

Pipeline Proyecto

Alumno: Apaza Apaza Nelzon Jorge

Curso: Tópicos en Ciencia de la Computación

1. Tema/área de interés:
Urban Data Analysis (Análisis Visual).
2. Fuente de datos:
El dataset proviene del VAST Challenge que es organizado por el IEEE VAST Symposium que es reconocida internacionalmente en el ámbito de la visualización y el análisis de datos. Cabe destacar que por límites de capacidad de almacenamiento limitado de mi unidad de Drive, solo estoy usando 10 gigas de las 22 gigas que es el tamaño original de los datos. Para que no se vea afectada la actividad que se desarrolla en este documento, se priorizan los archivos principales del reto en general.

1. Breve contexto:

¡En Engagement, Ohio, el futuro es ahora! Durante años, esta tranquila comunidad residencial fue una joya secreta en el corazón del estado. ¡Pero ahora el secreto ha salido a la luz y la gente no camina, sino que corre para reclamar su pedazo de este paraíso!







Anticipándose a un rápido crecimiento, la ciudad de Engagement, Ohio (EE. UU.) está realizando un ejercicio de planificación urbana participativa para comprender el estado actual de la ciudad e identificar oportunidades para su desarrollo futuro. Alrededor de 1000 residentes representativos de esta ciudad de tamaño modesto han aceptado proporcionar datos utilizando la aplicación de planificación urbana de la ciudad, la cual registra los lugares que visitan, sus gastos y sus compras, entre otras cosas. A partir de estos voluntarios, la ciudad contará con datos que ayudarán en sus importantes esfuerzos de revitalización comunitaria, incluyendo cómo asignar una gran subvención de renovación urbana que han recibido recientemente. Como experto en análisis visual, te has unido al equipo de planificación urbana para dar sentido a los datos proporcionados por estos residentes.

2. Análisis rápido de la base de datos (luego están las preguntas):

Revisión, recolección y limpieza de tus datos.

- Los datos estan almacenados de la siguiente manera:

Carpetas y archivos:

Nombre	Propietario
 Activity Logs	 yo
 Journals	 yo
 Attributes	 yo

...

Proyecto - Base de Dat...

Activity Logs

Tipo

Personas

Modificado

Fuente

Nombre	Propietario
.DS_Store	yo
ParticipantStatusLogs10.csv	yo
ParticipantStatusLogs9.csv	yo
ParticipantStatusLogs8.csv	yo
ParticipantStatusLogs7.csv	yo
ParticipantStatusLogs6.csv	yo
ParticipantStatusLogs5.csv	yo
ParticipantStatusLogs4.csv	yo
ParticipantStatusLogs3.csv	yo
ParticipantStatusLogs2.csv	yo
ParticipantStatusLogs1.csv	yo

...

Proyecto - Base de Dat...

Journals

Tipo

Personas

Modificado

Fuente

Nombre	Propietario
.DS_Store	yo
TravelJournal.csv	yo
SocialNetwork.csv	yo
FinancialJournal.csv	yo
CheckinJournal.csv	yo

...

Proyecto - Base de Dat...

Attributes

Tipo

Personas

Modificado

Fuente

Para determinar el formato de los archivos, se revisó cada uno con un editor de texto simple:

ParticipantStatusLogs1.csv: Bloc de notas

Archivo

Edición

Formato

Ver

Ayuda

```
timestamp,currentLocation,participantId,currentMode,hungerStatus,sleepStatus,apartmentId,available
2022-03-01T00:00:00Z,POINT (-2724.6277665310454 6866.2081834436985),0,AtHome,JustAte,Sleeping,926,
2022-03-01T00:00:00Z,POINT (-1526.9372331431534 5582.2951345645315),1,AtHome,JustAte,Sleeping,928,
2022-03-01T00:00:00Z,POINT (-1360.9905987829304 2108.804385379679),2,AtHome,JustAte,Sleeping,291,1
2022-03-01T00:00:00Z,POINT (-1558.517200825967 5600.664347152427),3,AtHome,JustAte,Sleeping,1243,1
2022-03-01T00:00:00Z,POINT (976.2409614204214 4574.575079082071),4,AtHome,JustAte,Sleeping,194,-68
2022-03-01T00:00:00Z,POINT (-1525.6957374012197 1994.5285187115571),5,AtHome,JustAte,Sleeping,243,
2022-03-01T00:00:00Z,POINT (1795.1297501295278 3238.4053705049837),6,AtHome,JustAte,Sleeping,183,1
2022-03-01T00:00:00Z,POINT (-1023.8165705255449 1578.3713681439597),7,AtHome,JustAte,Sleeping,97,6
2022-03-01T00:00:00Z,POINT (616.2956028633527 2274.8909931311796),8,AtHome,JustAte,Sleeping,321,54
```

