EDA Satisfacción en la Unión Europea

1. Contexto

Para este proyecto he creado una situación hipotética en la que una empresa de Estados Unidos, a la que he llamado Better Life, Inc. se pone en contacto conmigo y me encarga que analice un conjunto de datos de la satisfacción en la UE. Esta empresa presta servicios en Estados Unidos y en Canadá a personas que no están del todo satisfechas con su vida, les ayuda a reconducirla desde diferentes aspectos, como pueden ser el aspecto laboral, las relaciones con otras personas, desde pareja, familia, amigos, etc. y otro tipo de aspectos como cambiar los hábitos perjudiciales que no llenan a las personas por otros beneficiosos tanto para su salud física como para su salud mental. De esta forma, la empresa ayuda a las personas a estar más satisfechas con sus vidas, lo cual va muy ligado a la salud mental.

Con la problemática de la salud mental a nivel mundial, Better Life, Inc. ha experimentado un gran crecimiento en las regiones en las que tiene presencia y está buscando expandirse a Europa, por este motivo, me ha contratado para que analice unos datos de una encuesta realizada por la Unión Europea que recoge la satisfacción con la vida en general de su población. Esta información se presenta por grupos de población en función de su edad, género, nivel de estudios, país y año (de 2021 a 2024). Me ha facilitado además unos datos que contienen la media y la mediana de los ingresos, también por grupos, pero esta vez sin tener en cuenta el nivel de estudios y otros que contienen el porcentaje de la población de cada país que no puede hacer frente a los gastos inesperados. Me ha pedido que junte todos los datos y los analice conjuntamente.

2. Hipótesis

Con esta situación ficticia, las hipótesis que me he fijado son las siguientes:

- Los jóvenes están menos satisfechos con su vida que las personas de mediana edad.
- Existe una fuerte correlación positiva entre la satisfacción y el nivel de ingresos.
- ¿Está más relacionado con la satisfacción la edad o el nivel de ingresos?
- ¿Afecta la estabilidad económica a la satisfacción de la población? (Por países)
- ¿Hay países, y dentro de esos países grupos de edad concretos, con buenas condiciones económicas, pero aun así poco satisfechos?

3. Análisis univariante

Lo primero que he hecho ha sido rellenar una tabla con información sobre las variables:

Columna/Variable	Descripción	Tipo_de_Variable	Importancia inicial	Nota
isced11	El nombre en castellano es Clasificación Internacional Uniforme de la Educación (CINE)	Categórica	2	Mostraré una leyenda con las categorías más abajo. En principio menos importante porque voy a centrarme más en la edad y los ingresos
genero	El género de los participantes: masculino, femenino o ambos	Categórica	2	En principio menos importante porque voy a centrarme más en la edad y los ingresos
edad	Grupos de edad de los participantes	Categórica	1	Permite analizar la satisfacción por la edad, una de las preguntas a las que busco dar respuesta
edad_ingresos	Columna que he creado yo para añadir los datos de ingresos al dataset de satisfacción			Me la quedo por si necesito consultarla, pero no la tendré en cuenta en el análisis
geo	Entidad geopolítica a la que corresponden los datos, son todo países menos uno que es "EU-27"	Categórica	1	Permite analizar la satisfacción según el país, importante, ya que busco quedarme con un país concreto y dentro de este país buscar un grupo objetivo
periodo	Año al que corresponden los datos: 2021-2024	Categórica	1	Los datos que me más me interesan son los de 2024, por ser los más actuales, pero los datos de los demás años me permitirán ver la evolución temporal
satisfaccion_general	Satisfacción de los participantes con la vida en general en una escala del 1 al 10 por nivel de estudios, edad y género. El valor que se presenta en esta columna es el valor medio de los participantes que pertenecen al mismo país, nivel de estudios, edad y género en el mismo año.	Numérica discreta	0	Es la variable central que da sentido a todo el análisis. Importante saber que son valores medios de los grupos que responden a la encuesta, no resultados por individuos aislados
media_ingresos	Media de ingresos en euros por edad y género (no tengo datos que tengan en cuenta el nivel de estudios)	Numérica continua	2	De momento no le doy gran importancia porque en principio voy a usar la mediana, que es más precisa
mediana_ingresos	Mediana de ingresos en euros por edad y género (no tengo datos que tengan en cuenta el nivel de estudios)	Numérica continua	0	Columna clave, ya que estoy buscando una buena oportunidad de mercado
no_afrontar_imprevistos(%)	Porcentaje de la población de cada país que no puede hacer frente a gastos inesperados	Numérica discreta/continua	1	Aunque no tenga información desglosada por grupos de población y solo del país en general, puede ser muy útil a la hora de decidir en qué país quiero establecer mi filial, ya que si hay una gran parte de la población que no puede hacer frente a gastos inesperados, probablemente no podrá permitirse tampoco un gasto extra para mejorar su satisfacción

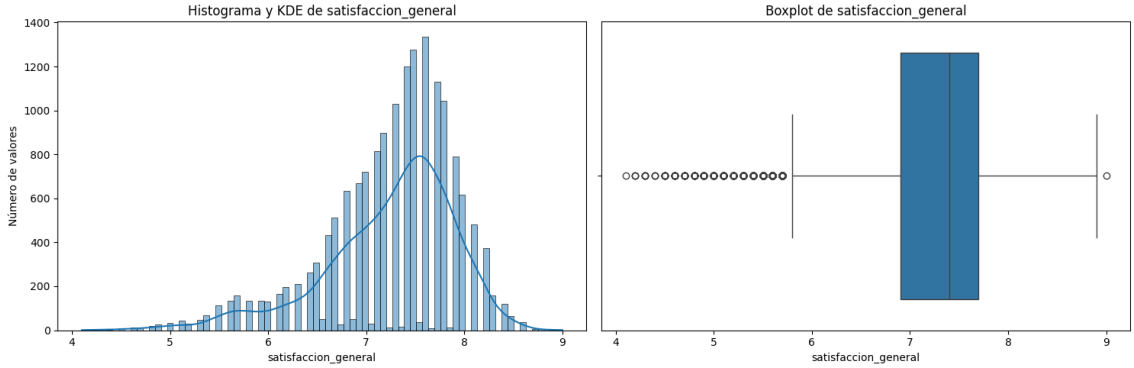
El caso de este dataset es muy particular. Al tener la satisfacción general en forma de valor medio para cada grupo determinado de personas en cada año, tenemos un número de valores muy similar para todas las categorías de las categóricas (con alguna diferencia por los registros nulos que hemos eliminado) y esto hace que no podamos obtener información del análisis univariante. Cuando podremos sacarles partido a estas variables será al analizar las variables numéricas en función de ellas.

