PRÁCTICA 2

1. Descripción del dataset. ¿Por qué es importante y qué pregunta/problema pretende responder?

El dataset en cuestión, contiene los datos correspondientes a las cifras de población de las comunidades autónomas y provincias españolas, junto a la total. Además, diferencia el sexo, lugar de procedencia, grupo quinquenal de edades, periodo de los datos y pobación que corresponde.

Vistos uno a uno los distintos campos del dataset:

- Edad: variable categórica que indica los rangos quinquenales de edad. Desde “De 0 a 4 años” hasta “100 y más años”.

- Sexo: variable categórica que indica el sexo de la población que corresponda. “Mujer”, “Hombre” o “Total”.

- Periodo: variable numérica que indica el año al que corresponde la medida de la población. Desde 2015 hasta 2020.

- Lugar: variable categórica que indica la procedencia donde se está realizando la medida. Desde el Total, comunidades y ciudades autónomas como Aragón o Melilla, y las provincias españolas.

- Nacimiento: variable categórica que indica el lugar de procedencia de la población que se está midiendo.

- Total: variable numérica que indica la población que corresponde.

Estos datos han sido recopilados del Instituto Nacional de Estadística de España y nos servirán para ver el envejecimiento en España, junto con las pirámides poblacionales para comparar población en cuanto a términos de edad y de localizaciones.

Una pirámide poblacional representa las características de una población en una localización concreta en función de rangos quinquenales de edad, para ver la distribución y comprar momentos y lugares distintos.

Con los datos de los que se parte se podrían hacer un montón de medidas y comparativas, pero nosotros nos vamos a centrar en el envejecimiento de España y las pirámides poblacionales de ciertas provincias y comunidades, más y menos pobladas, y la total par hacer estudios demográficos y equilibrio entre sexos.

Las preguntas a las que responderemos serán:

- ¿Es España un país viejo?
- ¿Hay equilibrio entre sexos en provincias más pobladas o menos pobladas?

2. Integración y selección de los datos de interés a analizar.

Para poder realizar el análisis que se ha detallado anteriormente, se van a leer los datos del fichero “population_spain_dataset.csv”, generado en la práctica anterior, y convertir en un dataframe para un manejo mejor de los datos:

```
df = pd.read_csv ('population_spain_dataset.csv')
df.tail()
```

	Lugar	Nacimiento	Edad	Sexo	Periodo	Total
3763579	52 Melilla	Resto de Países de Oceanía	100 y más años	Mujeres	2019	0
3763580	52 Melilla	Resto de Países de Oceanía	100 y más años	Mujeres	2018	0
3763581	52 Melilla	Resto de Países de Oceanía	100 y más años	Mujeres	2017	0
3763582	52 Melilla	Resto de Países de Oceanía	100 y más años	Mujeres	2016	0
3763583	52 Melilla	Resto de Países de Oceanía	100 y más años	Mujeres	2015	0

Una vez se han cargado en un dataframe, como en este caso no se va a utilizar el dato del origen de procedencia, se van a seleccionar las que correspondan a “Todas” las procedencias y se van a eliminar las demás, además de eliminar dicha columna una vez filtradas dichas filas:

```
df = df.loc[df['Nacimiento'] == 'Total']
del(df['Nacimiento'])
```

Como en los campos de ‘Sexo’ tenemos “Total”, “Hombre” y “Mujer” y el resultado de la suma de hombres y mujeres corresponde al total, se van a eliminar las filas que tengan ese total:

```
df = df.loc[df['Sexo'] != 'Total']
```

Con el campo de “edad” pasa algo similar, se tienen los grupos quinquenales además de un campo que tiene “todas las edades” que vamos a filtrar y quedarnos con el resto:

```
df = df.loc[df['Edad'] != 'Todas las edades']
# también hay algún espacio al principio, lo quitamos
df['Edad'] = df['Edad'].str.strip()
```

