
Práctica de Laboratorio: Análisis Exploratorio de Datos - Data Wrangling

Docente: Ana Maria Cuadros Valdivia

Alumno: Nelzon Apaza Apaza

CONJUNTO DE DATOS VAST CHALLENGE 2022

Breve Contexto:

1. Descripción (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.

1.1. Preguntas y observaciones realizadas en el proceso:

1.1.1 ¿Por qué Engagement fue una joya secreta del corazón del estado?:

Dato curioso en la vida real: "El Corazón de Ohio" (Ohio, the Heart of It All) es el nuevo eslogan oficial del estado de Ohio, utilizado para atraer visitantes y resaltar su ubicación central en Estados Unidos.

Además, normalmente Ohio suele hacer propaganda de diferentes lugares como Joyas Secretas. Cada ciudad se caracteriza por algo resaltante como historia del lugar, gastronomía, etc. Engagement (a pesar de ser ficticia) podría llegar a tener alguna característica importante que llamó la atención de todas las personas para querer vivir ahí.

1.1.2. ¿Cuál es el secreto que salió a la luz?

No se menciona, pero como mencioné anteriormente, Ohio suele hacer propaganda de diferentes lugares como Joyas Secretas. Cada ciudad se caracteriza por algo resaltante como historia del lugar, gastronomía, etc. Engagement (a pesar de ser ficticia) podría llegar a tener alguna característica importante que llamó la atención de todas las personas para querer vivir ahí.

1.1.3. Observación:

Al parecer los datos que se tomaron, fue antes de que lleguen más personas, debido al secreto revelado. Por ello se quieren anticipar.

1.1.4. ¿Por qué recibieron una gran subvención de renovación urbana?

Las ciudades que suelen recibir este apoyo, pueden presentar desafíos específicos como deterioro de infraestructura, pobreza, desempleo, degradación ambiental o necesitan revitalización económica(pero no se menciona directamente). Con ello, podríamos llegar a inferir el estado en el que se encuentra la ciudad.

1.1.5. Observación:

Efectivamente, necesitan conocer el estado actual de la ciudad, que incluso podría darnos a conocer el secreto que fue revelado, y así también se pueda identificar oportunidades de desarrollo a futuro.

2. Formulación del Problema / Preguntas

¿Cómo la base de datos podría servir para lograr que un grupo de personas establezcan estrategias de búsqueda de información colaborativamente?

- ¿Cual es mi objeto de estudio?

Se tiene los datos de la ciudad de Engagement, el objeto de estudio son los ciudadanos de dicha ciudad. Cabe resaltar que todo se verá con un enfoque a ser desarrollado de manera colaborativa.

- ¿Qué queremos saber?:

Para lograr lo colaborativo, direccionar a los usuarios a trabajar de manera colaborativa, necesito conocer qué interacciones realizaré, abstraer qué visualizaciones podría usar para expresar la información. Analizando también qué tan completa es la data para visualizarla o qué tan fácil es.

- **¿Cuáles son nuestras hipótesis?**
 - ¿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?

- Imagina 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?
- ¿Cuáles son nuestras métricas de éxito? (De las mismas hipótesis)
 - En las 2 primeras hipótesis tendrá que mostrarse un gráfico representando un mapa con lo pedido.
 - En la tercera hipótesis además de un mapa, tendrá que obtenerse datos que simulan el caso dado, sobre todo con los precios de los edificios que presentan precios en sus servicios.
- Recopilación de datos

En el contexto los datos fueron recopilados por la misma ciudad de Engagement. Y por otra parte, esta base de datos fue dada en el evento IEEE VIS 2022, organizado por la IEEE VIS Conference, un evento líder en visualización y análisis visual de datos.

- ¿Cómo se recopilaron los datos?

En el contexto, los datos fueron recolectados mediante un “aplicativo de planificación urbana”.

- ¿Cuándo se compilaron los datos?(La fecha afecta la actualidad y relevancia del análisis)

No se tiene una fecha específica en el contexto, que indique cuándo se compilaron los datos, pero al parecer estos fueron capturados antes de la venida masiva de más personas. Pero si revisamos “ParticipantStatusLogs1.csv” que representa a la recolección de datos del ciudadano participante cada cierto tiempo, se marca la primera fecha: “2022-03-01”.







- ¿La fuente es confiable o tiene autoridad?

Dentro del contexto, si es confiable, ya que es una fuente de datos gubernamental. Además de ello, el evento IEEE VIS 2022, organizado por la IEEE VIS Conference, es muy confiable, ya que es un evento líder en visualización y análisis visual de datos que organiza diferentes retos cada año.

3. Adquisición de Datos y Limpieza

- Almacenamiento de los datos:

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,-68
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 (-341.9635048000018 1772.5966329871515),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]",HighSchoolOrColle
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]",HighSchol
5,381,12.09382569,8:13:00 AM,4:13:00 PM,"[Monday,Sunday,Thursday,Tuesday,Saturday]",HighSchol
6,381,21.84618702,8:36:00 AM,4:36:00 PM,"[Monday,Tuesday,Wednesday,Thursday,Friday]",HighSchol
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
9,382,24.52833044,7:19:00 AM,3:19:00 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.

- File Encoding:

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/5TO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Activity Logs/ParticipantStatusLogs1.csv	ascii	1.0
rive/MyDrive/5TO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Activity Logs/ParticipantStatusLogs2.csv	ascii	1.0
rive/MyDrive/5TO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Activity Logs/ParticipantStatusLogs3.csv	ascii	1.0
rive/MyDrive/5TO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Activity Logs/ParticipantStatusLogs4.csv	ascii	1.0
rive/MyDrive/5TO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Activity Logs/ParticipantStatusLogs5.csv	ascii	1.0
rive/MyDrive/5TO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Activity Logs/ParticipantStatusLogs6.csv	ascii	1.0
rive/MyDrive/5TO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Activity Logs/ParticipantStatusLogs7.csv	ascii	1.0
rive/MyDrive/5TO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Activity Logs/ParticipantStatusLogs8.csv	ascii	1.0
rive/MyDrive/5TO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Activity Logs/ParticipantStatusLogs9.csv	ascii	1.0
rive/MyDrive/5TO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Activity Logs/ParticipantStatusLogs10.csv	ascii	1.0
rive/MyDrive/5TO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Attributes/Participants.csv	ascii	1.0
rive/MyDrive/5TO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Attributes/Restaurants.csv	ascii	1.0
rive/MyDrive/5TO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Attributes/Schools.csv	ascii	1.0
rive/MyDrive/5TO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Attributes/Employers.csv	ascii	1.0
rive/MyDrive/5TO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Attributes/Buildings.csv	ascii	1.0
rive/MyDrive/5TO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Attributes/Apartments.csv	ascii	1.0
rive/MyDrive/5TO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Attributes/Pubs.csv	ascii	1.0
rive/MyDrive/5TO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Attributes/Jobs.csv	ascii	1.0
rive/MyDrive/5TO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Journals/CheckinJournal.csv	ascii	1.0
rive/MyDrive/5TO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Journals/FinancialJournal.csv	ascii	1.0
rive/MyDrive/5TO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Journals/SocialNetwork.csv	ascii	1.0
rive/MyDrive/5TO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Journals/TravelJournal.csv	ascii	1.0

