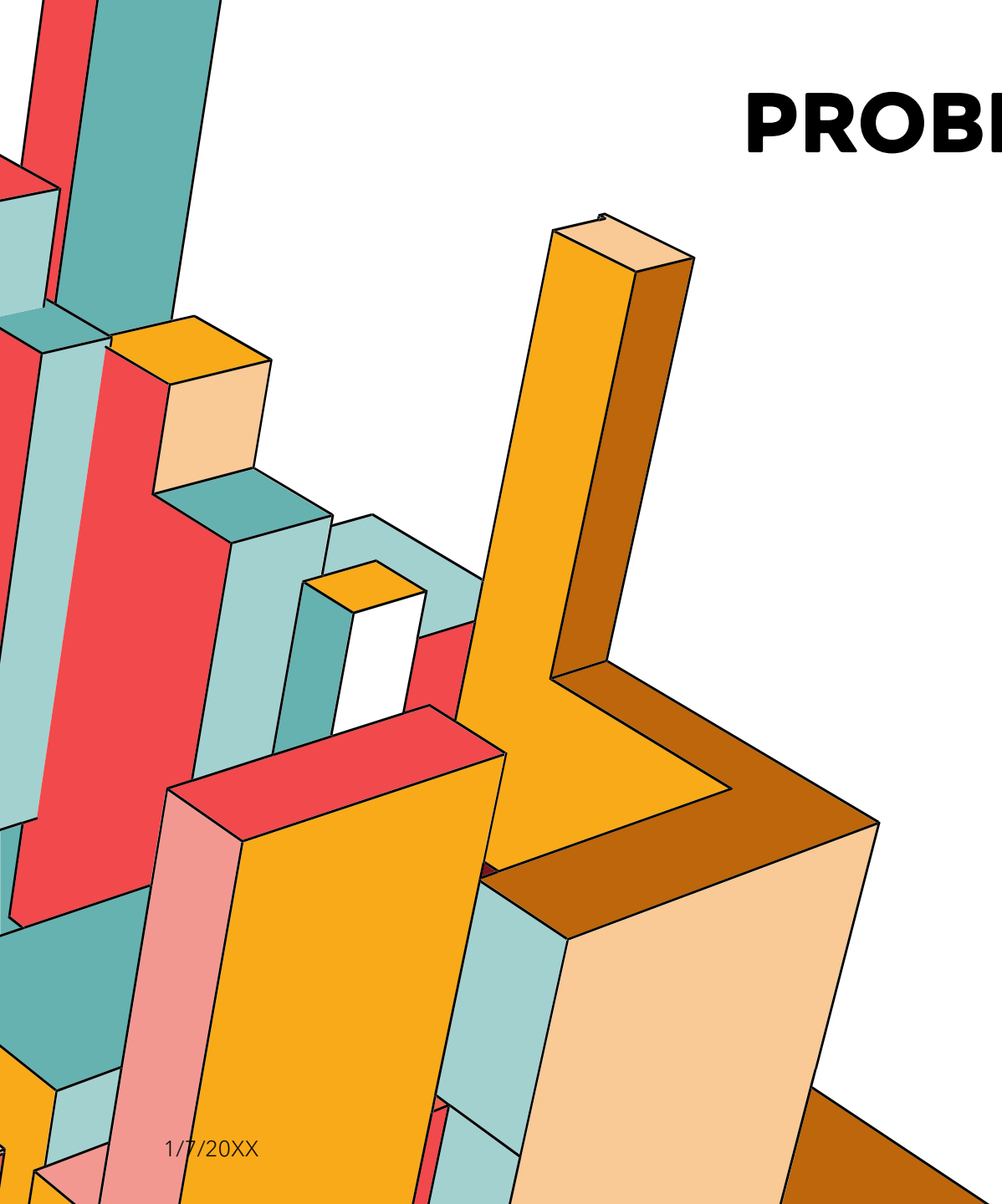


TIPOLOGÍA DE LOS DATOS

Se cuenta con información de diferentes años para 166 países acerca de indicadores macroeconómicos globales.

Variable	Identificador	Descripción
Prevalencia de VIH	HIVrate	Prevalencia estimada del VIH en 2009 % - (edades 15-49). Número estimado de personas que viven con el VIH por cada 100 habitantes del grupo de edad 15-49.
Uso de Internet	Internetuserate	Usuarios de Internet en 2010 (por 100 personas). Los usuarios de Internet son personas con acceso a la red mundial.
Esperanza de Vida	lifeexpectancy	Esperanza de vida al nacer en 2011 (años). El número promedio de años que viviría un recién nacido si los patrones de mortalidad actuales se mantuvieran igual.
Combustible per cápita	oilperperson	Consumo de petróleo per cápita en 2010 (toneladas por año y persona)
Tipo de Democracia	polityscore	Puntaje de democracia en 2009 (Polity). Puntaje de política general del conjunto de datos Polity IV, calculado restando un puntaje de autocracia de un puntaje de democracia. La medida de resumen de la naturaleza democrática y libre de un país. -10 es el valor más bajo, 10 el más alto.
Tasa de Suicidio	suicideper100TH	Suicidio, ajustado por edad, por 100.000 Mortalidad por lesiones autoinfligidas, por 100.000 habitantes estándar, ajustada por edad.
Uso de electricidad	relectricperperson	Consumo residencial de electricidad por persona (kWh).





