INFORME FINAL DE ANÁLISIS EXPLORATORIO DE DATOS DEL CONJUNTO DE DATOS VAST CHALLENGE 2022

1. **Hipótesis iniciales**:

1.1. Motivación:

El conjunto de datos VAST Challenge 2022 consiste en datos urbanos ficticios que emulan tareas y problemas de análisis visual del mundo real [He et al.]. Estos retos, incluyendo ediciones anteriores, también han sido utilizados en proyectos de analítica inmersiva colaborativa como de [Tong et al.], donde los participantes buscan resolver dichos retos.

En el trabajo de [Lee et al.] el análisis de datos urbanos se basa en al menos tres tareas directas, complementadas con actividades de exploración, buscando fomentar estrategias de búsqueda colaborativa entre sus usuarios. Así pues, mis hipótesis serán evaluadas para ver su complejidad al resolverlas y también para ver si estas direccionarían a los usuarios a trabajar de manera colaborativa.

1.2. Mis hipótesis:

- 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?
- Hipótesis 2:¿El nivel económico está relacionado con el grado de estudios y la ubicación geográfica?
- Hipótesis 3: Imagina que una familia de cuatro personas dispone de poco más de 3000 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?

1.3. Plan de análisis:

Describe qué pasos siguió para investigar las hipótesis.

Para investigar las hipótesis seguí los ciclos de vida de la ciencia de datos que fue explicada en el curso:

- Evaluación y análisis de contexto.
- Adquisición de datos, revisión y limpieza de estos.
- Conocer el dataset(aquí de paso ya iba resaltando las variables que me ayudarían con mis hipótesis).
- Arreglo e imputación de datos, transformaciones
- Observación y descubrimiento de patrones(algunos de ellos ya me daban la respuesta a las hipótesis).
- Llegando a la conclusión.

Cabe resaltar que constantemente regresaba entre los diferentes pasos que mencioné. Cumpliendose así lo indicado en teoría sobre el ciclo de vida de la ciencia de datos.

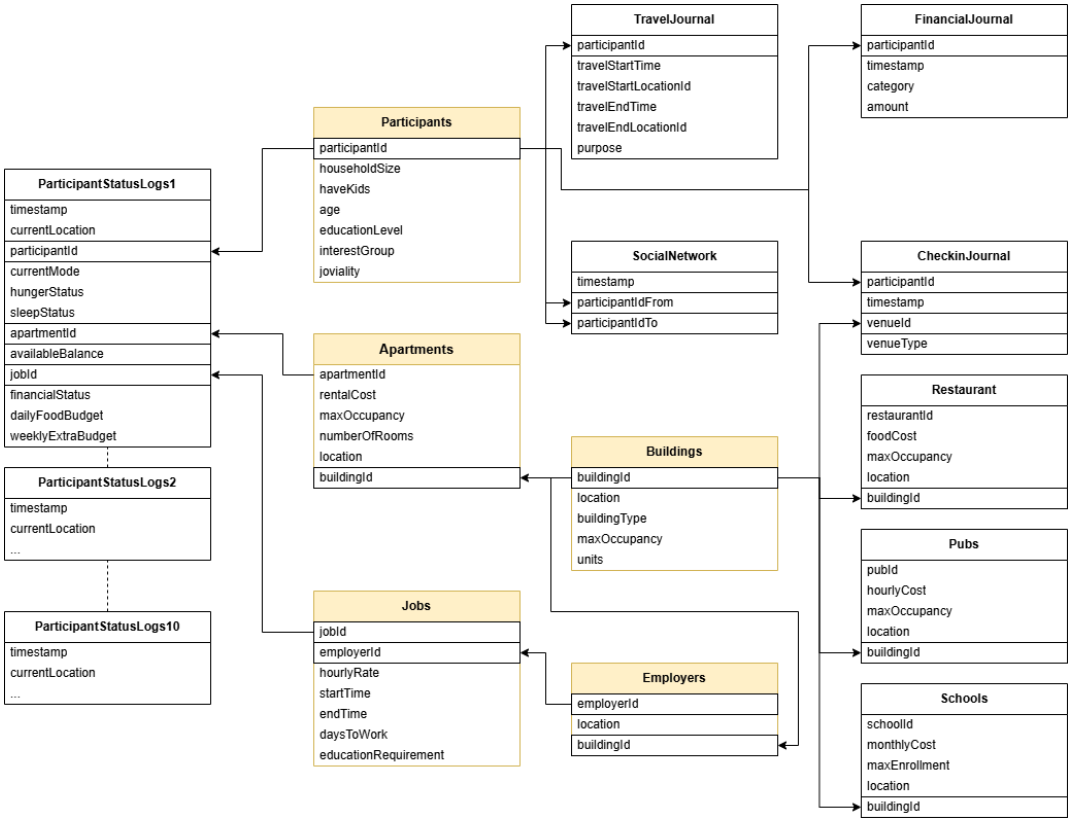
2. **Fuente de datos**:

2.1. **Fuente**:

La base de datos fue proporcionada en el IEEE VIS 2022, evento líder en visualización y análisis visual de datos, que organiza diferentes retos anuales a través de la IEEE VIS Conference. En el caso del VAST Challenge 2022 el tema central fue el Urban Data Analysis donde se emularon tareas y problemas de análisis visual del mundo real.
Todo el dataset, junto a sus variables, es de gran relevancia, ya que permite comprender: la demografía de la ciudad ficticia de Engagement, los patrones de vida cotidiana en la ciudad y la salud financiera de la misma [Het et al, Burmeister et al.].

2.2. **Descripción**:

Describe el conjunto de datos:



En este diagrama da un resumen del dataset que se tiene, qué tablas tienen variables que tienen una relación directa con otras tablas, e incluso este diagrama también nos ayuda a poder saber qué tablas y qué datos usaremos para poder responder las hipótesis planteadas.

A continuación par ver las tablas, atributos, registros, relación entre atributos. Solo se menciona aquí a los que nos ayudaron o fueron usados para resolver las hipótesis planteadas.

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.
- Registros: Un registro representa en sí a un participante individual, contiene sus datos demográficos (edad, número de personas en el hogar, si tiene hijos, nivel educativo, grupo de interés en redes sociales, jovialidad).

