



IIC2115 – Programación como Herramienta para la Ingeniería (II/2019)

Laboratorio 5 - Análisis de datos y tópicos avanzados

Objetivos

- Aplicar los contenidos de análisis y visualización para estudiar, graficar y predecir propiedades o relaciones que se pueden observar en un conjunto de datos.
- Utilizar la librería [Geopandas](#) para la visualización de mapas.
- Utilizar conceptos básicos de *Web Scraping* para extraer información de una página web.

Entrega

- **Lenguaje a utilizar:** Python 3.6
- **Lugar:** repositorio privado en GitHub. Recuerde incluir todo en una carpeta de nombre **L05**.
- **Entrega: Fecha a confirmar, cercano a fin de semestre**
- **Formato de entrega:** archivo python notebook (**.ipynb**) y archivo python (**.py**) con la solución de este enunciado. Los archivos deben estar ubicados en la carpeta **L05**. No se debe subir ningún otro archivo a la carpeta. Utilice múltiples celdas de texto y código para facilitar la revisión de su tarea. Los archivos **ipynb** y **py** deben contener la misma solución.
- **Descuentos:** se descontará 0.5 puntos por cada hora de atraso y fracción en la entrega final. Tareas que no cumplan el formato de entrega tendrán un descuento de 0.5 pts.
- **Tareas con errores de sintaxis y/o que generen excepciones serán calificadas con nota 1.0.**
- Si su laboratorio es entregado fuera de plazo, tiene 24 horas a contar de la fecha de entrega para responder el formulario de **entregas fuera de plazo** disponible en el Syllabus.

- Las discusiones en las *issues* del Syllabus en GitHub son parte de este enunciado.
- Dependiendo de las eventualidades y cierre de semestre podría publicarse una extensión de este enunciado.

Introducción

Nuevo estudio de los viajes de Santiago

Dado que usted ha entregado un excelente trabajo en el estudio solicitado por el MTT. Esta institución ha decidido contrarlo para un nuevo trabajo.

EOD-Santiago 2012

Recordemos que la EOD cuenta con una serie de Tablas que incluyen información relevante de viajes realizados en la ciudad de Santiago. También existe información relevante respecto de los hogares y personas asociadas a cada viaje. En la lista de a continuación, se incluye el nombre de cada Tabla en la base de datos y una breve descripción de su contenido.

Archivos de datos y Tablas de la EOD

En esta oportunidad, los datos de la EOD están disponibles en archivos CSV. Es decir, cada tabla estará contenida en un único archivo de nombre **NOMBRE_TABLA.csv** disponibles en la carpeta L05. Estos archivos de valores separados por “coma” o “punto y coma” podrán importarse a Python mediante **Pandas** y **Geopandas**. Las tablas corresponden a la mismas trabajadas en el Laboratorio 4, igualmente se detallan a continuación:

1. **Hogares:** En esta tabla se almacena toda la información referente a los hogares que participan en la encuesta. Entiéndase por hogar, al conjunto de personas que viven en un misma vivienda. Se puede encontrar información asociada a localización, ingreso del hogar, número de vehículos, bicicletas, etc.
2. **Vehículos:** En esta tabla se describen los vehículos presentes en los hogares de la encuesta. Incorpora información respecto al tipo de vehículo, marca, año, etc.
3. **Personas:** En esta tabla se almacena toda la información referente a las personas que participan de la encuesta. Se puede encontrar información asociada al hogar al que pertenecen, sexo, ingreso, ocupación, lugar de trabajo o estudios, etc.

4. **EdadPersonas:** Esta es una pequeña tabla que solo contiene información de la edad de las personas que participan en la encuesta.
5. **Viajes:** En esta tabla se almacena toda la información referente a las viajes encuestados por la EOD. Cada viaje es único y es realizado por una única persona (si dos personas viajan juntas, entonces habrán dos viajes). Se puede conocer la persona que realiza el viaje, su origen, destino, propósito, etc.
6. **DistanciaViaje:** Esta es una pequeña tabla que solo contiene información asociada a la distancia de cada viaje y si este fue imputado o no.
7. **ViajesDifusion:** En esta tabla se almacena toda la información referente al modo general utilizado en el viaje.
8. **Etapas:** En esta tabla se almacena toda la información referente a las etapas de los viajes encuestados por la EOD. Los viajes, presentes en la tabla Viajes, están compuestos por una o más etapas de viaje. Estas corresponden a las distintas fases realizadas por la personas para realizar un viaje. Se puede conocer el origen y destino de la etapa, así como también el tiempo de viaje, modo empleado, entre otras.

Tablas de parámetros

El detalle de las columnas de cada tabla de la EOD la puede consultar una vez cargados los datos en Pandas. Para entender el significado de algunas columnas, son necesarias una serie de tablas de parámetros que se dispone en la carpeta de este laboratorio. Por ejemplo, dentro la tabla **Viajes** existe una columna **ActividadDestino** que representa la actividad que desarrolla en el destino del viaje. Dentro de los datos de esta columna, solo hay enteros 1 y 8. Para saber que significa cada número, debemos buscar el archivo **ActividadDestino.csv** dentro de las tablas de parámetros y encontraremos los significados de cada valor.

Nuevas misiones del MTT

Para completar este nuevo requerimiento solicitado por el MTT usted deberá implementar una serie de misiones descritas más abajo. En ellas deberá utilizar las librerías **pandas**, **matplotlib**, **sklearn** **geopandas** y **bs4**.