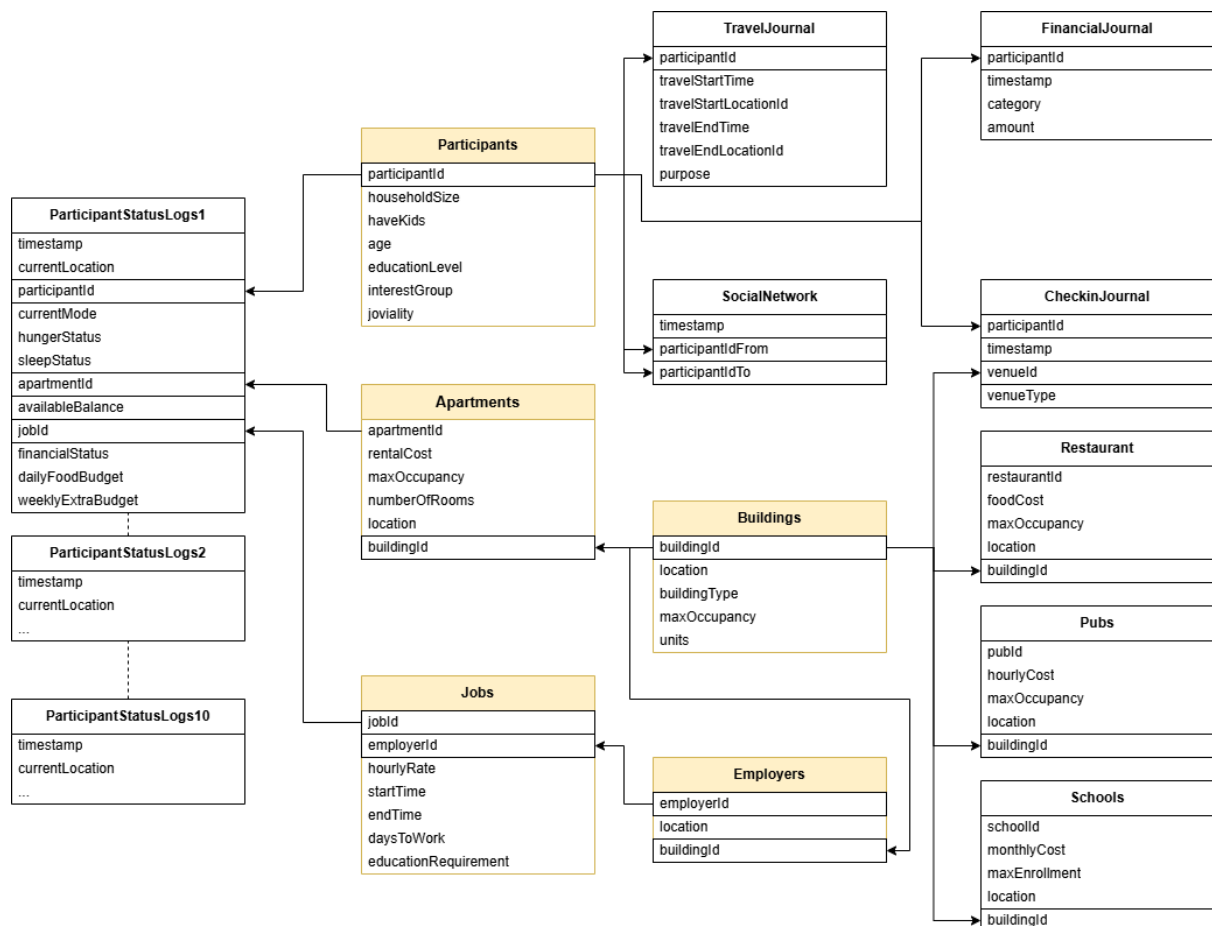
Resultado: Todos los archivos .csv tienen codificación ASCII (American Standard Code for Information Interchange), con una confianza del 100%.

- 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/5TO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Activity Logs/ParticipantStatusLogs1.csv	239419.32
/content/drive/MyDrive/5TO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Activity Logs/ParticipantStatusLogs2.csv	221062.57
/content/drive/MyDrive/5TO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Activity Logs/ParticipantStatusLogs3.csv	233192.95
/content/drive/MyDrive/5TO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Activity Logs/ParticipantStatusLogs4.csv	236015.18
/content/drive/MyDrive/5TO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Activity Logs/ParticipantStatusLogs5.csv	236957.12
/content/drive/MyDrive/5TO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Activity Logs/ParticipantStatusLogs6.csv	237644.42
/content/drive/MyDrive/5TO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Activity Logs/ParticipantStatusLogs7.csv	237054.42
/content/drive/MyDrive/5TO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Activity Logs/ParticipantStatusLogs8.csv	236753.26
/content/drive/MyDrive/5TO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Activity Logs/ParticipantStatusLogs9.csv	236739.76
/content/drive/MyDrive/5TO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Activity Logs/ParticipantStatusLogs10.csv	236728.95
/content/drive/MyDrive/5TO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Attributes/Participants.csv	43.53
/content/drive/MyDrive/5TO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Attributes/Restaurants.csv	1.28
/content/drive/MyDrive/5TO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Attributes/Schools.csv	0.33
/content/drive/MyDrive/5TO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Attributes/Employers.csv	13.53
/content/drive/MyDrive/5TO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Attributes/Buildings.csv	337.36
/content/drive/MyDrive/5TO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Attributes/Apartments.csv	97.72
/content/drive/MyDrive/5TO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Attributes/Pubs.csv	0.87
/content/drive/MyDrive/5TO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Attributes/Jobs.csv	130.9
/content/drive/MyDrive/5TO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Journals/CheckinJournal.csv	79030.51
/content/drive/MyDrive/5TO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Journals/FinancialJournal.csv	82646.44
/content/drive/MyDrive/5TO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Journals/SocialNetwork.csv	210246.61
/content/drive/MyDrive/5TO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Journals/TravelJournal.csv	314269.8

- Datos que tenemos: (Diagrama)



○ Archivos de estado de los participantes
(ParticipantStatusLogs1.csv ... ParticipantStatusLogs10.csv)

Archivo: /content/drive/MyDrive/5TO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Activity Logs/ParticipantStatusLogs1.csv

timestamp	currentLocation	participantId	currentMode	hungerStatus	sleepStatus	apartmentId	availableBalance	jobId	financialStatus	dailyFoodBudget	weeklyExtraBudget
2022-03-01T00:00:00Z	POINT (-2724.6277665310454 6866.2081834436985)	0	AtHome	JustAte	Sleeping	926.0	1286.519556	254	Stable	12	1104.302570
2022-03-01T00:00:00Z	POINT (-1526.9372331431534 5582.2951345645315)	1	AtHome	JustAte	Sleeping	928.0	860.574204	929	Stable	12	926.714377
2022-03-01T00:00:00Z	POINT (-1360.9905987829304 2108.804385379679)	2	AtHome	JustAte	Sleeping	291.0	1298.184541	348	Stable	16	848.802876
2022-03-01T00:00:00Z	POINT (-1558.517200825987 5600.664347152427)	3	AtHome	JustAte	Sleeping	1243.0	1180.641725	316	Stable	12	819.325405
2022-03-01T00:00:00Z	POINT (976.2409614204214 4574.575079082071)	4	AtHome	JustAte	Sleeping	194.0	-681.650588	177	Unstable	20	0.000000

- Las variables o atributos no están describiendo un objeto.
- Un registro representa en sí a un instante de tiempo en el que se capturó el estado de un participante, dando su coordenada geográfica(currentLocation), modo de actividad(currentMode), nivel de hambre/sueño, a qué apartamento pertenece (apartmentId), su saldo(availableBalance), su trabajo(jobId), y otros detalles económicos como dailyFoodBudget, weeklyExtraBudget, etc.