PROBLEMAS CON LOS DATOS

DATOS DUPLICADOS

En la consulta de la base de datos se encontraron 12 filas duplicadas que fueron eliminadas para mejorar la estimación.

DATOS FALTANTES

Para la gran mayoría de variables hay información faltante, siendo necesario investigar el dato con base en la información del Banco Mundial para completar (En particular, se completó la prevalencia de VIH, Personal Militar, Ingreso per cápita y el Índice de desarrollo Humano)

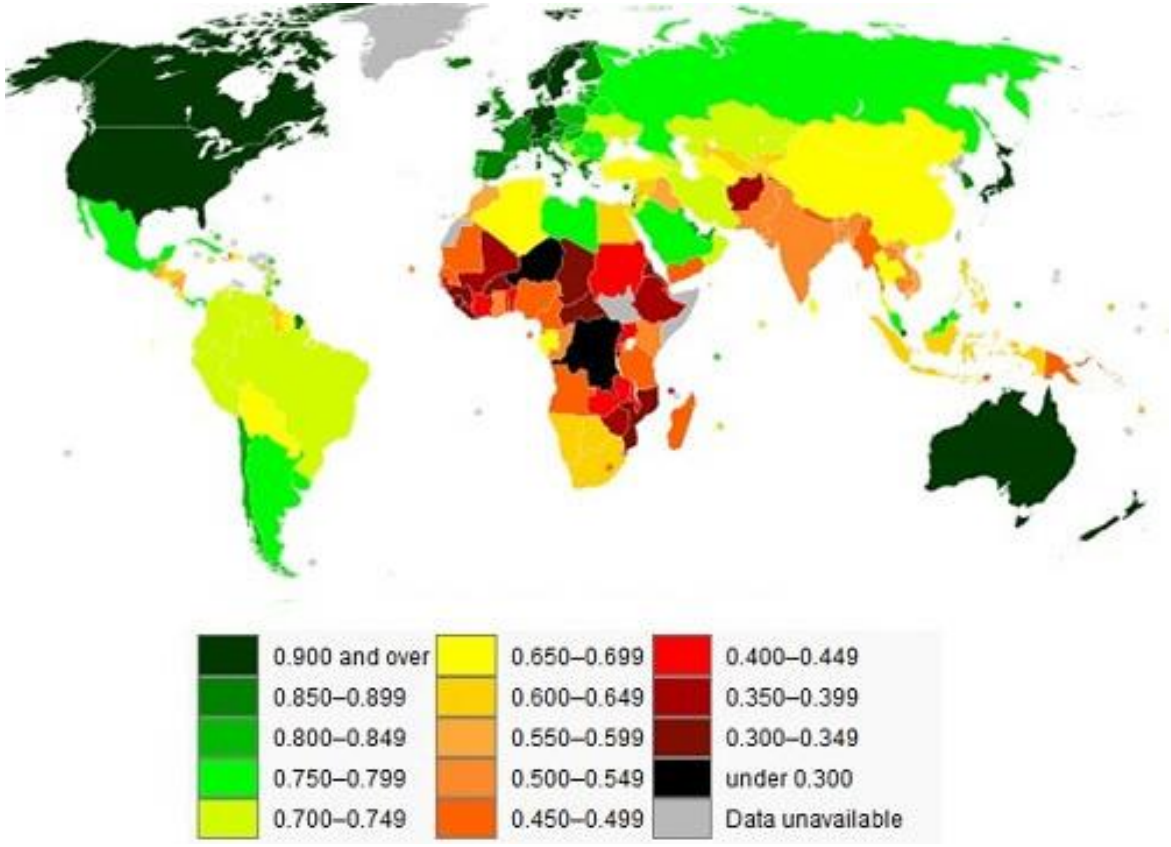
IMPUTACIÓN POR INVESTIGACIÓN

De la misma manera, para cuatro países se generó una imputación de Personal Militar a cero, basado en el tamaño del país y la poca o nula incidencia sobre el total de la información.

TRANSFORMACIONES LOGARÍTMICAS

De cara a mostrar un comportamiento suavizado de los datos del ingreso, se elaborará un modelo log-lin que captura la variación porcentual del ingreso con respecto a las variaciones absolutas de las variables exógenas presentadas.