Reglas de las misiones

Las misiones estarán divididas en dos partes. Cada parte reemplazará los contenidos de los laboratorios 5 y 6 respectivamente. Además, cada parte tendrá una asignación de puntaje independiente. Al responder las misiones, puede usar funciones de las librerías mencionadas u otras librerías de análisis de datos. Si utiliza nuevas librerías, debe dejarlas especificadas y justificadas en la celda posterior a uso (como celda de texto). Con el fin de facilitar la corrección, programe cada misión en un celda aislada (puede usar más de una celda por misión, pero no combinar desarrollos de varias misiones en una misma celda). Indique con un comentario al inicio de cada celda la misión que se trabaja en dicha celda. Cuando entregue su tarea procure **NO SUBIR LA BASE DE DATOS A GITHUB**

```
#Mision X PARTE Y
#acá va su desarrollo
#el output de la celda debe ser lo que se pida en cada misión
```

Donde “X” es el número de la misión a responder e “Y” es la parte de la misión. La respuesta debe poder visualizarse directamente en Python. A continuación se describen las misiones que deberá completar:

Misiones: Primera parte

- M1. Dado que has aceptado nuevo el trabajo impuesto por MTT, su primera misión será importar todas las tablas mencionadas mediante **pandas**. Para ello debe asegurar que su computador posee la librería pandas instalada. Aproveche el conocimiento que posee de los datos (por el laboratorio 4) para entregar información extra a la función de importación. Es decir, si los decimales están con punto o coma; o qué símbolo se utiliza como separador de datos. En esta misión se espera que cree un *dataframe* (objeto de **pandas**) para cada una de las tablas importadas. Con el objetivo de asegurar su importación, utilice el método *head* con el fin de mostrar las 5 primeras filas de datos. **Output esperado:** visualización de 5 filas de cada *dataframe*. (0.5 puntos)
- M2. ¿Cuántos viajes se realizan por propósito agregado (columna PropositoAgregado) del viaje? Utilice una función de **pandas** que le permita responder rápidamente esta pregunta. Con la información obtenida indique el porcentaje del total de viajes para cada propósito de forma genérica. Esto último quiere decir que, su código debe funcionar incluso si lo probamos con una porción de los datos. **Output esperado:** lista de tuplas con los porcentajes de viajes (propósito, porcentaje). (1.0 punto)

- M3. Construya un gráfico que permita identificar las comunas (basta con el código de la comuna para buscar) que poseen la mayor dispersión del ingreso del hogar (use el *dataframe* de hogares). **Output esperado:** Gráfico descrito en la misión. (0.8 puntos)
- M4. Investigue el uso de la función *loc* en pandas. Esta función le permite seleccionar un subconjunto de datos que cumplen con ciertas condiciones. Construya un nuevo *dataframe* a partir de la tabla Viajes. Utilice la función *loc* para obtener solo los viajes que se originan en una comuna en específico o llegan a una específico. **Output esperado:** Dos funciones, una que se llame “con_origen” y la otra “con_destino”, ambas deben recibir una comuna (como texto) y deberán retornar un *dataframe* solo con viajes que salen de esa comuna y la otra con viajes que llegan a esa comuna, respectivamente. (0.8 puntos)
- M5. Ahora que estas familiarizado con los filtros de información. Utilice la función **loc** para filtrar el *dataframe* de viajes y obtener solo los viajes que cuenten con información de las coordenadas de origen y de destino (ambas). Una vez obtenido, investigue sobre la función **apply** de pandas para crear una nueva columna llamada Distancia. Esta columna debe contener la distancia euclidiana entre el origen y el destino (utilice la fórmula de distancia entre dos puntos). **Output esperado:** Un *dataframe* de viajes filtrado y con una nueva columna de distancia obtenida de la forma indicada. Este *dataframe* será utilizado más adelante. (0.5 puntos)
- M6. Basado en el laboratorio 4, realice las siguientes consultas utilizando Pandas:
- (a) Construya un *dataframe* que incluya a todas las personas que participaron de la EOD, sean hombres y que posean licencia de conducir. (0.2 puntos)
 - (b) Tomando como base la respuesta obtenida de la parte (a). Construya un nuevo *dataframe* que entregue las personas que sean mayores de 30 y menores de 45. (0.2 puntos)
- M7. Basado en el *dataframe* resultante de la misión 5, construya tres gráficos (utilizando pandas y matplotlib) que muestren la relación entre la distancia media de viaje y tres categorías a elección. **Output esperado:** tres gráficos que muestren relación entre distancia de viaje promedio y tres variables diferentes **BONUS:** sea creativo y será premiado (0.5 puntos)
- M8. Llegó la hora de hacer algunas predicciones. Basándose en los capítulos **1.2.2.- Limpieza y depuración de los datos** y **1.2.3.- Construcción de modelos predictivos** de la materia del curso, su misión será realizar algunas predicciones. Usando la información presente, genere un modelo de

predicción del largo de viaje. Es libre de definir su modelo y variables a considerar, crear nuevas columnas, qué variables son independientes, cuáles son variables dependientes. **Output esperado:** Revisión de consistencia de datos a utilizar, posibles depuraciones y evaluación para completar datos faltantes. Además de la ejecución de un modelo predictivo que prediga el largo del viaje. (1.5 puntos)

Misiones: Segunda parte

Las misiones de la segunda parte estarán disponibles luego de la publicación de los contenidos respectivos. Estos serán publicados el 6 de Noviembre de 2019.

Corrección

Para la corrección de este laboratorio, se revisarán las respuestas entregadas para cada una de las misiones. Utilice múltiples celdas para facilitar su corrección. Pero recuerde no incluir más de una misión en la misma celda. Sea ordenado y claro en presentar sus respuestas.

Este laboratorio consta de dos partes independientes que darán origen a dos notas las que reemplazarán las notas de los L05 y L06 originales.

Política de Integridad Académica

“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.