Taller 2

En este documento se presentan los resultados del taller 2. Teniendo una sección por cada entregable que lo necesite.

# 1. Entendimiento de los datos

1\_Entendimiento\_de\_datos.ipynb

En esta sección se hace un reporte de entendimiento de datos inicial destacando el top 5 de características más importantes encontradas.

## Dimensiones del dataset

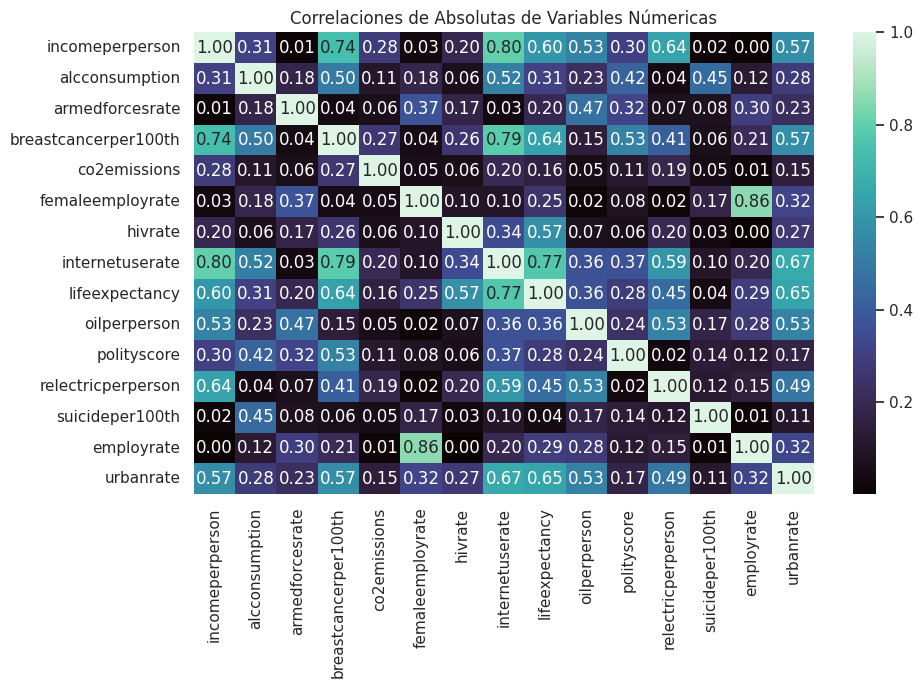
El dataset tiene 178 entradas de países con 15 columnas de indicadores y una para el nombre del país, donde todos los indicadores son numéricos. De estos indicadores se tiene como dato a destacar el hecho de que las siguientes columnas presentan más de un 7% de filas vacías:

* hivrate: 12.9% vacío.
* oilperperson: 62.3% vacío.
* relectricperperson: 20.7% vacío.

También se tienen países repetidos, como “Ireland” que aparece 2 veces.

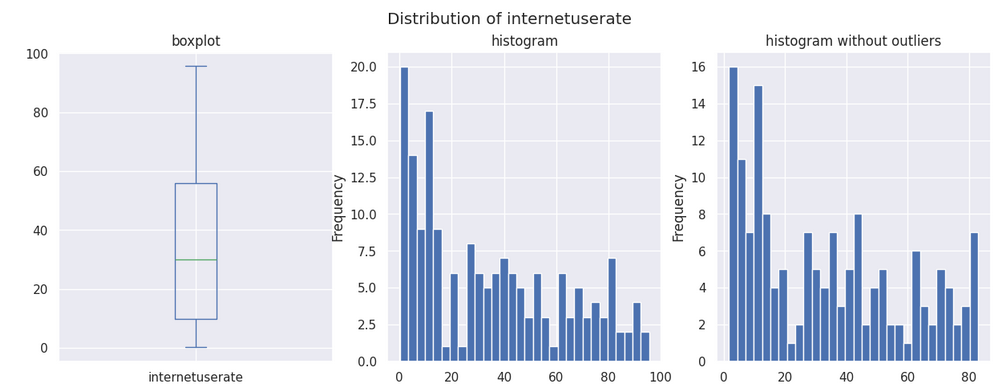
## Análisis univariado

Para este informe del análisis univariado se seleccionan los cinco indicadores que más correlación tienen con la variable de salida y se hace un análisis preliminar sin transformaciones.