A continuación, paso a las variables numéricas y lo primero que hago es visualizar una descripción de los datos de todas ellas:

	count	mean	min	25%	50%	75%	max	std	IQR	range	CV
satisfaccion_general	16860.0	7.242952	4.1	6.900	7.4	7.7	9.0	0.677186	0.800	4.9	0.093496
mediana_ingresos	16860.0	18858.729537	2426.0	9788.625	17560.0	26528.0	59956.0	10808.851002	16739.375	57530.0	0.573148
media_ingresos	16860.0	21124.499051	3218.0	10970.000	19764.0	29234.0	68404.0	11893.154362	18264.000	65186.0	0.563003
no_afrontar_imprevistos(%)	16860.0	29.795883	14.7	22.700	29.4	34.9	47.9	8.810363	12.200	33.2	0.295691

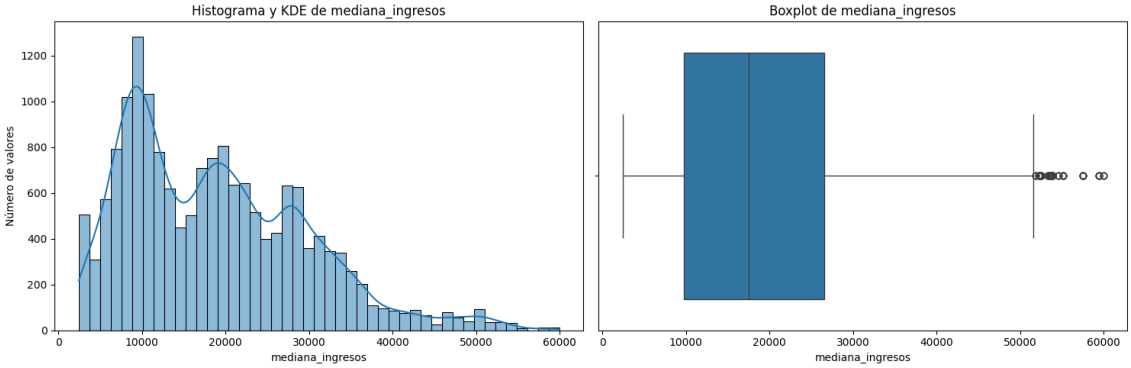
Las visualizaciones y conclusiones más relevantes de este análisis han sido:

Satisfacción general:



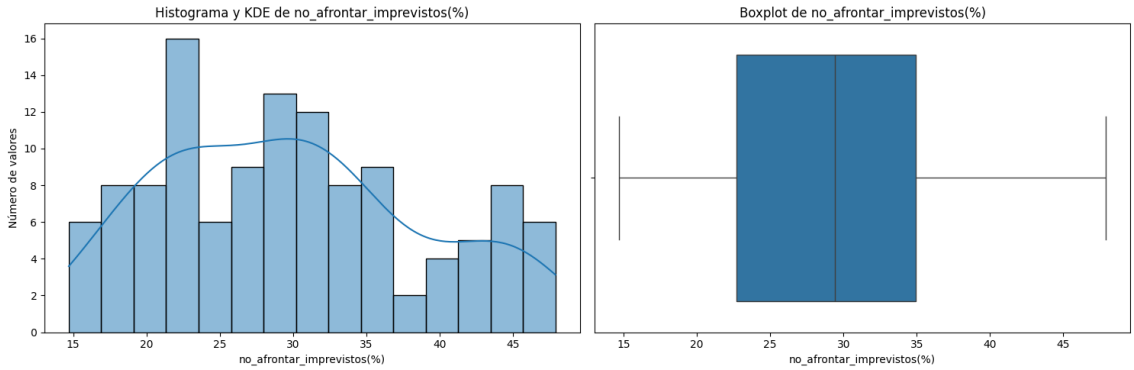
Es una variable que cuenta con bastante concentración de valores en la zona entre el 7 y el 8 y luego tenemos alguno outliers en la zona baja y uno en la zona alta. En general los valores que más nos interesan son los de la satisfacción considerada "baja", para definir cuáles son estos valores vamos a utilizar la mediana y el bigote. Vamos a crear una nueva columna categórica "categoria_satisfaccion" con los valores "baja" entre la mediana y el final del bigote, "muy baja" para los outliers y "alta" para el resto. Los valores que más nos interesan son los de satisfacción baja. Los de satisfacción alta nos interesan menos porque probablemente no necesites servicios para mejorarla y los de la satisfacción demasiado baja, al estar fuertemente ligada a factores económicos, probablemente tendrán también una economía que no sea muy atractiva para nosotros. De todas formas, más adelante analizaremos las categorías baja y muy baja en función de factores económicos.

Mediana ingresos:



La distribución es muy similar a la de "mediana_ingresos", ya que miden la misma magnitud, aunque la medida sea diferente, cabe destacar que los valores de "media_ingresos" son en general más altos que los de "mediana_ingresos", esto probablemente se debe a la dispersión de las muestras que se utilizaron para medir estas magnitudes, cuyos outliers por la parte alta hicieron que creciera la media. En nuestras variables pasa algo similar y es que, si nos fijamos en la tabla de funciones de agrupación, vemos que tanto para "media_ingresos" como para "mediana_ingresos" la media es mayor que la mediana y esto es debido a que tenemos outliers que destacan muy por arriba e inflan la media.