Con el campo de ‘Lugar’ debemos pensar que está el “total nacional”, todas las comunidades y las provincias. Nos vamos a quedar solamente con los datos de las provincias y el total nacional, ya que las pirámides de población la haremos de alguna provincia, cuando se quiera hacer de una comunidad se escogen las provincias que correspondan a cada comunidad y el total nacional se va a conservar ya que es solamente un dato y para calcular el envejecimiento se hará mucho más rápido en una sola consulta:

```
comun = ['01 Andalucía', '02 Aragón', '03 Asturias, Principado de', '04 Balears, Illes', '05
Canarias', '06 Cantabria', '07 Castilla y León',
'08 Castilla - La Mancha', '09 Cataluña', '10 Comunitat Valenciana', '11 Extremadura', '12
Galicia', '13 Madrid, Comunidad de', '14 Murcia, Región de',
'15 Navarra, Comunidad Foral de', '16 País Vasco', '17 Rioja, La', '18 Ceuta', '19 Melilla']
df = df.loc[~df['Lugar'].isin(comun)]
```

Se eliminan los números y espacios que haya en los nombres:

```
df['Lugar'] = df['Lugar'].str.strip()
```

Ahora se tienen los datos que se van a analizar posteriormente en los siguientes apartados.

3. Limpieza de los datos.

En este apartado, a partir del dataframe preparado con los datos a analizar que se ha realizado anteriormente, se va a ver si el dataset contiene elementos vacíos, ceros o valores extremos.

3.1. ¿Los datos contienen ceros o elementos vacíos? ¿Cómo gestionarías cada uno de estos casos?

En este caso, para saber si el dataframe contiene elementos vacíos se utilizan funciones de pandas que dan dicha información:

```
df.isna().sum()
```

```
Lugar      0
Edad       0
Sexo       0
Periodo    0
Total      0
dtype: int64
```

Se puede ver como no existen elementos vacíos dentro del set de datos. De haber datos vacíos, solamente tendría sentido en el campo de población ya que el resto son variables definidas previamente que son categóricas. El que un dato tenga el campo de población total vacío, indica que no se tiene sabiduría de dicho dato y que no podría obtenerse mediante ninguno de los métodos que conocemos, ya que no tendría sentido utilizar el método de los vecinos cercanos para obtener o predecir dicho valor por ejemplo. Por esto, se tendría que prescindir de dicho valor.

En cuanto a elementos nulos pasa algo similar que en el caso anterior pero en este caso no se tendría que prescindir, ya que el valor nulo en el campo de población sí que nos da información. El que el campo población sea 0, indica que para un lugar en concreto, un periodo, un grupo quinquenal y un sexo concreto no existe población, por lo que no se deben de eliminar.

3.2. Identificación y tratamiento de valores extremos.

De la misma forma que en el caso anterior, el que exista un valor extremo no se debe de tratar como algo a eliminar ya que tiene sentido para las conclusiones que se deben de sacar.

No existen valores extremos en este caso, pero de haberlos se deberían de dejar porque indican que para un lugar en concreto, un periodo, un grupo quinquenal y un sexo concreto existe un pico de población.

