Integración de la información:

Variables totales

En esta sección describiremos un poco el proceso con el cual se consiguió los datos necesarios para el proyecto, así como la forma en que se limpiaron y se integraron para su uso en la modelación. Cabe mencionar que la información recabada tiene una venta de tiempo del año 1998 al año 2018.

Los datos usados así como sus bases de datos respectivas, fueron:

* Ratio de desempleo, la base de datos original fue obtenida por Geographical Economic Data de ST. LOUIS FED.
* Hospitales en Estados Unidos, la base de datos original fue obtenida por Homeland Infrastructure Foundation.
* Rating de hospitales, la base de datos original fue creada por Centers for Medicare & Medicaid Services.
* Precio de inmuebles empresa Zillow. (county\_time\_series) y el dataset countycrosswalk.
* Escuelas públicas del país creado por Homeland Infrastructure Foundation.
* Ratio de crímenes obtenido por ICPSR.

Una vez descargados estos datasets, se hizo uso del lenguaje de programación Python, con el cual se desecharon las columnas que no eran necesarias. Al querer juntar ya todas las tablas, hubo un pequeño error al querer juntar la tabla de crimines, para la cual usamos una herramienta de perfilado y limpieza de datos, llamada Talend. Con esta herramienta exploramos de forma visual los datos, y nos dimos cuenta de ciertos detalles en algunas variables.

