**[20%] Entendimiento de los datos:** Genere un breve reporte de entendimiento de datos en donde destaque las dimensiones del dataset, los tipos de datos que contiene y el top 5 de los indicadores que considera más importantes para el ejercicio de regresión propuesto. Por cada indicador, incluya algunos elementos básicos de su comportamiento o distribución (análisis univariado). También incluya algunos elementos complementarios que permitan entender la relación entre pares de indicadores, el país o la región a la que pertenecen (análisis bivariado).

Los datos entregados contienen la siguiente forma:

0 country 178 non-null object

1 incomeperperson 175 non-null float64

2 alcconsumption 178 non-null float64

3 armedforcesrate 174 non-null float64

4 breastcancerper100th 177 non-null float64

5 co2emissions 176 non-null float64

6 femaleemployrate 177 non-null float64

7 hivrate 155 non-null float64

8 internetuserate 174 non-null float64

9 lifeexpectancy 178 non-null float64

10 oilperperson 67 non-null float64

11 polityscore 167 non-null float64

12 relectricperperson 141 non-null float64

13 suicideper100th 178 non-null float64

14 employrate 177 non-null float64

15 urbanrate 178 non-null float64

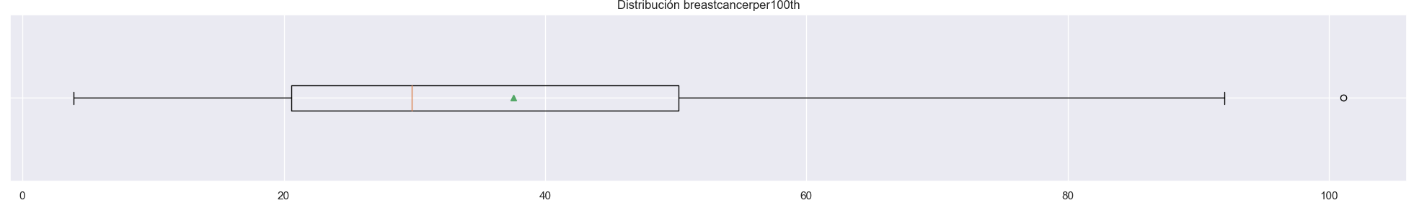
Tiene un total de 178 filas en la que cada una representa a los indicadores de un país en sí. Además, todas las filas son datos de números decimales.

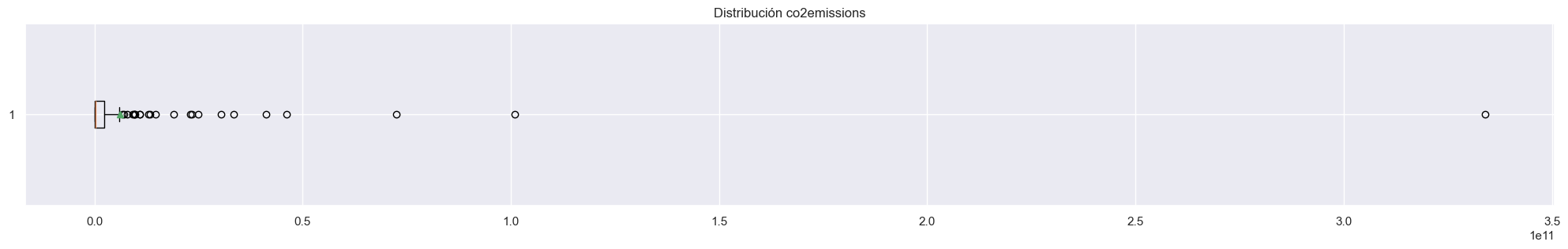
Es importante resaltar que todas las columnas o en su defecto indicadores tienen temporalidades distantes entre sí, es decir, el indicador del cáncer de seno está dado en el año 2002, mientras que el indicador de expectativa de vida está dado en el año.

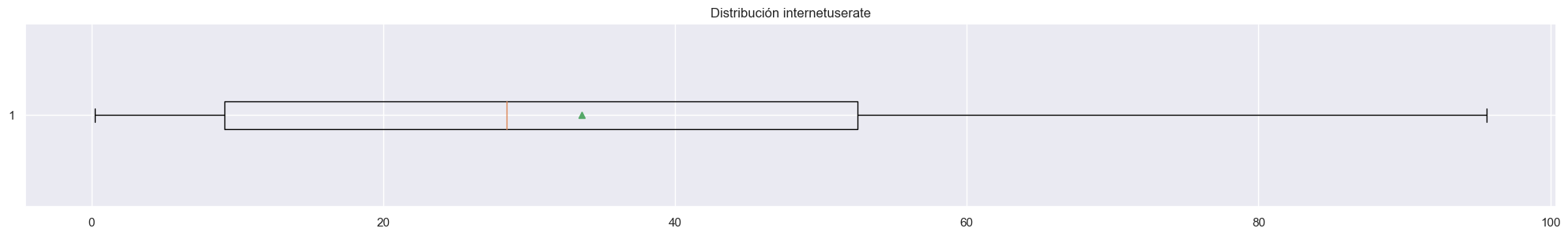
Los 5 indicadores que considero más importante de cara a la construcción de una regresión serían:

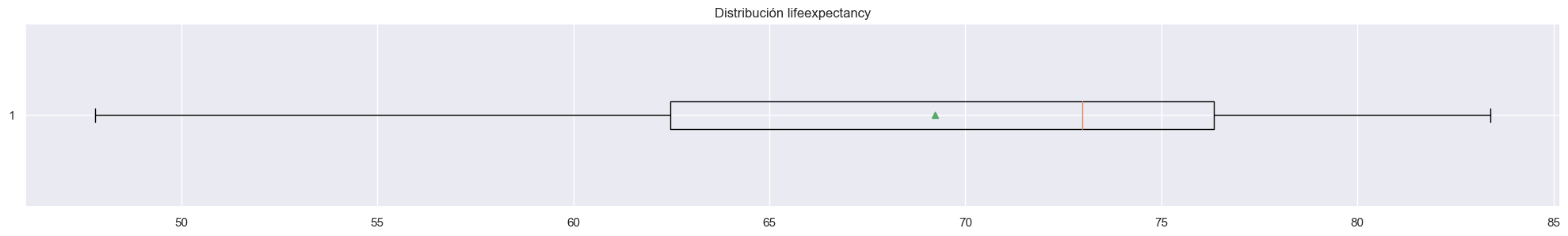
* breastcancerper100th
* co2emissions
* internetuserate
* lifeexpectancy
* relectricperperson
* babiesperwoman

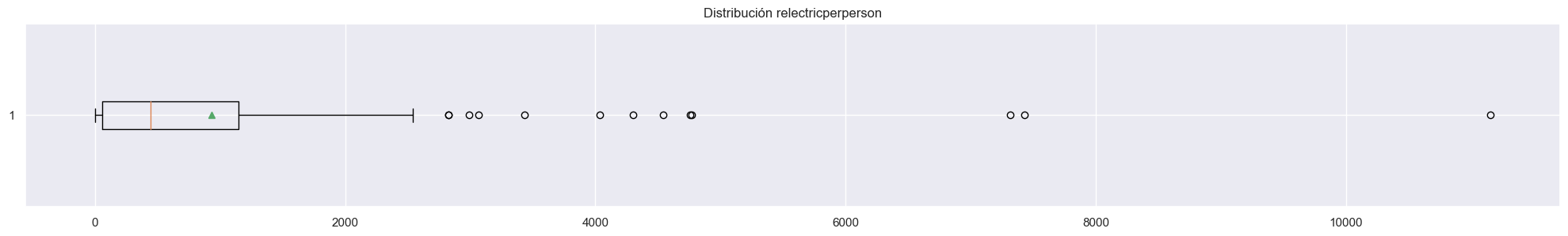
A continuación, se muestra la distribución que tienen estos indicadores:

