Proyecto Final de Ingeniería de Datos

A data-driven analysis of Argentinian wages and household economics in 2019

William Sebastián Cabrera Buitrago

Paola Andrea Cortés Montes

Andrés Felipe Miranda Mendoza

Sara Sofía Torres Méndez

Simón Vélez Castillo

Profesor: Javier Casas Salgado

Universidad del Rosario

Ingeniería de Datos

2024 - I

Descripción

Problema

Se dispone de un conjunto de datos que contiene información demográfica y socioeconómica de Argentina en el 2019, sujeta a ciertas restricciones y reglas. El objetivo es realizar un análisis detallado de estos datos para responder a una serie de preguntas relacionadas con los ingresos, la educación, el estado conyugal, la situación laboral y otros aspectos relevantes de la población. El enfoque principal está orientado a entender los distintos factores que influyen en la obtención de salarios elevados, identificando así las variables determinantes en el panorama socioeconómico del país.

Reglas

- Una persona puede tener solo un estado conyugal a la vez
- En cada hogar debe haber por lo menos un miembro.
- En cada una casa(id) puede haber un hogar o más.
- Cada hogar solo debe tener un jefe.
- Los años de escolaridad no pueden ser mayor a la edad de la persona.
- Todos los ingresos de las personas son no negativos.
- El sexo de las personas solo puede ser mujer o varón.
- El id de cada casa es único.
- Cada miembro debe tener un hogar.
- Cada hogar debe tener una casa (Id).

Preguntas

- ¿Cuáles son los ingresos de las personas las cuales han terminado sus estudios universitarios a comparación de aquellas personas que no la han completado?
- ¿Cómo compara la situación laboral de la gente que vive con sus padres versus quienes no?
- ¿En promedio cuantos ingresos de familias cuya cabeza de hogar sea mujer superan el salario mínimo en argentina durante el 2019?
- ¿Como se comparan los ingresos familiares y el salario entre universidad privada y pública?
- ¿Qué lugar de nacimiento (de los presentes en la encuesta) cuenta con los mayores salarios?

- ¿Las familias más grandes tienden a tener una mejor, igual o peor educación que las familias más pequeñas?
- ¿Las personas con ingresos familiares per cápita superiores a la media tienen salarios superiores a la media?

Análisis.

.

- Entidades:
    - Hogar (fuerte):se refiere al hogar de la familia encuestada sus atributos son dominio, comuna, id, casa_id_fk, ingresos_familiares, numero_hogar,calidad_ingresos familiares.
    - Miembro (Fuerte): es una persona de alguna de las familias encuestadas;sus atributos son id_miembro, numero_hogar_fk, sexo,edad,num_miembros,parentesco,situacion_sent,id_casa_fk
    - Tipo ingresos (Débil): Da información acerca de la situación economica de una persona;sus atributos son estado_ocup, cat_ocup, aporte_ingreso_fami, ingreso_tot_nolab, ingreso_tot_lab,calidad_ingre_ocup, calidad_tot_nolab, calidad_ingresos_totales,ingresos_totales, ingreso_per_capita_familiar, id2_miemmbro_pk
    - Educación (Débil): Da informacion del nivel academico de una persona; la sus atributos son años_Escolaridad, nivel_max_educativo,nivel_actual, sector_educativo, id1_miembro_pk, estado_educativo
    - Salud (Débil): aporta informacion medica de la personasus atributos son hijos_nacidos,cant_hijos_nacidos, afiliacion_salud, ligar_nacimiento, id_miembro_fk.
- Atributos:
    - **Id:**describe el id de cada casa y es un entero.
    - **nhogar**:este describe el número de hogar de cada casa correspondiente y es un entero.
    - **id_miembro**:es el id de cada miembro y es un entero.
    - **miembro**: describe el número de miembro correspondiente a cada hogar.
    - **comuna**:describe el dominio de cada casa (id)y este puede contener los datos de “Villas de la emergencia” o “Resto de la ciudad” y es un string.