Tablas - ParticipantStatusLogs*.csv:

Variable	Descripción
----------	-------------

timestamp	<ul style="list-style-type: none"> • Fecha y hora del registro • variable cuantitativa continua de tipo temporal (fecha/hora). • 1,667 valores únicos (desde 01-03-2022). • “2022-04-26T00:25:00Z” la granularidad es un instante puntual para cada participante. Pero podríamos mejor tomar magnitudes por semana tal vez.
currentLocation	<ul style="list-style-type: none"> • Ubicación geográfica actual del participante (coordenadas) • variable cuantitativa continua. • Indica un formato POINT (x y). • 23,741 valores únicos.
participantId	<ul style="list-style-type: none"> • Identificador único del participante. • variable cuantitativa discreta. • 1,011 valores únicos (0-1,010).
currentMode	<ul style="list-style-type: none"> • Estado o modo de actividad del participante (AtHome, AtWork, Transport, Shopping, Other). • variable cualitativa nominal.
hungerStatus	<ul style="list-style-type: none"> • Estado de hambre (JustAte, BecomingHungry, Hungry, Starving, BecameFull). • variable cualitativa ordinal si ordenamos por necesidad alimenticia.
sleepStatus	<ul style="list-style-type: none"> • Estado de sueño. • variable cualitativa nominal. • 3 estados: Awake, Sleeping, Drowsy.
apartmentId	<ul style="list-style-type: none"> • Identificador del apartamento donde vive el participante. • variable cuantitativa discreta. • 841 valores únicos (1-1,733). • Cada participante, en un instante, está asociado a un apartamento. Pertenece a un edificio (buildingId) que está en el archivo de edificios (Buildings.csv).
availableBalance	<ul style="list-style-type: none"> • Saldo disponible del participante. • variable cuantitativa continua. • Rango: [-681.65, 5,408.98]
jobId	<ul style="list-style-type: none"> • Identificador del trabajo del participante • variable cuantitativa discreta. • 1,190 valores únicos (1-1,326). • Puedo relacionarlo con la tabla de Jobs.csv.
financialStatus	<ul style="list-style-type: none"> • Estado financiero del participante • variable cualitativa ordinal • Tiene 3 niveles: Stable, Unstable, Critical. • Refleja la salud financiera del participante.
dailyFoodBudget	<ul style="list-style-type: none"> • Presupuesto diario para la comida. • variable cuantitativa discreta. • Presupuesto diario fijo asignado.
weeklyExtraBudget	Presupuesto semanal para gastos extras, Variable cuantitativa continua. 18,221 valores únicos. Rango: [0, 2,553.25]. Presupuesto adicional semanal (¿para gastos no esenciales?).

○ Archivo: Participants.csv

Archivo: /content/drive/MyDrive/5TO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Attributes/Participants.csv

participantId	householdSize	haveKids	age	educationLevel	interestGroup	joviality
0	3	True	36	HighSchoolOrCollege	H	0.001627
1	3	True	25	HighSchoolOrCollege	B	0.328087
2	3	True	35	HighSchoolOrCollege	A	0.393470
3	3	True	21	HighSchoolOrCollege	I	0.138063
4	3	True	43	Bachelors	H	0.857397

- Las variables o atributos están describiendo al objeto: Participante.
- Un registro representa en sí a un participante individual, contiene sus datos demográficos (edad, tamaño del hogar, si tienen hijos, nivel educativo, grupo de interés, jovialidad).

Tabla - Participants.csv:

Variable	Descripción
participantId	Identificador único del participante, variable cuantitativa discreta (identificador único). Tipo . Rango de 0 a 1010 (1011 valores únicos).
householdSize	Número de personas en el hogar, variable cuantitativa discreta. con 3 valores posibles: 1, 2 o 3 personas (sin decimales). Sin valores nulos. Media = 1.96 (≈ 2), moda = 3 (frecuente en datos previos), desviación = 0.79.
haveKids	Indica si tiene hijos (True/False), Variable cualitativa nominal (booleana). Dos valores: True (con hijos) o False (sin hijos).
age	Edad del participante, Variable cuantitativa discreta. Valores de 18 a 60 años (43 valores únicos).
educationLevel	Nivel educativo del participante, variable cualitativa ordinal. Con 4 categorías: Low, HighSchoolOrCollege, Bachelors, Graduate.
interestGroup	Grupo de interés (probablemente categorías), variable cualitativa nominal. Podrían ser grupos de interés o categorías arbitrarias.
joviality	Medida de alegría/positividad/estado de ánimo (valor numérico), variable cuantitativa continua. Rango [0.000204, 0.999234] (1011 valores únicos).

○ Archivo Apartments.csv

Archivo: /content/drive/MyDrive/5TO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Attributes/Apartments.csv

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	1259.10	4	3	POINT (2103.6301776944765 4266.932930123476)	397
5	411.50	1	4	POINT (7.0589743819342985 79.96163671849988)	628

- Las variables o atributos están describiendo al objeto: Departamento.
- Un registro representa en sí a un departamento individual, con datos como su costo de renta, capacidad, coordenadas de ubicación.

Tabla - Apartments.csv:

Variable	Descripción
apartmentId	Identificador único del apartamento, variable cuantitativa discreta (identificador único). Rango: 1 a 1733.
rentalCost	Costo de renta del apartamento, Variable cuantitativa continua. Rango: \$348.40 a \$1601.11.
maxOccupancy	Capacidad máxima de ocupantes, variable cuantitativa discreta. , 4 valores únicos: 1, 2, 3, 4 personas.
numberOfRooms	Número de habitaciones, variable cuantitativa discreta. , 4 valores únicos: 1, 2, 3, 4 habitaciones
location	Ubicación geográfica (coordenadas), Cuantitativa continua. Posible uso: geolocalización o agrupamiento por proximidad.
buildingId	Identificador del edificio que contiene el apartamento, variable cuantitativa discreta. , 468 valores únicos (rango: 2 a 1040).

○ **Archivo Buildings.csv**

Archivo: /content/drive/MyDrive/5TO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Attributes/Buildings.csv

buildingId	location	buildingType	maxOccupancy	units
1	POLYGON ((350.0638997002585 4595.665606173783,...	Commercial	NaN	NaN
2	POLYGON ((-1926.972613718425 2725.610686806701...	Residential	12.0	[481,498,534,652,818]
3	POLYGON ((685.6846002015491 1552.131491805318,...	Commercial	NaN	[382]
4	POLYGON ((-976.7845160060303 4542.38209636188,...	Commercial	NaN	NaN
5	POLYGON ((1259.3061988755617 3572.726728111263...	Residential	2.0	[231]

- Las variables o atributos están describiendo al objeto: Edificio.
- Un registro representa en sí a un edificio individual, su polígono, que es una manera de definir su ubicación en el mapa, que tipo de edificio es y cuantas unidades hay (al parecer se refiere a la cantidad de apartamentos o negocios que entiende).

Tabla - Buildings.csv:

Variable	Descripción
buildingId	Identificador único del edificio, variable cuantitativa discreta (identificador único). , 1042 valores únicos (1 a 1042).
location	Polígono que define la ubicación y forma del edificio, Cuantitativa continua. Cada edificio tiene una ubicación geográfica única definida por un polígono.
buildingType	Tipo de edificio (Commercial, Residential), variable cualitativa nominal. 3 categorías: Residential, Commercial, School.
maxOccupancy	Capacidad máxima de ocupación, variable cuantitativa continua. , 52

	valores únicos. Rango con datos: 1.0 a 418.0 personas.
units	Lista de unidades en el edificio (apartamentos, negocios), variable cualitativa (lista/array). 673 valores únicos.

○ **Archivo: Restaurant.csv**

Archivo: /content/drive/MyDrive/5TO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Attributes/Restaurants.csv

restaurantId	foodCost	maxOccupancy	location	buildingId
0	445	5.15	71 POINT (631.5130723031391 2001.4772026036535)	304
1	446	4.17	82 POINT (413.840000705876 1194.128694228948)	308
2	447	5.87	119 POINT (497.9967937001494 1624.515148185587)	58
3	448	4.07	98 POINT (698.2411158717262 4392.416668183332)	964
4	449	5.11	53 POINT (1407.7107695149243 4010.4574815269225)	181

- Las variables o atributos están describiendo al objeto Restaurante.
- Un registro representa en sí a un restaurante individual, con el costo de cada comida, la capacidad máxima de personas, la ubicación y el ID del edificio.

Tabla - Restaurant.csv:

Variable	Descripción
restaurantId	Identificador único del restaurante, variable cuantitativa discreta. Rango: 445 a 1805 (20 únicos).
foodCost	Costo promedio de comida, variable cuantitativa continua. Rango: \$4.07 a \$5.92 (20 valores únicos).
maxOccupancy	Capacidad máxima de ocupantes, variable cuantitativa discreta. Rango: 48 a 119 personas (17 valores únicos).
location	Ubicación geográfica (coordenadas), variable cualitativa. 20 valores únicos. Uso para geolocalización.
buildingId	Identificador del edificio que contiene el restaurante, variable cuantitativa discreta. Rango: 27 a 991 (20 únicos).