Atributos - Participants.csv:

Variable	Descripción
participantId	Identificador único del participante, variable cuantitativa discreta (identificador único) . Rango de 0 a 1010 (1011 valores únicos). Este es importante porque nos ayudará a identificar de manera correcta a los participantes en cada registro hecho en las tablas ParticipantsStatusLogs.
householdSize	Represneta al número de personas que se encuentran en el hogar, es una variable cuantitativa discreta, no puede llegar a ser decimal. Solo tiene 3 valores posibles: 1, 2 o 3 personas. Sin valores nulos. Esta nos puede ayudar con algún filtro de familias y su tamaño.
haveKids	Esta indica si tiene hijos, variable cualitativa nominal por ser booleana. Dos valores: True (con hijos) o False (sin hijos). Este atributo nos ayudará a definir quienes tienen familia con niños.
age	Edad del participante, variable cuantitativa discreta. Valores de 18 a 60 años, se tiene 43 valores únicos. Esta variable nos podría ayudar a filtrar a los participantes por su edad.
educationLevel	Nivel educativo del participante, variable cualitativa ordinal, por orden jerárquico. Con 4 categorías: Low, HighSchoolOrCollege, Bachelors, Graduate. Esta variable si la usamos para ver la relación que podría tener con el estado financiero del participante
interestGroup	Grupo de interés en las redes sociales, similar a grupos de whatsapp, variable cualitativa nominal porque no hay un orden lógico. Esta podría servirnos para poder conocer, por ejemplo, los intereses políticos u otros temas de los ciudadanos.
joviality	Medida de alegría/positividad/estado de ánimo del participante, variable cuantitativa continua, porque son decimales sus datos. El rango es [0.000204, 0.999234] (1011 valores únicos). Este podría

	servirnos para ver la salud mental de la persona o el estado emocional.

- **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 (-1558.517200825967 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, en si solo son registros dados cada cierto tiempo.
  - Registros: 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), estado o modo de actividad(currentMode) del particpante, nivel de hambre/sueño(sleepStatus), en qué apartamento se encuentra en ese momento(apartmentId), su saldo disponible en ese momento(availableBalance), su trabajo(jobId), estado financiero actual(financialStatus), presupuesto para comida diaria(dailyFoodBudget), presupuesto extra de la semana(weeklyExtraBudget), etc.

Atributos - ParticipantStatusLogs*.csv:

Variable	Descripción
timestamp	Fecha y hora en que se dio el registro, variable cuantitativa continua de tipo temporal (fecha/hora). La granularidad es un instante puntual “2022-04-26T00:25:00Z” para cada participante. Esto podría jugarnos a favor o en contra, debido a ello, el dataset es pesado, se tienen muchos registros cada cierto instante.
currentLocation	Ubicación geográfica actual del participante, el formato de POINT(x,y), son coordenadas [ESRI web], es una variable cuantitativa continua, porque tienen valores decimales, no finitos. Se tienen 23,741 valores únicos. Los datos x,y nos serán utiles para posicionarnos como en un mapa.
participantId	Identificador único del participante, variable cuantitativa discreta (identificador único) . Rango de 0 a 1010 (1011 valores únicos). Este es importante porque nos ayudará a identificar de manera correcta a los participantes por cada registro, también se encuentra en la tabla Participants.
currentMode	Estado o modo de actividad del participante (AtHome, AtWork, Transport, Shopping, Other). es una variable cualitativa nominal, no tienen un orden lógico. Este nos puede servir para filtrar

	participantes según su estado de actividad.
hungerStatus	Estado de hambre (JustAte, BecomingHungry, Hungry, Starving, BecameFull). Es una variable cualitativa ordinal porque tienen como un orden natural, si ordenamos por necesidad alimenticia. Esto nos puede servir para poder saber por ejemplo, el momento del día en que más se tiene hambre.
sleepStatus	Estado de sueño. variable cualitativa nominal ya que no se puede ordenar u organizar una secuencia. Tiene 3 estados: Awake, Sleeping, Drowsy. Este podría servirnos para poder saber en qué momentos suelen dormirse las personas, tal vez no solo de noche.
apartmentId	Identificador del apartamento donde vive el participante. variable cuantitativa discreta, tiene 841 valores únicos (1-1,733). Cada participante, en un instante, podría estar en un apartamento difrente. A la misma vez, este atributo pertenece a un edificio (buildingId) que está en el archivo de edificios (Buildings.csv).
availableBalance	Saldo disponible del participante, variable cuantitativa continua, ya que se presenta en decimales, es no finito y está dentro del rango: [-681.65, 5,408.98]. Esto podría servirnos para conocer cómo varía el saldo disponible de los participantes.
jobId	Identificador del trabajo del participante, es una variable cuantitativa discreta porque es un identificador único. Tiene 1,190 valores únicos. Puedo relacionarlo con la tabla de Jobs.csv y así saber el trabajo del participante.
financialStatus	Estado financiero del participante, variable cualitativa ordinal, ya que podemos establecer un orden jerárquico, se tienen 3 niveles: Stable, Unstable, Critical, Unknow. Con esto conoceremos la salud financiera del participante.
dailyFoodBudget	Presupuesto diario para la comida, es una variable cuantitativa discreta porque no tiene decimales y es finito. Con este dato podemos conocer el presupuesto diario fijo asignado en los participantes.
weeklyExtraBudget	Presupuesto semanal para gastos extras, es una variable cuantitativa continua porque contiene decimales. Tiene 18,221 valores únicos y sus valores están en el rango: [0, 2,553.25]. Con esta variable podríamos conocer el presupuesto adicional semanal (¿para gastos no esenciales?) de los participantes.

- **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.

- Registros: Un registro representa en sí a un departamento individual, con datos como su costo de renta, capacidad, número de habitaciones, coordenadas de ubicación y en qué edificio se encuentra.

