



IIC2115 – Programación como Herramienta para la Ingeniería (I/2023)

Laboratorio 4

Aspectos generales

- **Objetivo:** evaluar individualmente el aprendizaje sobre extracción de datos web y bases de datos en Python, a través de la construcción de una serie de tareas asociadas a la extracción de datos y estructuración de BBDD.
- **Lugar de entrega:** jueves 7 de julio a las 23:59 hrs. en repositorio privado.
- **Formato de entrega:** el archivo Python Notebook (**L4.ipynb**) con la solución del laboratorio en el buzón de tareas. Es requerimiento de formato el utilizar múltiples celdas de texto y código para la construcción de la solución. Laboratorios que no cumplan el formato de entrega tendrán un descuento de 0,5 pts. El ayudante se reserva el derecho de penalizar hasta 1 punto dependiendo del orden y legibilidad de su código. Recuerde responder el formulario asociado al laboratorio, ya que será tomado como parte del formato de entrega.
- **Entregas atrasadas:** El descuento por atraso se realizará de acuerdo a lo definido en el programa del curso. Si su laboratorio es entregado fuera de plazo, tiene hasta el **viernes 8 de julio a las 11:59 AM** para responder el formulario de **entregas fuera de plazo** disponible en el buzón de tareas.
- **Issues:** Las discusiones en las *issues* del Syllabus que sean relevantes para el desarrollo del laboratorio, serán destacadas y se considerarán como parte de este enunciado. En esta ocasión el uso de librerías será libre.
- **Laboratorios con errores de sintaxis y/o que generen excepciones en todas las ejecuciones** serán calificados con **nota 1.0**.

Introducción

Tras rescatar a su profesor de un fatídico destino en el desafío previo, el ha decidido recompensarlo con un desafío final. En este desafío, usted tendrá que estructurar una base de datos en con datos de propiedades, gente interesada en comprar y arrendar y propietarios, El objetivo es utilizar estos datos para comprender mejor el mercado de viviendas y poder vincular propietarios con posibles interesados. Para este desafío, puede utilizar las herramientas que ha aprendido durante el curso y cuenta con tres archivos CSV:

1. **phones.csv** Contiene una columna de identificador de personas (propietarios y compradores) y su número telefónico. Durante el enunciado se referirá tanto a personas interesadas en comprar como arrendar como interesados o compradores de manera indistinta.
2. **compradores.csv** Contiene una columna de identificador de personas, presupuesto, tipo de operación que busca (arrendar o comprar) y cantidad de cuartos mínimos que desea.
3. **propiedades.csv** Contiene columna de identificador, de propiedades, identificador del dueño de la propiedad y distintos atributos de esta.

1. Generación de Base de datos (2.5 pts)

- a) Estructure una base de datos con los datos disponibles, con una tabla para cada csv. Identifique las llaves primarias y llaves foráneas.
- b) Calcule para cada comuna el promedio, la desviación estándar y la mediana del costo por metro cuadrado. Al menos una de estas variables deberá ser obtenida mediante una consulta de SQL.
- c) Genere una función que ejecute una consulta a la base de datos que para cada interesado que entregue las posibles propiedades de su interés (que estén en su presupuesto, tenga la cantidad de habitaciones mínimas que requiere, etc...).
- d) Para cada comprador, escoja la propiedad con mayor cantidad de metros dentro de las propiedades que cumplan con los requisitos mínimos (en base a la consulta generada anteriormente). Genere en la base de datos una tabla de "transacciones" que contenga el número telefónico de comprador, de propietario, los id correspondientes y el id de la propiedad. Puede descartar los compradores que no encuentren ni una propiedad de interés.