INCLUSIÓN DE VARIABLES ADICIONALES

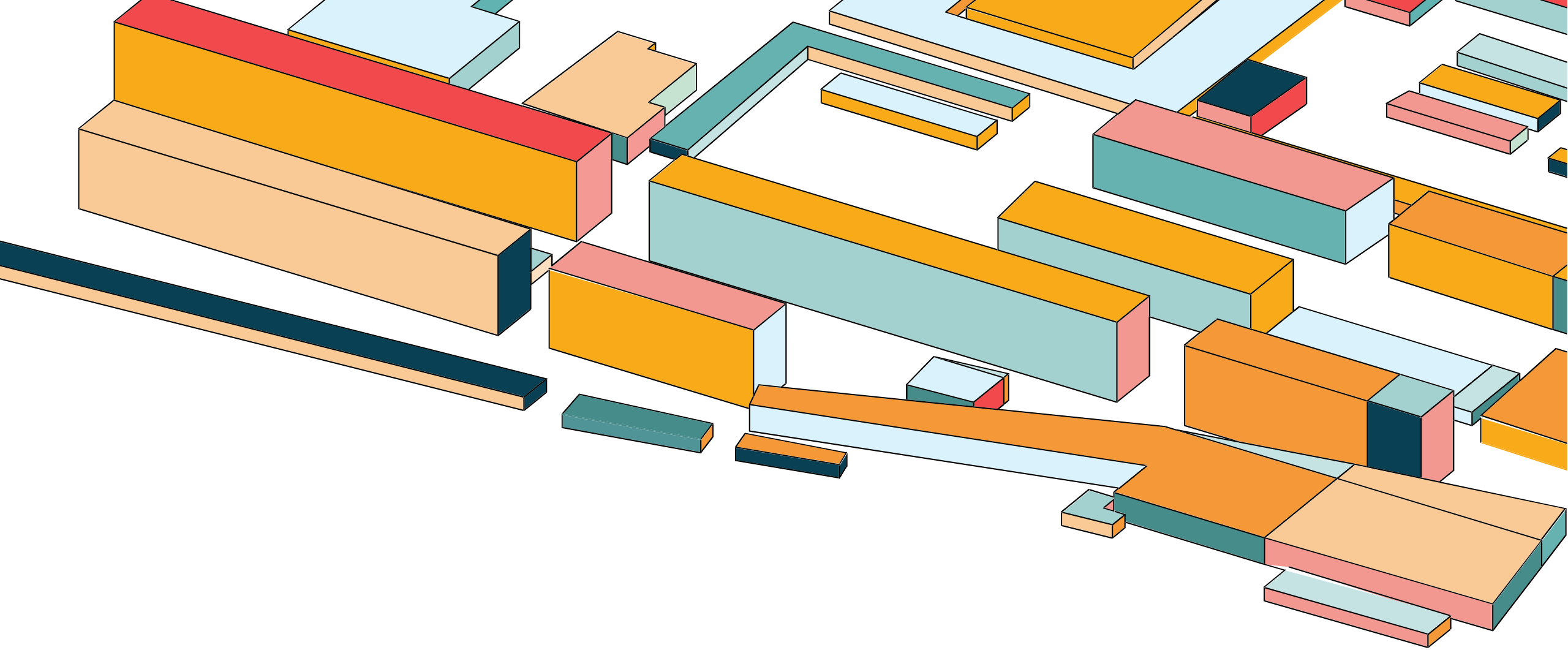


Fuente: World Data Bank (2023)

Se incluyó el Índice de Desarrollo Humano como indicador del nivel de desarrollo de los países

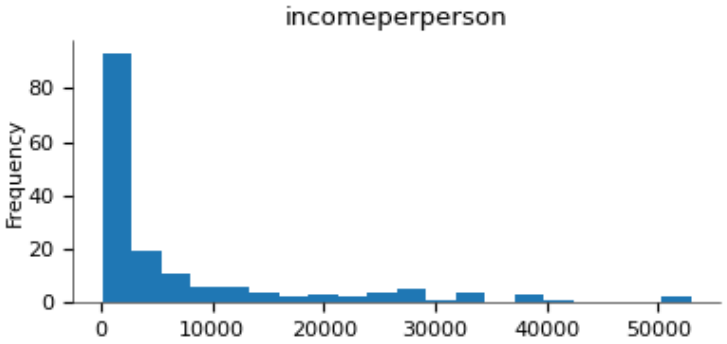
Se incluyó el campo de Región para clasificar en una manera agregada a los diferentes países

Se manejan 4 niveles de IDH y se cuenta con la columna numérica (Enfoque recomendado)