4. Análisis de los datos.

4.1. Selección de los grupos de datos que se quieren analizar/comparar (planificación de los análisis a aplicar).

Para este trabajo, se ha decidido seleccionar la provincia de Madrid, la provincia de Zaragoza, la comunidad autónoma de Extremadura y los datos de España para calcular las pirámides de población correspondientes y el envejecimiento.

```
# Madrid 2020
df_madrid = df.loc[(df['Lugar']=='Madrid') & (df['Periodo']==2020)]
df_mad = pd.DataFrame({'Edad': list(df_madrid.loc[(df_madrid['Sexo']=='Mujeres')]['Edad']),
                       'Hombres': list(df_madrid.loc[(df_madrid['Sexo']=='Hombres')]['Total']),
                       'Mujeres': list(df_madrid.loc[(df_madrid['Sexo']=='Mujeres')]['Total'])})


# Zaragoza 2020
df_zaragoza = df.loc[(df['Lugar']=='Zaragoza') & (df['Periodo']==2020)]
df_zgz = pd.DataFrame({'Edad':
list(df_zaragoza.loc[(df_zaragoza['Sexo']=='Mujeres')]['Edad']),
                       'Hombres': list(df_zaragoza.loc[(df_zaragoza['Sexo']=='Hombres')]['Total']),
                       'Mujeres': list(df_zaragoza.loc[(df_zaragoza['Sexo']=='Mujeres')]['Total'])})
```

```
# Extremadura 2020
df_extremadura = df.loc[((df['Lugar']=='Cáceres') | (df['Lugar']=='Badajoz')) &
(df['Periodo']==2020)]
edades = list(df_extremadura.loc[(df_extremadura['Sexo']=='Mujeres') &
(df_extremadura['Lugar']=='Badajoz')]['Edad'])
total_mujeres_ext = df_extremadura.loc[(df_extremadura['Sexo']=='Mujeres') &
(df_extremadura['Lugar'].isin(['Badajoz', 'Cáceres']))]
total_hombres_ext = df_extremadura.loc[(df_extremadura['Sexo']=='Hombres') &
(df_extremadura['Lugar'].isin(['Badajoz', 'Cáceres']))]
mujeres_ex = []
hombres_ex = []
for edad in edades:
   mujeres_ex.append(total_mujeres_ext.loc[total_mujeres_ext['Lugar'].isin(['Cáceres',
'Badajoz']) & (total_mujeres_ext['Edad'] == edad)]['Total'].sum())
   hombres_ex.append(total_hombres_ext.loc[total_hombres_ext['Lugar'].isin(['Cáceres',
'Badajoz']) & (total_hombres_ext['Edad'] == edad)]['Total'].sum())
df_ext = pd.DataFrame({'Edad': edades,
                'Hombres': hombres_ex,
                'Mujeres': mujeres_ex)
```

Para el caso de españa se va a aplicar la fórmula del envejecimiento, que corresponde con la proporción de mayores de 64 años= ((Población >64 años / Población total) x100), por lo que los datos serán:

```
# España 2015-2020
df_espana = df.loc[(df['Lugar']=='Total Nacional')]
periods = list(df_espana['Periodo'].unique())
edades_mas_64 = ['De 65 a 69 años', 'De 70 a 74 años', 'De 75 a 79 años', 'De 80 a 84
años', 'De 85 a 89 años', 'De 90 a 94 años', 'De 95 a 99 años', '100 y más años']
dict_env = {}
for per in periods:
        total_pob = df_espana.loc[df_espana['Periodo'] == per]['Total'].sum()
        total_pob_mas_64 = df_espana.loc[df_espana['Edad'].isin(edades_mas_64) &
(df_espana['Periodo'] == per)]['Total'].sum()
        env = (total_pob_mas_64/total_pob)*100
        dict_env[per] = env
# dataframe pirámide 2020
df_esp = pd.DataFrame({'Edad': list(df_espana.loc[(df_espana['Sexo']=='Mujeres') &
(df_espana['Periodo'] == 2020)]['Edad']),
                'Hombres': list(df_espana.loc[(df_espana['Sexo']=='Hombres') &
(df_espana['Periodo'] == 2020)]['Total']),
                'Mujeres': list(df_espana.loc[(df_espana['Sexo']=='Mujeres') &
(df_espana['Periodo'] == 2020)]['Total'])})
```

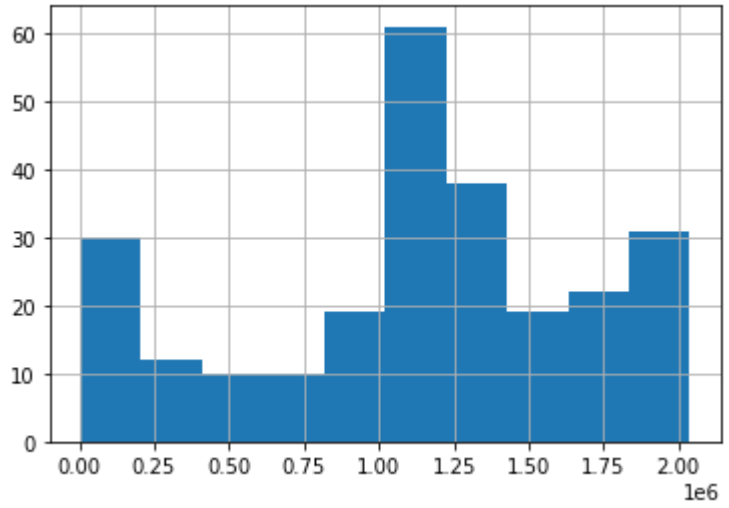
En este punto, ya tenemos los datos preparados para realizar el estudio.