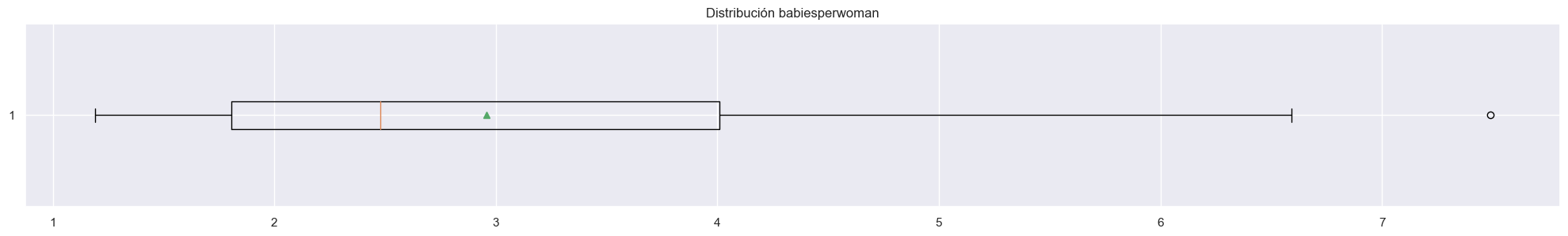






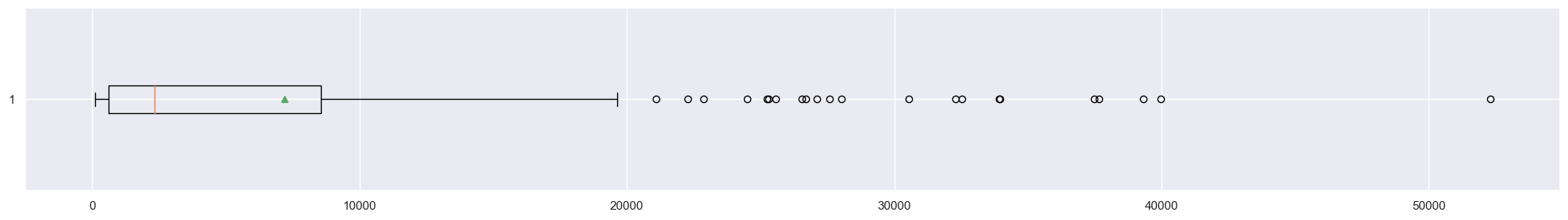




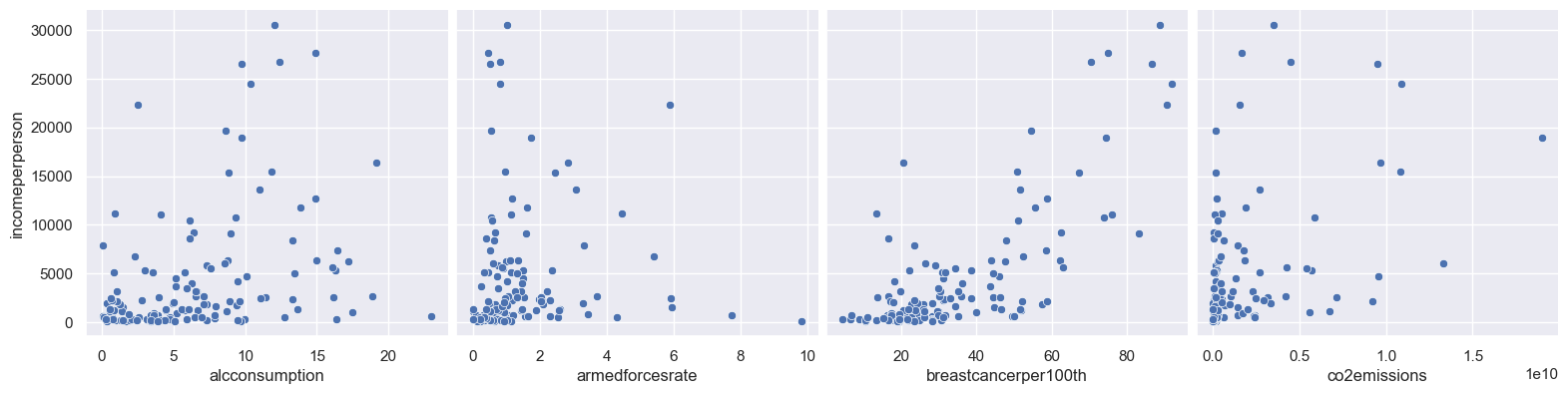


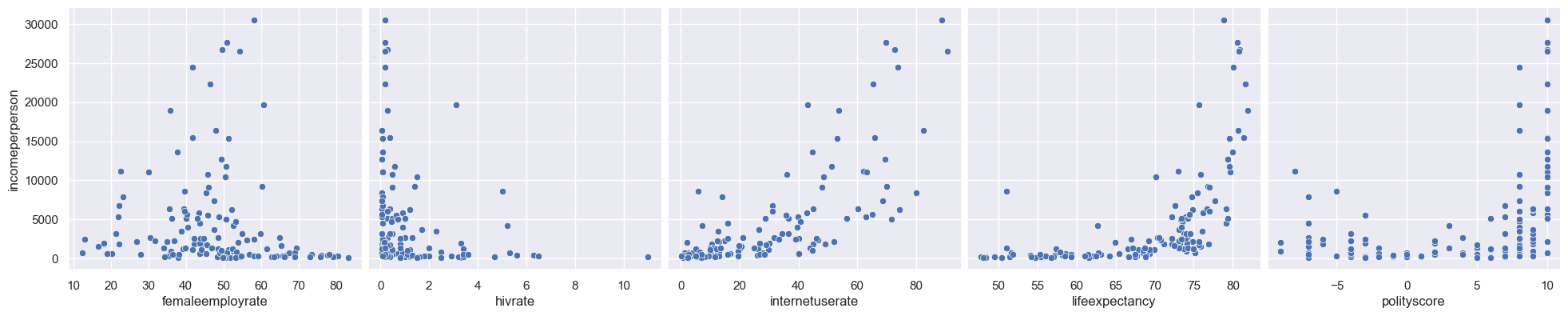
Se pueden observar que las variables de co2emissions y relectricperson tienen outliers bastante grandes lo cuál va a ser importante de cara al siguiente paso de preparación de los datos donde se podrá considerar si eliminar estos datos o no de cara a la construcción del modelo de regresión lineal.

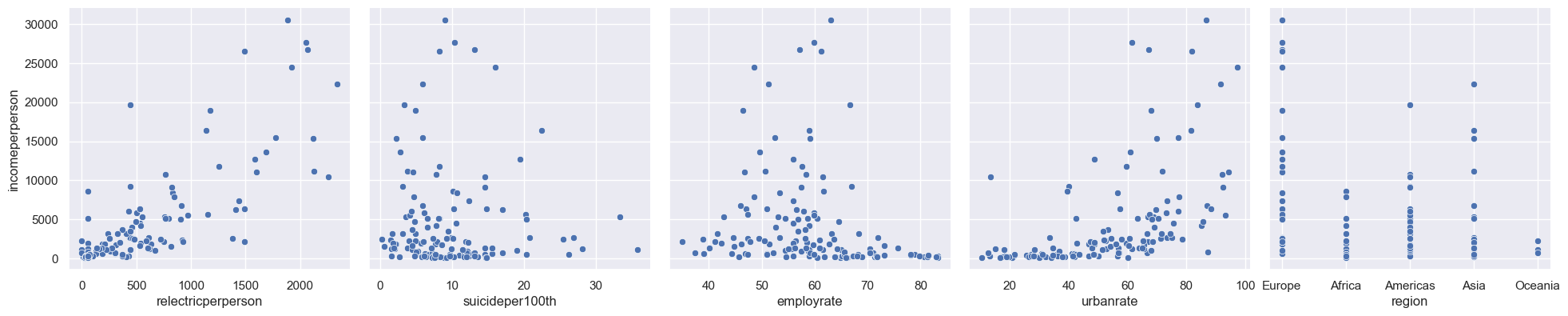
Adicionalmente, se puede ver que la variable objetivo incomeperperson tiene un outlier bastante grande, específicamente el país de Luxemburgo.