Apartments.csv: Bloc de notas

Archivo

Edición

Formato

Ver

Ayuda

```
apartmentId,rentalCost,maxOccupancy,numberOfRooms,location,buildingId
1,768.16,2,4,POINT (1077.6979444315298 648.4427163702453),340
2,1014.55,2,1,POINT (-185.9292838076562 1520.3270983045118),752
3,1057.39,4,3,POINT (2123.0141855392585 5126.753457243003),639
4,1259.1,4,3,POINT (2103.6301776944765 4266.932930123476),397
5,411.5,1,4,POINT (7.0589743819342985 79.96163671849988),628
6,859.58,3,2,POINT (2250.85490611142 5251.3368306902885),533
7,982.11,3,4,POINT (486.8811262316384 2251.1260599901484),61
8,980.05,4,1,POINT (1233.4547558395932 1768.6111384755895),360
9,433.45,1,3,POINT (1274.2715913565519 1163.5051209752276),251
10,1104.33,3,4,POINT (-1697.0303105735857 1239.0301787057274),512
11,960.16,3,2,POINT (-341.9635948004018 1772.5966399871515),922
```

Jobs.csv: Bloc de notas

Archivo

Edición

Formato

Ver

Ayuda

```
jobId,employerId,hourlyRate,startTime,endTime,daysToWork,educationRequirement
0,379,10,7:46:00 AM,3:46:00 PM,"[Monday,Tuesday,Wednesday,Thursday,Friday]",HighSchoolOrColl
1,379,22,21763336,7:31:00 AM,3:31:00 PM,"[Monday,Tuesday,Wednesday,Thursday,Friday]",Bachelor
2,380,10,8:00:00 AM,4:00:00 PM,"[Monday,Tuesday,Wednesday,Thursday,Friday]",Bachelors
3,380,15,31207064,7:39:00 AM,3:39:00 PM,"[Monday,Tuesday,Wednesday,Thursday,Friday]",Bachelor
4,381,21,35540929,7:53:00 AM,3:53:00 PM,"[Monday,Tuesday,Wednesday,Thursday,Friday]",HighSch
5,381,12,09382569,8:13:00 AM,4:13:00 PM,"[Monday,Sunday,Thursday,Tuesday,Saturday]",HighSch
6,381,21,84618702,8:36:00 AM,4:36:00 PM,"[Monday,Tuesday,Wednesday,Thursday,Friday]",HighSch
7,381,18,71319307,7:47:00 AM,3:47:00 PM,"[Monday,Tuesday,Wednesday,Thursday,Friday]",Low
8,382,13,37043245,7:45:00 AM,3:45:00 PM,"[Monday,Tuesday,Wednesday,Thursday,Friday]",Bachelor
9,382,24,52833044,7:10:00 AM,3:10:00 PM,"[Monday,Tuesday,Wednesday,Thursday,Friday]",Graduate
```

Como se ve en las imágenes, incluyendo a los archivos faltantes, todos son de **formato Delimitado** usando como caracter específico a la **coma(,)**, que suele usarse en **archivos con extensión csv**. Además de ello, la primera línea en todos los archivos da a conocer el nombre de las columnas.

- **Encoding de mis archivos:**

Se revisó la fuente de la base de datos y no nos brinda una documentación donde se tenga la codificación que se usó. Tal y como indica el libro leído en clases, inferimos con Chardet:

archivo	Codificación	Confianza
rive/MyDrive/STO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Activity Logs/ParticipantStatusLogs1.csv	ascii	1.0
rive/MyDrive/STO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Activity Logs/ParticipantStatusLogs2.csv	ascii	1.0
rive/MyDrive/STO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Activity Logs/ParticipantStatusLogs3.csv	ascii	1.0
rive/MyDrive/STO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Activity Logs/ParticipantStatusLogs4.csv	ascii	1.0
rive/MyDrive/STO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Activity Logs/ParticipantStatusLogs5.csv	ascii	1.0
rive/MyDrive/STO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Activity Logs/ParticipantStatusLogs6.csv	ascii	1.0
rive/MyDrive/STO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Activity Logs/ParticipantStatusLogs7.csv	ascii	1.0
rive/MyDrive/STO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Activity Logs/ParticipantStatusLogs8.csv	ascii	1.0
rive/MyDrive/STO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Activity Logs/ParticipantStatusLogs9.csv	ascii	1.0
rive/MyDrive/STO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Activity Logs/ParticipantStatusLogs10.csv	ascii	1.0
rive/MyDrive/STO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Attributes/Participants.csv	ascii	1.0
rive/MyDrive/STO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Attributes/Restaurants.csv	ascii	1.0
rive/MyDrive/STO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Attributes/Schools.csv	ascii	1.0
rive/MyDrive/STO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Attributes/Employers.csv	ascii	1.0
rive/MyDrive/STO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Attributes/Buildings.csv	ascii	1.0
rive/MyDrive/STO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Attributes/Apartments.csv	ascii	1.0
rive/MyDrive/STO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Attributes/Pubs.csv	ascii	1.0
rive/MyDrive/STO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Attributes/Jobs.csv	ascii	1.0
rive/MyDrive/STO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Journals/CheckinJournal.csv	ascii	1.0
rive/MyDrive/STO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Journals/FinancialJournal.csv	ascii	1.0
rive/MyDrive/STO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Journals/SocialNetwork.csv	ascii	1.0
rive/MyDrive/STO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Journals/TravelJournal.csv	ascii	1.0

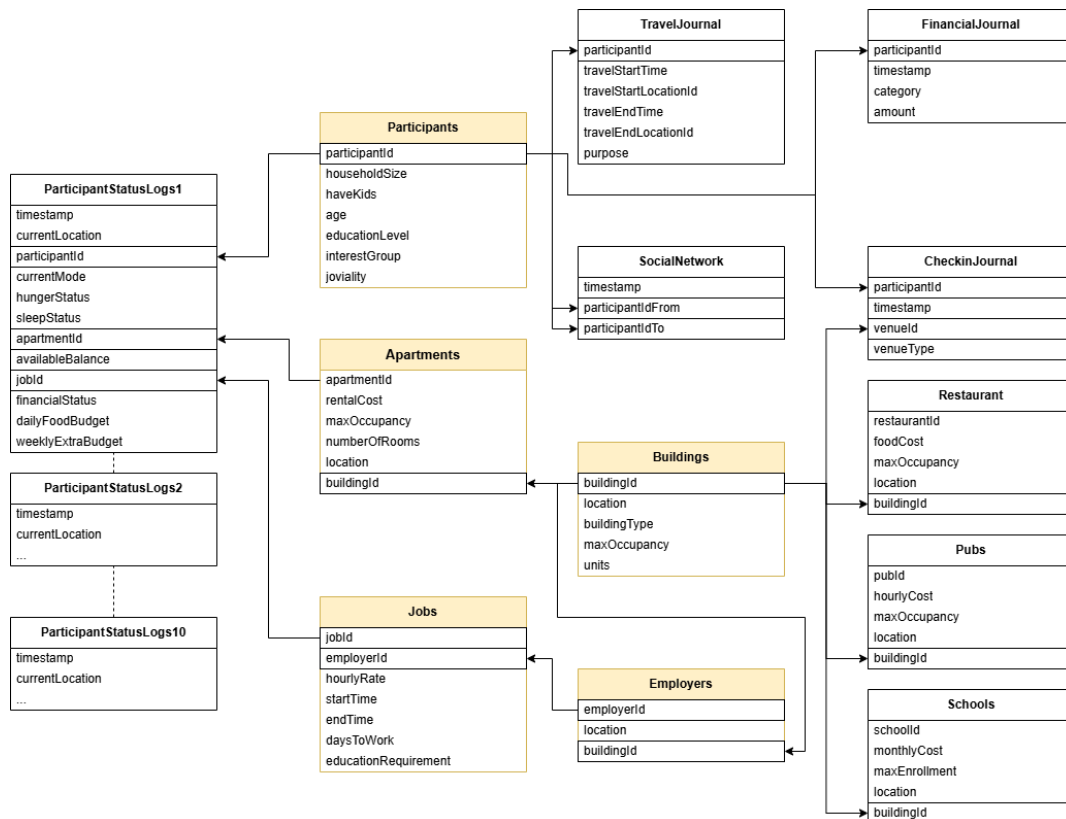
Resultado: Todos los archivos .csv tienen codificación ASCII (American Standard Code for Information Interchange), con una confianza del 100%.