### internetuserate (Pearson 0.80)

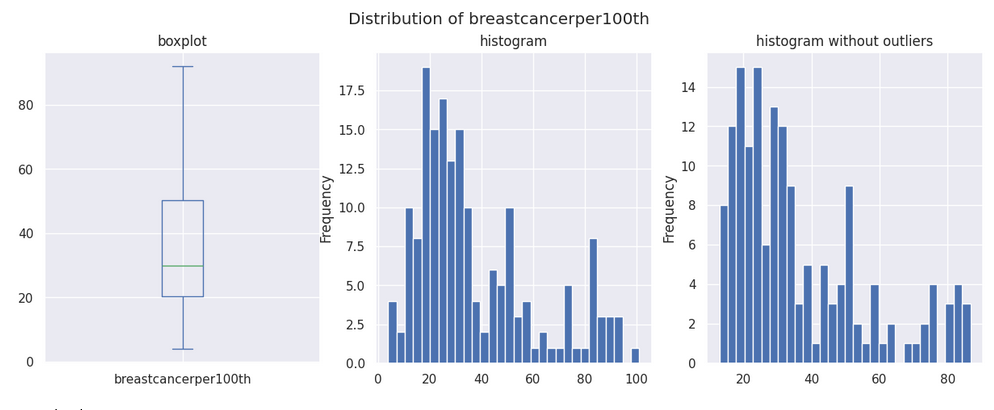
Este indicador establece cuantas personas de cada 100 tienen acceso a internet, siendo un indicador del avance tecnológico y de infraestructura del país, tiene una correlación de 80% con la variable de salida incomeperperson.



Al observar la distribución se observa que tiene una gran concentración de valores debajo de 20, y no hay valores que se consideren anómalos en términos de negocio. Como hay valores concentrados debajo de 20 y el resto está encima de este valor, se tiene una mediana muy cercana a 30.

### breastcancerper100th (Pearson 0.74 donde la correlación sin valor absoluto también es positiva)

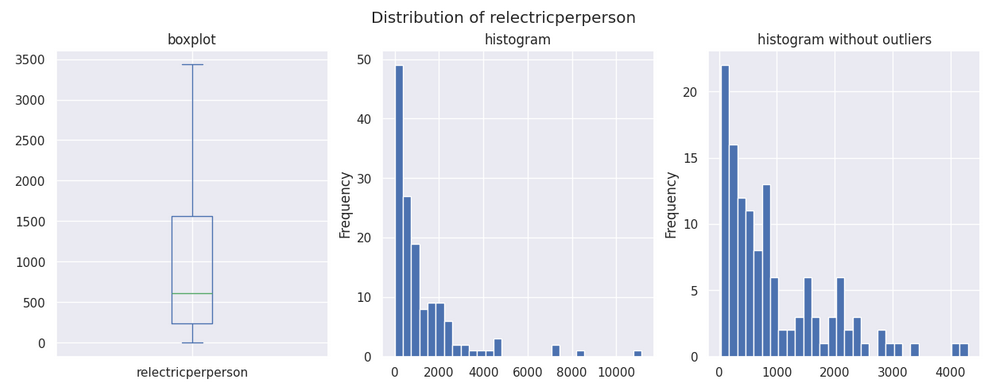
Se esperaría que un país con alto PIB también tenga bajos índices de cáncer debido a la mejor calidad de los servicios de salud. No obstante, como menciona [World Cancer Research Fund International](https://www.wcrf.org/cancer-trends/cancer-rates-human-development-index/) los países desarrollados suelen tener mayor porcentaje de la población debido a factores como el alcoholismo y el sedentarismo que hacen que una persona aumente de peso y por ende sea más propensa a tener cáncer.



