**Extracción de datos de Centros Educativos**

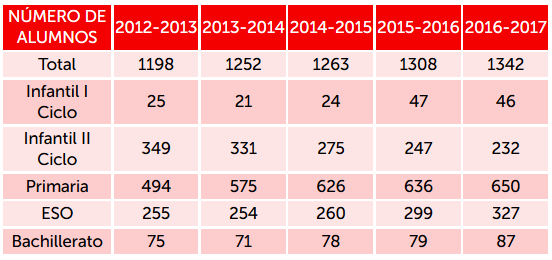
Una de las aportaciones que vimos más atractivas para nuestra idea de negocio fue la de incorporar los datos de los colegios e institutos para dar más valor a lo que nuestros futuros clientes necesiten en la búsqueda de un piso, más allá del precio, metros cuadrados o ubicación.

Vimos que esto podía ayudar en la decisión de elegir una vivienda para aquellos compradores con perfiles de familias con hijos o que esperan tenerlos, de entre 36 y 45 años, ya que valoran mucho la ubicación con colegios en la zona (\* ver referencia)

El origen de los datos ha sido la página de la Comunidad de Madrid en la que están todos los centros educativos tanto públicos como privados de la comunidad.

<http://www.madrid.org/wpad_pub/run/j/MostrarConsultaGeneral.icm>

En esta web, además de incluir datos del centro, etapas educativas que incluye, opciones lingüísticas, etc, aporta además datos estadísticos y gráficas sobre número de alumnos, procesos de admisión a comienzo de curso, titulación según el grado en comparativa con la media de la comunidad, pruebas de conocimientos y destrezas indispensable y, para los centros que incluyen educación secundaria estadísticas sobre la prueba de acceso a la universidad.



Ejemplo de tabla de número de alumnos de un centro

\* Ref. <http://www.fotocasa.es/blog/compraventa/perfil-del-comprador-vivienda-2016>



Ejemplo de gráfica de titulación de un centro

El problema que hemos tenido es que la información que permite bajarse como fichero de datos es un listado con todos los centros con su información básica, pero los datos estadísticos que nos parecían interesantes y que podían aportar más decisión a la hora de seleccionar un centro, y por consiguiente influir en la toma de decisión en la compra de una vivienda, no estaban disponibles. Nos pusimos en contacto con la consejería de Educación de la CM para preguntar si esta información estaba disponible en fichero que pudiésemos utilizar para nuestro propósito, pero nos dijeron que no era posible.

Por tanto para obtener estos datos hemos recurrido al scrapeado de la web de todos los centros.

El diseño de esta web se estructura en una serie de pestañas por cada centro, y los datos que nos interesan están en

* Datos Estadísticos:
  + Tabla con número de alumnos por año y nivel de enseñanza
  + Tabla con datos de admisión por año (admitidos/no admitidos)
* Resultados Académicos:
  + Tabla con la nota media por año de la Titulación en educación Secundaria del centro y la nota media general en la comunidad de Madrid
  + Tabla con la nota media por año de la Prueba de conocimientos y destrezas indispensables (CDI) del centro y la nota media general en la comunidad de Madrid
  + Tabla con la nota media por año de la Prueba externa en centros bilingües del centro y la nota media general en la comunidad de Madrid
  + Tabla con el promedio de la calificación en la fase general por año de la Prueba de acceso a la universidad (PAU) del centro y la nota media general en la comunidad de Madrid

Para esto hemos creado un proceso con Python (*Scrap\_colegios.py*) que con la utilidad webdriver de la librería *selenium* hemos accedido de forma iterativa, partiendo del código de centro que tenemos en el excell general, a todos los centros y por cada uno a cada pestaña para recoger estos datos e ir añadiendo cada tabla a un fichero txt.

Además en el mismo programa incluimos una función para geolocalizar cada colegio ya que en el excell solo tenemos la dirección del centro.