- **Tamaño archivos:**

Todos los archivos .csv expresados en Kibibytes (KiB - 1024). Se usó la biblioteca “os”. A continuación tenemos el tamaño de todos los archivos.

Nombre del archivo	Tamaño (KiB)
/content/drive/MyDrive/STO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Activity Logs/ParticipantStatusLogs1.csv	239419.32
/content/drive/MyDrive/STO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Activity Logs/ParticipantStatusLogs2.csv	221062.57
/content/drive/MyDrive/STO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Activity Logs/ParticipantStatusLogs3.csv	233192.95
/content/drive/MyDrive/STO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Activity Logs/ParticipantStatusLogs4.csv	236815.18
/content/drive/MyDrive/STO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Activity Logs/ParticipantStatusLogs5.csv	236957.12
/content/drive/MyDrive/STO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Activity Logs/ParticipantStatusLogs6.csv	237644.42
/content/drive/MyDrive/STO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Activity Logs/ParticipantStatusLogs7.csv	237854.42
/content/drive/MyDrive/STO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Activity Logs/ParticipantStatusLogs8.csv	236753.26
/content/drive/MyDrive/STO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Activity Logs/ParticipantStatusLogs9.csv	236739.76
/content/drive/MyDrive/STO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Activity Logs/ParticipantStatusLogs10.csv	236728.95
/content/drive/MyDrive/STO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Attributes/Participants.csv	43.53
/content/drive/MyDrive/STO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Attributes/Restaurants.csv	1.28
/content/drive/MyDrive/STO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Attributes/Schools.csv	0.33
/content/drive/MyDrive/STO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Attributes/Employers.csv	13.53
/content/drive/MyDrive/STO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Attributes/Buildings.csv	337.36
/content/drive/MyDrive/STO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Attributes/Apartments.csv	97.72
/content/drive/MyDrive/STO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Attributes/Pubs.csv	0.87
/content/drive/MyDrive/STO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Attributes/Jobs.csv	130.9
/content/drive/MyDrive/STO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Journals/CheckinJournal.csv	79030.51
/content/drive/MyDrive/STO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Journals/FinancialJournal.csv	82646.44
/content/drive/MyDrive/STO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Journals/SocialNetwork.csv	210246.61
/content/drive/MyDrive/STO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Journals/TravelJournal.csv	314269.8

- **Datos que tenemos: (Diagram)**



PREGUNTAS PROPUESTAS A RESOLVER:

1. ¿Qué problemas identifican en el dataset? (Dar 2)

Analizando el dataset que está orientado hacia lo que busca el proyecto, se hizo las siguientes preguntas:

- ¿En qué vecindario (edificio/apartamento) vive la mayor cantidad de familias con niños y qué características hacen que esa área sea tan popular para ellos?
- ¿El nivel económico está relacionado con el grado de estudios y la ubicación geográfica?
- Imaginando que una familia de cuatro personas dispone de poco más de 3 000 USD al mes y al menos un miembro con educación secundaria. ¿En qué edificio o conjunto de apartamentos les recomendarías vivir y por qué esa opción es la mejor?

Pero en el camino para resolver dichas preguntas, como en el Pipeline de Hubway, nos encontramos con algunos problemas, 2 de ellos son:

- Al observar "ParticipantStatusLogs*", hay algunos archivos que en apartmentId los tiene como tipo de almacenamiento "float", pero en verdad debería ser "int".
 - Debido a ello, convertimos la columna 'apartmentId' a tipo entero,
 - apartmentId es un identificador discreto y nominal (nunca debería tener decimales).

```

Tipos de datos por columna:
timestamp          object
currentLocation    object
participantId      int64
currentMode        object
hungerStatus       object
sleepStatus        object
apartmentId        int64
availableBalance    float64
jobId              int64
financialStatus     object
dailyFoodBudget    int64
weeklyExtraBudget   float64
dtype: object
  
```

- b. En la tabla de apartments "location" es una variable cualitativa nominal codificada en una estructura tipo texto, con contenido que representa coordenadas.

apartmentId	rentalCost	maxOccupancy	numberOfRooms	location	built
0	1	768.16	2	4	POINT (1077.6979444315298 648.4427163702453)
1	2	1014.55	2	1	POINT (-185.9292838076562 1520.3270983045118)
2	3	1057.39	4	3	POINT (2123.0141855392585 5126.753457243003)
3	4	1259.10	4	3	POINT (2103.6301776944765 4266.932930123476)
4	5	411.50	1	4	POINT (7.0589743819342985 79.96163671849988)