Limpieza de la información

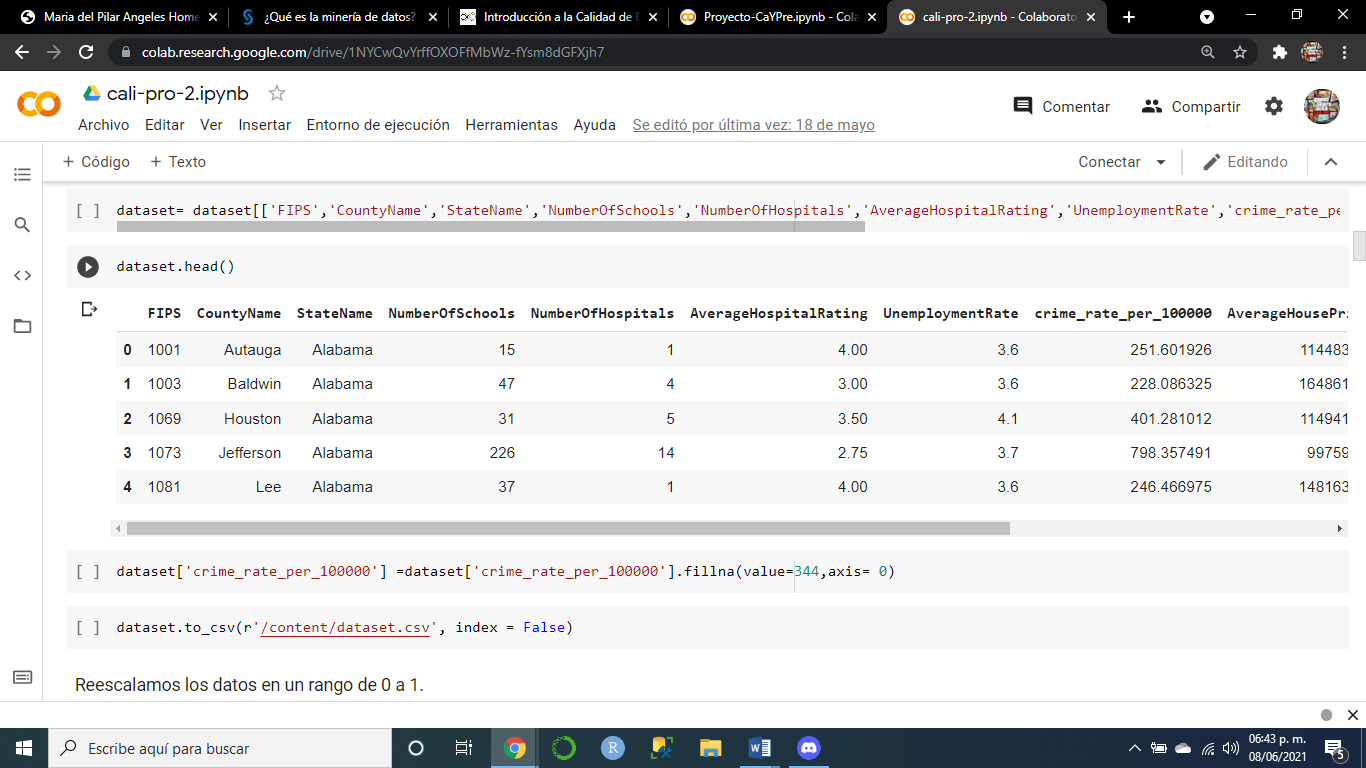
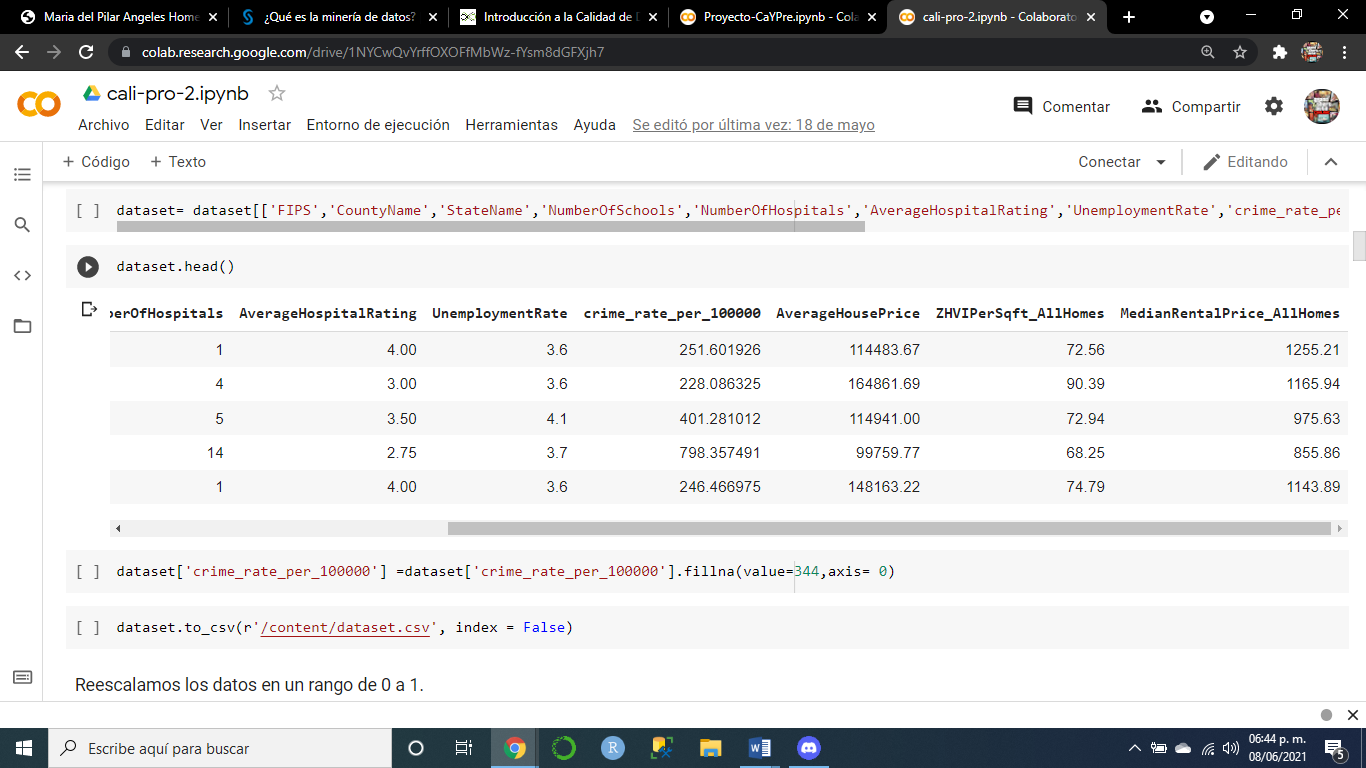
Como ya se mencionó anteriormente en esta parte usamos la herramienta Talend Preparation, la cual es bastante fácil de usar y contiene bastantes técnicas de limpieza y transformación, que nos servirán para el proyecto.

Una vez cargado los datos a Talend nos dimos cuenta que una variable tenía entradas inválidas, después de investigar un poco detectamos el error, el cual era básicamente que en algunos números hacía falta un digito, generando todo el problema anterior.

Posteriormente, notamos que los valores flotantes tenían muchos decimales, así que decidimos dejarlos hasta dos puntos decimales. Luego también observamos que cierta columna (Rating de hospitales) tenía valores nulos, los cuales completamos con la moda de cada condado.