Personas de cada país que no tienen capacidad económica para afrontar gastos inesperados:

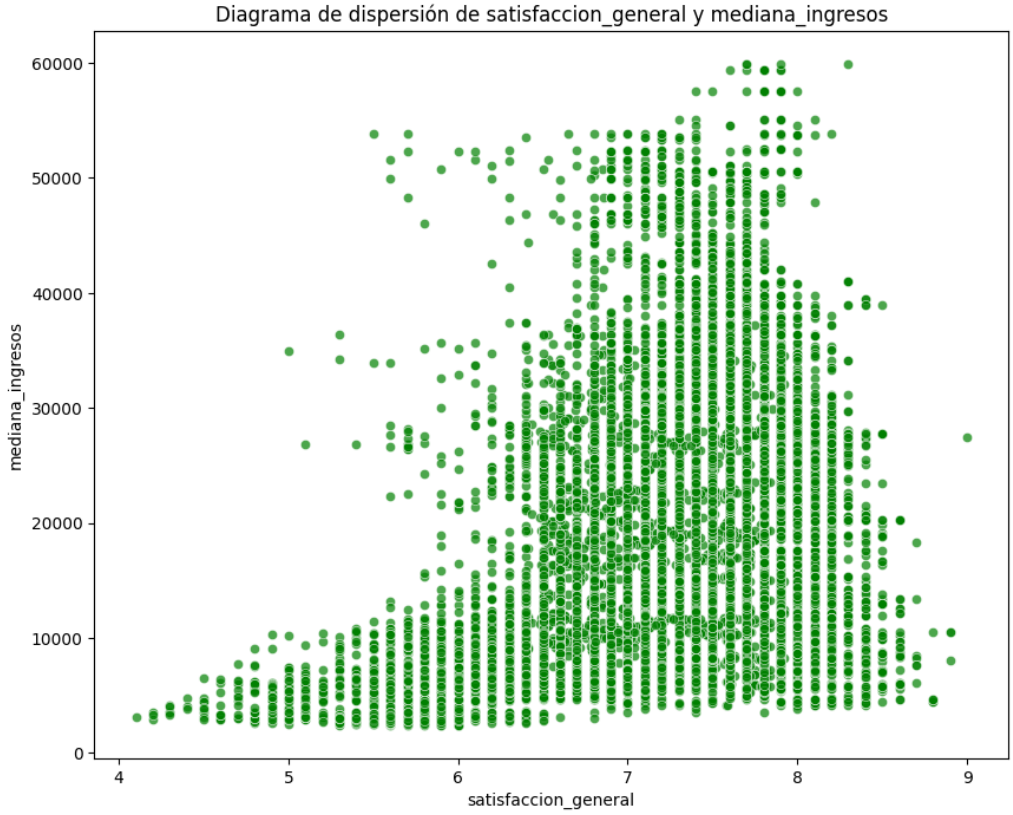


Vemos que es una variable con bastante dispersión, aunque presenta una acumulación algo mayor en la parte baja. Esta parte es la que más nos interesa, ya que cuanta más población pueda hacer frente a gastos imprevistos, también será mayor la cantidad de gente que pueda permitirse los servicios que ofrece la empresa. Siempre hay que tener en cuenta los ingresos, ya que, si mucha gente puede permitirse hacer frente a gastos inesperados, pero los ingresos son muy bajos, quizá la rentabilidad no va a ser muy alta. Como se puede apreciar en el diagrama de cajas, la mediana está más o menos en el centro del rango de valores que tenemos y, al no ser un rango demasiado grande, vamos a crear una columna categórica únicamente con las categorías "bajo" y "alto" utilizando la mediana como punto de referencia.

4. Análisis bivariante y multivariante

Visualizaciones y conclusiones interesantes del análisis bivariante y multivariante:

Satisfacción general y mediana de ingresos:

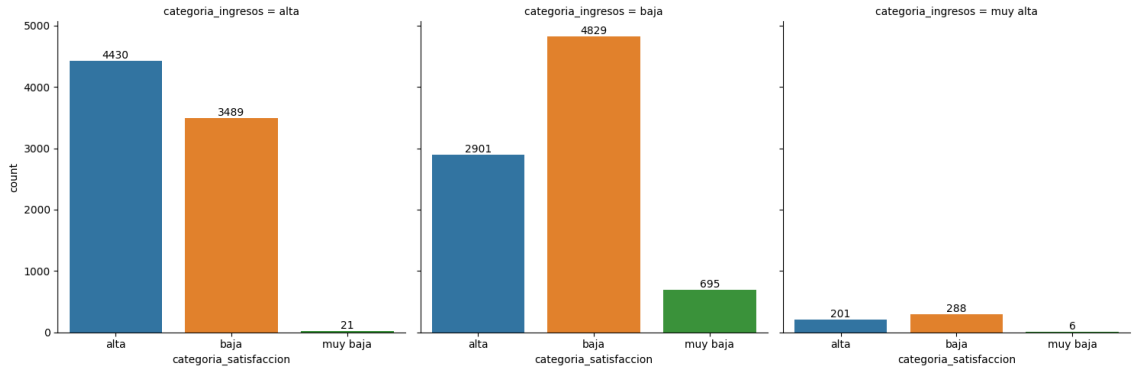


Con la visualización de los histogramas y diagramas de cajas de "satisfaccion_general" y "mediana_ingresos" ya se puede observar que la distribución de ambas es muy diferente y que los valores de "mediana_ingresos" se agrupan mucho más en la parte baja que los de "satisfaccion_general", por lo que podemos suponer que no hay una fuerte correlación lineal positiva entre estas dos variables, en todo caso negativa. Con el diagrama de dispersión comprobamos que existe una cierta correlación positiva, ya que los valores de ingresos tienden a crecer a medida que aumenta la satisfacción, aunque no parece demasiado fuerte. Si bien se observa una línea cuyos ingresos crecen de manera clara a medida que aumenta su satisfacción (y algunas otras líneas con esta tendencia, pero no se ven tan claramente por la acumulación de valores) y para los valores más bajos de satisfacción los valores de ingresos son también bajos, al seguir avanzando por el eje x, en el que se representa la satisfacción, vemos que hay valores de ingresos que no crecen mucho, estos son los valores que menos nos interesan. Los resultados más atractivos, por el contrario, son aquellos cuyos ingresos son altos, pero su satisfacción se mantiene baja, es decir, quedan por encima de la línea que mencionaba antes. Vamos a comprobar la correlación entre las dos variables con el coeficiente de Pearson.

	satisfaccion_general	mediana_ingresos
satisfaccion_general	1.000000	0.329677
mediana_ingresos	0.329677	1.000000

El valor que nos da este coeficiente muestra lo que ya anticipábamos al ver el diagrama de dispersión, existe correlación entre las dos variables, pero no es muy fuerte.