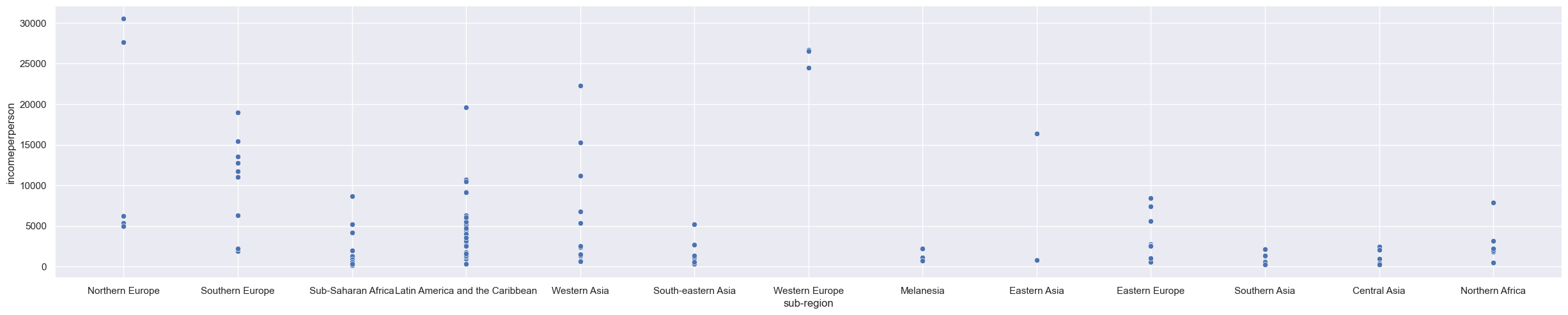


Ahora vamos a ver como se relacionan todas las variables con respecto a la variable objetivo incomeperperson (Análisis bivariado):









Se ve que las variables que se pusieron en el top 5 de este análisis tienen tendencias lineales ya sea positivas o negativas con respecto a nuestra variable objetiva. Además, es interesante la diferencia del incomeperperson según las subregiones, se ve que que el norte de Europa y el oeste de Europa son los países con la mayor cantidad de países con un PIB per cápita alto. Por otro lado, se podría decir que existen países latinoamericanos que tienen un mejor indicador que los países del este de Europa. Finalmente, las subregiones con menor PIB per cápita son los países ubicados en el sur de Asia, Asia central y Melanesia.

2. **[25%] Preparación de datos:** Realice un análisis de calidad de datos teniendo en cuenta las diferentes dimensiones de calidad y determine las estrategias más apropiadas para su limpieza. Como estrategia de imputación de datos, se recomienda revisar el portal de Gapminder en donde se encuentra mayor detalle de la información contenida en el dataset. Complementariamente, evalúe la posibilidad de complementar el dataset con información categórica relativa al continente o a la región a la que pertenece, si es un país desarrollado o subdesarrollado, entre otros.

En primer lugar, se encontraron países duplicados en la tabla, por lo tanto, se hizo la limpieza respectiva para limpiar como tal cada uno de los países duplicados. Además, se quitaron los países que tenían nulo la variable objetivo porque deja de ser información relevante de cara a la construcción de una regresión lineal.

Ahora, viendo los diferentes nulos en ciertas columnas y en especial la columna relectricperperson que es importante para el análisis debido a una distribución parecida a lineal con la variable anterior, se va a agregar para cada país la región y la subregión a la que pertenece para que así se pueda hacer imputación de datos mediante la mediana de cada dato, pero por subregión para así tener unos datos lo más reales posibles. Además, se añadieron datos como territorio en km2 (2023), tierra en km2 (2023), agua en km2 (2023), % de agua (2023), total de población (2023), densidad poblacional (2023) y el indicador de natalidad bebes por mujer (2010).

Algo que no se mencionó al principio de este documento es que se decidió eliminar el indicador oilperperson ya que este solo lo tenían 67 países de 178.

Los nulos de la variable armedforcesrate se llenaron con 0 porque se verificó con información real que los países con nulos eran países sin fuerzas armadas como tal.

Todas las demás columnas con valores se llenaron con la mediana de la sub-región de cada páis en cuestión, estas columnas son: breastcancerper100th, co2emissions, femaleemployrate, hivrate, internetuserate, polityscore, relectricperperson y employrate.

