UNIVERSITAT OBERTA DE CATALUNYA

MÁSTER CIENCIA DE DATOS

INTEGRANTES: DIEGO CASTILLO CARRIÓN

CARLOS HERNÁNDEZ MARTÍNEZ

PRÁCTICA 1: WEB SCRAPING

ABRIL 2019

1. **Contexto. Explicar en qué contexto se ha recolectado la información. Explique-por qué el sitio web elegido proporciona dicha información.**

Hemos elegido el sitio web properati.com porque se requiere realizar la siguiente investigación y resolver los objetivos planteados.

Análisis descriptivo de los anuncios de ventas de casas en el Distrito Metropolitano de Quito - Ecuador.

Objetivos:

- Calcular el precio aproximado de las casas en venta según el área.
- Determinar la variación de los precios en las casas anunciadas según los sectores o barrios en donde se ubican.
- Establecer precios máximos y mínimos de las casas en venta según el número de habitaciones.

El sitio web elegido, actualmente consta de las variables necesarias para poder realizar el análisis anteriormente descrito.

Con las variables definidas se puede cumplir los objetivos propuestos.

Actualmente la página web properati.com es una de las más conocidas y demandadas del país, se encarga de publicar anuncios sobre compra venta y alquiler de varios inmuebles a nivel nacional y actualmente esta página o empresa ofrece una aplicación móvil que facilita la navegación y día a día sigue creciendo el número de personas que accede a ella, para publicar sus bienes inmuebles que pretenden vender.

2. **Definir un título para el dataset. Elegir un título que sea descriptivo.**

Anuncios de ventas de casas en el Distrito Metropolitano de Quito - Ecuador.

3. **Descripción del dataset. Desarrollar una descripción breve del conjunto de datos que se ha extraído (es necesario que esta descripción tenga sentido con el título elegido).**

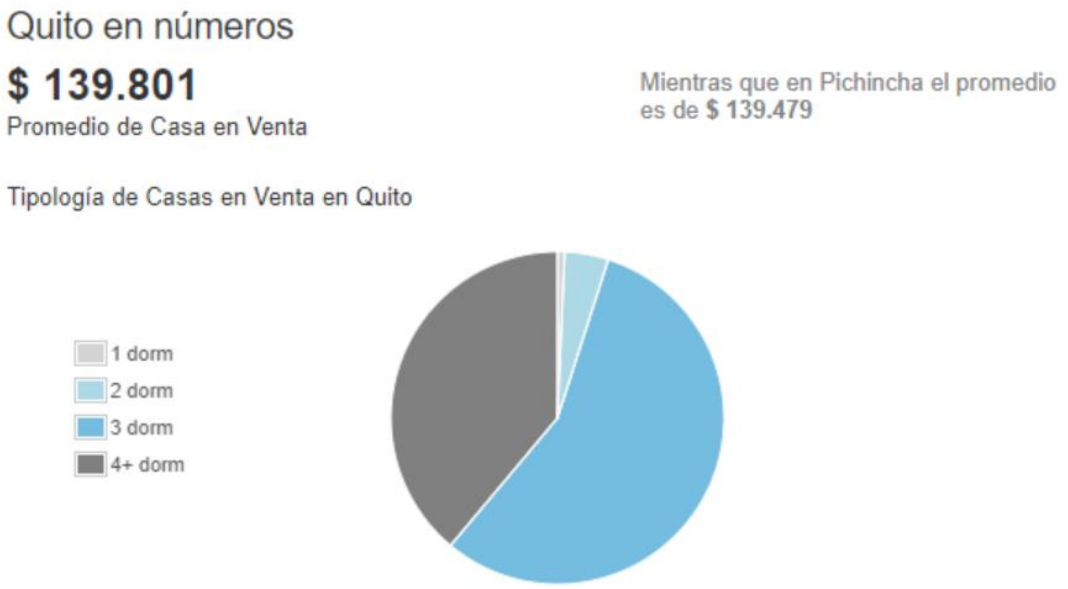
El conjunto de datos contiene la información capturada desde el sitio web: properati.com. El Dataset contiene la información más relevante de los anuncios

de ventas de bienes inmuebles para el Distrito Metropolitano de Quito en Ecuador. El principal objetivo del Dataset es responder preguntas acerca del precio de los inmuebles según el área, número de habitaciones, ubicación, etc.

Unos de los inconvenientes del dataset puede ser que presenta los anuncios de solo una página web, para análisis más relevantes se debería utilizar información obtenida desde diferentes sitios web de anuncios de ventas de bienes inmuebles. otro inconveniente es que el dataset no presenta importación más desagregada de los anuncios como por ejemplo el número de baños, si cuenta o no con patio, si está ubicado dentro de un conjunto residencial y si el inmueble cuenta con servicios básicos.

Se debería realizar un tratamiento de la información antes de realizar el análisis. por ejemplo, se debería realizar la eliminación de caracteres especiales del campo descripción, eliminación del signo de dólar ($) del campo precio, eliminación del símbolo de metros cuadrados del campo área.