Atributos - Apartments.csv:

Variable	Descripción
apartmentId	Identificador único del apartamento, variable cuantitativa discreta (identificador único). sus valores en el rango 1 a 1733. Este nos conecta con las tablas ParticipantStatusLogs*, así podemos saber en qué apartamento y edificio se encuentra el participante.
rentalCost	Costo de renta del apartamento, Variable cuantitativa continua porque sus valores se encuentran en decimales, el rango observado es de $348.40 a $1601.11. Esta variable nos podría ser de ayuda para hacer un filtro de apartamentos y su costo, incluso incluir el presupuesto que tenga el participante.
maxOccupancy	Capacidad máxima de ocupantes, variable cuantitativa discreta ya que es en sí un conteo, tiene 4 valores únicos: 1, 2, 3, 4 personas. Esta variable nos ayudaría a conocer por ejemplo el apartamento adecuado para alguna familia.
numberOfRooms	Número de habitaciones que tiene el apartamento, variable cuantitativa discreta porque es un conteo, tiene 4 valores únicos: 1, 2, 3, 4 habitaciones. Esto podría también ayudarnos a conocer por ejemplo el apartamento adecuado para alguna familia.
location	Ubicación geográfica (coordenadas POINT(x,y)[ESRI web]), Cuantitativa continua, ya que tienen valores decimales no finitos. Se podría usar para hacer filtros de distancia entre edificios o entre una ubicación específica.
buildingId	Identificador del edificio que contiene el apartamento, es una variable cuantitativa discreta porque es en sí un único valor, tiene 468 valores únicos. Esta nos unirá con la tabla de Buildings. Así podríamos ubicarnos mejor.

- **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...	Residental	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...	Residental	2.0	[231]

- Las variables o atributos están describiendo al objeto: Edificio.
- Registros: Un registro representa en sí a un edificio individual, su locación(ubicación en el mapa pero con POLYGON[Wonsang et al.]), qué tipo de edificio es, cuál es capacidad 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). Esta variable está presente en varias tablas, pero nos ayuda a conocer la ubicación de negocios y apartamentos.
location	POLYGON[Wonsang et al.] define la ubicación más precisa y forma del edificio, es una variable cuantitativa continua por sus valores decimales no finitos. Cada edificio tiene una ubicación geográfica única definida por un polígono.
buildingType	Tipo de edificio (Commercial, Residental), es una variable cualitativa nominal porque no podemos poner un orden lógico o alguna secuencia. Tiene 3 categorías: Residental, Commercial, School. Esto nos puede ayudar en filtros por tipo.
maxOccupancy	Capacidad máxima de ocupación, variable cuantitativa discreta, porque es un conteo. Pero falta transformación de float a int, porque los valores siempre son exactos, como 418.0 -> 418. El rango de datos va de 1 a 418 personas.
units	Lista de unidades en el edificio (apartamentos, negocios), variable cualitativa (ya que es una lista/array). Tienen 673 valores únicos. Esto podría servirnos para saber cuán ocupados están ciertos edificios por estar en alguna ubicación especial tal vez.

- **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.
- Registros: 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 en el que se encuentra.

Atributos - Restaurant.csv:

Variable	Descripción
restaurantId	Identificador único del restaurante, variable cuantitativa discreta porque, como ya se mencionó es un valor único en sí.
foodCost	Costo promedio de comida, es una variable cuantitativa continua, ya que contiene valores decimales. Sus valores están en el rango: $4.07 a $5.92 (tiene 20 valores únicos). Esto nos podría ayudar al tratar de conocer el pago que daría un grupo de participantes en cierto intervalo de tiempo.

maxOccupancy	Capacidad máxima de ocupantes, variable cuantitativa discreta porque es un conteo de personas. Rango: 48 a 119 personas (17 valores únicos).
location	Ubicación geográfica (coordenadas POINT(x,y)[ESRI web]), variable cuantitativa continua porque contiene valores decimales de geolocalización, con. 20 valores únicos. Necesita transformarse a coordenadas X,Y.
buildingId	Identificador del edificio que contiene el restaurante, variable cuantitativa discreta porque es un valor único. 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.
- Registros: 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.

Atributos - Pubs.csv:

Variable	Descripción
pubId	Identificador único del pub, variable cuantitativa discreta, porque es un valor único. Rango: 442 a 1800 (12 valores únicos).
hourlyCost	Costo por hora en el pub, variable cuantitativa continua porque los datos están en decimales, no finitos. Rango: $6.42 a $14.84 (12 valores únicos).
maxOccupancy	Capacidad máxima de ocupantes, variable cuantitativa discreta porque es un conteo. Los valores van en el rango: 60 a 96 personas (10 valores únicos).
location	Ubicación geográfica (coordenadas POINT(x,y)[ESRI web]), variable cuantitativa continua porque contiene valores decimales de geolocalización. Necesita transformarse a coordenadas X,Y.
buildingId	Identificador del edificio que contiene el pub, variable cuantitativa discreta porque es un valor único. 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

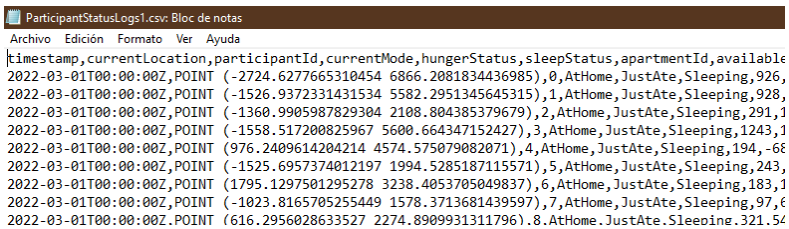
- Las variables o atributos están describiendo al objeto Pub.
- Registros: Un registro representa en sí a un colegio individual, con el costo de por mes, la capacidad máxima de estudiantes, la ubicación y el ID del edificio.

Atributos - Schools.csv:

Variable	Descripción
schoolId	Identificador único de la escuela (cuantitativo discreto porque es unico). Rango: 0 a 1350 (4 únicos). Nota: Hay una escuela con ID=0, no es razón de eliminar, porque si tiene los demás datos, es algo normal de la data.
monthlyCost	Costo mensual por estudiante (cuantitativo continuo, está en decimales y es no finito). Rango: $12.81 a $91.14. Esto nos ayudará a calcular presupuestos para participantes
maxEnrollment	Capacidad máxima de estudiantes (cuantitativo discreto, ya que es un conteo). El rango va desde 242 a 418 alumnos. Esto nos puede ayudar a saber si hay muchos niños o pocos niños por la zona.
location	Ubicación geográfica (coordenadas POINT(x,y)[ESRI web]), variable cuantitativa continua porque contiene valores decimales de geolocalización, con 4 valores únicos. Necesita transformarse a coordenadas X,Y.
buildingId	Identificador del edificio (cuantitativo discreto, porque es único). Rango: 123 a 943 (4 únicos).