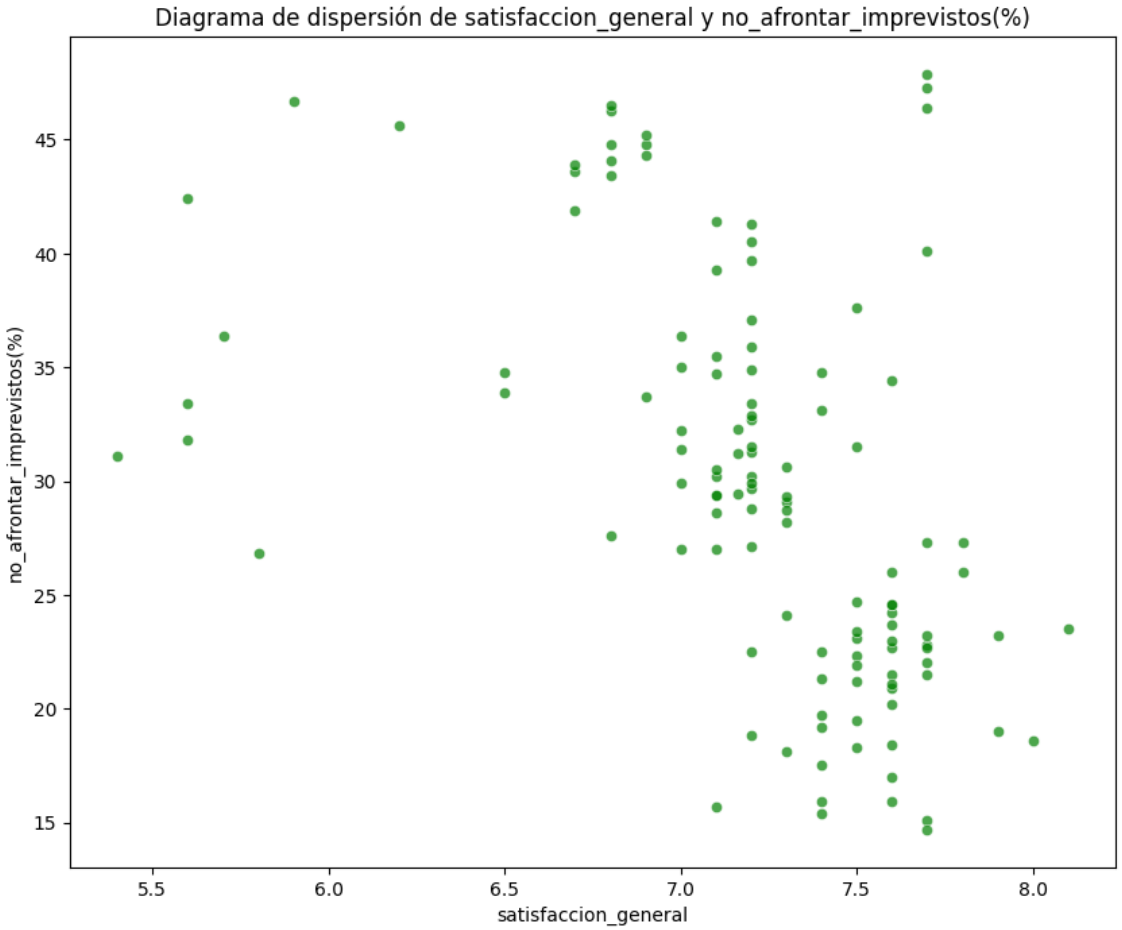
Análisis de las categóricas creadas a partir de estas dos variables:



Ya empezábamos a encontrar resultados potentes al analizar el diagrama de dispersión que nos indicaban que había grupos cuyos ingresos eran elevados y su satisfacción no lo era tanto, pero con estos gráficos de barras y la tabla de contingencia nos queda más claro todavía. En la categoría de ingresos muy alta, aunque tenemos pocos valores, podemos ver que ya mayoría de los grupos tienen una satisfacción baja y algunos, muy pocos, incluso muy baja. Estos son los grupos más interesantes. También en la categoría de ingresos alta tenemos un número abundante de grupos con satisfacción baja. Más adelante será interesante ver si hay algún país en el que se concentren en mayor medida

estos grupos a la hora de ver cuál es el país idóneo para establecer la filial de la empresa. Vamos a analizar ahora la capacidad de los ciudadanos de cada país para afrontar gastos inesperados, que es otro indicador que nos ayuda a determinar si existen países en los que la satisfacción no está tan ligada al aspecto económico.

Satisfacción general y dificultad para afrontar imprevistos:



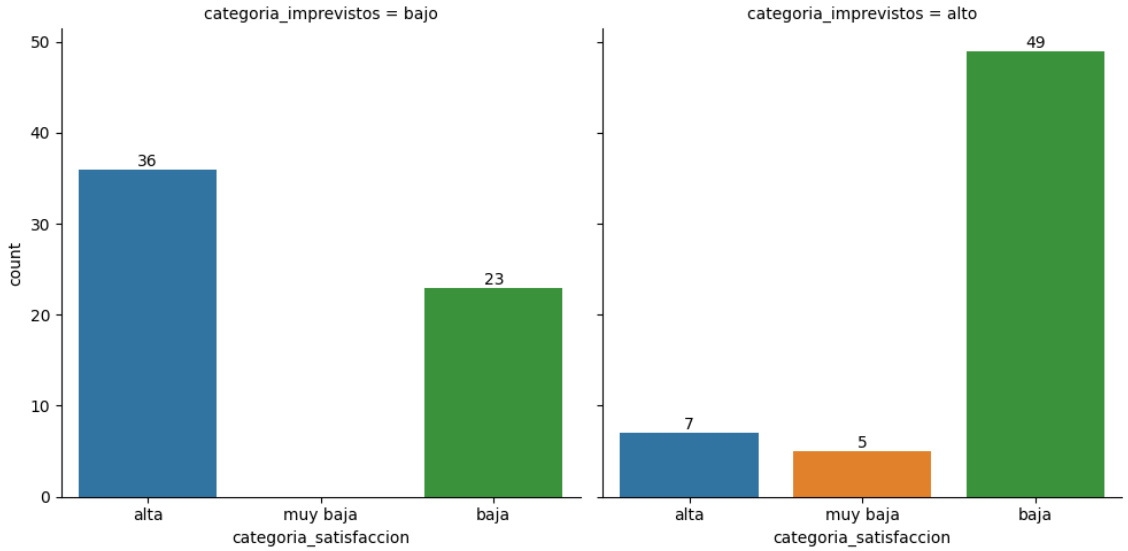
Con la visualización individual del histograma y KDE, así como del diagrama de caja de las variables "satisfaccion_general" y "no_afrontar_imprevistos(%)" podemos ver que sus distribuciones no son muy parecidas, siendo más dispersa y uniforme la variable "no_afrontar_imprevistos(%)", aunque sus valores tienen una mayor concentración en la parte baja. Esto nos hace pensar que la correlación entre estas dos variables no será demasiado fuerte, igual que en el caso anterior, y que es posible que exista una correlación lineal negativa, lo cual sería esperable, es decir, la satisfacción aumenta a medida que disminuye la incapacidad para hacer frente a imprevistos financieros. Al observar el diagrama de dispersión, podemos apreciar que existe una relación lineal negativa que se confirma esta hipótesis: ninguno de los países con valores más bajos de incapacidad para hacer frente a gastos inesperados tiene valores de satisfacción extremadamente bajos, aunque estos países no tienen tampoco los valores más altos. Esto es algo positivo de cara a nuestro análisis, ya que hay países en los que la población, según esta estadística podría permitirse los servicios de la empresa y que pueden