○ **Archivo: Pubs.csv**

pubId	hourlyCost	maxOccupancy	location	buildingId
0	442	8.281103	64 POINT (964.4380231713202 3991.603473784208)	556
1	443	6.417435	64 POINT (1809.880173357865 4339.172426035451)	29
2	444	12.581806	84 POINT (770.4279044387976 932.5852003214752)	1012
3	892	11.642905	96 POINT (-1524.9573211662105 3815.271490114369)	502
4	893	14.840473	79 POINT (-1608.766411449925 3886.4924784954583)	164

- Las variables o atributos están describiendo al objeto Pub.
- Un registro representa en sí a un Pub(como bar) individual, con el costo de por hora, la capacidad máxima de personas, la ubicación y el ID del edificio.

Tabla - Pubs.csv:

Variable	Descripción
----------	-------------

publd	Identificador único del pub, variable cuantitativa discreta. Rango: 442 a 1800 (12 valores únicos).
hourlyCost	Costo por hora en el pub, variable cuantitativa continua. Rango: \$6.42 a \$14.84 (12 valores únicos).
maxOccupancy	Capacidad máxima de ocupantes, variable cuantitativa discreta. Rango: 60 a 96 personas (10 valores únicos).
location	Ubicación geográfica (coordenadas), variable cualitativa. 12 valores únicos (posible uso para geolocalización).
buildingId	Identificador del edificio que contiene el pub, variable cuantitativa discreta. Rango: 29 a 1012 (12 únicos).

○ **Archivo: Schools.csv**

	schoolId	monthlyCost	maxEnrollment	location	buildingId
0	0	12.812445	242	POINT (-376.7505037068263 1607.9843212558562)	662
1	450	91.143514	418	POINT (-2597.447677094323 3194.1547530883445)	943
2	900	38.005380	394	POINT (-2539.1584040534744 6556.0323181733565)	262
3	1350	73.197852	384	POINT (-4701.462928834322 5141.762936081409)	123

- Las variables o atributos están describiendo al objeto Pub.
- Un registro representa en sí a un colegio individual, con el costo de por mes, la capacidad máxima de estudiantes, la ubicación y el ID del edificio.

Tabla - Schools.csv:

Variable	Descripción
schoolId	Identificador único de la escuela (cuantitativo discreto). Rango: 0 a 1350 (4 únicos). Nota: Hay una escuela con ID=0.
monthlyCost	Costo mensual por estudiante (cuantitativo continuo). Rango: \$12.81 a \$91.14 (4 únicos). ¡Variación extrema!
maxEnrollment	Capacidad máxima de estudiantes (cuantitativo discreto). Rango: 242 a 418 alumnos (4 únicos).
location	Ubicación geográfica (coordenadas). 4 valores únicos, incluyendo puntos con coordenadas negativas.
buildingId	Identificador del edificio (cuantitativo discreto). Rango: 123 a 943 (4 únicos).

○ **Archivo: Jobs.csv**

Archivo: /content/drive/MyDrive/5TO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Attributes/Jobs.csv

jobId	employerId	hourlyRate	startTime	endTime	daysToWork	educationRequirement
0	379	10.000000	7:46:00 AM	3:46:00 PM	[Monday,Tuesday,Wednesday,Thursday,Friday]	HighSchoolOrCollege
1	379	22.217633	7:31:00 AM	3:31:00 PM	[Monday,Tuesday,Wednesday,Thursday,Friday]	Bachelors
2	380	10.000000	8:00:00 AM	4:00:00 PM	[Monday,Tuesday,Wednesday,Thursday,Friday]	Bachelors
3	380	15.312071	7:39:00 AM	3:39:00 PM	[Monday,Tuesday,Wednesday,Thursday,Friday]	Bachelors
4	381	21.355409	7:53:00 AM	3:53:00 PM	[Monday,Tuesday,Wednesday,Thursday,Friday]	HighSchoolOrCollege

- Las variables o atributos están describiendo al objeto: Puesto de Trabajo.
- Un registro representa en sí a un puesto de trabajo individual, con un id de quien es empleador, salario por hora, hora de inicio y fin de la jornada, días de trabajo y requisitos de educación para el puesto.

Tabla - Jobs.csv:

Variable	Descripción
jobId	Identificador único del trabajo, variable cuantitativa discreta (identificador único). , 1,328 valores únicos (0 a 1,327)
employerId	Identificador del empleador, variable cuantitativa discreta. , 253 valores únicos (379 a 1,797).
hourlyRate	Salario por hora, variable cuantitativa continua. Rango: \$10.00 a \$100.00.
startTime	Hora de inicio de la jornada, variable cuantitativa continua de tipo temporal (hora de inicio).
endTime	Hora de fin de la jornada, variable cuantitativa continua de tipo temporal (hora de fin).
daysToWork	Días de trabajo (lista de días de la semana), variable cualitativa ordinal (lista orden de días).
educationRequirement	Requisito educativo para el trabajo, variable cualitativa ordinal. 4 niveles: HighSchoolOrCollege, Bachelors, Graduate, Low. Jerarquía educativa clara.

○ **Archivo: Employers.csv**

employerId	location	buildingId
0	379 POINT (-1849.997168394888 1744.6010147106394)	823
1	380 POINT (41.51783767879146 418.7264799744545)	154
2	381 POINT (877.2786575380362 1358.5441805909259)	279
3	382 POINT (670.3987400004884 1584.4743462106067)	3
4	383 POINT (829.9556783260775 2163.4803049897623)	146

- Las variables o atributos están describiendo al objeto: Empleador.
- Un registro representa en sí a un empleador individual.

Tabla - Employers.csv:

Variable	Descripción
employerId	Identificador único del empleador (cuantitativo discreto). Rango: 379 a 1797 (253 únicos).
location	Ubicación geográfica (coordenadas). 250 valores únicos (3 duplicados). Uso clave para mapeo de clusters laborales.
buildingId	Identificador del edificio (cuantitativo discreto). Rango: 3 a 1041 (185 únicos). Algunos edificios albergan múltiples empleadores

○ **Archivo: CheckinJournal.csv**

Archivo: /content/drive/MyDrive/5TO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Journals/CheckinJournal.csv

participantId	timestamp	venueId	venueType
619	2022-03-01T05:35:00Z	1798	Pub
15	2022-03-01T05:50:00Z	1798	Pub
23	2022-03-01T05:55:00Z	894	Pub
699	2022-03-01T06:00:00Z	1798	Pub
876	2022-03-01T06:00:00Z	1804	Restaurant

- Las variables o atributos no están describiendo un objeto.
- Un registro representa en sí a un check-in único (como registro) de un participante en un local específico (registro de ubicación).

Tabla - CheckinJournal.csv:

Variable	Descripción
participantId	Identificador del participante, variable cuantitativa discreta (ID del participante). Rango: 0–1010
timestamp	Fecha y hora del check-in, variable cuantitativa continua de tipo temporal (fecha y hora del registro).
venueId	Identificador del lugar visitado, variable cuantitativa discreta (ID del lugar visitado). Rango: 0–1,805 (1,114 IDs únicos).
venueType	Tipo de lugar (Pub, Restaurant, etc.), variable cualitativa nominal (tipo de lugar). Tipo: object. 4 categorías: Apartment, Pub, Restaurant, Workplace.

○ **Archivo: TravelJournal.csv**

Archivo: /content/drive/MyDrive/5TO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Journals/TravelJournal.csv

participantId	travelStartTime	travelStartLocationId	travelEndTime	travelEndLocationId	purpose	checkInTime	checkOutTime	startingBalance	endingBalance
23	2022-03-01T05:20:00Z	532.0	2022-03-01T05:55:00Z	894	Recreation (Social Gathering)	2022-03-01T05:55:00Z	2022-03-01T06:00:00Z	851.223425	850.197491
876	2022-03-01T05:50:00Z	NaN	2022-03-01T06:00:00Z	1804	Eating	2022-03-01T06:00:00Z	2022-03-01T06:05:00Z	2071.779647	2065.864612
902	2022-03-01T06:05:00Z	NaN	2022-03-01T06:10:00Z	1801	Eating	2022-03-01T06:10:00Z	2022-03-01T06:15:00Z	2115.790341	2110.280842
919	2022-03-01T06:00:00Z	NaN	2022-03-01T06:10:00Z	1802	Eating	2022-03-01T06:10:00Z	2022-03-01T06:15:00Z	2120.433773	2115.351009
154	2022-03-01T05:55:00Z	NaN	2022-03-01T06:10:00Z	446	Eating	2022-03-01T06:10:00Z	2022-03-01T06:15:00Z	2246.247936	2242.074591

- Las variables o atributos no están describiendo un objeto.