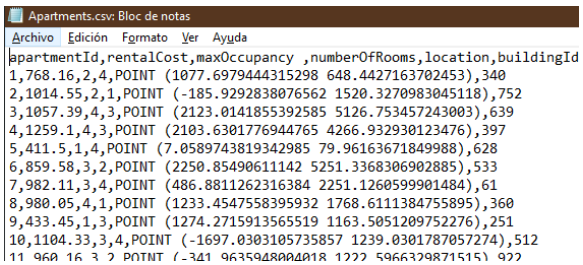
2.3. Formato:

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


ParticipantStatusLogs1.csv: Bloc de notas

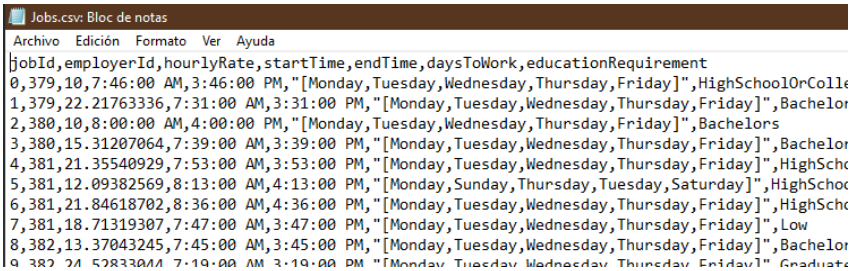
Archivo Edición Formato Ver Ayuda

```
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 1222.596632987151 5),922
```



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.

2.4. Transformaciones y limpieza de datos:

Las transformaciones que se aplicaron fueron a las coordenadas POINT y POLYGON, en el primer caso se creaba 2 columnas extra con el nombre X,Y, que eran tomadas como coordenadas, no se eliminó o daño el dataset original porque ello podría dañar toda la información. En el caso de POLYGON, se transformó en un objeto `shapely.Polygon`, shapely crea un objeto geométrico en un espacio cartesiano genérico. Las coordenadas se tratan como valores numéricos puros (como un gráfico de matplotlib cualquiera), sin referencia a la Tierra[Burmeister et al., Wonsang et al]. Luego se tuvieron transformaciones muy simples como en el caso de maxOccupancy de Buildings, donde se cambio de tipo de variable cuantitativa continua, a discreta.

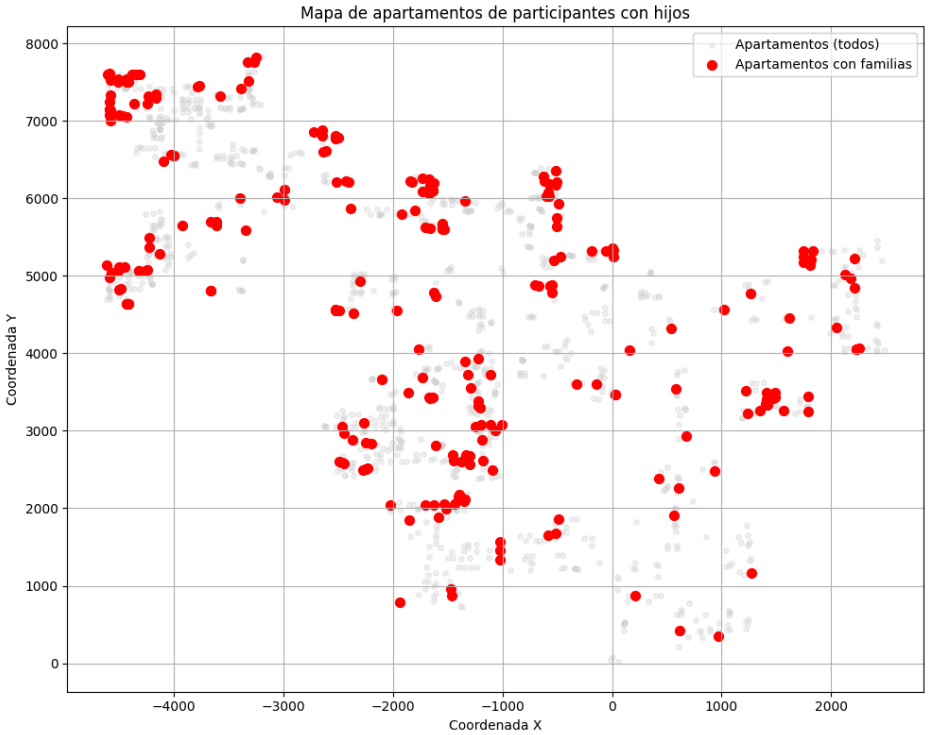
Además de ello, en la parte donde se dan categorías como financialStatus en ParticipantStatusLogs se tuvieron que agrupar las categorías para detectar mejor a los participantes y su nivel de educación. Y también se tuvo que crear tablas aparte sin dañar el dataset original. Para así poder tener la información necesaria para resolver nuestras hipótesis.

3. Exploración:

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?

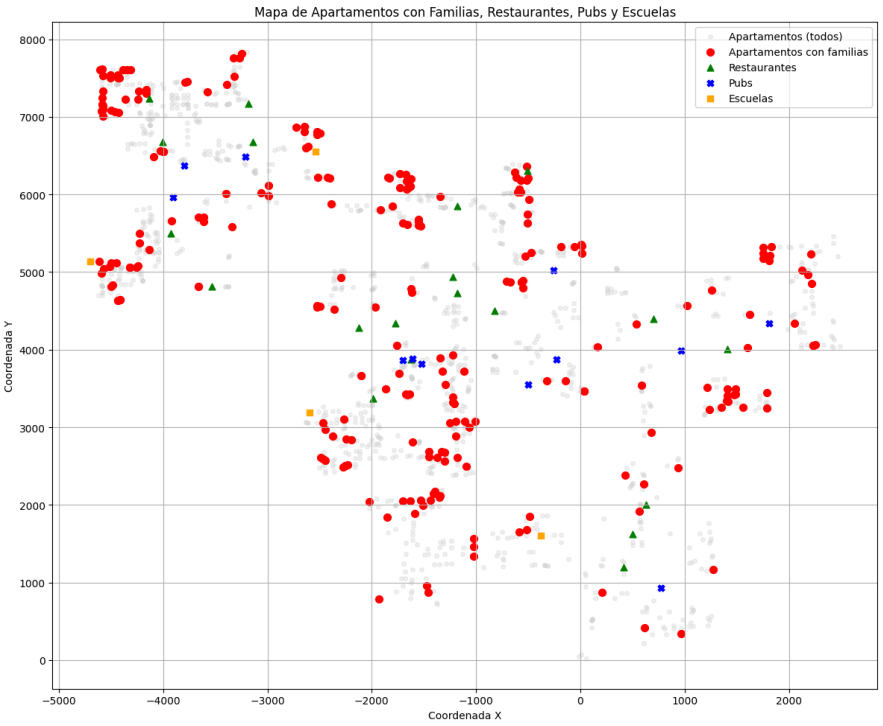
1. Filtrar participantes que tienen por lo menos un niño: En Participants.csv mire sus columnas y tipos de datos. Noté la columna haveKids que es booleana (True/False). Siempre que haveKids sea True, es una familia, porque quiere decir que tienen uno o más hijos.
2. Determinar el ‘apartamento principal’ de cada persona: Como los ParticipantStatusLogs muestran cambios de apartamento, conté cuántas veces aparece cada persona en cada apartmentId. Decidí que el apartamento en el que más veces apareciera sería su “apartamento principal”

