**RETO: Proyecto de Data Science en Python - Caso: Análisis de la Facturación de Agua en 3 zonas de Donosti | Amara3, Orioventa1 y Orioventa5**

**Descripción del proyecto:**

Por medio de este análisis de los datos intentaremos responder a preguntas específicas relacionadas al tema de la base de datos, el cual nos muestra datos de 3 zonas de la ciudad de Donosti (Amara y Orioventa) y el consumo de agua en sus respectivas casas. En concreto, el dataset nos muestra las zonas de Amara 3 y dos zonas de Orioventa, (Orioventa 1 y Orioventa 5) y a partir del análisis de estos datos podremos extraer algunas conclusiones acerca de cada zona.

Primero llevaremos a cabo la fase de preparación y la filtración de los datos para comparar estas tres zonas de Donosti y a continuación, haremos cálculos matemáticos por medio de fórmulas estadísticas conocidas para las medidas de posición (media, moda, quartiles) y medidas de dispersión (la desviación estándar). Haremos uso de la librería Numpy para todo ello.

Finalmente, visualizaremos los datos mediante gráficos que nos proporcionan las librerías Pandas, MatplotLib y Seaborn que nos ayudarán a poder llegar a información de valor durante el análisis de los datos. Haremos al final una pequeña interpretación de los datos extraidos.

**Librerías:**

* Pandas
* Numpy
* Matplotlib.pyplot
* Seaborn

**Principales preguntas finales a responder:**

1. ¿Cuántas viviendas tenemos en cada zona?
2. ¿Cuántos días han transcurrido de media entre las dos lecturas? ¿Y en cada zona?
3. ¿De qué zona son las 10 viviendas con menos tiempo entre lecturas? ¿Y dentro de Orioventa (1 y 5)?
4. Ver en qué años se han dado de alta las viviendas.
5. ¿Qué porcentaje tienen los tipos de lectura en el total?
6. ¿Cuánta agua consumen las viviendas por día?
7. ¿Cuánta agua consumen las viviendas por día en cada zona?
8. ¿De qué zona son las viviendas que más consumen?
9. ¿Cuál es el año de alta que más se repite en el total?
10. Calculamos la correlación entre el año de alta y el consumo por día.

**1.pregunta:**

AMA-3 4716

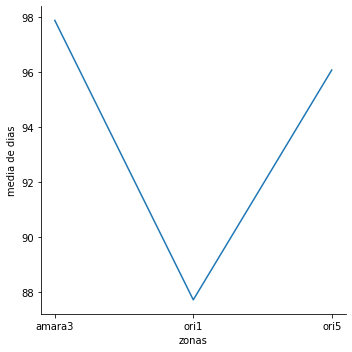
ORI-5 1498

ORI-1 692

**2.pregunta:**

Media de días toda la muestra: 96 días

Media de cada zona (gráfico 0)



**3.pregunta:**

5359 ORI-1

1257 AMA-3

3725 AMA-3

680 AMA-3

1466 AMA-3

2022 AMA-3

3232 AMA-3

817 AMA-3

3986 AMA-3

790 AMA-3

Dentro de Orioventa:

5359 ORI-1

4743 ORI-1

6134 ORI-5

6135 ORI-5

5514 ORI-5

5633 ORI-5

5631 ORI-5

5597 ORI-5

5609 ORI-5

5547 ORI-5

**4.pregunta:**

[1976,

1978,

1980,

1985,

1987,

1988,

1989,

1990,

1991,

1992,

1993,

1994,

1995,

1996,

1997,

1998,

1999,

2000,

2001,

2002,

2003,

2004,

2005,

2006,

2007,

2008,

2009,

2010,

2011,

2012,

2013,

2014,

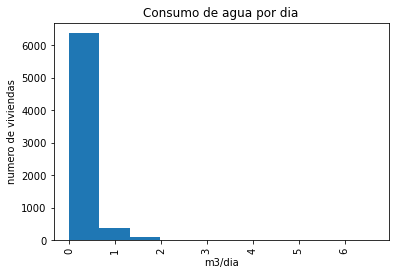
2015]