- Cada fila o registro es un viaje completo de un participante

Tabla - TravelJournal.csv:

Variable	Descripción
participantId	Identificador del participante, variable cuantitativa discreta (identificador de participante). , 1011 valores únicos (0 a 1010).
travelStartTime	Hora de inicio del viaje, variable cuantitativa continua de tipo temporal (fecha/hora).
travelStartLocationId	Identificador de ubicación de inicio, variable cuantitativa discreta. 1,114 valores únicos (1 a 1,805).
travelEndTime	Hora de fin del viaje, variable cuantitativa continua de tipo temporal (fecha/hora).
travelEndLocationId	Identificador de ubicación de destino, variable cuantitativa discreta. 1,114 valores únicos (0 a 1,805).
purpose	Propósito del viaje, variable cualitativa nominal. 5 categorías: Work/Home Commute, Eating, Recreation (Social Gathering), Shopping, Other.
checkInTime	Hora de check-in en el destino, variable cuantitativa continua de tipo temporal. Coincide con travelEndTime
checkOutTime	Hora de check-out del destino, variable cuantitativa continua de tipo temporal.
startingBalance	Saldo inicial antes del viaje, variable cuantitativa continua. , 1,664,729 valores únicos. Rango: [-681.65, 240,494.70].
endingBalance	Saldo final después del viaje, variable cuantitativa continua. , 1,664,581 valores únicos. Rango: [-640.71, 240,838.80].

○ **Archivo: FinancialJournal.csv**

Archivo: /content/drive/MyDrive/5TO_AÑO_1ER_SEMESTRE/Tópicos en Ciencia de Datos/Proyecto - Base de Datos/Journals/FinancialJournal.csv

participantId	timestamp	category	amount
0	2022-03-01T00:00:00Z	Wage	2472.507559
0	2022-03-01T00:00:00Z	Shelter	-554.988622
0	2022-03-01T00:00:00Z	Education	-38.005380
1	2022-03-01T00:00:00Z	Wage	2046.562206
1	2022-03-01T00:00:00Z	Shelter	-554.988622

- Las variables o atributos no están describiendo al objeto.
- Un registro es en sí una transacción única(de ingreso o de gasto) de un participante y la cantidad de dinero que mueve.

Tabla - FinancialJournal.csv:

Variable	Descripción
participantId	Identificador del participante, variable cuantitativa discreta (identificador de participante). Rango: 0 a 1010, un participante

	puede hacer múltiples transacciones.
timestamp	Fecha y hora de la transacción, variable cuantitativa continua de tipo temporal (fecha y hora de transacción).
category	Tipo de transacción, variable cualitativa nominal (tipo de transacción). 6 categorías: Wage (ingresos), Shelter, Education, RentAdjustment, Food, Recreation.
amount	Cantidad de dinero (positivo para ingresos, negativo para gastos), variable cuantitativa continua (monto de transacción). Rango: -1,562.726 (egresos) a 4,096.526 (ingresos).

○ **Archivo: SocialNetwork.csv**

	timestamp	participantIdFrom	participantIdTo
0	2022-03-01T00:00:00Z	173	180
1	2022-03-01T00:00:00Z	178	183
2	2022-03-01T00:00:00Z	178	185
3	2022-03-01T00:00:00Z	180	173
4	2022-03-01T00:00:00Z	183	178

- Las variables o atributos no están describiendo un objeto.
- Un registro representa en sí a una interacción social puntual en un momento exacto, con un participante que envía y otro que recibe.

Tabla - SocialNetwork.csv:

Variable	Descripción
timestamp	Fecha y hora de la interacción (formato ISO UTC). 451 valores únicos, desde 2022-03-01 hasta otro período no especificado. El día con más actividad: 2022-07-25 (21,426 interacciones).
participantIdFrom	ID del participante que inicia la interacción (cuantitativo discreto). Rango: 0 a 1,010 (963 únicos). ¿El ID=0 es un usuario sistema o error?
participantIdTo	ID del participante que recibe la interacción (cuantitativo discreto). Mismo rango y valores únicos que participantIdFrom.

- Limpieza de datos:

Viendo el número de filas total de cada archivo:

```

Conteo de filas en Activity Logs:
ParticipantStatusLogs1.csv: 1625339 filas
ParticipantStatusLogs2.csv: 1489143 filas
ParticipantStatusLogs3.csv: 1571687 filas
ParticipantStatusLogs4.csv: 1591403 filas
ParticipantStatusLogs5.csv: 1598101 filas
ParticipantStatusLogs6.csv: 1603924 filas
ParticipantStatusLogs7.csv: 1601317 filas
ParticipantStatusLogs8.csv: 1598477 filas
ParticipantStatusLogs9.csv: 1598751 filas
ParticipantStatusLogs10.csv: 1599356 filas

Total de observaciones en ParticipantStatusLogs1-10: 15877498

Conteo de filas en Attributes:
Participants.csv: 1011 filas
Restaurants.csv: 20 filas
Schools.csv: 4 filas
Employers.csv: 253 filas
Buildings.csv: 1042 filas
Apartments.csv: 1517 filas
Pubs.csv: 12 filas
Jobs.csv: 1328 filas

Conteo de filas en Journals:
CheckinJournal.csv: 2100635 filas
FinancialJournal.csv: 1856330 filas
SocialNetwork.csv: 7482488 filas
TravelJournal.csv: 2099656 filas

```

Verificando duplicados de participantId en Activity Logs, ya que si tengo uno repetido sería algo muy inesperado, pero felizmente no hay repetidos:

```

✓ ParticipantStatusLogs1.csv: sin duplicados
✓ ParticipantStatusLogs2.csv: sin duplicados
✓ ParticipantStatusLogs3.csv: sin duplicados
✓ ParticipantStatusLogs4.csv: sin duplicados
✓ ParticipantStatusLogs5.csv: sin duplicados
✓ ParticipantStatusLogs6.csv: sin duplicados
✓ ParticipantStatusLogs7.csv: sin duplicados
✓ ParticipantStatusLogs8.csv: sin duplicados
✓ ParticipantStatusLogs9.csv: sin duplicados
✓ ParticipantStatusLogs10.csv: sin duplicados

```

Verificando ahora la unicidad de participantId en Attributes/Participants.csv:

```

Verificación de unicidad en Participants.csv:
✓ No hay participantes repetidos

```

4. Análisis Exploratorio de Datos (EDA) y Visualización

- **HIPÓTESIS 1: ¿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?**

Los pasos a seguir para resolver la hipótesis y también evaluar la complejidad de este, se propone los siguientes pasos **(en toda esta metodología, se dará a conocer las diferentes modificaciones o limpieza de los datos que se necesitaron para conseguir resolver la pregunta):**

1. Hacer un filtro del número de participantes que tienen niños, considero que cuando un participante tiene por lo menos un niño, sería una familia.
2. Si nos damos cuenta en los registros de los "ParticipantStatusLogs*", hay varios participantes que cambian de apartmentId, no sabemos si solo están ahí por mientras, así que tendríamos que contar la cantidad de veces que están en un apartamento y la mayoría de veces que se encontraba en un apartamento, ese será su apartamento.
3. Teniendo ahora a los participantes con sus respectivos apartamentosId, pero ahora necesitamos saber, quienes son familia. Los mismos participantes que vivan en el mismo apartamento los consideraré familia.

4. En una gráfica como de un mapa, ubicar con colores resaltantes o figuras los apartamentos de dichas familias, según la ubicación de sus apartamentos.
5. Mostrar restaurantes, pubs y otros en el mismo mapa y observar las ubicaciones en las que estas se encuentran.
6. Sacar nuestras conclusiones.