3. Emparejar familias con su apartamento principal: Hice un “merge” entre la lista de participantes con niños y la tabla de “apartamento principal”. Así vi cuántas familias efectivamente tenían un apartamento asignado.
4. Agrupar participantes del mismo apartamento como una sola familia: A cada participante con niños le asigné familiaId = apartmentId_principal. De ese modo, todos los que compartían un apartamento contaban como una familia. Luego conté cuántas familias distintas había (un apartamento = una familia).
5. Ubicar los apartamentos de familias en un mapa: Leí la tabla Apartments.csv, transformé la columna de texto “POINT (x y)” en dos variables numéricas (x y y) y filtré solo los edificios donde vivían familias. En paralelo, dejé el resto de los apartamentos en gris claro como referencia.



Vi que los edificios donde hay familias con niños se concentran principalmente en la parte superior izquierda del mapa y, en menor medida, en el centro. Esto ya me daba zonas de interés.

6. Incluir restaurantes, pubs y escuelas en el mismo mapa: Para entender por qué esas zonas atraían a familias, saqué las tablas de Restaurants, Pubs y Schools, les extraje sus coordenadas x,y y las dibujé sobre el mismo fondo de apartamentos

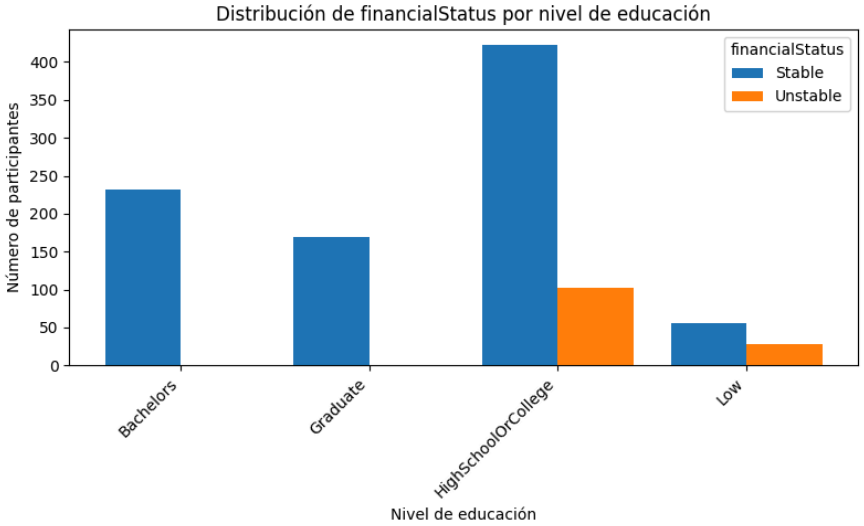


Confirmé que los edificios rojos (familias) están rodeados de varios cuadrados naranjas (escuelas) y triángulos verdes (restaurantes). En la parte superior izquierda hay al menos dos escuelas muy cerca y tres o cuatro restaurantes a menos de 100 metros. Esa cercanía explica por qué las familias con niños prefieren esa zona.

HIPÓTESIS 2:

1. Elegir los niveles educativos: Primero, abrí la tabla de Participants y me fijé en la columna educationLevel. Vi que tenía categorías como HighSchoolOrCollege, Bachelors, Masters, Graduated y algunos valores vacíos. Convertí esa columna a tipo “categoría” para facilitar el manejo y luego conté cuántos participaban en cada nivel.
    a. Además noté que la mayor parte de la gente estaba en HighSchoolOrCollege, mientras que muy pocos eran Graduated.
2. Sacar el último estado financiero de cada participante: Revisé todos los archivos ParticipantStatusLogs* y observé la columna financialStatus. Había unos cuantos NaN, así que rellené esos con “Unknown” y convertí todo a “categoría”. Luego ordené cada participante por fecha y me quedé con el registro más reciente para quedarse con su financialStatus final (“Stable”, “Unstable”, “Unknown”).
    a. Vi que la mayoría mantenía el mismo estado financiero todo el tiempo, pero existía un grupo pequeño que cambió de “Stable” a “Unstable” o viceversa en algún momento.
3. Relacionar educación con estado financiero: Junté (merge) la tabla de participantes (con su educationLevel) y la tabla que tenía el último financialStatus de cada uno. Después hice una tabla de contingencia que

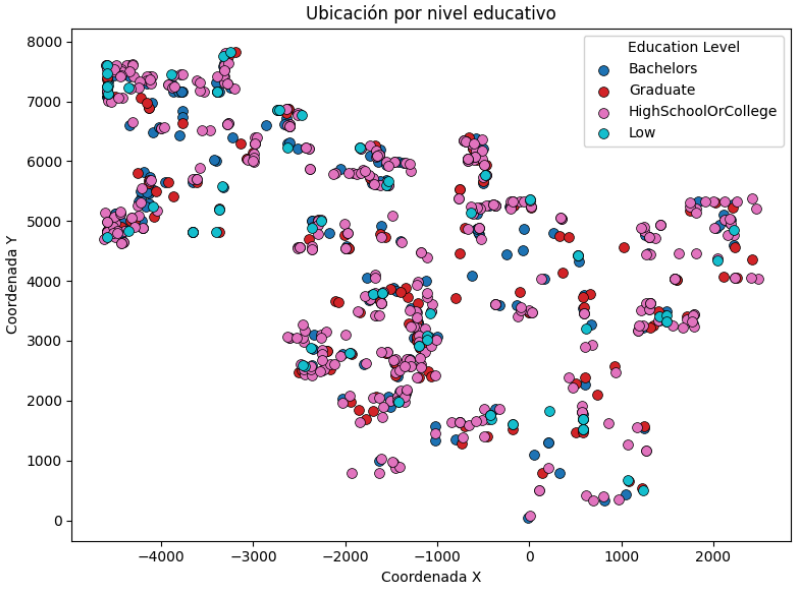
daba el número de participantes de cada nivel educativo según su estado financiero.