Analytical Base Table (ABT)

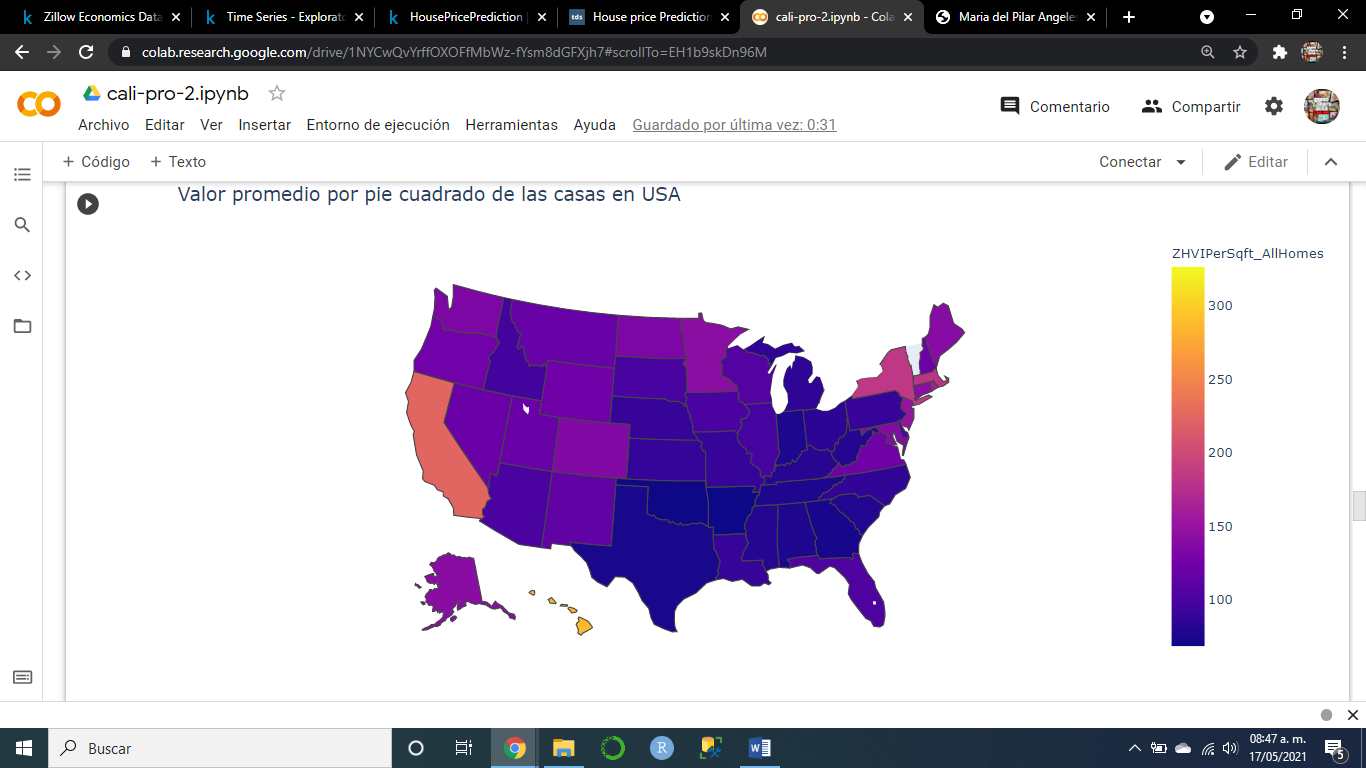
Una vez terminada la limpieza, nos pasamos otra vez a Python, donde ya tendremos la base de datos completa, con la cual se hará la modelación. Aunque primero haremos una pequeña exploración de unos estadísticos relevantes de nuestros datos.

Veamos la tabla final, después de hacer todo este proceso:

Exploración

En esta parte visualizaremos algunos estadísticos interesantes.

Para tener una idea general del valor de las casas, sin tomar en cuenta el tamaño de ellas, tenemos una variable que nos dice el valor por pie cuadrado de los inmuebles, dándonos información intrínseca del valor de las casas.

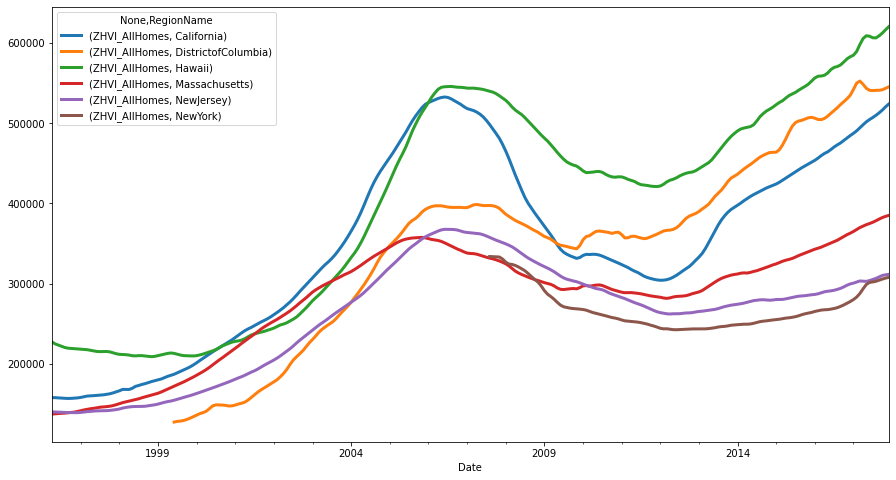
**

**Observaciones de los resultados:**

* Los estados al sureste del país tienen menor valor comparados con los estados del noreste.
* Hawái tiene el valor más grande de todos los estados, seguido de california.
* En general, los estados del norte tienen un mayor valor que los del sur.