4.2. Comprobación de la normalidad y homogeneidad de la varianza.

Como comparar todos los datos no sería correcto, ya que tenemos desde provincias pequeñas hasta el total de España, hemos decidido comprobar la normalidad para los datos que corresponden al total de España.

Para comprobar la normalidad de los datos, pintamos un histograma:

df_espana['Total'].hist()



Se puede ver como los datos de población totales para el caso de España no siguen una distribución normal.

La media y la varianza del conjunto son:

df_espana['Total'].describe()

```
count    2.520000e+02
mean     1.114895e+06
std      5.720857e+05
min      2.975000e+03
25%      8.507642e+05
50%      1.179642e+06
75%      1.507398e+06
max      2.037837e+06
Name: Total, dtype: float64
```

Como se ve tras usar el método describe(), la media total de España es 1.11e+06, mientras que la varianza tiene un valor de 5.72e05

4.3. Aplicación de pruebas estadísticas para comparar los grupos de datos. En función de los datos y el objetivo del estudio, aplicar pruebas de contraste de hipótesis, correlaciones, regresiones, etc. Aplicar al menos tres métodos de análisis diferentes.

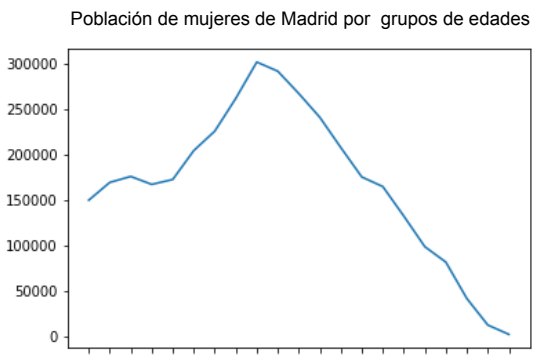
El primer método de análisis que realizaremos será un estudio estadístico de las pirámides poblaciones de la comunidad de Madrid (muy poblada) con la comunidad de Extramadura (menos poblada).

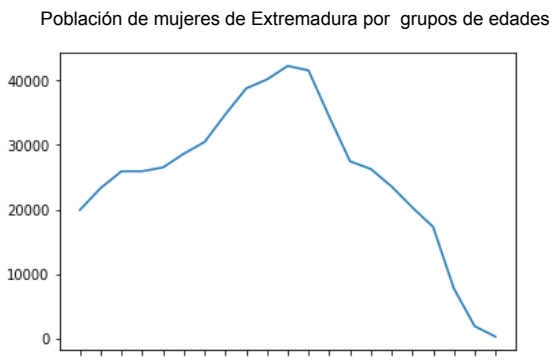
df_madrid['Total'].describe()
df_extremadura['Total'].describe()

	Madrid	**Extremadura**
mean	161425.904762	12666.511905
std	85693.977597	7250.616108
min	431	30
25%	107255	8138.25
50%	173438.5	12453.5
75%	215262.25	17005.5
max	301202	26499

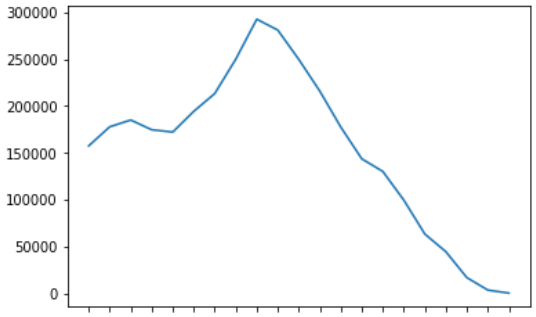
Podemos ver que tanto en las variables de media, min y max nos indican que la población de Madrid es muy superior a la población de Extremadura, lo cual nos sirve para realizar un análisis posterior comparando estos dos tipos de localidades.