2. Expansión de la base de datos (2.5 pts)

- (a) Complemente la tabla de propiedades de la base de datos con propiedades extraídas de portales de anuncios como portal inmobiliario o zoom inmobiliario. Para esto, los datos relevantes a extraer serán compras y arriendo de unidades ya sean nuevas o usadas de casas y departamentos, velando por tener las específicamente las mismas columnas de la tabla propiedades. Apunte a obtener al menos todos los anuncios disponibles en el sitio para la Región Metropolitana (puede ser para todo Chile).
- (b) Limpie la base de datos eliminando outliers de alguna manera (puede ser algo sencillo como sacar los cuantiles extremos) o registros poco racionales (precio y metros cuadrados cercanos a 0). Enfóquese principalmente en la columna de metros y de precio. Esto puede realizarlo utilizando pandas antes de pasar los datos a la base de datos.
- (c) Recuerde que el id de cada propiedad deberá ser única, utilice la id del anuncio mas una letra al principio que identifique el portal (para asegurar id única, Z para Zoom y P para Portal) y elimine los anuncios repetidos (ya que la tabla de propiedades es una muestra de una extracción de estos mismos datos ejecutada previamente)
- (d) Si opta por utilizar portal inmobiliario, puede imputar la comuna a cada anuncio utilizando geopandas y el shape de comunas de la biblioteca del congreso nacional.

3. Consultas finales (1 pts)

- (a) Calcule para cada comuna el promedio, la desviación estándar y la mediana del costo por metro cuadrado con la base de datos actualizada. Al menos una de estas variables deberá ser obtenida mediante una consulta de SQL. Además, calcule la variación porcentual de cada uno de estos indicadores con respecto al mismo calculo realizado en la parte 1.

Bonus (0.5 pts extra)

Realice la parte 2 y parte 3 del ejercicio tanto con Portal Inmobiliario como con Zoom Inmobiliario.

Corrección

Es importante que deje todas las celdas de su trabajo ejecutadas antes de subir el archivo, de lo contrario se le aplicará un descuento de 0,5 ptos. al puntaje total. Para la corrección de este laboratorio, se revisarán los

procedimientos desarrollados para responder las diferentes misiones propuestas. Se espera que las respuestas incluyan análisis y visualizaciones que permitan justificar las decisiones tomadas. **Como principio director para las misiones, piense siempre cómo resolverlas utilizando los tópicos utilizados en este capítulo, por sobre los introducidos en capítulos anteriores.**

Política de Integridad Académica

Los alumnos de la Escuela de Ingeniería deben mantener un comportamiento acorde al Código de Honor de la Universidad:

“Como miembro de la comunidad de la Pontificia Universidad Católica de Chile me comprometo a respetar los principios y normativas que la rigen. Asimismo, prometo actuar con rectitud y honestidad en las relaciones con los demás integrantes de la comunidad y en la realización de todo trabajo, particularmente en aquellas actividades vinculadas a la docencia, el aprendizaje y la creación, difusión y transferencia del conocimiento. Además, velaré por la integridad de las personas y cuidaré los bienes de la Universidad.”

En particular, se espera que mantengan altos estándares de honestidad académica. Cualquier acto deshonesto o fraude académico está prohibido; los alumnos que incurran en este tipo de acciones se exponen a un procedimiento sumario. Ejemplos de actos deshonestos son la copia, el uso de material o equipos no permitidos en las evaluaciones, el plagio, o la falsificación de identidad, entre otros. Específicamente, para los cursos del Departamento de Ciencia de la Computación, rige obligatoriamente la siguiente política de integridad académica en relación a copia y plagio: Todo trabajo presentado por un alumno (grupo) para los efectos de la evaluación de un curso debe ser hecho individualmente por el alumno (grupo), sin apoyo en material de terceros. Si un alumno (grupo) copia un trabajo, se le calificará con nota 1.0 en dicha evaluación y dependiendo de la gravedad de sus acciones podrá tener un 1.0 en todo ese ítem de evaluaciones o un 1.1 en el curso. Además, los antecedentes serán enviados a la Dirección de Docencia de la Escuela de Ingeniería para evaluar posteriores sanciones en conjunto con la Universidad, las que pueden incluir un procedimiento sumario. Por “copia” o “plagio” se entiende incluir en el trabajo presentado como propio, partes desarrolladas por otra persona. Está permitido usar material disponible públicamente, por ejemplo, libros o contenidos tomados de Internet, siempre y cuando se incluya la cita correspondiente.