también necesitarlos, por no tener unos niveles de satisfacción muy altos. Para comprobar que existe esta correlación vamos a ver su coeficiente de Pearson:

	satisfaccion_general	no_afrontar_imprevistos(%)
satisfaccion_general	1.000000	-0.481568
no_afrontar_imprevistos(%)	-0.481568	1.000000

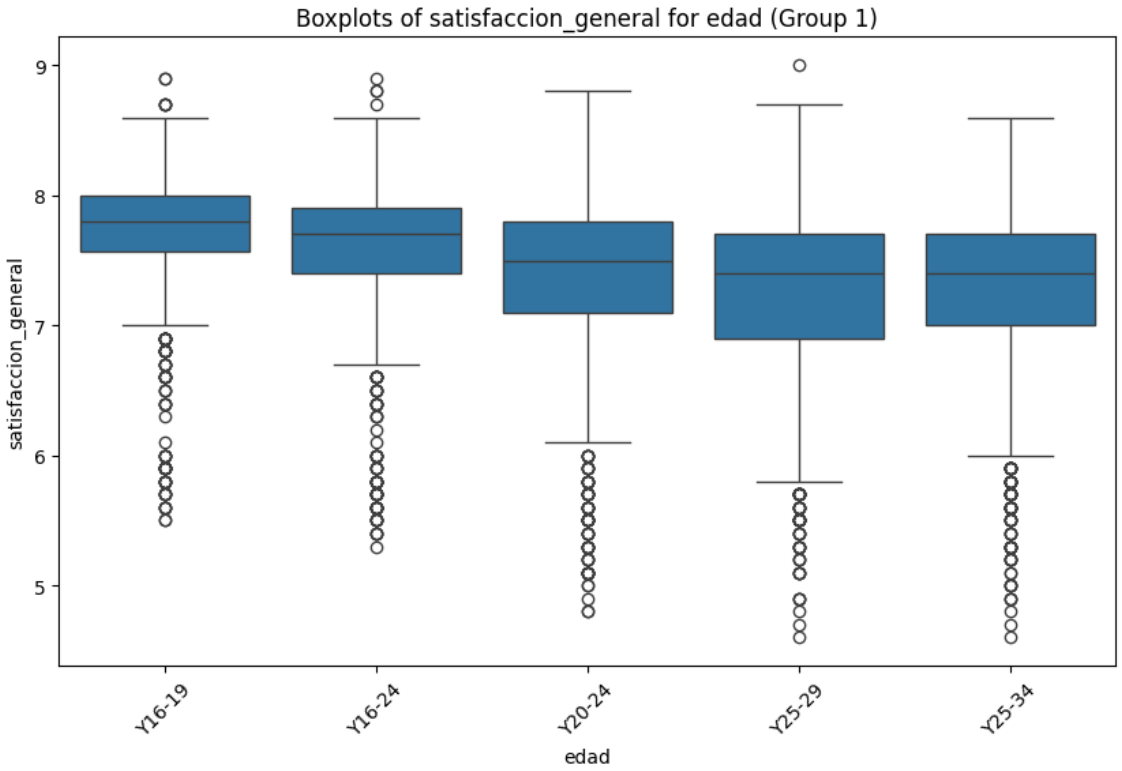
El valor que nos da este coeficiente muestra lo que ya anticipábamos al ver el diagrama de dispersión, existe correlación lineal negativa entre las dos variables.

Tabla con las variables categóricas creadas a partir de estas numéricas:



En esta tabla ya vemos más claramente que tenemos un grupo de país-año (aunque el año que nos interesa sea 2024) en el que puede estar nuestro público objetivo, con un porcentaje bajo de personas que no pueden permitirse los gastos inesperados y una satisfacción baja (ninguno muy baja).