El siguiente análisis que vamos a hacer es comparar las poblaciones de Madrid y Extremadura por grupos de edades para ver qué tipo de gráficas hacen tanto en mujeres como en hombres.

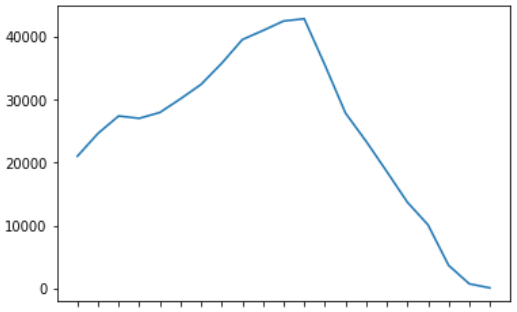




Población de hombre de Madrid por grupos de edades



Población de hombre de Extremadura por grupos de edades

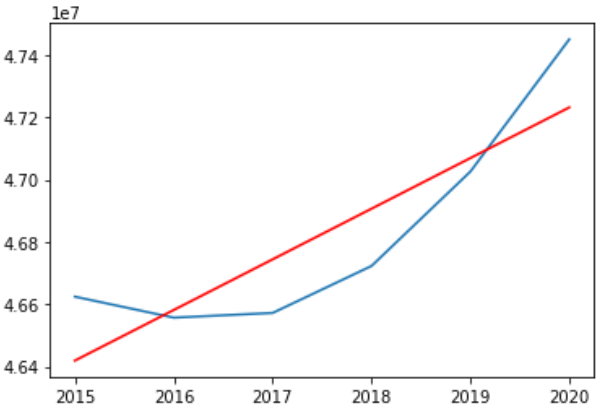


	Madrid (Mujeres)	Extremadura (Mujeres)	Madrid (Hombres)	Extremadura (Hombres)
Madrid (Mujeres)	1	0.96640392	0.98386117	0.97778231
Extremadura (Mujeres)	0.96640392	1	0.92147595	0.98104048
Madrid (Hombres)	0.98386117	0.92147595	1	0.95992307
Extremadura (Hombres)	0.97778231	0.98104048	0.95992307	1

Podemos ver que en todos los casos encontramos una correlación muy alta, con un valor muy cercano a 1, lo cual nos indica que todas las pirámides poblacionales son muy similares entre sí, independientemente del sexo o la región que estemos evaluando.

Por último, vamos a realizar un análisis de regresión sobre los datos de la población total de España por años. Queremos analizar si la tendencia que se tiene es creciente o decreciente con una regresión lineal.