4. **Presentar una imagen o esquema que identifique el dataset visualmente**



5. **Contenido. Explicar los campos que incluye el dataset, el periodo de tiempo de los datos y cómo se ha recogido.**

Los campos contenidos en el dataset son:

Descripción: Descripción del anuncio detallando las características del inmueble por parte del vendedor, tipo de dato string.

Precio: Valor en Dólares del Inmueble, tipo de dato integer.

Tipo: Variables que describe si inmueble es casa o departamento, tipo de dato string.

Ubicación: Dirección del Inmueble, tipo de datos string.

Fecha de publicación: Fecha en la que se publicó el anuncio de venta del inmueble, tipo de dato date

Área: Área (en metros cuadrados) del inmueble, tipo de dato integer.

Núm. habitaciones: Número de habitaciones con las que cuenta el inmueble, tipo de dato integer

La recolección de la información se realizó el día 06.04.2018. Estos datos se recogieron mediante técnicas de Web scraping utilizando la librería BeautifulSoup sobre el entorno Jupiter.

Por motivo de no saturar la página web de peticiones únicamente se capturo el resultado de las primeras 20 páginas de anuncios.

6. **Agradecimientos. Presentar al propietario del conjunto de datos. Es necesario incluir citas de investigación o análisis anteriores (si los hay).**

Properati

Es una plataforma web y mobile de propiedades que nació para cambiar la forma en que se venden y arriendan inmuebles en Latinoamérica. Quienes busquen un nuevo hogar o quieran invertir en propiedades encontrarán en Properati, además de las ofertas disponibles, valiosa información para tomar las mejores decisiones: promedios de precios, características de los barrios, comparaciones, etc.

Properati también ofrece una propuesta novedosa para las inmobiliarias o agentes que quieran vender una propiedad, ya que el modelo de negocios se basa en entregar contactos de calidad. Al pagar sólo por contacto recibido y no por banners o pop-ups, los incentivos entre el vendedor, el usuario y Properati quedan alineados, dando por resultado un sitio limpio y con una interfaz amigable.

Actualmente Properati está online en Argentina, Colombia, Ecuador, Perú y Uruguay, y en todos los países realizó acuerdos con las inmobiliarias, agentes y constructoras más importantes para publicar sus propiedades. También están disponibles las versiones Android e iOS de Properati, con funcionalidades especialmente diseñadas para dispositivos móviles, como la búsqueda de propiedades cerca de la ubicación actual y una navegación simple y clara.

Actualmente no hemos encontrado investigaciones o análisis anteriores.

7. **Inspiración. Explique por qué es interesante este conjunto de datos y qué preguntas se pretenden responder.**

Este conjunto de datos es interesante por su funcionalidad, a futuro nos va a permitir cumplir con varios objetivos propuestos, a la vez tienen variables para realizar cálculos matemáticos, estadísticos.

Anteriormente planteamos algunos objetivos que se pueden alcanzar con este trabajo, a continuación, está la explicación para poder realizarlos.

Objetivo 1. Para calcular el precio en metros cuadrados vamos a filtrar por "ubicación" esta variable nos indica si está en el norte, sur, centro o valles, con esta información vamos a determinar el valor del metro cuadrado en función al "precio" y el "área".

Objetivo 2. Vamos a sectorizar los precios a través de la variable ubicación, calculamos la media del precio de cada ubicación especifica.

Objetivo 3. Se filtrará los resultados para calcular el precio máximo y mínimo en función de la variable número de habitaciones.

8. **Licencia. Seleccione una de estas licencias para su dataset y explique el motivo de su selección.**

La licencia seleccionada es Released Under CC0: Public Domain License, el motivo de la selección es una licencia de derecho de autor y código abierto, al ser un trabajo practico hemos pensado que es importante que esta información se pueda estudiar, compartir con la finalidad de que muchas usuarios tengan acceso y se puedan beneficiar

9. **Código. Adjuntar el código con el que se ha generado el dataset, preferiblemente en Python o, alternativamente, en R.**

`web_scraping/src/scraper.py`

10. Dataset. Presentar el dataset en formato CSV

```
web_scraping/data/casas.csv
```

Contribuciones	Firma
Investigación previa	**DCC, CHM**
Redacción de las respuestas	**DCC, CHM**
Desarrollo código	**DCC, CHM**

Recursos

● Subirats, L., Calvo, M. (2018). Web Scraping. Editorial UOC.

● Masip, D. El lenguaje Python. Editorial UOC.

Tipología y ciclo de vida de los datos Práctica 1 pág 2

● Lawson, R. (2015). Web Scraping with Python. Packt Publishing Ltd. Chapter 2.

Scraping the Data.

● Simon Munzert, Christian Rubba, Peter Meißner, Dominic Nyhuis. (2015).

Automated Data Collection with R: A Practical Guide to Web Scraping and Text

Mining. John Wiley & Sons.

Internet

https://guides.github.com/activities/hello-world.

https://www.properati.com.ec