Las subregiones tenidas en cuenta son las siguientes: Northern Europe, Southern Europe, Sub-Saharan Africa,Latin America and the Caribbean, Western Europe,Western Asia, South-eastern Asia, Melanesia, Eastern Asia,Eastern Europe, Southern Asia, Central Asia, Northern Africa, Australia and New Zealand, Northern America.

Un último detalle es que para la subregión de Melanesia ningún país tenía información sobre la variable relectricperperson por lo que se llenó con 0 este parámetro, ya que se hizo la respectiva investigación y se encontró que estos países no tienen energía residencial durante el día como tal.

Luego de todo el proceso anterior mencionado, quedó un dataset de la siguiente manera:

0 country 163 non-null object

1 incomeperperson 163 non-null float64

2 alcconsumption 163 non-null float64

3 armedforcesrate 163 non-null float64

4 breastcancerper100th 163 non-null float64

5 co2emissions 163 non-null float64

6 femaleemployrate 163 non-null float64

7 hivrate 163 non-null float64

8 internetuserate 163 non-null float64

9 lifeexpectancy 163 non-null float64

10 polityscore 163 non-null float64

11 relectricperperson 163 non-null float64

12 suicideper100th 163 non-null float64

13 employrate 163 non-null float64

14 urbanrate 163 non-null float64

15 country\_mod 163 non-null object

16 name 163 non-null object

17 region 163 non-null object

18 sub-region 163 non-null object

19 poblacion 163 non-null float64

20 babiesperwoman 163 non-null float64

21 total\_areakm2 163 non-null float64

22 tierra\_area\_km2 163 non-null float64

23 agua\_area\_km2 163 non-null float64

24 % water 163 non-null float64

25 densidadpoblacional 163 non-null float64

Quedaron un total de 163 países luego de la limpieza de datos.

Luego, se consideró el asunto de los datos outliers de ciertas variables lo cual hizo que se limpiara el dataset de la siguiente manera:

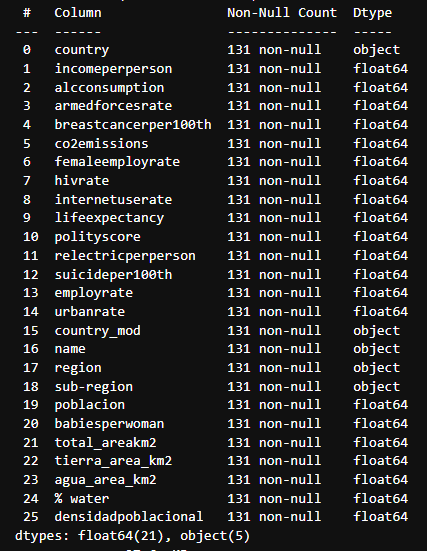
df\_visualizar= archivo\_taller2\_fil.loc[archivo\_taller2\_fil["incomeperperson"] <=archivo\_taller2\_fil["incomeperperson"].quantile(.95)]

df\_visualizar= df\_visualizar.loc[df\_visualizar["co2emissions"] <= df\_visualizar["co2emissions"].quantile(.95)]

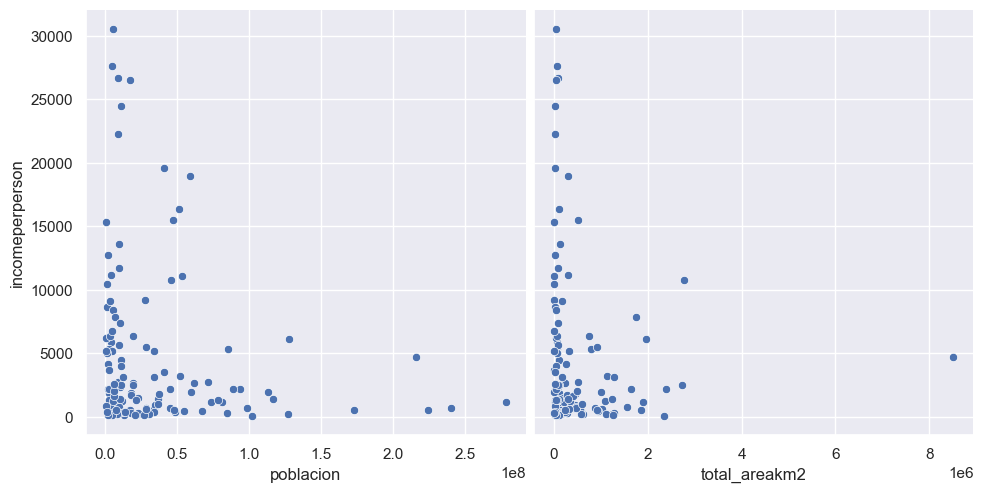
df\_visualizar= df\_visualizar.loc[df\_visualizar["hivrate"] <= df\_visualizar["hivrate"].quantile(.95)]