Regresión lineal de población en España por años

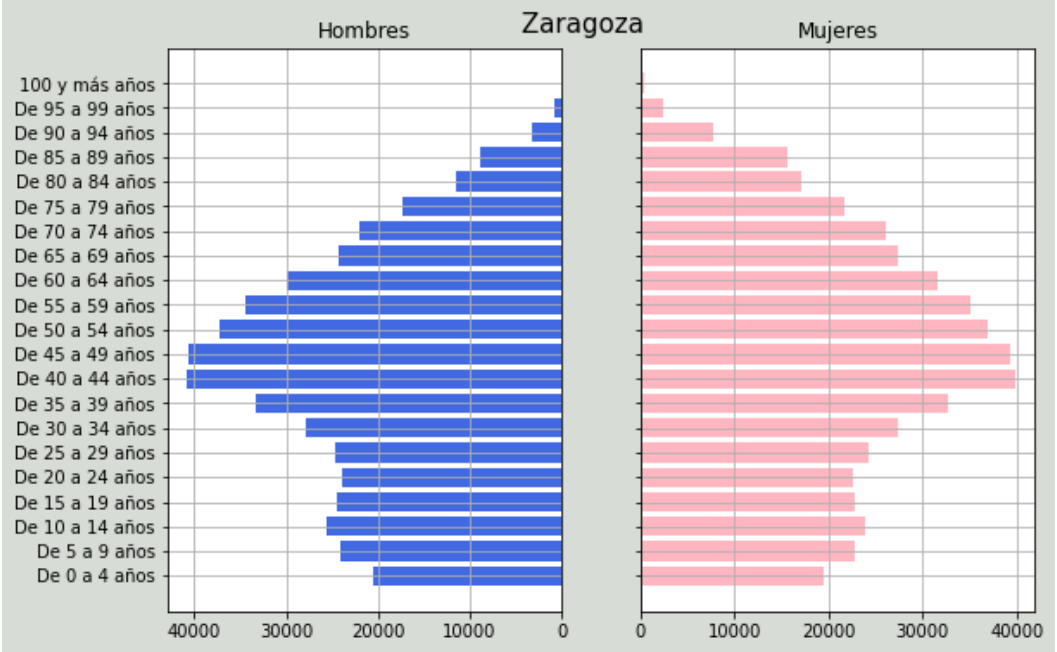


Tras aplicar la regresión lineal, podemos observar que la población española va en aumento.

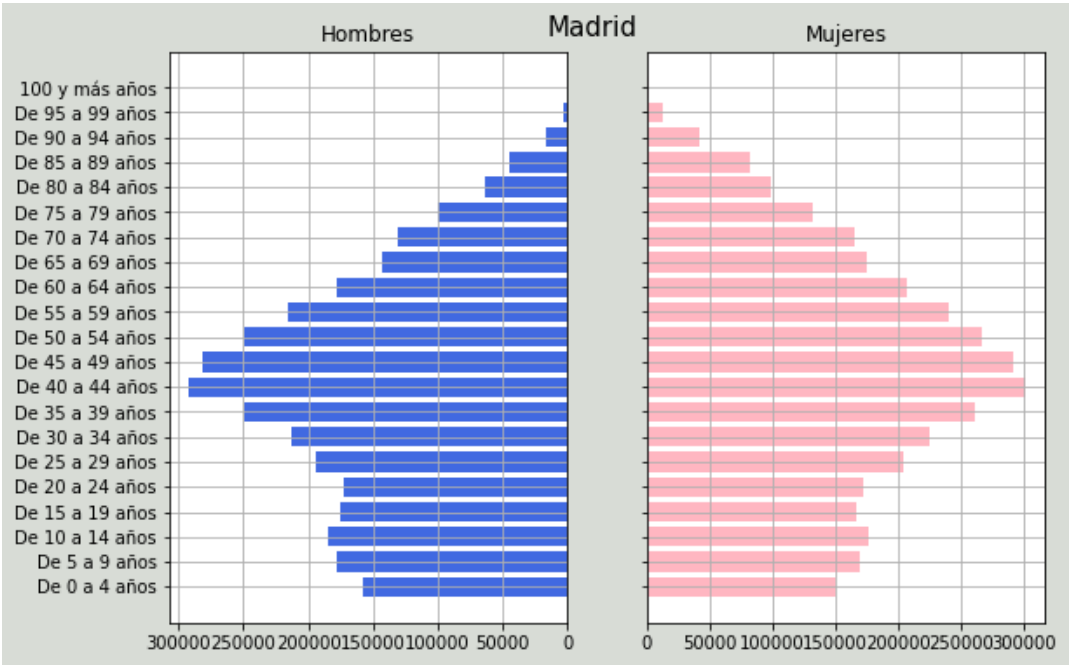
5. Representación de los resultados a partir de tablas y gráficas.

