Sistemas de Gestión de Bases de Datos
$1^{er.}$ cuatrimestre 2019

Guía Práctica 2
Procesamiento de tablas

Ingeniería Informática
Universidad Nacional de Avellaneda

1. Introducción

Los objetivos de esta práctica son:

- Comprender las estructuras de datos para el procesamiento de tablas
- Procesar archivos csv
- Describir estadísticamente un conjunto de datos
- Visualizar resultados mediante gráficos de barras y scatterplots

La bibliografía utilizada es:

1. *Data Science From Scratch*, Joel Grus, O'Reilly
2. *Python for Data Analysis*, Wes McKinney, O'Reilly

2. Actividades preliminares

En esta práctica van a trabajar con la librería "pandas" que provee estructuras de datos convenientes para procesar tablas (tipo Excel) en python.

Se requiere instalar los siguientes módulos de python (en caso de que no estén instalados): "matplotlib", "pandas" y "ipython".

2.1. Actividades

En el capítulo 5 del libro 2 ("Getting started with pandas") realizar de modo interactivo (mediante ipython) las actividades de las siguientes secciones:

1. Introduction to pandas Data Structures (pág. 112-120)
2. Indexing, selection, and filtering (pág. 125-128)
3. Function application and mapping (pág. 132-133)
4. Summarizing and Computing Descriptive Statistics (pág. 137-139)
5. Unique Values, Value Counts, and Membership (pág. 141-142)
6. Filtering Out Missing Data (pág. 143-145)

3. Ejercicios

El sitio `http://www.properati.com.ar` publica información sobre alquiler y venta de inmuebles.

El archivo "properati-AR-2018-02-01-properties-sell.csv" contiene información sobre venta de inmuebles registrado en febrero de 2018.

3.1. Ejercicio 1

Escribir programas en python para:

1. Calcular el valor medio de los deptos 2 ambientes en Capital Federal
2. Hacer un gráfico de barras por cantidad de ambientes en Capital Federal quitando los outliers
3. Hacer un gráfico de barras horizontal de los 10 barrios con mayor cantidad de publicaciones de deptos. de 2 ambientes en Capital Federal

3.2. Ejercicio 2

Para aquellas propiedades de Capital Federal que tengan información geográfica se pide escribir un programa para hacer un scatterplot de las propiedades que difieran a lo sumo en 0.05 grados en latitud y longitud respecto al centro geográfico de la ciudad. Obs.: obtener las coordenadas del centro de la ciudad de modo aproximado con googlemaps).

3.3. Ejercicio 3

En el cap. 5 (pág. 96-102) del libro 1 se explican las medidas de tendencia central y de dispersion para descibir estadísticamente un conjunto de datos. En este ejercicio la idea es utilizar algunas de esas medidas para realizar especulaciones fundamentadas sobre el conjunto de datos.

Las cinco ciudades con mayor población del país son:

1. Buenos Aires
2. Cordoba
3. Rosario
4. La Plata
5. Mar del Plata

3.3.1.

Considerando solamente los deptos. de 3 ambientes, escribir un programa que grafique un boxplot de los precios de esos deptos. de las 5 ciudades mencionadas.

3.3.2.

Basandose en el gráfico anterior, responder a las siguientes preguntas:

1. ¿Cuál es la ciudad con mayor costo de vida?
2. ¿Cuál es la ciudad más equitativa?
3. Proponer algunos argumentos por los cuales sería incorrecto deducir las dos respuestas anteriores del conjunto de datos que estamos utilizando

4. Referencias

- Para importar archivos csv como Data Frames en pandas ver las primeras páginas del capítulo 6 ("Data Loading, Storage, and File Formats") del libro 2.
- Para ver la funcionalidad de ploteo en pandas ver el capítulo 8 ("Plotting and Visualization") sección Plotting Functions in pandas"del libro 2.
- Más sobre visualización:
  - scatterplots se puede ver en el capítulo 3 de libro 1
  - `https://pandas.pydata.org/pandas-docs/stable/user_guide/visualization.html`
  - `https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.boxplot.html`