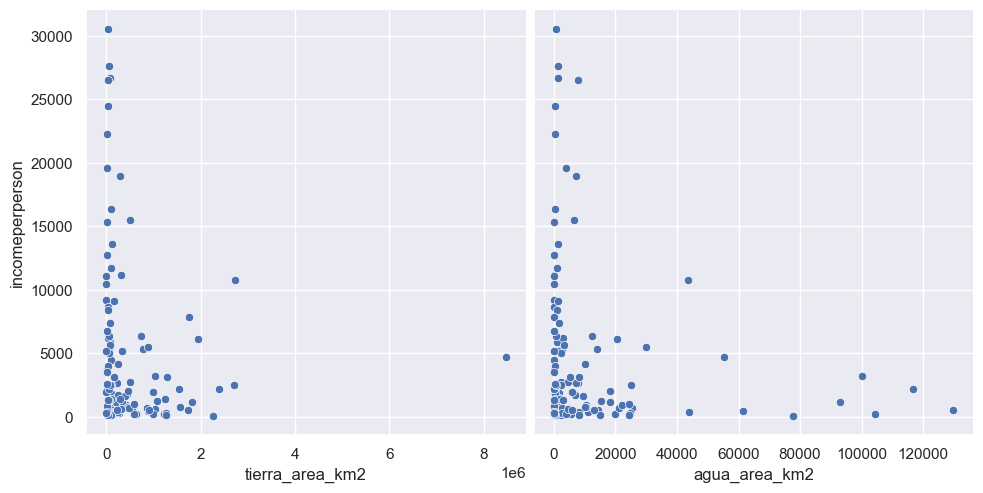
df\_visualizar= df\_visualizar.loc[df\_visualizar["relectricperperson"] <= df\_visualizar["relectricperperson"].quantile(.95)]

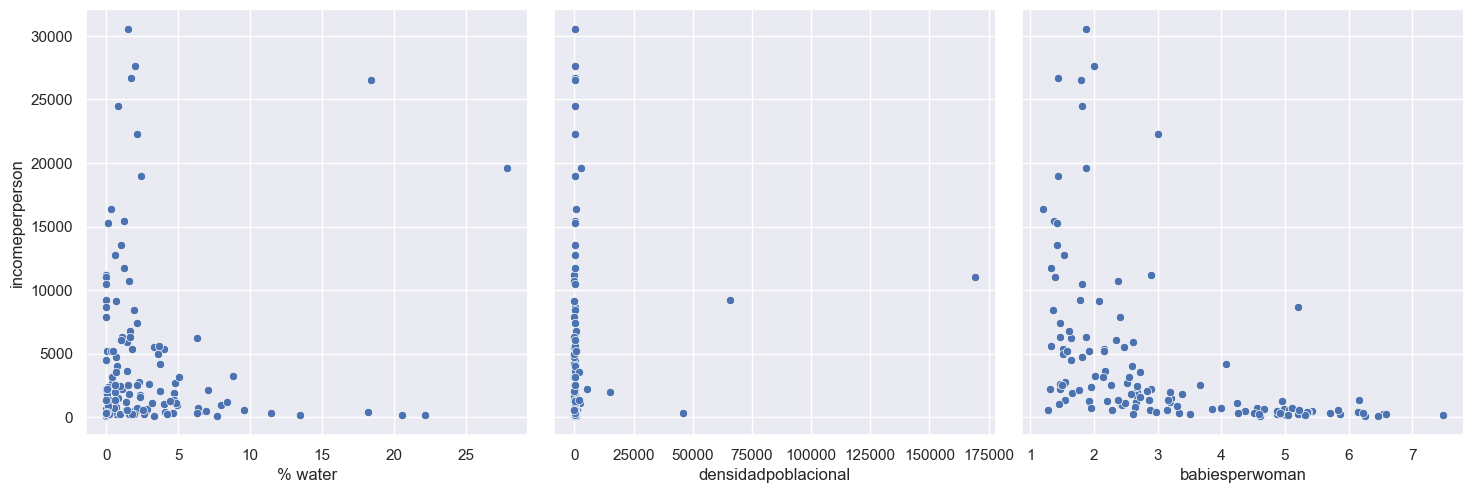
Lo cuál significa que se quitaron ciertos datos y el dataset quedó finalmente con un total de 131 países.



Los análisis de las variables nuevas fue el siguiente:







La única variable interesante es la variable babiesperwoman que tiene una tendencia lineal negativa con respecto a nuestra variable objetivo.

3. **[30%] Entrenamiento del modelo de regresión:** Utilizando los indicadores que considere más relevantes, además de las variables categóricas construidas, entrene un modelo de regresión lineal que prediga el PIB per cápita de un país. No olvide realizar las transformaciones de datos necesarias y la validación de supuestos para asegurar no solo un buen ajuste del modelo sino también una correcta interpretación de los coeficientes.

En primer lugar, para realizar la regresión con los datos ya limpios y pre-procesados, se decide seguir los siguientes pasos:

* Estandarización de variables.
* Creación de los datasets de entrenamiento y prueba.
* Escoger las variables más importantes para el modelo.
* Creación del modelo de regresión lineal.
* Análisis del modelo creado.
* Análisis del error del modelo en sus predicciones.