**5. pregunta:**

R: 0.924703

E: 0.075297

**6. pregunta: gráfico 1**

****

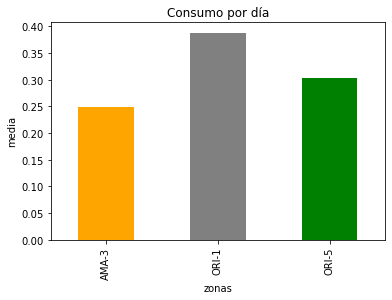
**La media de consumo** es: 0.2750063180334733 m3/día

En **Amara3**: 0.2496340905248967 m3/día

En **Ori1**: 0.3883962167507313 m3/día

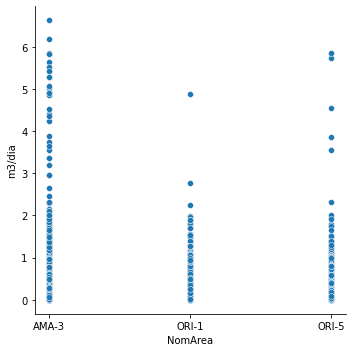
En **Ori5**: 0.3025027232525022 m3/día

Gráfico 2:



**7. pregunta:**

**Gráfico 3:**



**8. pregunta:**

2784 AMA-3

3555 AMA-3

4990 ORI-1

1267 AMA-3

2752 AMA-3

4999 ORI-1

3554 AMA-3

5010 ORI-1

2756 AMA-3

5013 ORI-1

**9. pregunta:**

2003

**10. pregunta:**

-0.038227604836558106

**INTERPRETACIÓN DE LOS DATOS:**

En primer lugar vemos que la cantidad de las viviendas de Amara 3 son de mucha mayor cantitad que las de OrioVenta ya que son más que el doble, aunque esto no tiene mayor importancia ya que nosotros trabajaremos con el promedio de los datos.

Después la media de los días entre lecturas es de 96 días y vemos que en este caso cuando comparamos las zonas, Oriventa 1 tiene 10 días menos (87 días) que Amara 3 (97 días) y Orioventa 5 (96 días), el cual es un dato a destacar como se percibe en el gráfico 0 pero desconocemos la causa.

Hemos visto igualmente que la vivienda con menos tiempo entre lecturas es de Orioventa1 mientras que las 9 siguientes coincide que son de Amara 3.

Dentro de Orioventa pasa algo similar. Mientras que las dos primeras son de Orioventa 1, las siguientes con menor duración son de Orioventa 5.

A continuación, como información adicional hemos visto que las viviendas se dieron de alta entre los años 1976 y 2015.

También hemos visto que en la última lectura la mayoría de las lecturas han sido de tipo “reales”, algo más que una novena parte.

A continuación, hemos pasado al análisis del consumo de agua como dato transformado y vemos que la mayoría de las observaciones consumen entre 0 y 1 m3/día, como nos indica el gráfico 1. Llegamos a la misma conclusión extrayendo la media de cada barrio también, pero vamos a verlo a continuación con más precisión en el gráfico 2.

En este gráfico, llegamos a la parte importante del análisis ya que analizamos por separado la media de cada zona y vemos que en Amara 3 el consumo baja considerablemente. En Amara 3 el consumo medio es de 0,24 y en Orioventa 5 es de 0,38, lo cual nos indica una diferencia a tener en cuenta. En el próximo gráfico vemos los datos de cada zona mucho más esparcidos y nos percatamos que la cantidad de observaciones en Amara 3 es considerablemente mayor. Es de reseñar que en 6 de las 10 primeras casas con mayor consumo son de Amara 3 y 4 de Orioventa1. No tenemos ninguna casa de Orioventa 5 en esta clasificación.

Cómo últimos datos, 2003 es el año que más viviendas se dieron de alta y hay una correlación negativa mínima (-0.03) entre el año de factura y el consumo por día, por lo parece no haber relación entre estas variables.

Para más información sobre el análisis hemos sacado las medidas de posición y dispersión comentadas al inicio del informe, que se pueden contemplar en el script del proyecto. Otro dato interesante es que la cantidad de consumo por día que más se repite es 0, lo cual nos indica que una cantidad importante de casas no han sido habitadas.