En este punto se van a representar los datos anteriormente seleccionados:

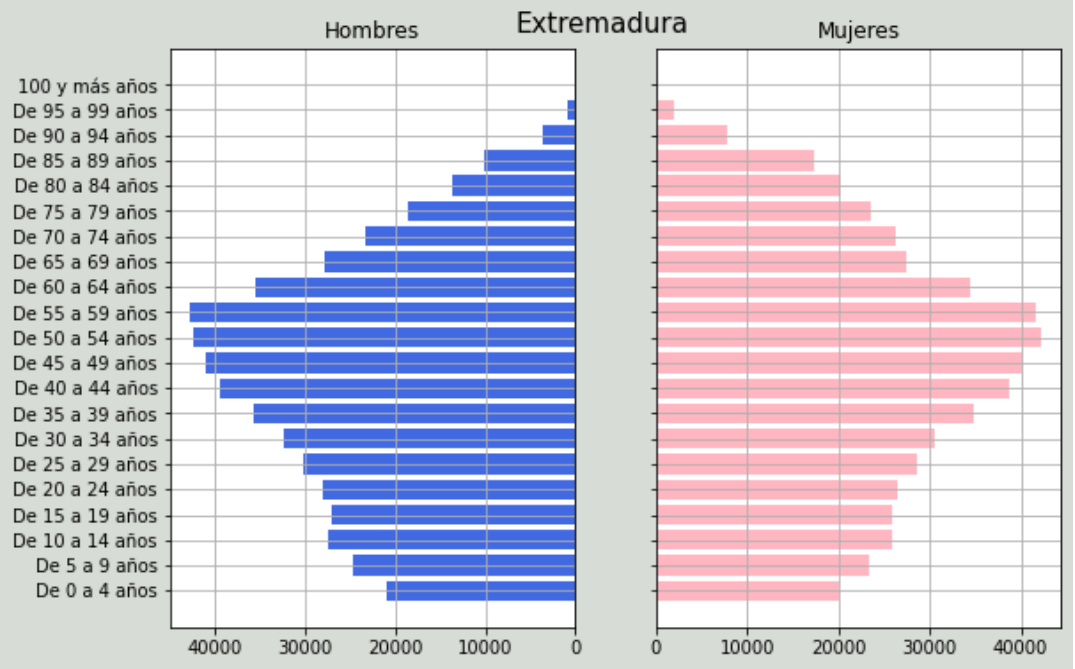
- Pirámide poblacional de la provincia de Zaragoza en 2020



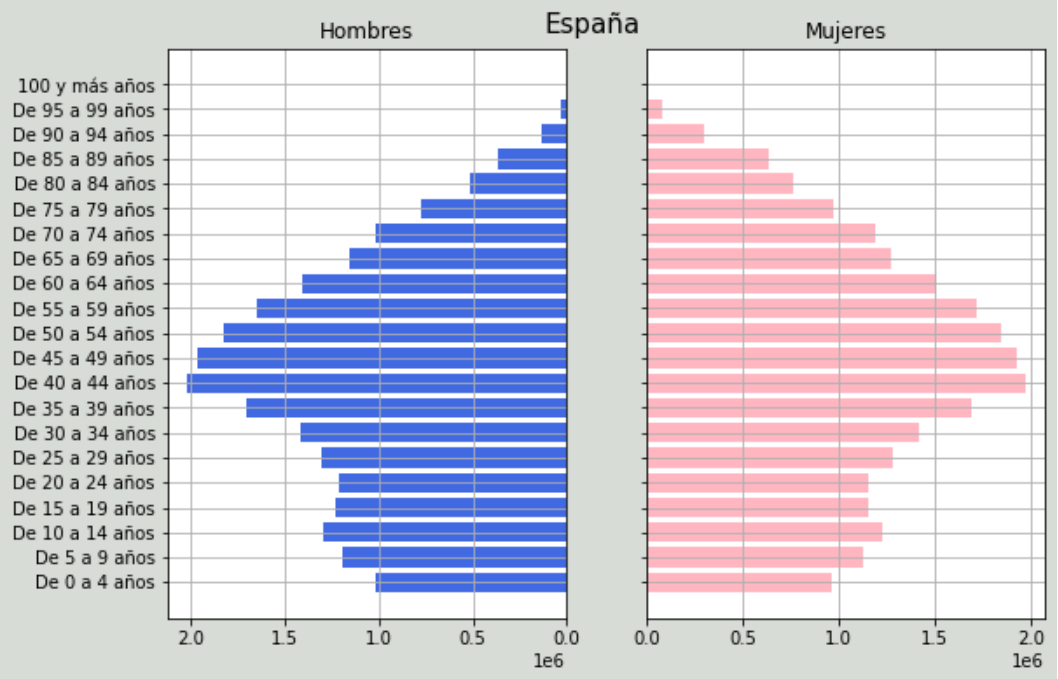
- Pirámide poblacional de la provincia de Madrid (Comunidad autónoma) en 2020