Veamos ahora como se ha desarrollado el precio de las casas a lo largo de los años para el top de estados:

Serie de tiempo, precio de inmuebles  
 (En USD)

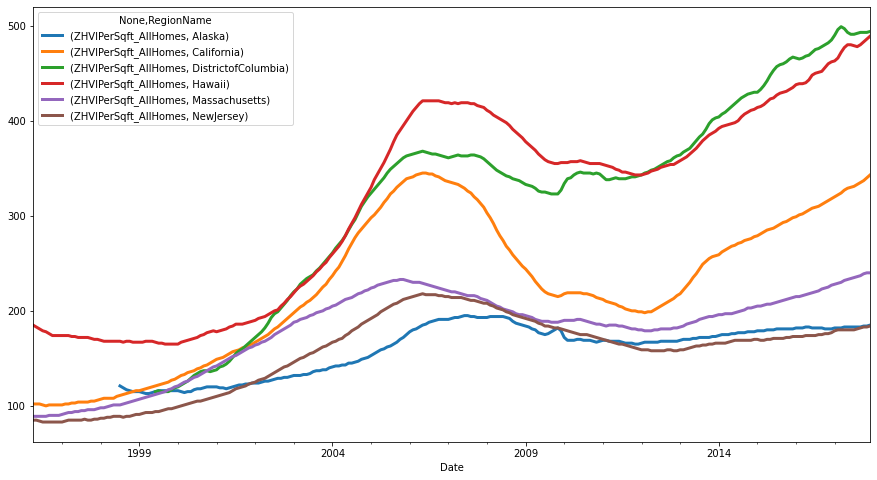


**Observaciones de los resultados:**

* Sepuede ver cómo quedó marcada la crisis inmobiliaria (2007-2008) en el precio de los inmuebles a lo largo de esos años. Después del 2012 se empieza a recuperar los precios para posteriormente iniciar una tendencia alcista (hasta la próxima crisis).
* Al estado de New Jersey le ha costado más recuperarse de esa crisis, vemos como antes de la crisis tenía una muy buena tendencia alcista pero después del 2008 su tendencia ha sido plana.

A continuación veremos algo muy parecido al gráfico anterior, pero en esta ocasión veremos reflejado el valor por pie cuadrado

Serie de tiempo, precio por pie cuadrado  
 (EN USD)

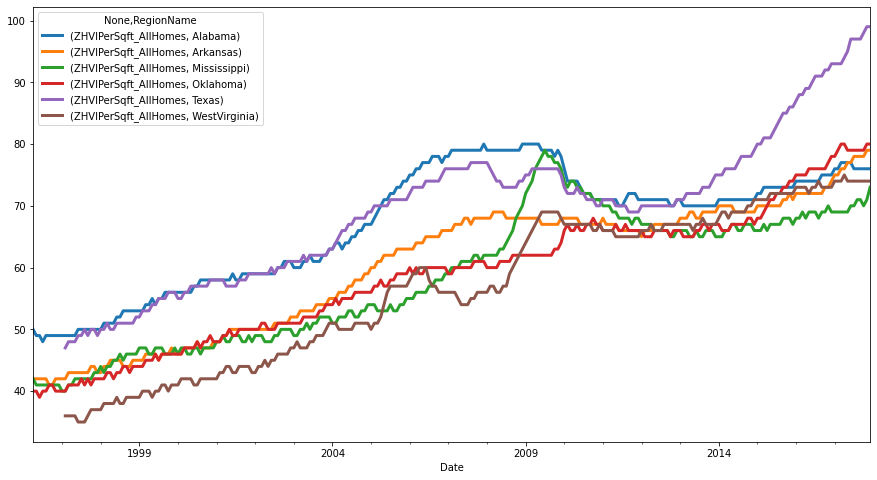


**Observaciones de los resultados:**

* Notamos un patrón similar al gráfico anterior.
* El caso de New Jersey nos llamó la atención, otra vez. Antes de la crisis tenía toda la pinta de que se iba a posicionar en los primero lugares, pero una vez pasada la crisis no se pudo recuperar de ella.
* También notamos que California fue bastante castigada por la crisis.

Por último veamos la tendencia del valor por pie cuadrado para el top de los estados más baratos.

Serie de tiempo, precio por pie cuadrado  
 (EN USD)



**Observaciones de los resultados:**

* En este gráfico ocurre todo lo contario, en vez de bajar por la crisis tienen un considerable aumento. Esto se debe a que las personas dejaron de comprar en los estados caros y se fueron a vivir a los estados más “económicos”.