Se crean los datasets de prueba y entrenamiento:

X = df\_visualizar[variables\_x]

Y = df\_visualizar["incomeperperson"]

X\_train, X\_test, Y\_train, Y\_test = train\_test\_split(X, Y, test\_size=0.2, random\_state=101)

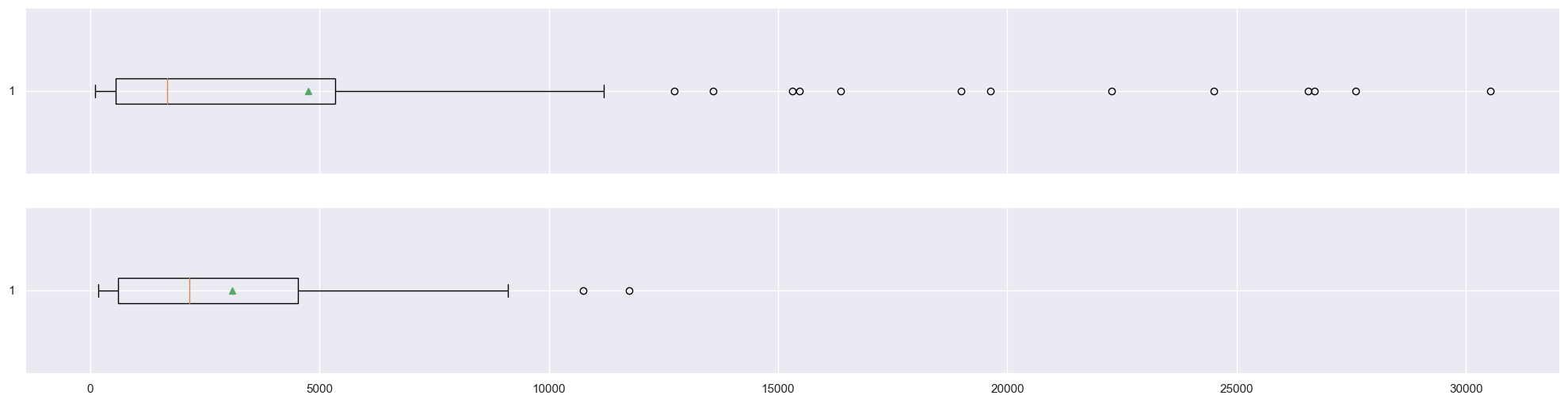
Se estandarizan los datos con la librería StandardScaler de la siguiente manera:

scaler = StandardScaler()

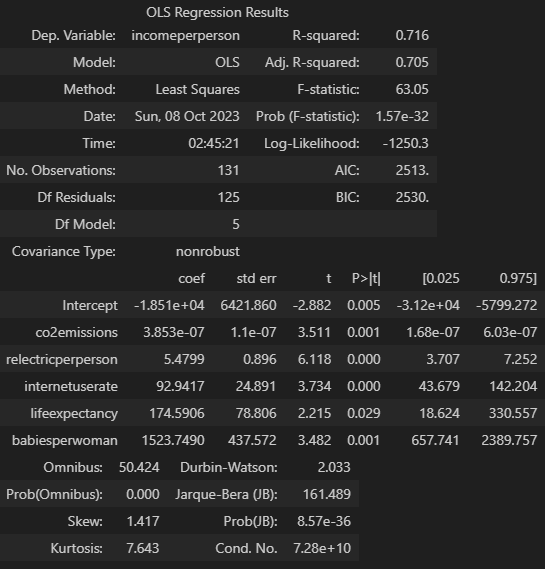
X\_train\_scaled = scaler.fit\_transform(X\_train)

X\_test\_scaled = scaler.transform(X\_test)

La distribución de la variable objetivo en ambos subdatasets fue la siguiente:



Ahora, para escoger las variables que se van a usar se realiza primero una construcción de un modelo OLS para ver la significancia estadística de las variables:



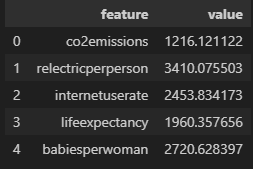
Vemos que tienen una alta significancia estadística las variables de internetuserate, co2emissions, relectricperperson,lifeexpectancy y babiesperwoman. Por lo tanto, estas cinco variables serán las escogidas para la construcción de la regresión lineal deseada. Por otro lado, se tiene que tener en cuenta que quizás una que otra variable como breastcancerper100th era muy significante pero debido a su alta colinealidad con la variable internetuserate se descartó, por lo tanto, se eligieron las cinco variables que podrían generar recomendaciones importantes y que cumplían con las hipótesis estadísticos.