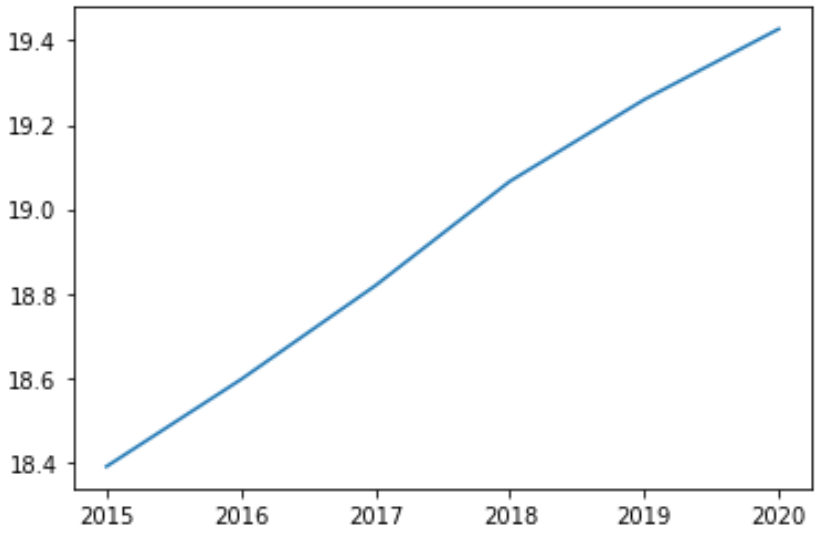
- Pirámide poblacional de la Comunidad autónoma de Extremadura en 2020



- Pirámide poblacional de España en 2020



- Envejecimiento de España de 2015 a 2020 medido en porcentaje:



6. Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema?

Vistas las gráficas anteriores donde se representaban las pirámides de población que corresponden a los distintos lugares podemos concluir que:

- Todas la pirámides pintadas anteriormente tienen la misma forma tanto si es una provincia, comunidad o país. Se puede ver que la mayoría de la población se encuentra en edades comprendidas entre los 35 y 60 años tanto en hombres como en mujeres de una forma similar, lo único que cambia es la cantidad total de habitantes de cada una de las localizaciones.

- En la gráfica que corresponde al envejecimiento, se puede ver como el porcentaje va de un 18,4% en 2015 a un 19,4% en 2020. Recordemos que en envejecimiento es el porcentaje de población de más de 64 años en función de la total, con los datos anteriores se puede ver que con el paso de los años el porcentaje de envejecimiento es mayor.

Por último, dando respuesta a las preguntas que se han planteado en el ejercicio 1:

- ¿Es España un país viejo?
  Sí. En los últimos 5 años, el porcentaje de envejecimiento ha ido en aumento pasando de un 18.4% hasta un 19,4%, lo que supone un aumento de un punto porcentual. Podríamos decir entonces que España es un país viejo y sigue una tendencia ascendente en cuanto a este tema.

- ¿Hay equilibrio entre sexos en provincias más pobladas o menos pobladas?
  Sí, podemos ver que en las pirámides poblacionales mostradas encontramos unas formas parecidas, sin destacar algún cambio en función del sexo. Vemos también que independientemente de si la provincia está más o menos poblada, encontramos una forma regresiva en las pirámides de población.

7. Código

Adjuntado en el fichero clean_data.py dentro de github.

Anexo Contribuciones

Investigación previa	M.N.P.G, M.S.P
Redacción de las respuestas	M.N.P.G, M.S.P
Desarrollo del código	M.N.P.G, M.S.P