Problemas encontrados en el proceso:

- Si nos damos cuenta en los registros de los "ParticipantStatusLogs*", hay varios participantes que cambian de apartmentId, no sabemos si solo están ahí por mientras, así que tendríamos que contar la cantidad de veces que están en un apartamento y la mayoría de veces que se encontraba en un apartamento, ese será su apartamento.
- Quiero ver los tipos de datos de participantLogs*, al observar, hay algunos archivos que en apartmentId los tiene como tipo de datos float, pero en verdad debería ser int. Debemos asegurarnos de cambiar ello, para antes de juntar todo en uno solo.
 - 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

```

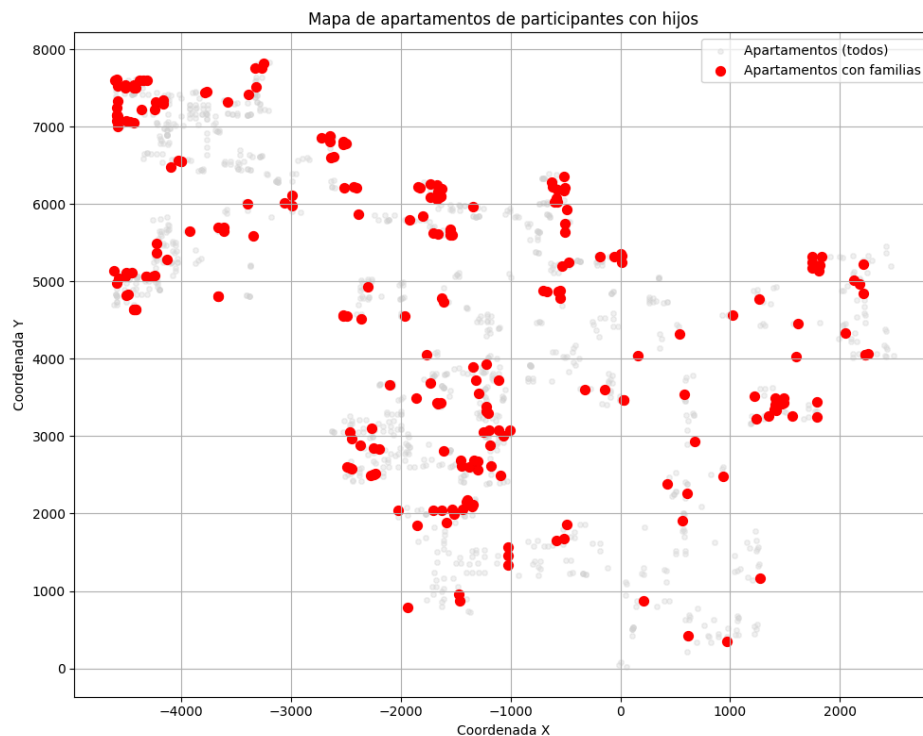
- Necesitamos saber, quienes son familia. Los mismos participantes que vivan en el mismo apartamento los consideraré 1 familia.

```

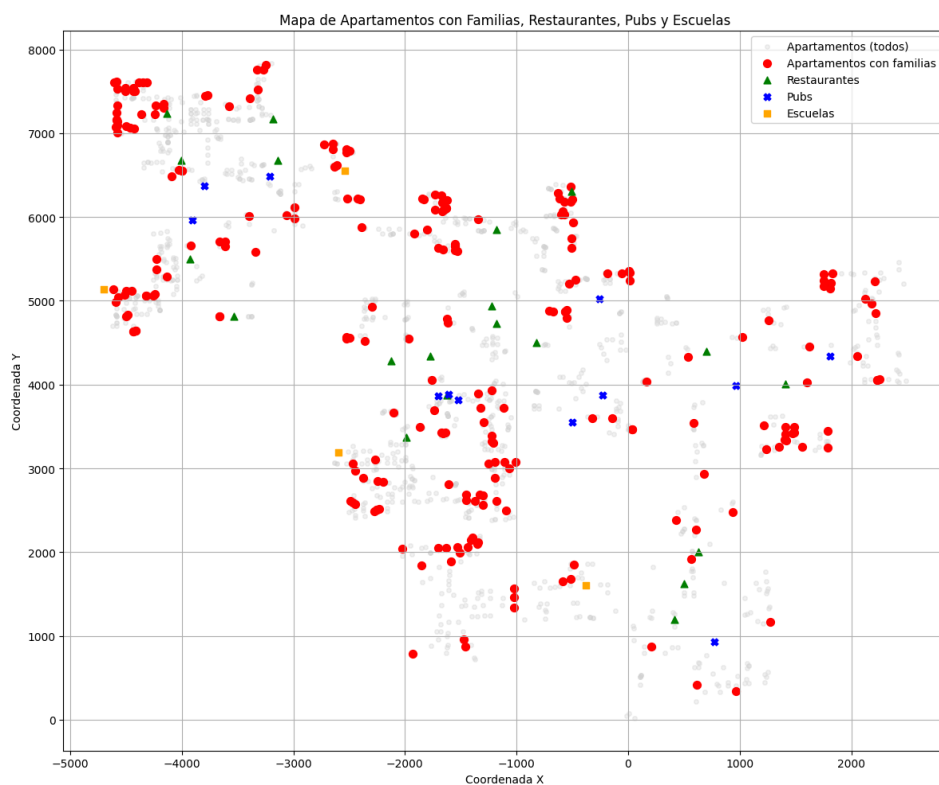
participantId  apartmentId_principal  familiaId
0              0                    926      926
1              1                    928      928
2              2                    291      291
3              3                    1243     1243
4              4                    136      136
Número total de familias (apartamentos únicos con al menos un participante con hijos): 247

```

- "location" es una variable cualitativa nominal codificada en una estructura tipo texto (cadena de caracteres), con contenido que representa coordenadas. (ASI LO PUSIERON)
- 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.
- Primera gráfica:



- Ahora vamos a mostrar restaurantes, pubs y schools en el mismo mapa, así podremos observar las ubicaciones en las que estas se encuentran.



Respondiendo a la pregunta: ¿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?

- La mayor cantidad de familias con niños viven en los edificios que se ubican en la parte superior izquierda y casi en la parte central de la ciudad. Las

características del por qué estos se agrupan más en estos lugares es porque se tiene mayor presencia tanto de escuelas (principalmente porque tienen niños), restaurantes y pubs.

- **HIPÓTESIS 2: ¿El nivel económico está relacionado con el grado de estudios y la ubicación geográfica?**

Los pasos a seguir para resolver la hipótesis y también evaluar la complejidad de este, se propone los siguientes pasos **(en toda esta metodología, se dará a conocer las diferentes modificaciones o limpieza de los datos que se necesitaron para conseguir resolver la pregunta):**

1. Filtrar a los participantes por su grado de estudio, osea separarlos por grupos.
2. Ya los grupos separados anteriormente, ahora extraemos su nivel económico o sus ganancias.
3. Plasmamos esto en una gráfica (plots, de acuerdo al tipo de variable que se tenga). Y observamos si hay una relación entre estos 2 factores (grado de estudios y nivel económico).
4. Ahora escogemos al grupo que tienen ganancias altas y bajas, y los plasmamos en el mapa, para ver si la ubicación geográfica también está relacionado con estos 2 factores.

Problemas encontrados en el proceso:

- Veo niveles de educación están presentes, de paso veo si hay valores faltantes o inconsistencias.
- Dado que educationLevel es nominal y no habrá operaciones numéricas sobre ella, convertirla a category facilita la memoria y deja claro al pipeline que es variable categórica y clasificar esta columna como category permite que, en etapas posteriores (visualización), se trate como cualitativa.

