ANEXO: DATOS ANÓMALOS

Respecto a los datos anómalos hemos identificado 6 tipos diferentes, aunque hemos eliminado 8 tipos por decisiones tomadas en común que explicaremos más adelante.

En los 12 meses encontramos los siguientes datos anómalos:

- Negativos
- Nulos
- Nan (Casilla vacía)
- Mes siguiente
- Negativos (KWH/15 MIN)
- Atípicos

Negativos

Se tratan de datos que aparecen negativos en la columna KWH. Recordemos que en esta columna se parte de un número inicial y las siguientes filas son iguales o mayores a este. Esto quiere decir que se va acumulando el consumo. Por este motivo creamos una nueva columna que contenga simplemente lo que se consume cada 15 minutos.

Tener en la columna KWH datos negativos carece de sentido, puesto que como hemos explicado se parte de un número y los siguientes deben de ser iguales o mayores. Cabe añadir, que encontramos escasos errores de este tipo. Estos resultados se deben a que ciertos sensores dejan de contar adecuadamente. Hay que recordar que los sensores miden el consumo mediante una fórmula formada por varios medidores. Es decir, el resultado reflejado puede ser la suma del sensor 1 y 2 menos el sensor 3. En este caso si el 1 y el 2 dejan de funcionar el resultado nos daría negativo.

Nulos

Son datos cuyo valor en la casilla correspondiente a la columna KWH es 0. Como hemos visto anteriormente, en esta columna partimos de un número (mayor que 0) y lo siguientes son iguales (si no se consume nada) o mayores (si consume). Por consecuencia, hemos decidido eliminar todos aquellos que sean nulos puesto que se presentan como datos anómalos. Al igual que ocurre con los negativos, se nos muestran en la base de datos en una cantidad escasa. El mes que más contiene es enero con 163 aunque el resto varían entre 0 y 67. Hay varios meses que no tienen errores de este tipo. La mayoría de los que no tienen son los más recientes. Esto puede ser debido a una mejora de los sensores.

Nan (Casilla Vacía)

En ocasiones los sensores no transmiten ninguna información y esto se traduce en casillas vacías. Este tipo de datos anómalos 'Nan' (Not a number), aparecen con más frecuencia respecto a los negativos y nulos. Enero es el mes que más contiene con 31.183 datos de esta clase. El resto varía entre 1.000 y 12.000. **Nos damos cuenta de que los meses más recientes tienen menos frecuencia de estos datos**.

Mes siguiente

Otro dato anómalo presente en todas las bases de datos es el correspondiente a la fecha del mes siguiente. Es decir, consideramos que es información duplicada. Un ejemplo que nos permite entenderlo podría ser que, partiendo del base de datos de noviembre, tendríamos un

error en cada edificio con fecha 01/12/2018 y hora 00:00. Decimos que este dato no es correcto puesto que lo tendremos de la misma forma en la base de datos de diciembre por lo que lo eliminamos para no tener información repetida. Se produce por cada edificio por lo que tenemos entre 137 y 143 errores por cada mes.

Negativos (KWH/15 MIN)

Como hemos redactado anteriormente, hemos creado una columna que nos muestra lo que se consume cada 15 minutos. En esta columna encontramos datos negativos que no deberían de estar. Esto se produce por error en los medidores. Que en esta columna encontremos un número negativo implica que si partíamos en la columna de KWH de un número si el siguiente es menor se mostrará en la nueva columna como negativo. Es anómalo puesto que como hemos dicho si partimos de un número, el siguiente debe ser igual o mayor. Explicado de otra forma, no puede existir un consumo negativo. Encontramos en torno a 300 por cada mes. El mes que más contiene es octubre con 737.

Atípicos

En la columna creada observamos que la media es muy superior en todas las bases de datos con la mediana, lo que se puede traducir a la existencia de datos atípicos que modifican el promedio. Teniendo en cuenta que la estamos midiendo el consumo cada 15 minutos, carecería de sentido encontrarnos con datos excesivamente grandes. Además, explicaremos como hemos procedido a la eliminación e identificación de estos datos. Esto lo trataremos con más detalle posteriormente. Cabe añadir que los meses más recientes no presentan casi datos de este tipo por lo que creemos que se debe a una mejor de los sensores. No obstante, adelantamos que en la gran mayoría de casos el consumo o es nulo o es muy pequeño por lo que podemos afirmar que cantidades excesivas son incorrectas y no corresponden a la realidad. Además, al crear una nueva columna, cada vez que cambia de edificio el número de partida de la columna KWH es diferente, es decir, la resta en este caso puede dar lugar a un número muy grande o a un número negativo.