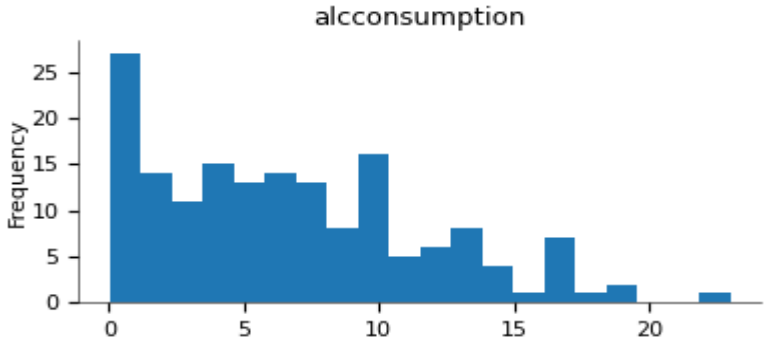


ANÁLISIS EXPLORATORIO DE DATOS

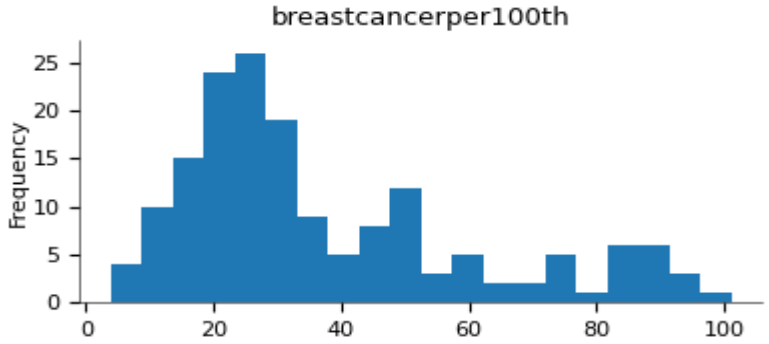
ANÁLISIS GRÁFICO



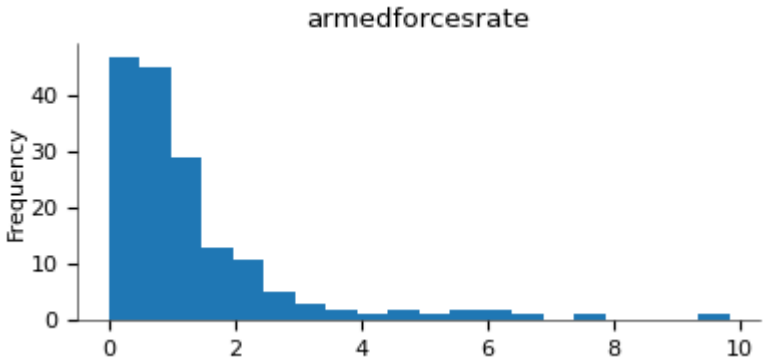
Ingreso Promedio: $7,424 USD
Distribución sesgada a la derecha



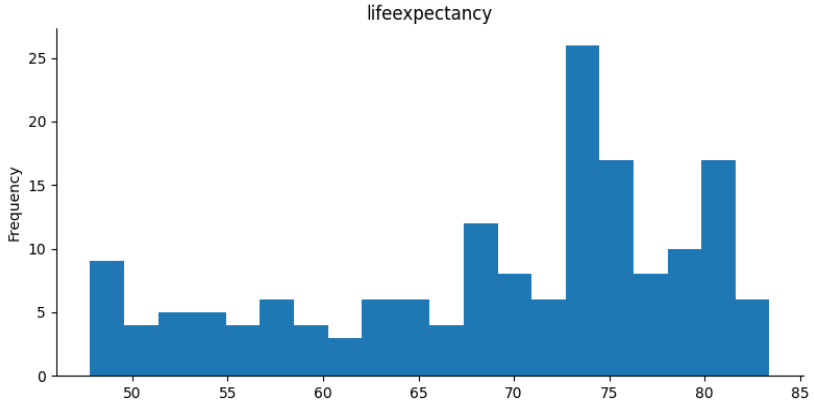
Consumo de Alcohol Promedio: 6,66 litros
Distribución sesgada a la derecha.



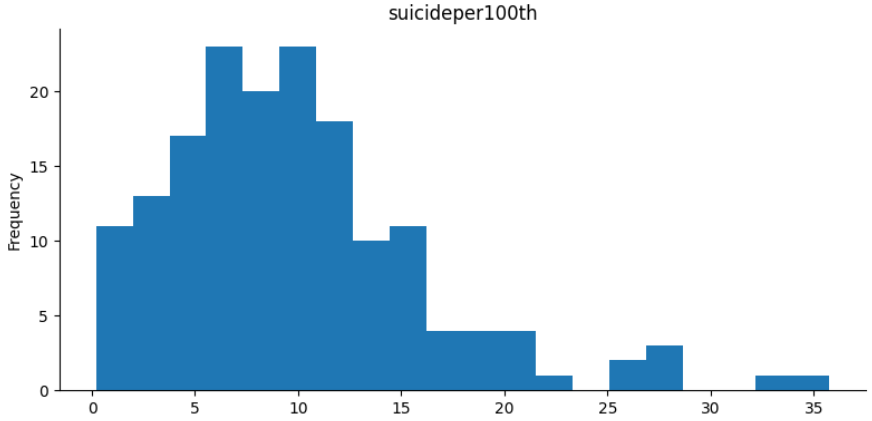
% Cáncer de Seno: 37,3%
Distribución sesgada a la derecha.



Porcentaje promedio de Personal Militar: 1,32%
Distribución sesgada a la derecha

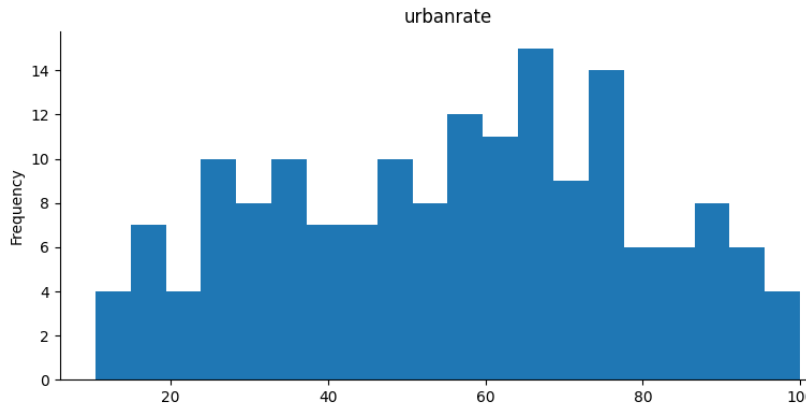


Esperanza de vida promedio: 69,10 años
Distribución sesgada a la izquierda.