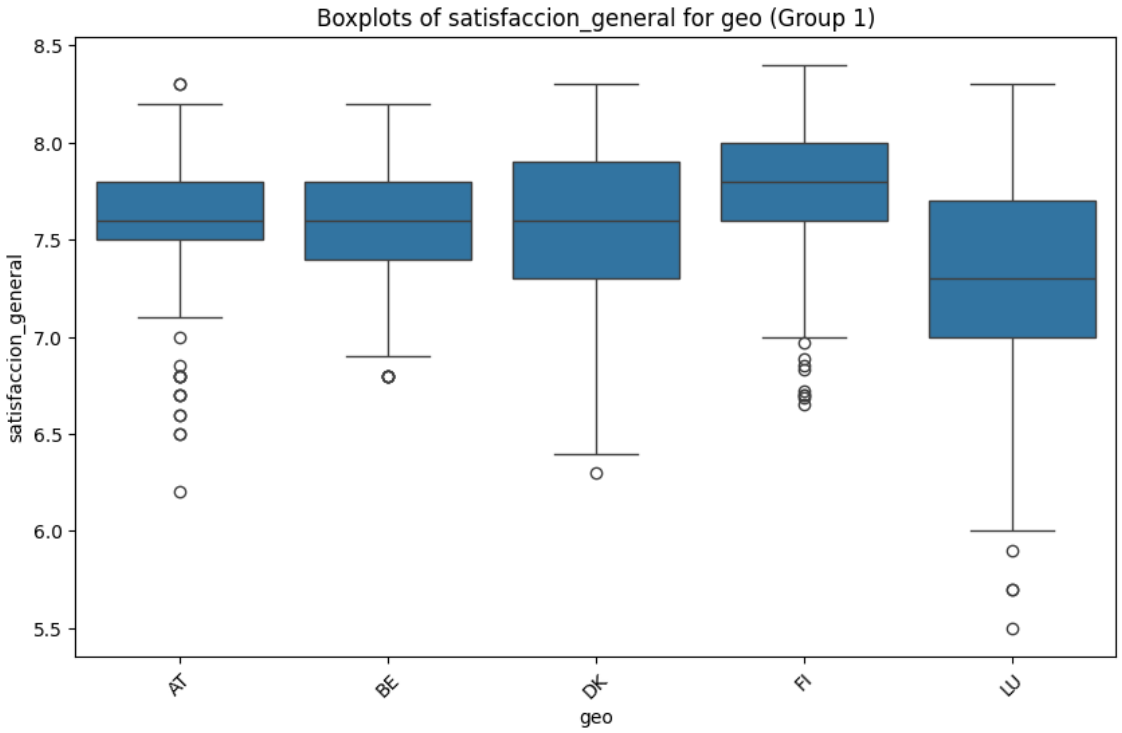
Edad y satisfacción general:

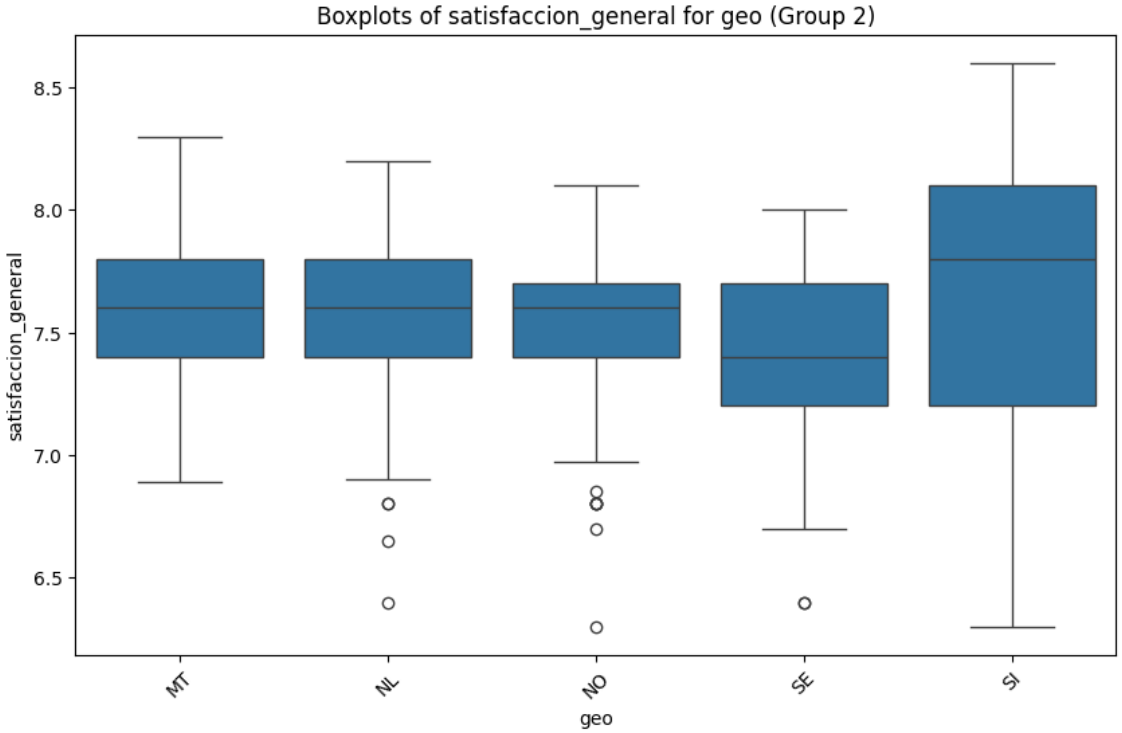


Aquí ya solo tenemos los datos que nos interesan para descartar nuestra hipótesis y parece claro que cuanto más joven es la población mayor es su satisfacción, los valores son también menos dispersos y se concentran todos en la parte alta, no tienen ningún outlier excesivamente bajo.

País y satisfacción general:

País, a pesar de ser una variable categórica tiene 30 valores diferentes, por lo que, para acotar, directamente vamos a quedarnos con países con un porcentaje de población bajo que no puede afrontar imprevistos y con ingresos altos (es decir, los países que nos interesan desde el punto de vista económico) y vamos a ver su satisfacción en 2024.