Al observar la distribución de los datos se tiene que hay una gran concentración entre 20 y 40 casos por cada 100,000 mujeres en los países presentes y no existen errores de entrada en los datos. Asimismo, no hay países donde no haya mujeres sin cáncer de seno, teniendo un mínimo de 3.9.

### relectricperperson (Pearson 0.64)

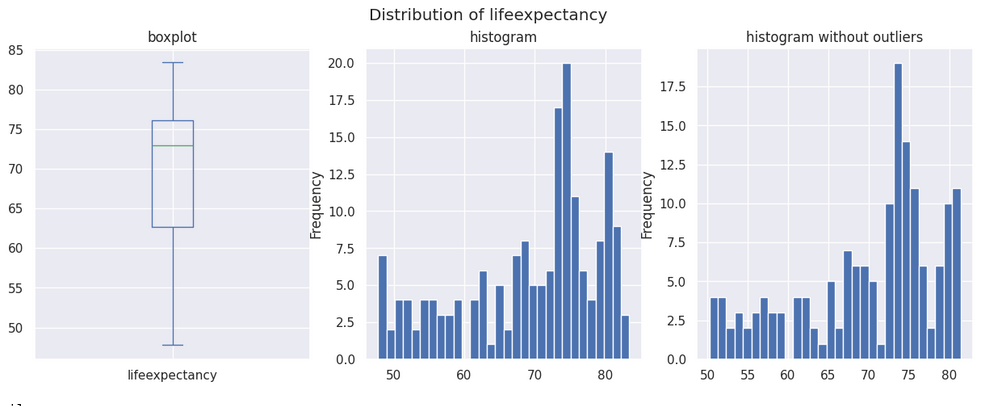
De manera similar a internetuserate, el consumo eléctrico residencial por persona en kWh en un año es un indicador del desarrollo tecnológico de la infraestructura del país, como la infraestructura de internet es un subconjunto de la de electricidad, se da el hecho de que muchos más países tienen mejor infraestructura eléctrica que internet y a su vez la correlación que se tiene es más baja con incomeperperson, con un valor de 0.64.



Al observar cómo se distribuye, se tiene un percentil 50 de 614 kWh y uno 75 de 1566 kWh, indicando una distribución altamente sesgada a valores “cercanos” a cero. También se tienen valores “extremos” por encima de 4000 kWh.

### lifeexpectancy (Pearson 0.60)

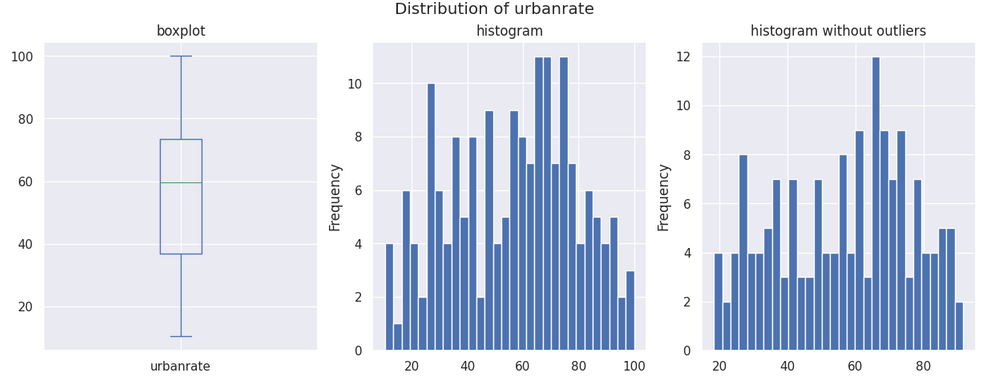
De manera similar a breastcancerper100th, se espera que lifeexpectancy resuma una buena calidad de vida de los ciudadanos, teniendo mayores valores para países con mayor PIB. En este caso, si se cumple la expectativa y se tiene una correlación de Pearson de 0.60.



A diferencia de las variables anteriores, esta se concentra hacia los valores mayores, teniendo un valor mínimo de 47 y una mediana de 72.