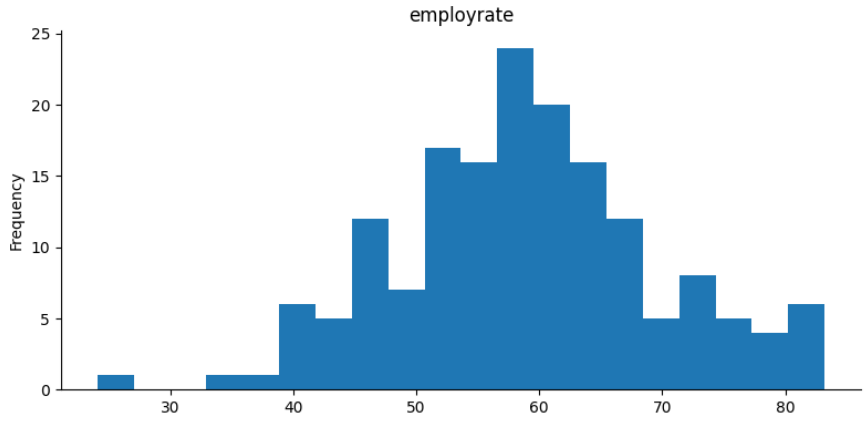


Tasa promedio de suicidio: 9,8%
Distribución sesgada a la izquierda.

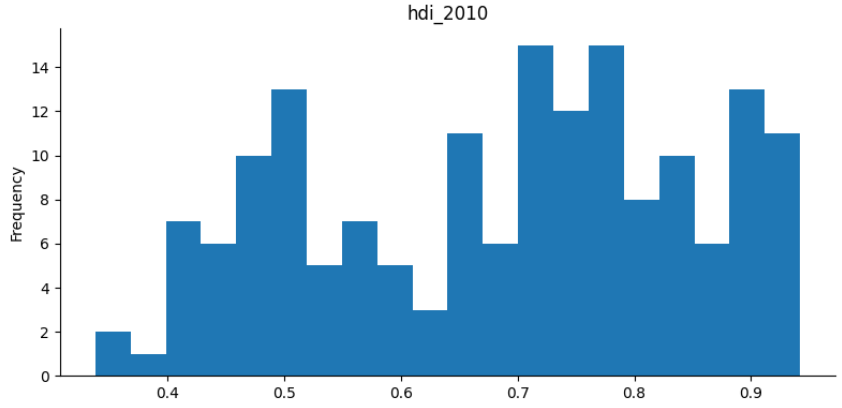
ANÁLISIS GRÁFICO



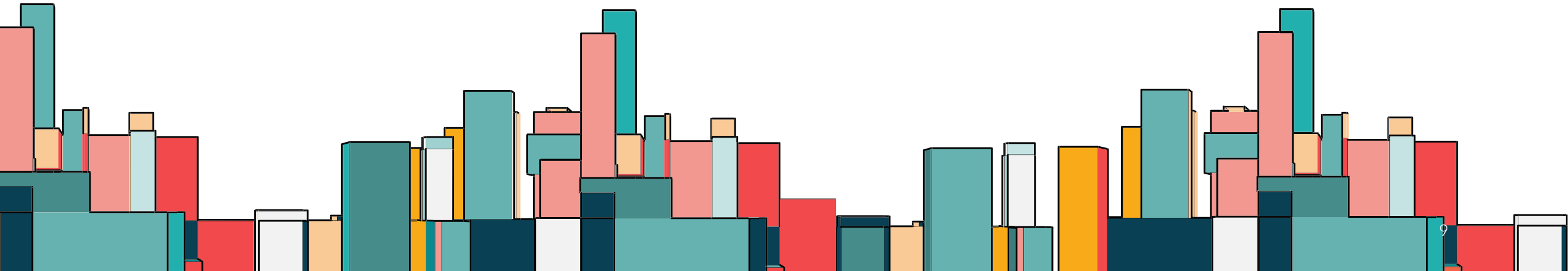
Porcentaje promedio de Urbanismo: 55,9%
Distribución ligeramente sesgada a la izquierda