```
0
participantId int64
householdSize int64
haveKids bool
age int64
educationLevel category
interestGroup object
joviality float64
dtype: object
```

```
Nivel 'Bachelors': 232 participantes
Nivel 'Graduate': 170 participantes
Nivel 'HighSchoolOrCollege': 525 participantes
Nivel 'Low': 84 participantes
```

- Ya los grupos separados anteriormente, ahora extraemos su nivel económico o sus ganancias.

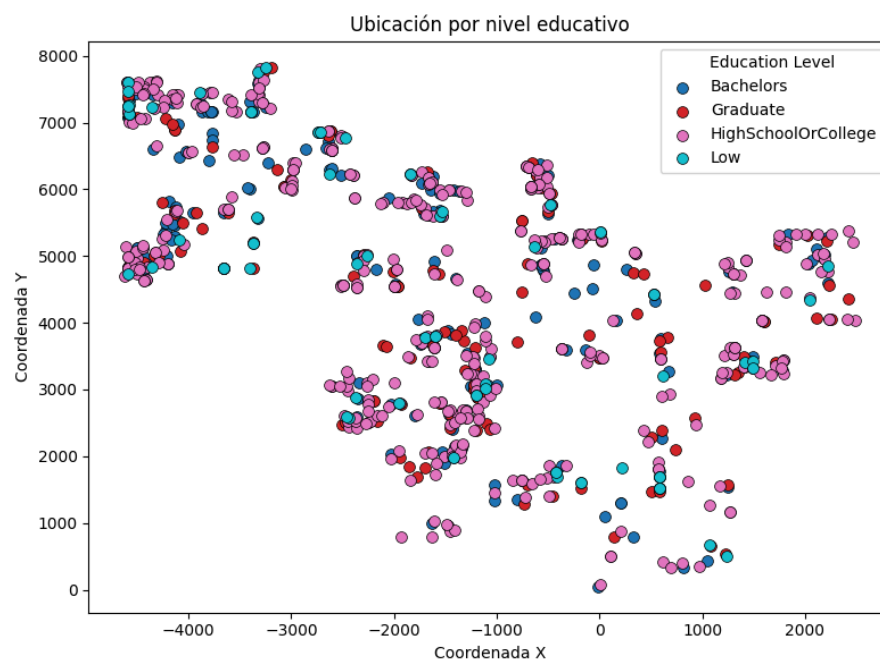
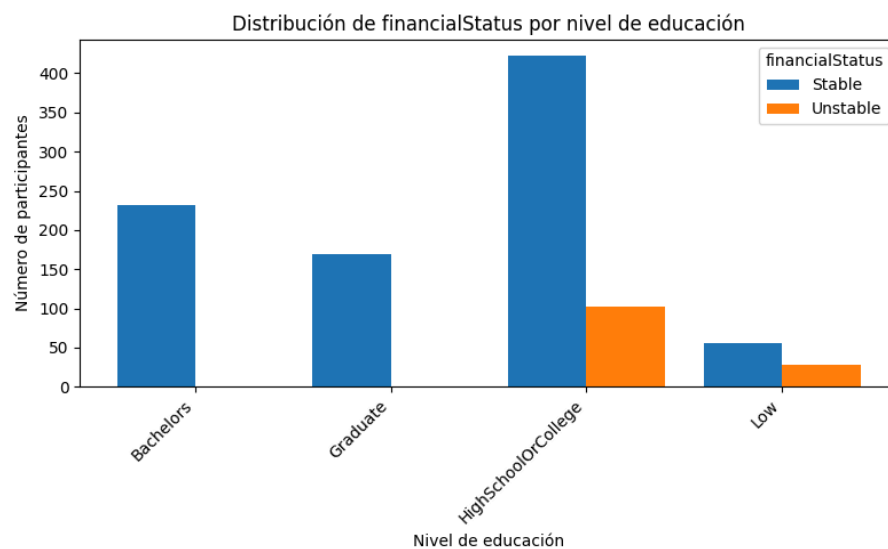
```
Categorías encontradas en 'financialStatus': ['Stable' 'Unstable' 'Unknown' nan]
Cantidad de valores faltantes en 'financialStatus': 1
```

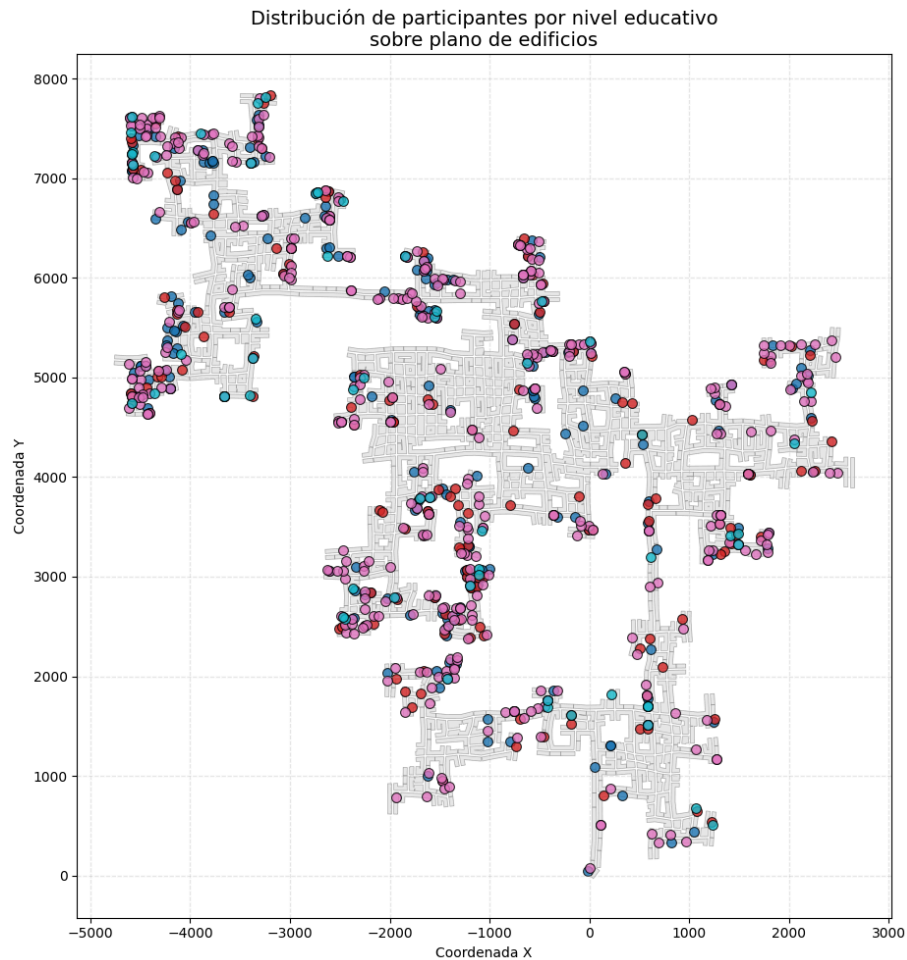
- Como se tiene un valor con nan, es lógico que pongamos como Unknown, creamos una nueva columna que impute ese NaN a 'Unknown' y la convertimos en categoría de paso.

Forma del DataFrame final: (1011, 3)

	participantId	educationLevel	financialStatus
0	0	HighSchoolOrCollege	Stable
1	1	HighSchoolOrCollege	Stable
2	2	HighSchoolOrCollege	Stable
3	3	HighSchoolOrCollege	Stable
4	4	Bachelors	Stable

- Graficamos:





Respondiendo a la pregunta: ¿El nivel económico está relacionado con el grado de estudios y la ubicación geográfica?

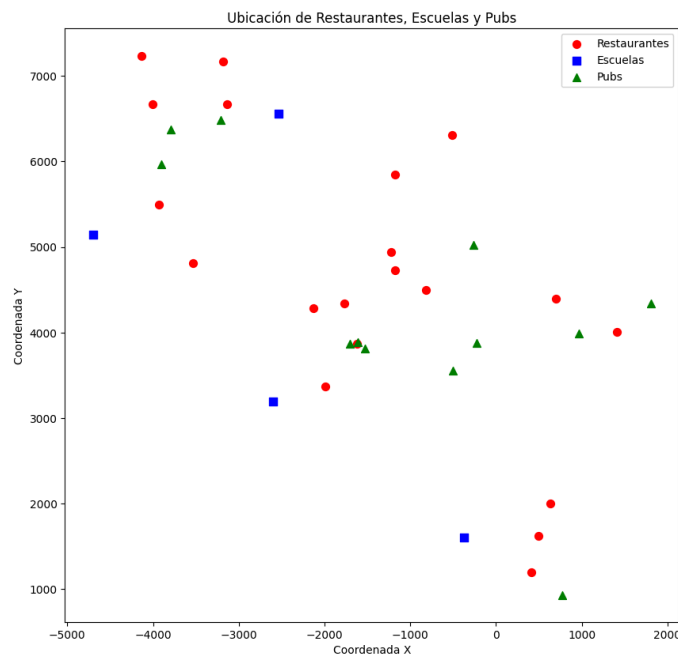
- En el gráfico de barras de la distribución de financialStatus por nivel de educación, ahí tenemos que el estado financiero de "Estable" es liderado por HighSchoolOrCollage con más de 4000 participantes, pero también lidera en "Inestable". Dando a conocer que se podría dar cambios entre estos 2 estados para dichos participantes. Quienes no presentan el estado de "Inestable", son los que tienen grados de Bachiller o son Graduados.
- Mientras que si observamos su ubicación en el mapa, podemos apreciar que los de nivel HighSchoolOrCollage, se encuentran esparcidos por casi toda la ciudad, mientras que se tiene más presencia por el centro de la ciudad a los participantes con nivel de estudio Bachiller y Graduado.
- Si recordamos el estado financiero de los participantes, los de nivel de estudio HighSchoolOrCollage, son quienes tienen un nivel de "Estable", lo cual puede que mueva diferentes negocios en las zonas donde mayormente habitan, junto con los demás participantes que tienen nivel financiero estable. Como en el centro de la ciudad y en la esquina superior izquierda del mapa.
- **HIPÓTESIS 3:** Imagina que una familia de cuatro personas dispone de poco más de 3 000 USD al mes. ¿En qué edificio o conjunto de apartamentos les recomendarías vivir y por qué esa opción es la mejor?

Los pasos a seguir para resolver la hipótesis y también evaluar la complejidad de este, se propone los siguientes pasos(en toda esta metodología, se dará a conocer las diferentes modificaciones o limpieza de los datos que se necesitaron para conseguir resolver la pregunta):