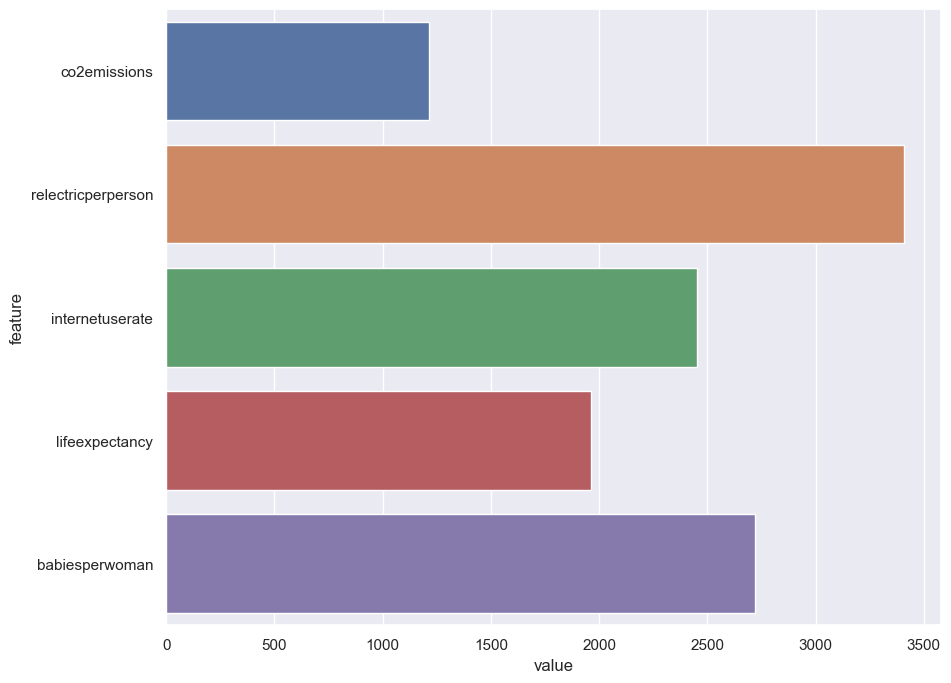
Luego, se crea el modelo correspondiente:

model = LinearRegression()

model.fit(X\_train\_scaled, Y\_train)

Posteriormente, se muestran los valores obtenidos de cada una de las variables:





En el siguiente paso después de tener el modelo creado se procede a evaluarlo para ver su desempeño con datos nuevos, entonces se calcula tanto el MAE como el RMSE para ambos conjuntos de datasets (training y test):

MAE train: 2368.641669695693

MAE test: 2346.061037295371

RMSE train: 3520.7988355234056

RMSE test: 2877.991820756402

Además, obtenemos el valor del R2 de la regresión:

R2 score: 0.7385780347033711

Finalmente, vamos a evaluar el error del modelo a la hora de predecir con nuevos datos:

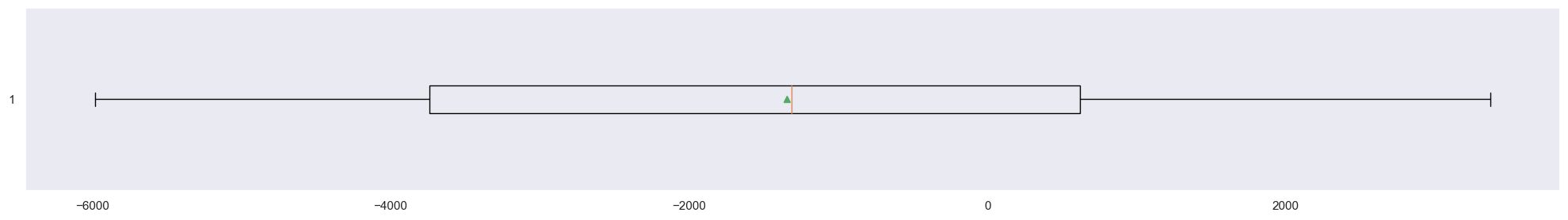
* Primero, vemos la distribución que tiene la diferencia entre la variable objetivo real – valor predicho:

plt.figure(figsize = (25, 3))

plt.boxplot((Y\_test - preds\_test), showmeans=True, vert=False)

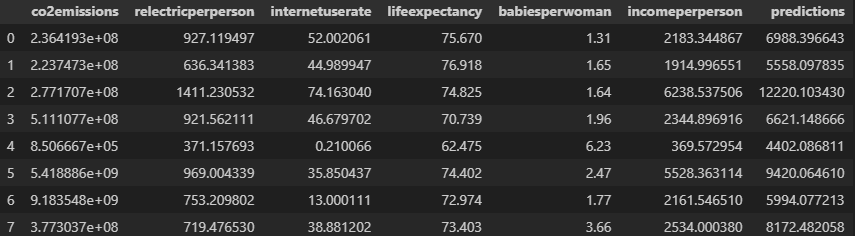
plt.grid()

plt.show()



Se puede observar que no suceden diferencias tan grandes en general y no hay outliers tan grandes. Además, que importante resaltar que su media y la mediana están casi en el mismo valor alrededor de -1300, que indicaría que en promedio las predicciones del PIB per cápita se pasan por un valor de 1300.

* Ahora definimos una condición que indique los peores casos, en este caso ponemos un valor de diferencia de 3500 PIB per cápita y solo hubo 8 casos de 27 posibles que cumplieron:



4. [**25%] Generación de resultados:** Redacte un informe ejecutivo o una presentación corta en donde muestre los hallazgos o insights más importantes en términos de entendimiento de los datos, métricas de error e interpretación del modelo. Responda a la pregunta: dado que un alto PIB puede representar un buen comportamiento de pago de un país, a partir de las relaciones más importantes que, según el modelo de regresión, existe entre el PIB y otros indicadores, ¿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?