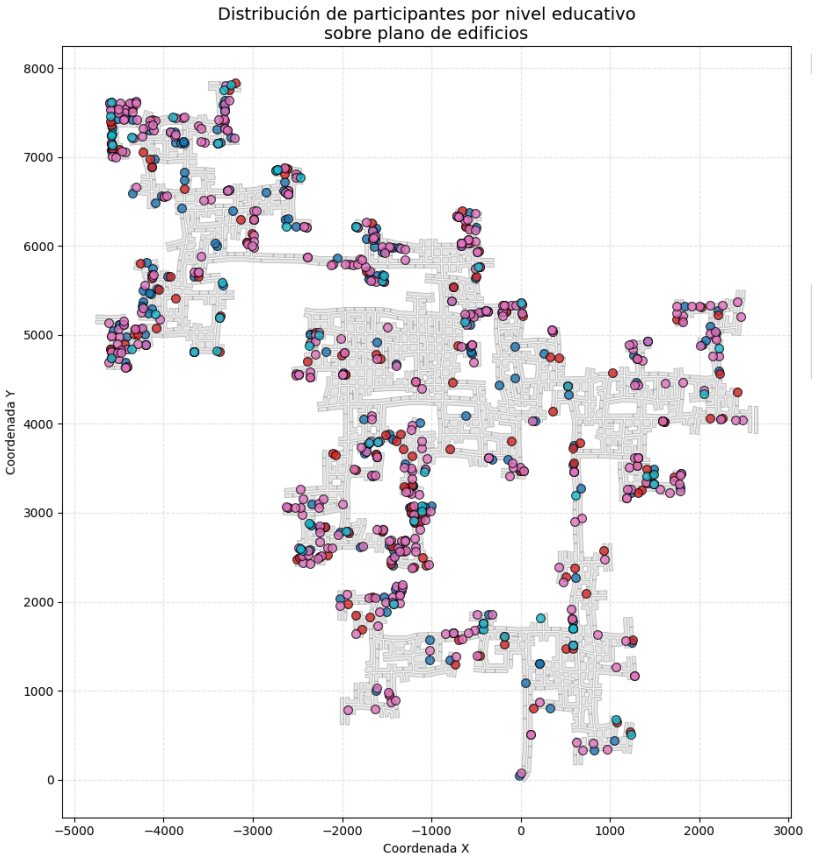


En HighSchoolOrCollege había más de 4 000 participantes con Stable, pero también ese mismo grupo lideraba en Unstable. Eso me hizo pensar que muchos de esos participantes(o quienes sólo terminaron la prepa) oscilan entre estar bien y estar mal.

En cambio, quienes tenían Bachelors o eran Graduated casi no aparecían en Unstable; la mayoría estaban en Stable. Eso sugiere que a mayor grado, menor probabilidad de tener problemas de dinero.

4. Checar ubicación geográfica según nivel educativo: Para ver si la educación también se agrupaba en ciertas zonas, abrí el CSV de Apartments, transformé cada “POINT (x y)” en coordenadas numéricas, y uní esas x, y con la tabla que ya tenía participantId, educationLevel y apartmentId.

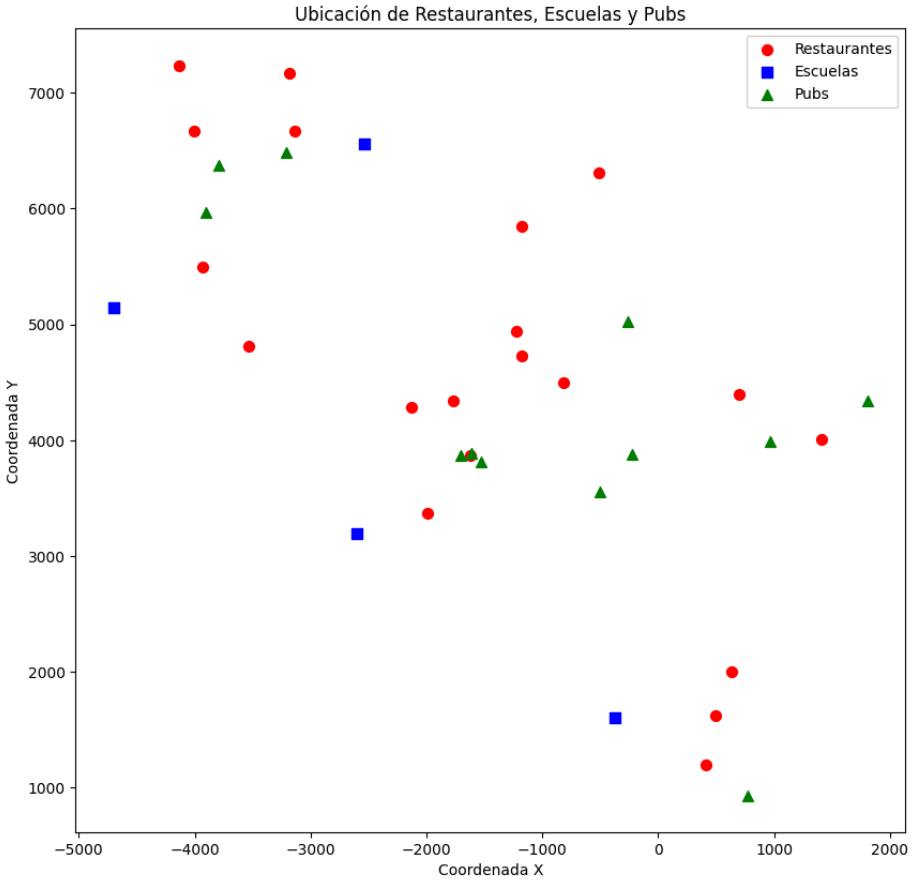




Los de HighSchoolOrCollege estaban prácticamente dispersos en toda la ciudad, sin un foco claro. Sin embargo, los que tenían Bachelors o eran Graduated se veían más cerca del centro de la ciudad (concentración en una zona más acotada).