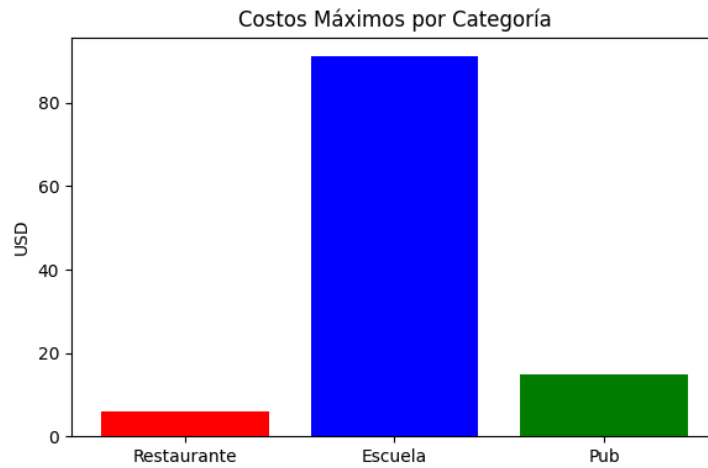
1. Ubicar en el mapa los distintos edificios que se tienen como restaurantes, escuelas, etc.
2. Una vez se tienen las ubicaciones de estos edificios, observamos los precios de estos, tomando en cuenta los precios más altos de cada uno, para poder hacer una suma y no pasar la cantidad indicada en la pregunta.
3. Además de ello, como es una familia, es muy probable que se tenga niños, considerando que tener cerca una escuela, también sería un factor importante.
4. Sacamos nuestras conclusiones.

Problemas encontrados en el proceso:

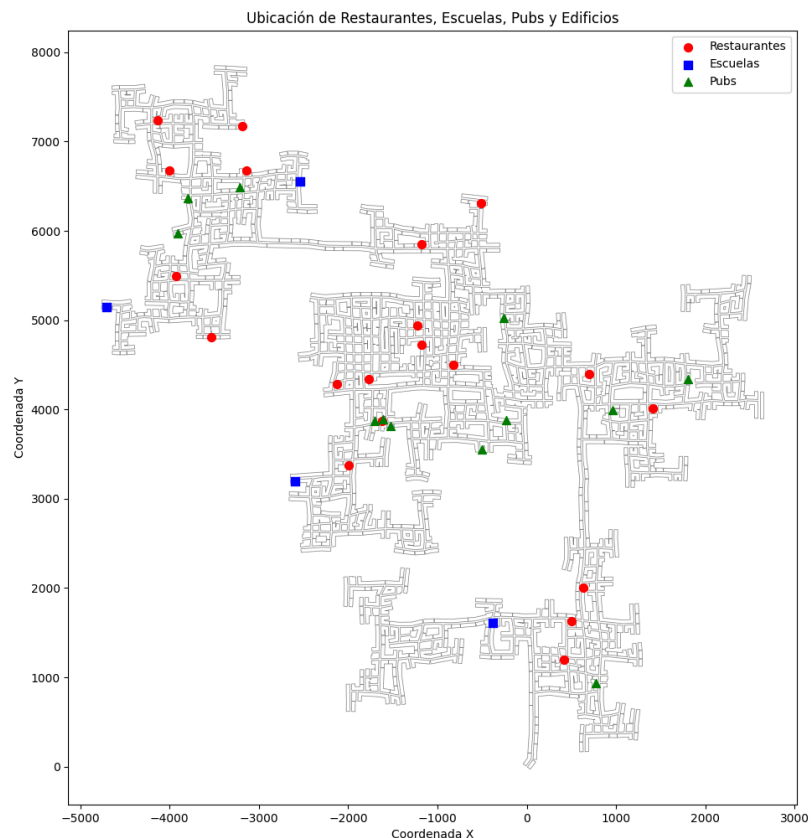
- Convertir la columna 'location' de WKT a geometría y a partir de ello se grafica como mapa:



- Se obtuvo:
 - Costo máximo restaurante (foodCost): 5.92 USD
 - Costo máximo escuela (monthlyCost): 91.14351385 USD
 - Costo máximo pub (hourlyCost): 14.84047265 USD
 - Suma de los costos máximos: 111.9039865 USD



- Cálculo mensual:
 - Costo máximo por comida (foodCost): 5.92 USD
 - Costo mensual máximo por estudiante (monthlyCost): 91.14 USD
 - Costo máximo por hora en pub (hourlyCost): 14.84 USD
 - > Supongo el siguiente caso
 - Cada persona come 1 vez al día → 4×30 comidas al mes.
 - 2 niños asisten a la escuela y pagan mensualidad.
 - 2 adultos van al pub 8 veces al mes, 1 hora cada vez.
 - > Gasto Mensual Estimado
 - Restaurante: 710.40 USD
 - Escuela: 182.29 USD
 - Pub: 237.45 USD
 - Total: 1130.13 USD



Respondiendo a la pregunta: Imagina que una familia de cuatro personas dispone de poco más de 3 000 USD al mes. ¿En qué edificio o conjunto de apartamentos les recomendarías vivir y por qué esa opción es la mejor?

- Debido a que la familia tienen un poco más de 3000 dólares, si vemos los cálculos realizados en el caso anterior:
 - Cada persona come 1 vez al día, osea 4×30 comidas al mes.
 - 2 niños asisten a la escuela y pagan mensualidad.
 - 2 adultos van al pub 8 veces al mes, 1 hora cada vez.Con pago máximo.

- Gasto Mensual Estimado:
 - Restaurante: 710.40 USD
 - Escuela: 182.29 USD
 - Pub: 237.45 USD
 - Total: 1130.13 USD

Podemos concluir, que la familia no gastaría más de su presupuesto al mes, incluso si pagaran los precios más elevados. Además, considerando que es posible que la familia tenga niños, es preferible que vivan cerca de una escuela. Si observamos el gráfico, se tienen 4 escuelas en total, pero deberían estar cerca a una escuela. También es primordial, en segundo lugar, estar cerca de restaurantes. Donde si nos dirigimos hacia la parte superior izquierda del mapa, tenemos los edificios más adecuados para que ellos puedan vivir sin ningún problema, teniendo al alcance 4 restaurantes, 3 bares y 2 escuelas.