IIC2433 — Minería de Datos

Desafío: correlación y limpieza de datos

Trabajaremos con los datos del sistema nacional de información municipal, del gobierno de Chile, y una planilla de datos preliminares del plebiscito 2020[1]
Tienes acceso a tres fuentes de datos.

- Un archivo csv llamado poblacion.csv, que tiene la población de las comunas.
- Un archivo csv llamado presupuesto.csv, con el presupuesto de las comunas.
- Una archivo csv llamad `Resultados_Pleb.csv`, que contiene la votación agregada de cada comuna para el plebiscito de Octubre.

La idea es hacer un análisis de datos en torno a la siguiente problemática:

> ¿Cómo se relaciona el presupuesto de las comunas con su votación en el plebiscito?

1. Tareas a realizar

Ten cuidado: puede ser que los datos estén sucios, pues son llenados por humanos. Es tu responsabilidad limpiar esos datos: para las siguientes tareas no debes tomar en cuenta datos nulos o que no han sido reportados (aunque eso signifique dejar fuera algunas comunas).

1.1. Manejo de datos

1. Importa los datos desde los archivos .csv
2. . Muy posiblemente nos va a convenir trabajar con un dataframe grande, que tenga en cada fila toda la información que conocemos para cada comuna. Crea este dataframe (lo típico es usar el comando merge de Pandas).
3. ¿Cuántas comunas tiene Chile?

[1]Todavía no tenemos la versión oficial. Esta fue compilada por Vergara Perucich, Francisco & Greene, Ricardo & Correa Parra, Juan & Aguirre-Nuñez, Carlos & Cancino Contreras, Francisca. (2020). Cartografías del apruebo: Análisis preliminar del plebiscito para cambio constitucional, Chile 2020. 10.13140/RG.2.2.24281.34408.

1.2. Correlación entre habitantes y presupuesto

1. Crea un gráfico de puntos para visualizar la correlación entre el presupuesto de las comunas y la cantidad de habitantes de esa comuna.

2. Observa que la correlación es ligeramente positiva (como es de esperar). Sin embargo, hay un outlier cuyo presupuesto se escapa. Puedes ver qué comuna es?

1.3. Descartando ciertos outliers

Ahora refinemos un poco la tabla de comunas que vamos a utilizar.

1. Genera un nuevo dataframe que contenga todas las comunas, **excepto** i) El outlier que identificaste en el apartado anterior, y ii) todas las comunas que tengan 5000 habitantes o menos. Deberían haber 286 comunas.

1.4. Correlación entre presupuesto y votación del apruebo

1. Crea un gráfico de puntos para visualizar la correlación entre el presupuesto de las comunas y el porcentaje de votos apruebo (sobre el total de votos) en esas comunas. Anota el coeficiente de correlación.

2. Ahora visualiza el mismo gráfico, pero utilizando solo aquellas comunas que filtraste en el apartado anterior. Vuelve a anotar el coeficiente de correlación.

3. ¿Como cambia la correlación? ¿Por qué crees que pasó eso?

2. Detalles administrativos

Junto con subir tu control al cuestionario en siding, debes *autoreportar* tu nivel de avance en la planilla excel de autoreporte (busca el link en la página del curso). Los niveles posibles son **L** (de logrado, cuando completaste todo o casi todo), **P** (de parcial, cuando trabajaste un poco pero no alcanzaste a ver todo), o **N** (de no logrado, cuando no hiciste mucho más que leer el notebook).