El proceso de eliminación de estos datos se ha repetido varias veces en los 6 primeros meses y en los otros 6 hemos encontrado menos y no ha hecho falta. La frecuencia en los primeros es superior a 140 mientras que en el resto es inferior a 10.

Dedicaremos un apartado para hablar del tratamiento de los atípicos.

Estas son las 6 clases que hemos encontrado de datos anómalos. No obstante, hemos eliminado dos más:

- Nulos (KWH /15 MIN)

- Clima

Nulos (KWH /15 MIN)

Estos datos no son en su totalidad datos anómalos, pero hemos decidido eliminarlos por los siguientes motivos:

- No podemos distinguir entre aquellos que son correctos y que significan que no se ha consumido nada en esos 15 minutos y entre aquellos incorrectos que nos aparecen como nulos por fallo de los sensores.
- Son muchos, por lo que la eliminación de estos simplifica el trabajo y reduce su tiempo.

- Nos marcamos como objetivo más representativo e interesante para la universidad mostrar el consumo total, por lo que la eliminación de nulos no influye.
- Dejar estos datos nos daría un promedio mucho menor al real puesto que se trata de miles de datos, que además incluyen algunos anómalos.

Clima

Respecto a los datos que corresponden con clima hemos decidido eliminarlos por los siguientes motivos:

- Nuestro objetivo es analizar el gasto como conjunto, saber los edificios que gastan más y poder llegar a conclusiones a partir de ello. Por este motivo descartamos estos datos, ya que no nos resultan útiles para alcanzar nuestra meta.
- La eliminación de estos facilita el trabajo puesto que reduce la base de datos. Encontramos alrededor de 150.000 datos de clima por cada mes.
- En los datos 'ALL' están incluidos los de clima.

Hemos creado un Excel que contiene toda la información sobre estos datos de cada mes.

La eliminación de todos estos reduce todos los meses a casi la mitad de los datos iniciales y amplía la comodidad para poder trabajar a la misma vez que nos acerca a la realidad de forma significativa.

Además, es importante saber que antes de eliminarlos teníamos 143 bloques de edificios y después tenemos alrededor de 85. Esto se debe a que existían bloque con datos únicamente de 'CLIMA'.

A continuación está insertado el código base que hemos utilizado en cada mes para eliminar los valores anómalos. Aunque algunos parámetros pueden variar en algunos meses, el código implementado ha sido el siguiente:

```
#!/usr/bin/env python
# coding: utf-8
import pandas as pd
df = pd.read_csv('nuevo-04.csv', sep=';', encoding = 'ISO-8859-1')
df.head()
df.count()

nulos = df.loc[df['KWH'] == 0]
negativos = df.loc[df['KWH'] < 0]
noHayNada = df.loc[df.isnull().any(axis='columns')]
datosFecha = df.loc[df['FECHA_LC'] == '01/05/2019']
datosClima = df.loc[df['USO_ENERGIA'] == 'CLIMA']
negativosAcumulativos = df.loc[df['KWH/15 MIN'] < 0]
nulosAcumulativos = df.loc[df['KWH/15 MIN'] == 0]
```

```
df.drop(nulos.index, inplace = True)

df.drop(negativos.index, inplace = True)

df.drop(noHayNada.index,inplace = True)

df.drop(datosFecha.index,inplace = True)

df.drop(datosClima.index,inplace = True)

df.drop(negativosAcumulativos.index,inplace = True)

df.drop(nulosAcumulativos.index,inplace = True)


df.count()


df.to_csv('AbrConAti.csv', sep = ';')
```

En este proceso, como se puede observar, hemos usado la librería Pandas de Python. Además, también hemos guardado el número de valores de cada tipo que hemos eliminado. Esta parte del código no aparece en el párrafo anterior, ya que no tiene relación en cuanto a cómo se han utilizado los datos.