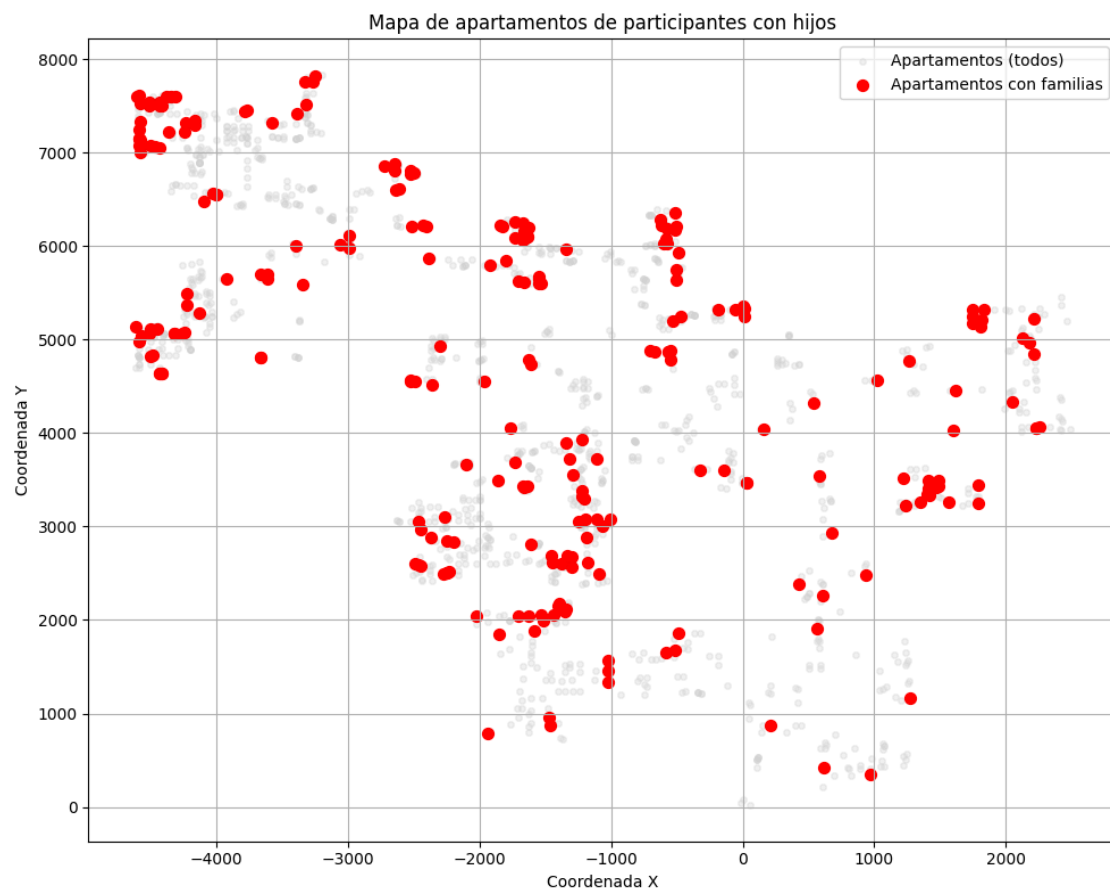
- Su forma original "POINT (x y)" no puede ser procesada numéricamente ni graficada sin transformación. Además la ubicación en el espacio geográfico tiene significado métrico: permite comparar distancias, calcular cercanías, concentraciones, proximidades, etc.
- Así que nos conviene transformar a dos variables cuantitativas continuas: x y.

```
def parse_point(point_str):
    # elimino las palabras "POINT (" y el otro parentesis ")" para extraer "x y"
    coords = point_str.replace("POINT (", "").replace(")", "").split()
    return float(coords[0]), float(coords[1])

Aplico la función para extraer 'x' y 'y'

[ ] apartments[['x', 'y']] = apartments['location'].apply(lambda loc: pd.Series(parse_point(loc)))
```