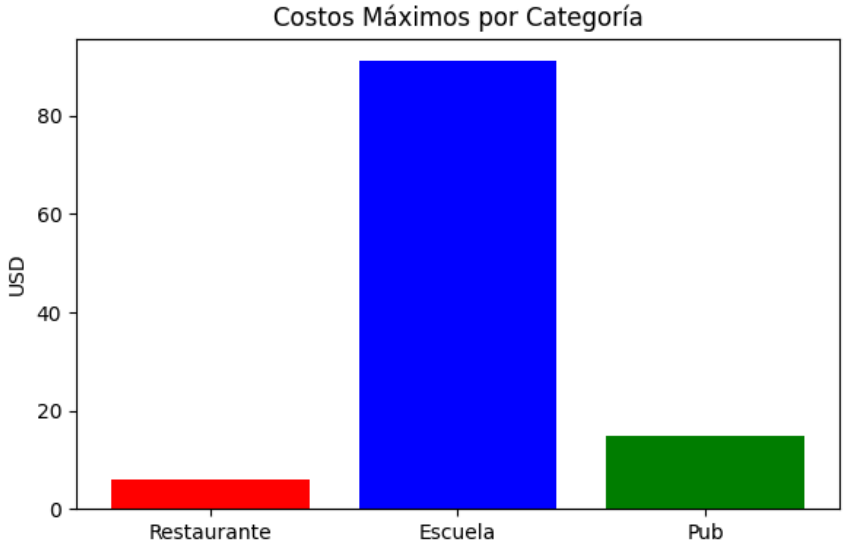
HIPÓTESIS 3:

1. Ubicación de restaurantes, escuelas, pubs y edificios: Primero marqué en un mismo mapa todas los restaurantes, escuelas y pubs.



Se pudo apreciar que muchas escuelas estaban concentradas hacia la esquina superior izquierda y un poco en el centro. Los restaurantes y pubs se esparcen por el alrededor de esas mismas zonas, pero la gran mayoría de pubs están alejados de las escuelas.

2. Costos máximos y gasto mensual estimado: Luego identifiqué los precios más altos registrados en cada categoría:

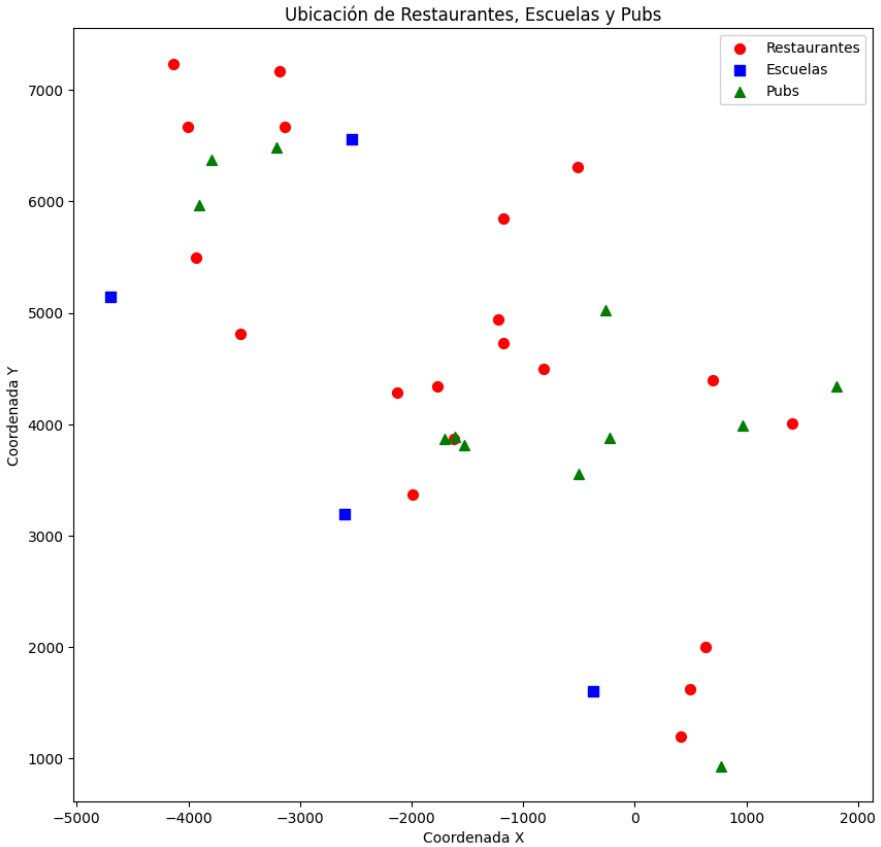


Después supuse:

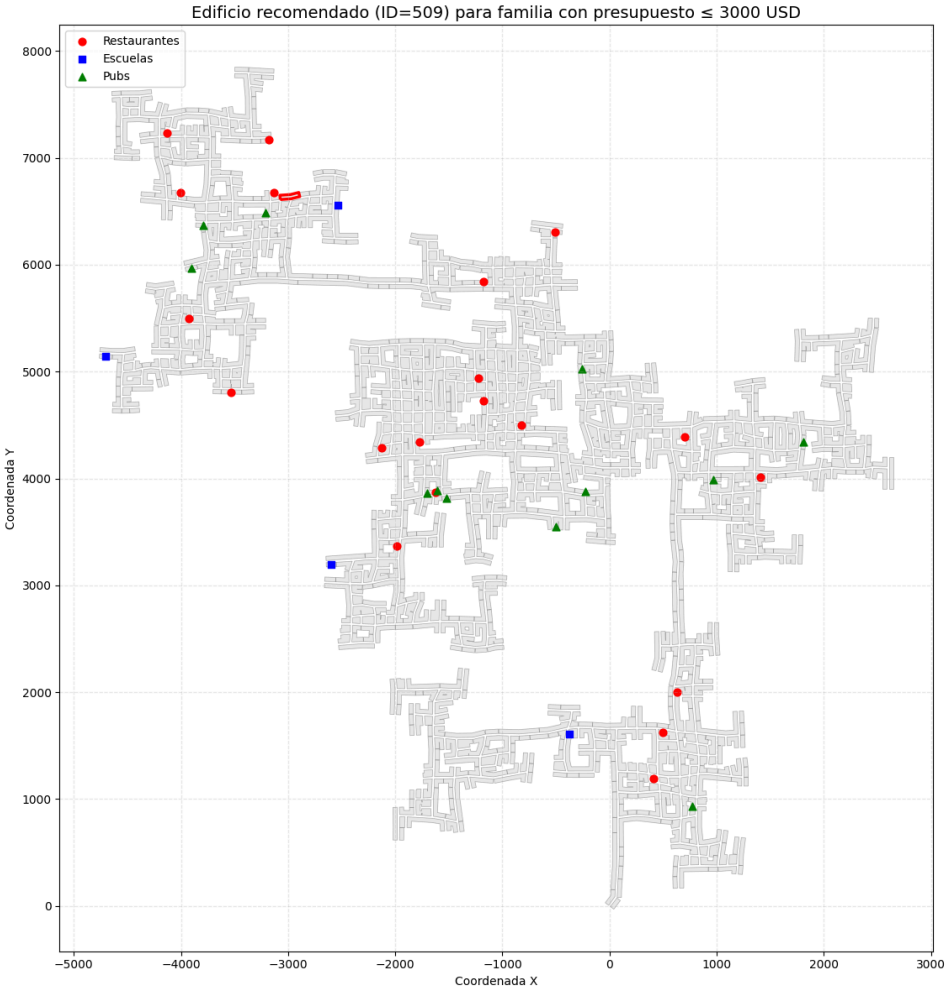
```
Costo máximo por comida (foodCost): 5.92 USD
Costo mensual máximo por estudiante (monthlyCost): 91.14 USD
Costo máximo por hora en pub (hourlyCost): 14.84 USD

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
```