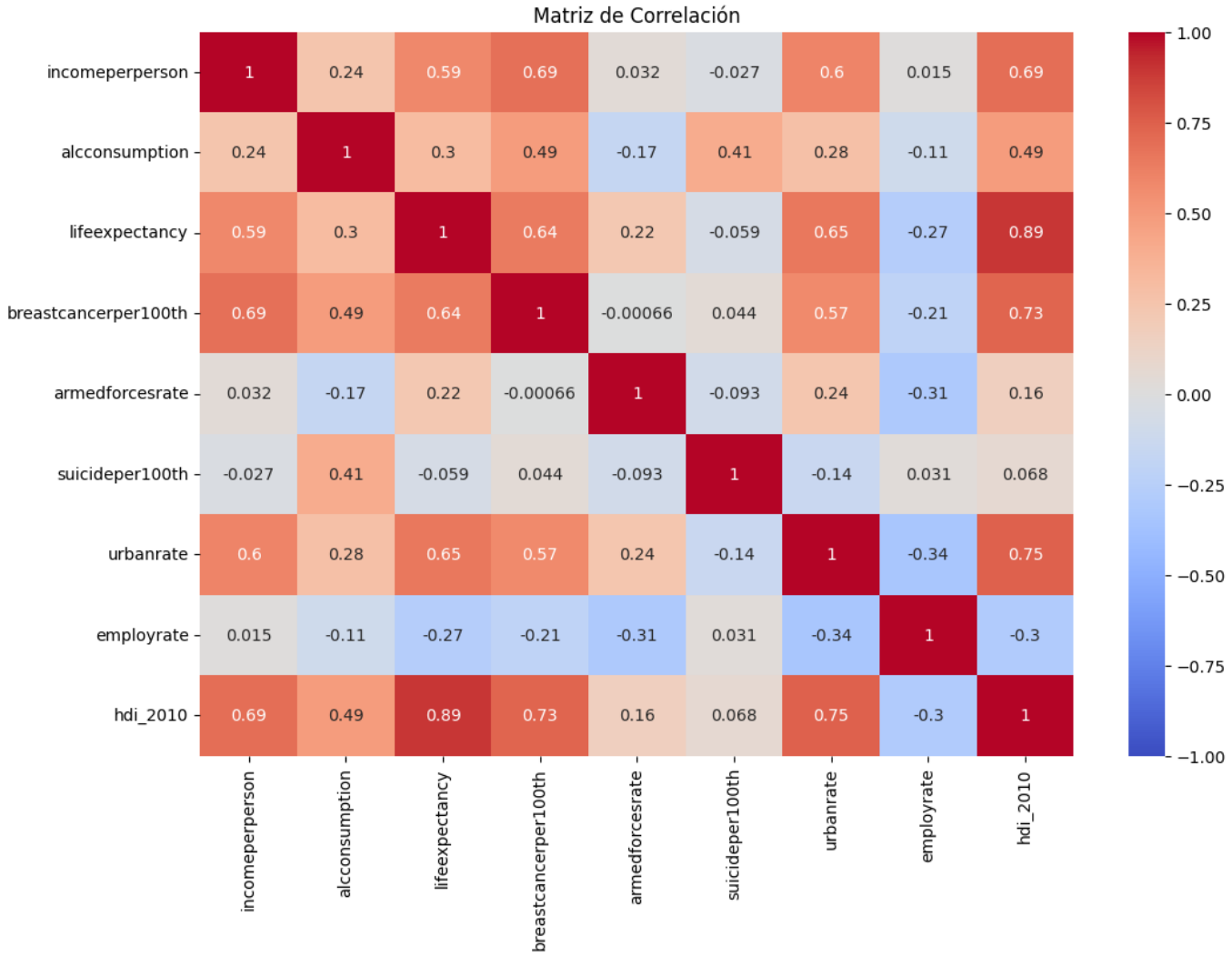
Tasa Promedio de Empleo: 58,97%
Distribución sesgada a la izquierda.



Índice Promedio de desarrollo Humano: 0,68
Distribución sesgada a la izquierda.



MATRIZ DE CORRELACIÓN



Una vez generado el alistamiento de datos se escogieron las variables más relevantes y con el menor nivel de información faltante. Así, se tiene:

Se observan valores importantes en la esperanza de vida y el IDH (0,89) (Posible multicolinealidad)

Correlaciones positivas entre el ingreso per cápita y:
-Consumo de Alcohol (0,24)
-Esperanza de Vida (0,59)
-Cáncer de Seno (0,69)
-IDH (0,69)

Correlaciones débiles o negativas entre el ingreso per cápita y:
-Tasa de Suicidios (-0,027)
-Tasa de Empleo (0,015)

Relaciones curiosas como:
Empleo vs Urbanismo (-0,34)
Empleo vs Personal Militar (-0,31)

Posibles relaciones espurias:
Personal Militar vs Cáncer de Seno (-0,00066)

ALGORITMO DE REGRESIÓN LINEAL

Se usó la metodología de OLS (Mínimos Cuadrados Ordinarios) para la ecuación

$$Incomeperperson = \beta_0 + \beta_1 alcconsumption + \beta_2 lifeexpectancy + \beta_3 breastcancerper100th + \beta_4 armedforcesrate + \beta_5 suicideper100th + \beta_6 urbanrate + \beta_7 employrate + \beta_8 hdi_{2010} + \varepsilon$$

Variable Endógena

Ingreso Per Cápita (y)

Variables Exógenas

Consumo de Alcohol (x1)
Esperanza de Vida (x2)
Cáncer de Seno (x3)
Fuerzas Armadas (x4)
Tasa de Suicidio (x5)
Tasa de Urbanismo (x6)
Tasa de Empleo (x7)
Índice de Desarrollo Humano (x8)

Término de Error

Representa la diferencia entre el valor observado de la variable dependiente y el valor predicho por el modelo de regresión (ε)

RESULTADOS RELEVANTES

OLS Regression Results

Dep. Variable:	y	R-squared:	0.854
Model:	OLS	Adj. R-squared:	0.847
Method:	Least Squares	F-statistic:	114.9
Date:	Sun, 08 Oct 2023	Prob (F-statistic):	1.30e-61
Time:	22:18:11	Log-Likelihood:	-152.95
No. Observations:	166	AIC:	323.9
Df Residuals:	157	BIC:	351.9
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.2855	0.566	4.036	0.000	1.167	3.404
x1	-0.0421	0.013	-3.134	0.002	-0.069	-0.016
x2	-0.0229	0.012	-1.935	0.055	-0.046	0.000
x3	0.0093	0.003	2.871	0.005	0.003	0.016
x4	-0.0442	0.036	-1.221	0.224	-0.116	0.027
x5	-0.0092	0.009	-1.017	0.311	-0.027	0.009
x6	0.0119	0.004	3.409	0.001	0.005	0.019
x7	0.0080	0.005	1.601	0.111	-0.002	0.018
x8	8.8178	0.942	9.362	0.000	6.957	10.678

Omnibus:	7.023	Durbin-Watson:	1.827
Prob(Omnibus):	0.030	Jarque-Bera (JB):	9.633
Skew:	0.232	Prob(JB):	0.00810
Kurtosis:	4.085	Cond. No.	2.29e+03

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.29e+03. This might indicate that there are strong multicollinearity or other numerical problems.