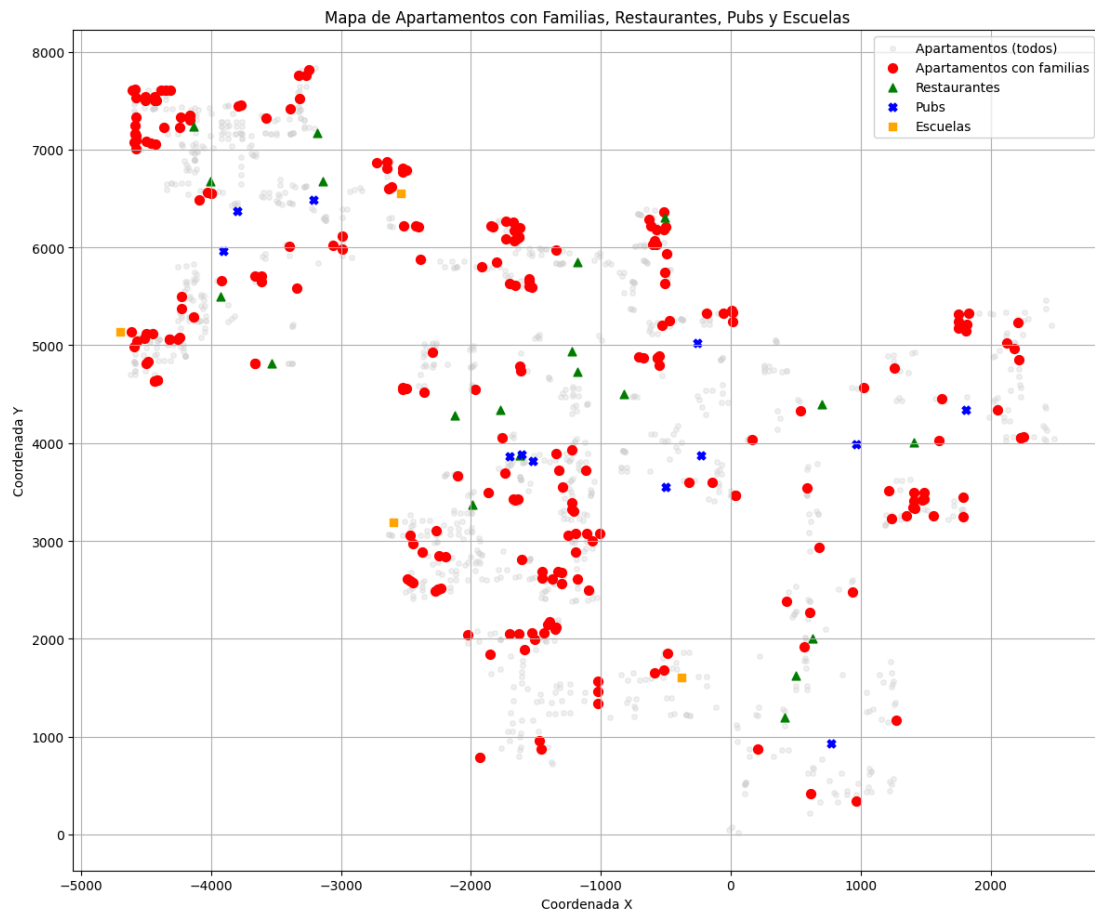
2. ¿Qué descubrieron al analizar los datos? (DAR 1)



Se descubrió que existen zonas preferidas para vivir para las familias que se tienen en el dataset, ellos prefieren tanto la parte céntrica como la parte de la esquina superior izquierda,

viendo que las demás zonas no viven muchas de ellas. Esto nos da una nueva pregunta ¿Por qué estas personas(familias), prefieren dichas zonas? ¿Acaso habrá algún local o actividad que los atrae?

3. ¿Qué reflejan los patrones de tendencia? (DAR 2)



- Las zonas que mayor número de habitantes en general tienen son aquellas que tienen mayor presencia de 3 tipos de edificios: restaurantes, escuelas y pubs. Esto incluso responde a nuestra pregunta propuesta en la pregunta anterior, donde se ve que las familias también se inclinan a buscar dichas zonas.
 - Se puede apreciar que hay zonas donde hay presencia de un mayor número de restaurantes, pero poca o ninguna presencia de escuelas o pubs. Estas zonas no son muy habitadas, demostrando que las personas prefieren estar en lugares donde puedan acceder fácilmente a distintos lugares.
4. ¿Cómo es afectado el comportamiento humano (movilidad) en relación con la geografía?
- En la imagen de abajo, e incluso revisando las anteriores, el movimiento de las personas se da en torno a la ubicación de los principales edificios: restaurantes, escuelas, y pubs. Donde resalto a los colegios, ya que alrededor de estos, la movilidad de las personas es mayor. Y si se tiene a los 3 tipos de edificios que acabo de mencionar, el movimiento es mayor en dichas zonas.