Incluso con los precios más altos, la familia gastaría menos de la mitad de su presupuesto mensual, así que tienen "margen" para escoger un apartamento con renta prudente y todavía pagar restaurantes, escuelas y pubs sin pasarse del presupuesto.

3. Elegir el edificio ideal según distancias y los otros edificios como restaurantes, bares, etc. El segundo criterio fue que la familia quiera estar cerca de una escuela (por los niños), de al menos un par de restaurantes y quizá un pub para los adultos. Calculé la distancia mínima entre ese polígono de "edificio" y el punto de cada escuela. Guardé el menor valor. Lo mismo para restaurantes y pubs. Sumé esas tres distancias mínimas para cada edificio y obtuve una "distancia combinada".



```
Edificio recomendado (distancia mínima combinada):
buildingId          509.000000
dist_escuela        361.645493
dist_restaurante     63.226191
dist_pub            183.256274
suma_dist           608.127958
Name: 508, dtype: float64
```



Así pude obtener el edificio adecuado para vivir, ya que su presupuesto no sobrepasa los límites y además está bastante cerca de restaurantes, bares y escuelas.

4. Conclusión:

Respondiendo a la 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?

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

Respondiendo a la HIPÓTESIS 2: ¿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 obsevamos 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 particpantes, 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.

Respondiendo a la 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?

- 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 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 pubs y 2 escuelas. Además en el último gráfico ubicamos el mejor edificio para ellos tomando en cuenta la distancia mínima hacia restaurante, escuela y bares.

Conclusiones:

Al terminar este análisis exploratorio de datos, siento que no solo aprendí a manipular datasets algo complejos, sino también a pensar como un verdadero científico de datos, a la vez me di cuenta que no es del todo fácil.. Me di cuenta de que detrás de cada hipótesis hay una historia que se puede revelar con paciencia, lógica y herramientas adecuadas. El trabajo con los datos del VAST Challenge 2022 me permitió integrar conocimientos técnicos

con el pensamiento crítico, y me mostró cómo la visualización y el análisis pueden ayudar a responder preguntas sociales importantes, aunque se trate de una ciudad ficticia.

También entendí que en el análisis de datos no todo es lineal: muchas veces tuve que volver sobre mis pasos, reevaluar decisiones y ajustar enfoques. Esto, lejos de ser una falla, es parte natural del proceso. Cabe resaltar que considero que aún continuo en una etapa de aprendizaje, ya que el campo de la ciencia de datos es muy extenso.

Anexos:

- PDF colab : ENLACE COLAB
- DASHBOARD: Dashboard
- Vídeo DASHBOARD: VIDEO

Referencias

He, Edward & Tolessa, Daniel & Suh, Ashley & Chang, Remco. (2022). Analysis Without Data: Teaching Students to Tackle the VAST Challenge. 10.48550/arXiv.2211.00567.

W. Tong et al., "Towards an Understanding of Distributed Asymmetric Collaborative Visualization on Problem-solving," 2023 IEEE Conference Virtual Reality and 3D User Interfaces (VR), Shanghai, China, 2023, pp. 387-397, doi: 10.1109/VR55154.2023.00054. keywords: {Performance evaluation;Visualization;Three-dimensional displays;Collaboration;Prototypes;Virtual reality;User interfaces;asymmetric collaborative visualization;virtual reality;data visualization;problem solving},

B. Lee, X. Hu, M. Cordeil, A. Prouzeau, B. Jenny and T. Dwyer, "Shared Surfaces and Spaces: Collaborative Data Visualisation in a Co-located Immersive Environment," in IEEE Transactions on Visualization and Computer Graphics, vol. 27, no. 2, pp. 1171-1181, Feb. 2021, doi: 10.1109/TVCG.2020.3030450.
keywords: {Data visualization;Collaboration;Three-dimensional displays;Task analysis;Virtual environments;Operating systems;Two-dimensional displays;Immersive analytics;collaboration;virtual reality;qualitative study;multivariate data},

Burmeister, J., Liao, J., Yang, J., Wei, Q., & Wang, K. (2022). Visual Analytics for Urban Data Analysis. IEEE Visual Analytics Science and Technology (VAST Challenge Workshop) 2022. https://doi.org/10.24406/publica-1591

Song, Wonsang & Lee, Jae Woo & Schulzrinne, Henning. (2011). Polygon Simplification for Location-Based Services Using Population Density. IEEE International Conference on Communications. 1-6. 10.1109/icc.2011.5963369.

Environmental Systems Research Institute (ESRI). (s. f.). Agregar datos de coordenadas X-Y como una capa. ArcGIS Desktop. Recuperado el [fecha en que accediste, p. ej., 5 de junio de 2025] de https://desktop.arcgis.com/es/arcmap/latest/map/working-with-layers/adding-x-y-coordinate-data-as-a-layer.htm