### Urbanrate (Pearson 0.57)

De acuerdo con [Our World in Data](https://ourworldindata.org/grapher/urbanization-vs-gdp), se tiene que, entre más porcentaje de la población viva en zonas urbanas, mayor sera el PIB del pais, siendo esto causado por el enfoque de las economías en productos y servicios más avanzados que los agrarios y de transformación.



Como se observa, la distribución es bastante uniforme entre 20% y 100%, con picos entre 60% y 80% y una mediana en 60%.

## Análisis bivariado

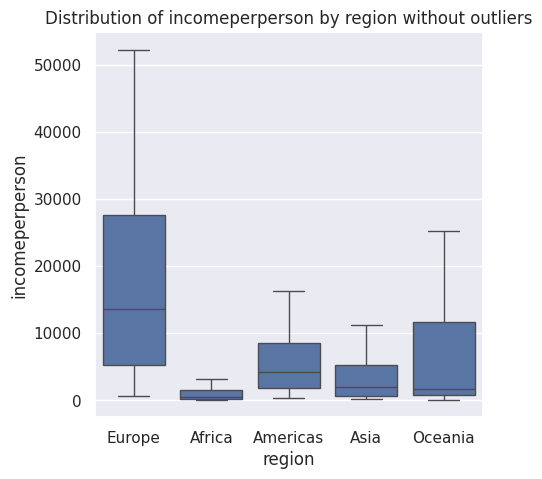
### Correlaciones entre variables numéricas

Para el analisis bivariado se tiene la correlacion de las variables presentada en el heatmap del análisis univariado, donde eliminando el valor absoluto también se pueden destacar las siguientes correlaciones:

* breastcancerper100th vs internetuserate: 0.78 (países con mayor población que use internet también tienen mayor proporción de mujeres con cáncer de seno).
* Breastcancerper100th vs lifeexpectancy: 0.63 (entre mayor espectativa de vida, mayor proporción de mujeres con cáncer de seno).
* armedforcesrate vs femaleemployrate: -0.36 (entre más porcentaje de la población trabaje en las fuerzas armadas, menos porcentaje de las mujeres se encuentra empleado).

### Comportamiento de variable de salida contra categóricas

Se tiene un comportamiento en el cual ciertos grupos de países presentan un PIB superior a otros, como se observa en el siguiente diagrama de cajas donde África tiene un máximo de PIB sin outliers inferior a la mediana del PIB de las demás regiones; y Europa tiene un percentil 25 superior a la mediana de las demás regiones siendo ambos extremos contrarios de riqueza.

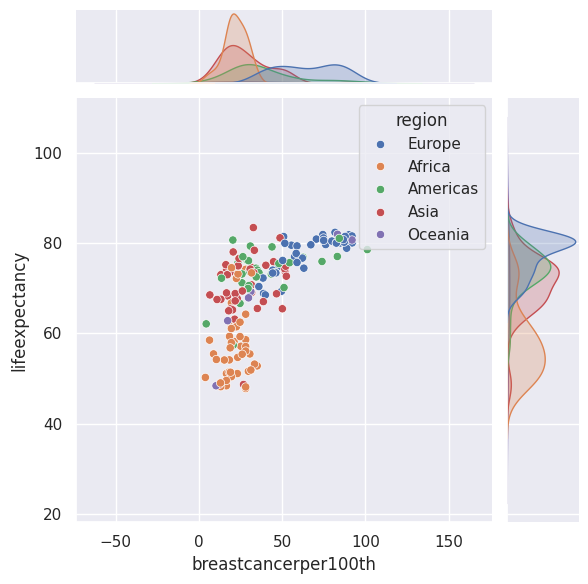


Respecto de la situación de desarrollo, los países en desarrollo presentan un máximo sin outliers menor el percentil 25 de los países en desarrollo respecto del PIB.



### Relacion entre variables Top separadas por región

Visualizando las correlaciones entre variables numéricas del top 5, se tiene como dato a resaltar el comportamiento de breastcancerper100th, donde esta variable a pesar de tener que comportarse de manera inversamente proporcional a los otros indicadores top, se comporta de manera opuesta, como se ve en la siguiente comparación con lifeexpectancy:



Por otra parte, África siempre se encuentra en los valores más bajos de los indicadores, incluyendo breastcancerper100th.