Fuente: Elaboración Propia

Se realizó una transformación logarítmica para poder generar un modelo que tenga validados la mayor parte de supuestos del modelo.

El R cuadrado es de 0,85, esto es, la variabilidad del ingreso per cápita es explicada por las variables predictoras en un 84,1%.

El R cuadrado ajustado es de 84,7% lo cual indica que hay una buena relación entre la variabilidad del modelo y la variación de las variables exógenas.

Se observan relaciones negativas en las primeras cinco variables, mientras relaciones positivas en las últimas tres.

Sin embargo, se observa que hay valores p mayores a 0,05, con lo que se procede a hacer un ajuste de variables por parsimonia y para genera mejores resultados de cara a la significancia de los estadísticos.

Adicionalmente, el modelo refleja multicolinealidad fuerte, principalmente asociada por la alta correlación entre la esperanza de vida y el IDH.

REGRESIÓN AJUSTADA (1/2)

OLS Regression Results

Dep. Variable:	y	R-squared:	0.841
Model:	OLS	Adj. R-squared:	0.837
Method:	Least Squares	F-statistic:	212.1
Date:	Sun, 08 Oct 2023	Prob (F-statistic):	4.54e-63
Time:	22:28:01	Log-Likelihood:	-160.34
No. Observations:	166	AIC:	330.7
Df Residuals:	161	BIC:	346.2
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.8384	0.232	7.938	0.000	1.381	2.296
x1	-0.0341	0.012	-2.831	0.005	-0.058	-0.010
x2	-0.0811	0.036	-2.282	0.024	-0.151	-0.011
x3	0.0141	0.003	4.152	0.000	0.007	0.021
x4	8.0441	0.523	15.380	0.000	7.011	9.077

Omnibus:	7.207	Durbin-Watson:	1.829
Prob(Omnibus):	0.027	Jarque-Bera (JB):	9.325
Skew:	0.270	Prob(JB):	0.00944
Kurtosis:	4.028	Cond. No.	676.

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Fuente: Elaboración Propia

Con base en los resultados arrojados por el modelo anterior, se realizó una corrección asociada a la supresión de variables no significativas.

Con un escenario más ajustado, se observa una similitud importante en el R cuadrado. Los índices de bondad del modelo AIC y BIC dan valores altos, lo que indica que el modelo tiene una bondad de ajuste pertinente.

Los valores p de las variables que se quedaron en el análisis son significativas a un grado de confianza del 95%, manteniendo todo lo demás constante.

El test Omnibus cercano a cero sugiere que la distribución no sigue una distribución normal. La prueba Jarque Bera también lo sugiere.

El test Durbin-Watson tiene un valor cercano a 2, por lo que no hay evidencia de presencia de correlación.

Los datos cuentan con una distribución sesgada hacia la derecha, pero moderada. De la misma manera, la curotsis indica que la distribución de los residuos tiene unas colas más pesadas que una distribución normal.

REGRESIÓN AJUSTADA (2/2)

OLS Regression Results

Dep. Variable:	y	R-squared:	0.841
Model:	OLS	Adj. R-squared:	0.837
Method:	Least Squares	F-statistic:	212.1
Date:	Sun, 08 Oct 2023	Prob (F-statistic):	4.54e-63
Time:	22:28:01	Log-Likelihood:	-160.34
No. Observations:	166	AIC:	330.7
Df Residuals:	161	BIC:	346.2
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.8384	0.232	7.938	0.000	1.381	2.296
x1	-0.0341	0.012	-2.831	0.005	-0.058	-0.010
x2	-0.0811	0.036	-2.282	0.024	-0.151	-0.011
x3	0.0141	0.003	4.152	0.000	0.007	0.021
x4	8.0441	0.523	15.380	0.000	7.011	9.077

Omnibus:	7.207	Durbin-Watson:	1.829
Prob(Omnibus):	0.027	Jarque-Bera (JB):	9.325
Skew:	0.270	Prob(JB):	0.00944
Kurtosis:	4.028	Cond. No.	676.

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Fuente: Elaboración Propia

Interpretando los coeficientes, se tienen las siguientes inferencias, manteniendo todo lo demás constante.