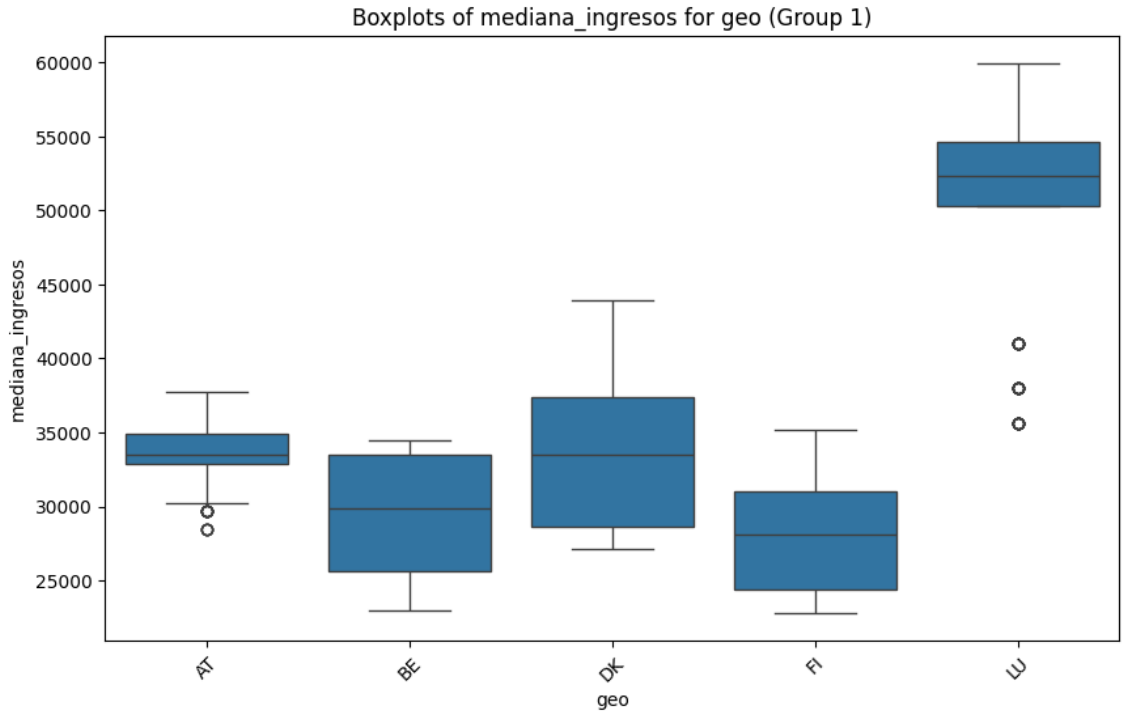


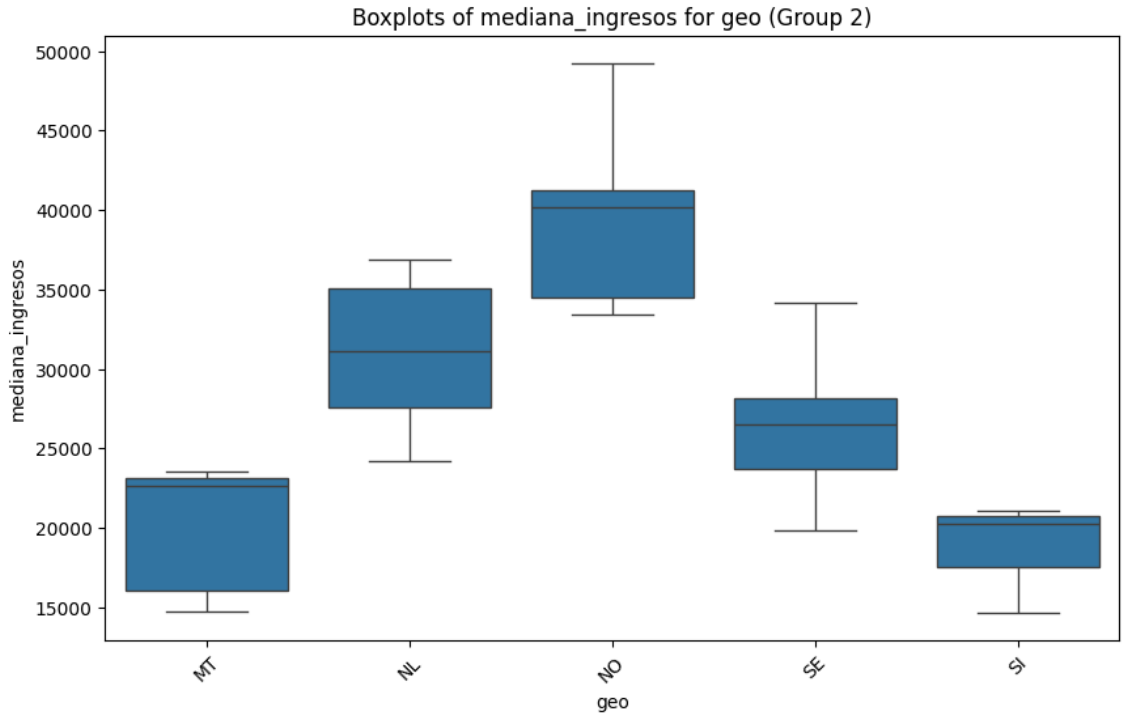


Con la visualización de los diagramas de caja de este grupo seleccionado de países con una buena situación desde el punto de vista económico obtenemos unos resultados muy potentes. A priori podría parecernos que los más interesantes son los que tienen la categoría de satisfacción baja, pero al visualizar la dispersión de los valores vemos que es así en el caso de Luxemburgo, ya que es el que tiene los ingresos más altos y tiene una dispersión de valores grande, por lo que tiene grupos de población con una satisfacción bastante baja. Sin embargo, debido a su dispersión de valores de

satisfacción, Dinamarca (categoría satisfacción: alta) y Eslovenia (categoría satisfacción: alta) pueden ser mejores candidatos para establecer una filial que Suecia (categoría satisfacción: baja), ya que los primeros tienen un rango de satisfacción que llega hasta más abajo y el último está menos disperso y no tiene valores tan bajos.

País y mediana de ingresos:

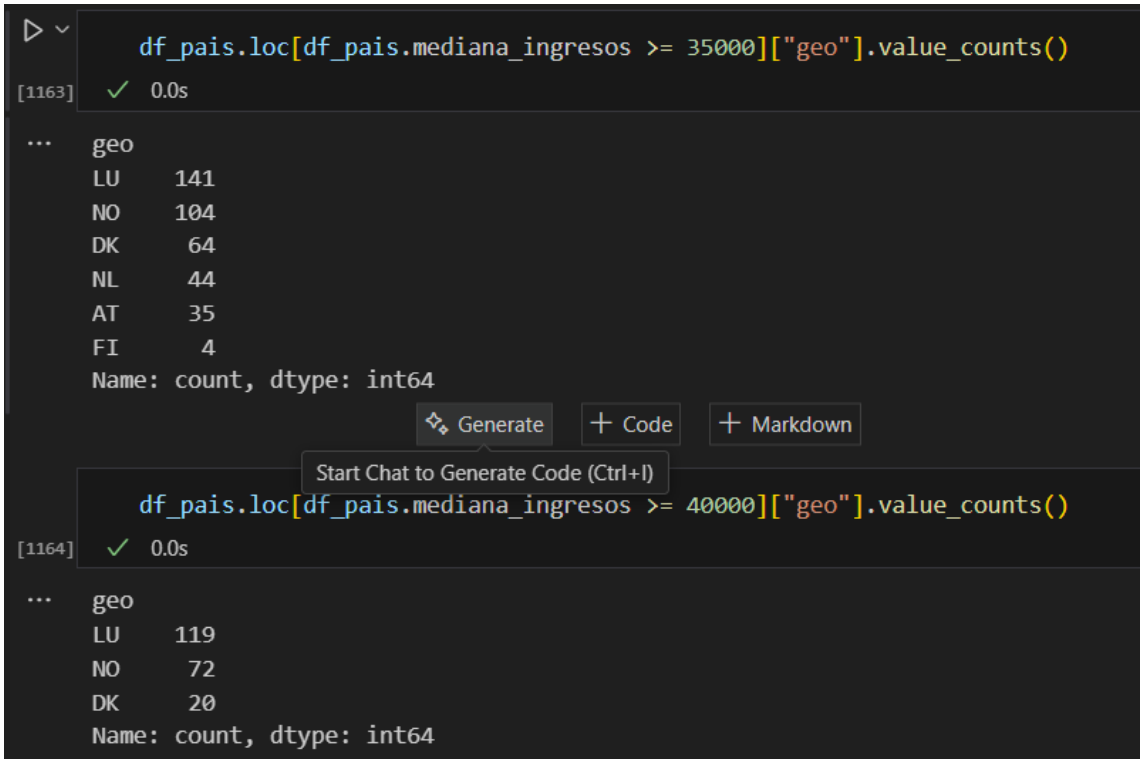




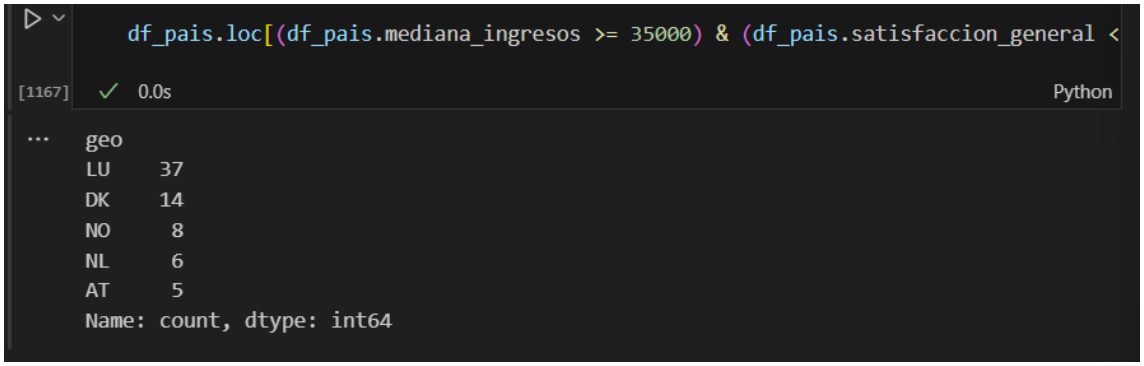
La variable "mediana_ingresos" en función del país es en general una variable poco dispersa y con pocos outliers, lo que quiere decir que los valores de cada país están concentrados. Los países con los valores más interesantes en cuanto a ingresos son,

por este orden, Luxemburgo, Noruega, Dinamarca, Países Bajos y Austria. Al ver que sus ingresos son bastante bajos en relación con los demás países de la lista, podemos descartar a Eslovenia. Si no encontrásemos un candidato mejor, Austria y Países Bajos pueden ser buenos candidatos, ya que también tienen grupos con valores bajos de satisfacción.

Comprobamos lo observado con números:



Viendo los datos numéricos parece claro que tenemos dos candidatos claros al país idóneo para que la empresa abra su filial, ya que Luxemburgo y Dinamarca son dos países con bastantes valores de satisfacción baja y además ambos están en el top 3 de países con ingresos más altos. Ahora falta ver cuántos de los grupos con satisfacción baja se encuentran entre los que tienen unos ingresos altos. Aunque podemos anticipar que si consideramos como umbral de ingresos los 35000€ en Luxemburgo todos los grupos con satisfacción baja van a estar entre los de ingresos altos.



Viendo los datos, el segundo mejor candidato podría estar entre Noruega y Dinamarca, ya que Noruega tiene también bastantes valores de satisfacción baja y con ingresos altos, pero claramente es Luxemburgo el país que más tiene.

Dentro de Luxemburgo vemos los grupos que menor satisfacción tienen:

	isced11	genero	edad	edad_ingresos	periodo	satisfaccion_general
14775	TOTAL	M	Y35-49	Y25-49	2024-01-01	6.8
13335	TOTAL	F	Y35-49	Y25-49	2024-01-01	6.9
16215	TOTAL	T	Y35-49	Y25-49	2024-01-01	6.9
13455	TOTAL	F	Y50-64	Y50-64	2024-01-01	7.0
14535	TOTAL	M	Y25-34	Y25-49	2024-01-01	7.1
13095	TOTAL	F	Y25-34	Y25-49	2024-01-01	7.1
12855	TOTAL	F	Y20-24	Y16-24	2024-01-01	7.1
12975	TOTAL	F	Y25-29	Y25-49	2024-01-01	7.1
15975	TOTAL	T	Y25-34	Y25-49	2024-01-01	7.1

Los tres grupos que están entre los 35 y los 49 años son los que menor satisfacción tienen. Como además sus ingresos son altos, este será el principal grupo objetivo de la empresa Better Life en Europa: los habitantes de Luxemburgo entre 35 y 49 años.

5. Conclusiones

Este dataset tiene muchos datos interesantes, que permiten llegar a conclusiones muy útiles y que nos ha permitido responder las hipótesis iniciales que habíamos planteado y encontrar un lugar para que Better Life pueda establecerse e iniciar su andadura en Europa.

Tanto las variables económicas como la edad están relacionadas con la satisfacción de la población en Europa. La que tiene una relación más fuerte con la satisfacción es la edad.

Las variables de carácter financiero tienen cierta relación con la satisfacción, pero no es demasiado fuerte y es precisamente eso lo que abre oportunidades para la empresa, encontrando grupos de población que puedan necesitar sus servicios y que a su vez hagan posible una rentabilidad alta.

El país en el que se concentran más grupos de este tipo es Luxemburgo, un país pequeño, pero con unos ingresos y un nivel de vida muy altos, aunque según nuestros datos no son proporcionales a la satisfacción de sus habitantes.