# 2. Preparación de los datos

2\_Preparacion\_de\_datos.ipynb

En el notebook anterior se añadieron las siguientes columnas al dataframe original aparte de región para tener información adicional del país:

* Independent: Indica si el país es independiente o no.
* Region: continente en donde el país esta ubicado.
* Subregion: Región dentro del continente donde el país está ubicado.
* Landlocked: Si el país no tiene acceso al mar.
* Area: Area del territorio continental del país en km2.
* Is\_developed: Si el pais es desarrollado o no.

Una vez completada esta informacion se encontraron las siguientes correcciones con base en determinadas dimensiones de calidad. Como se puede observar la mayoria de soluciones de conformidad limitan el rango de valores de la variable, aunque esta se podría usar sin estos cambios y por eso no se consideran problemas de consistencia. En la siguiente tabla se listan los problemas y soluciones aplicadas. En general, como los datos son de 2008, se tiene un alto problema de temporalidad si se quisiera aplicar el modelo de regresión a la actualidad.

|  |  |  |  |
| --- | --- | --- | --- |
| Columna | Dimensión de calidad | Nivel | Solución |
| country | Consistencia | Tabla | Eliminar las filas con países duplicados para evitar “confundir” al regresor |
| incomeperperson | Completitud | Atributo | Llenar vacíos con el promedio de la región |
| incomeperperson | Conformidad | Atributo | Clipping de valores muy altos por IQR |
| alcconsumption | Conformidad | Atributo | Clipping de valores muy altos por IQR |
| armedforcesrate | Completitud | Atributo | Llenar vacíos con el promedio de la región |
| armedforcesrate | Conformidad | Atributo | Clipping de valores muy altos por IQR |
| breastcancerper100th | Completitud | Atributo | Llenar vacíos con el promedio de la región |
| co2emissions | Completitud | Atributo | Llenar vacíos con el promedio de la región |
| co2emissions | Conformidad | Atributo | Clipping de valores muy altos por IQR |
| femaleemployrate | Completitud | Atributo | Llenar vacíos con el promedio de la región |
| hivrate | Completitud | Columna | Eliminar columna |
| internetuserate | Completitud | Atributo | Llenar vacíos con el promedio de la región |
| oilperperson | Completitud | Columna | Eliminar columna |
| relectricperperson | Completitud | Columna | Eliminar columna |
| polityscore | Completitud | Atributo | Llenar vacíos con el promedio de la región |
| suicideper100th | Conformidad | Atributo | Clipping de valores muy altos por IQR |
| employrate | Conformidad | Atributo | Llenar vacíos con el promedio de la región |
| independent | Conformidad | Columna | Convertir valor a booleano |
| landlocked | Conformidad | Columna | Convertir valor a booleano |
| area | Conformidad | Atributo | Clipping de valores muy altos por IQR |
| Is\_developed | Conformidad | Columna | Convertir valor a booleano |

# 3. Entrenamiento de modelo

3\_Entrenamiento\_del\_modelo\_de\_regresion.ipynb

Para entrenar el modelo de regresión se sigue un proceso en que se tienen los siguientes pasos:

1. Separar columnas en tipos numéricos y categóricos.
2. Eliminar variables altamente colineales (employerate).
3. Evaluar linealidad de columnas numéricas y mantener aquellas que tengan una correlación significativa luego de calcular correlación con y sin transformación logarítmica:
   1. "alcconsumption",
   2. "breastcancerper100th",
   3. "co2emissions",
   4. "internetuserate",
   5. "lifeexpectancy",
   6. "polityscore",
   7. "urbanrate"
4. Evaluacion del comportamiento normal del error y homocedasticidad de su varianza.
5. Transformacion de region y subregion con onehotencoder.

De estos pasos se concluye que las variables presentes sirven para entrenar un modelo de regresión después de cumplir todos los lineamientos de la validación de supuestos. Para ver en detalle se recomienda revisar el notebook correspondiente.

# 4. Generación de resultados

En la presentación adjunta se presentan los insights del taller 2.