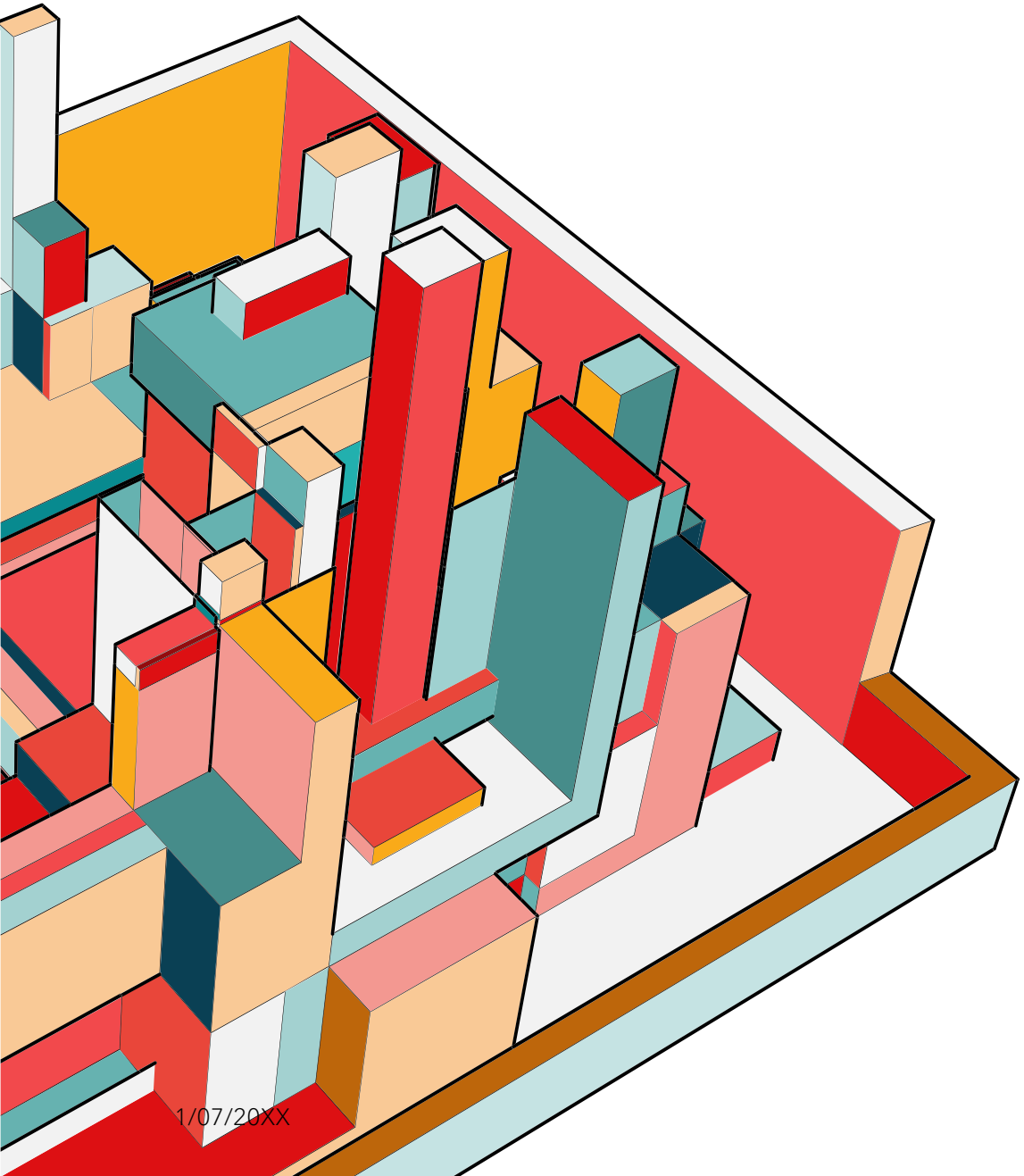
-Por cada litro de alcohol adicional(x1) de consumo disminuye, en promedio, un 3,41% el valor esperado del ingreso per cápita.

-Por cada uno por ciento adicional de fuerzas militares, se estima que en promedio el ingreso per cápita disminuye 8,11%.

-Por cada uno por ciento adicional en la tasa de urbanismo de un país, se estima que en promedio hay un alza del 1,41% en el ingreso per cápita.

-El ritmo promedio de crecimiento del ingreso per cápita es de 1,83%.

-Por cada punto adicional que tiene un país en el índice de desarrollo humano, se estima que en promedio el ingreso per cápita crece un 804%. Es importante establecer que, si bien el indicador del coeficiente es muy grande, es una variable significativa positiva que aporta al R cuadrado, razón por la cual su interpretación debe ser basada en su significancia más no en su magnitud.



¿QUÉ CONJUNTO DE POLÍTICAS PÚBLICAS RECOMENDARÍA IMPLEMENTAR, A PARTIR DE LA PREMISA DE QUE LA MEJORA EN ESTAS ÁREAS INDICARÍA AL BANCO MUNDIAL QUE EL PAÍS ES ESTABLE, ESTÁ EN UNA TRAYECTORIA DE DESARROLLO SOSTENIBLE Y TIENE LA CAPACIDAD DE ADMINISTRAR Y REEMBOLSAR PRÉSTAMOS DE MANERA EFECTIVA?

En primer lugar, es importante entender el IDH y el compendio de indicadores que contempla, con lo que acciones dirigidas a su mejoramiento son claves para la generación de resultados positivos en la producción nacional.

En segundo lugar, si bien las fuerzas armadas tienen impactos negativos en la producción, debe hacerse un foco importante en la sustitución del gasto público hacia sectores más productivos de la economía.

Tercero, políticas educativas de ampliación de cobertura y calidad pueden tener impacto positivo en el reconocimiento de los efectos negativos del alcohol y contribuir a mejores indicadores. Vale la pena resaltar que esta variable puede ser espuria, y se requieren mayores detalles para poder verificar la dependencia entre ambas variables.

La urbanización de las ciudades es clave para el mejoramiento del ingreso per cápita, por lo que actividades asociadas al mejoramiento de las ciudades capitales e intermedias tienen un impacto positivo en la generación de ingresos para la economía.

MUCHAS GRACIAS

Jonnatan Triana

Ciencia de Datos Aplicada

j.trianaa@uniandes.edu.co