- **sexo**:desribe el sexo binario de cada miembro y solo recibo los datos de “Mujer” y “Varón” y es un string.
- **edad**: describe la edad de un miembro en años y este siempre es un número no negativo, es un entero.
- **situacion_conyugal**: describe el estado civil de cada miembro y este puede contener los datos de “Soltero/a”, “Viudo/a”, “Unido/a”, “No corresponde”, “Divorciado/a”; este es un string.
- **Dominio**:el lugar donde se encuentra cada hogar y es un stsring que contiene “Resto de la Ciudad” o “Villas de emergencia”
- **parentesco**: describe el rol de la persona en el hogar y es un string.
- **num_miembro_padre**: describe el número del miembro del padre si vive en el hogar, si no vive en el hogar le otorgamos un -1 y si no corresponde -2, el tipo de dato es string.
- **num_miembro_madre**: describe el número de miembro de la madre si vive en el hogar, si no vive en el hogar le otorgamos un -1 y si no corresponde -2, el tipo de dato es string.
- **estado_ocupacional**: contiene los datos de “Inactivo”, “Ocupado” y “Desocupado” y es un string.
- **Cat_ocupacional:** contiene la información de si la persona es o no asalariada, as un string
- **calidad_ingreso_lab**: Indica la calidad de los ingresos de un individuo, lo que podría incluir aspectos como la estabilidad del empleo, la regularidad de los ingresos, entre otros. Es un string.
- **calidad_ingreso_familiares**: Indica la calidad de los ingresos de un hogar, lo que podría incluir aspectos como la estabilidad del empleo, la regularidad de los ingresos, entre otros. Es un string.
- **ingreso_total:** Representa el ingreso total percibido por un individuo en un período determinado. Es un entero
- **calidad_ingresos_nolab:** Similar a calidad_ingreso, _lab pero específicamente relacionado con ingresos no laborales, como rentas, intereses, etc. Es string.
- **ngreso_total_nolab:** El ingreso total derivado de fuentes no laborales. Es entero.
- **ingreso_total_lab:** El ingreso total derivado de fuentes laborales. Es entero.
- **calidad_ingresosTotales:** Es la calidad de los ingresos laborales y no laborales. Es un string.
- **ingresos_familiares:** Representa el total de ingresos percibidos por una familia en un período determinado. Es un entero.

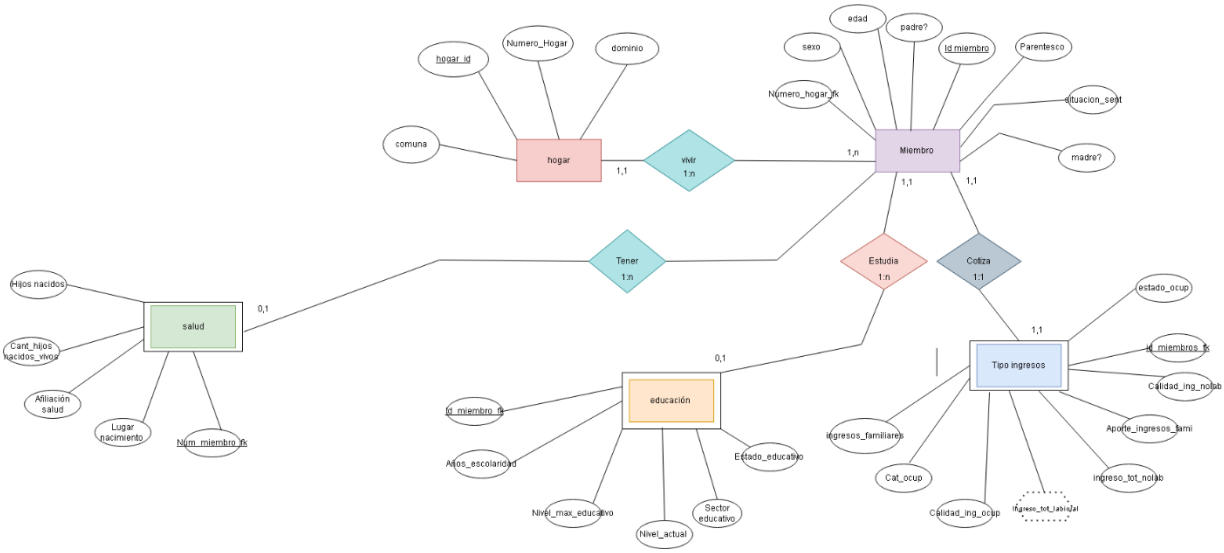
- **ingresos_per_capita_familiar**: Los ingresos totales de la familia divididos por el número de miembros de la familia. Es un entero.
- **años_escolaridad**: describe la cantidad de años que ha estudiado la persona y es un entero.

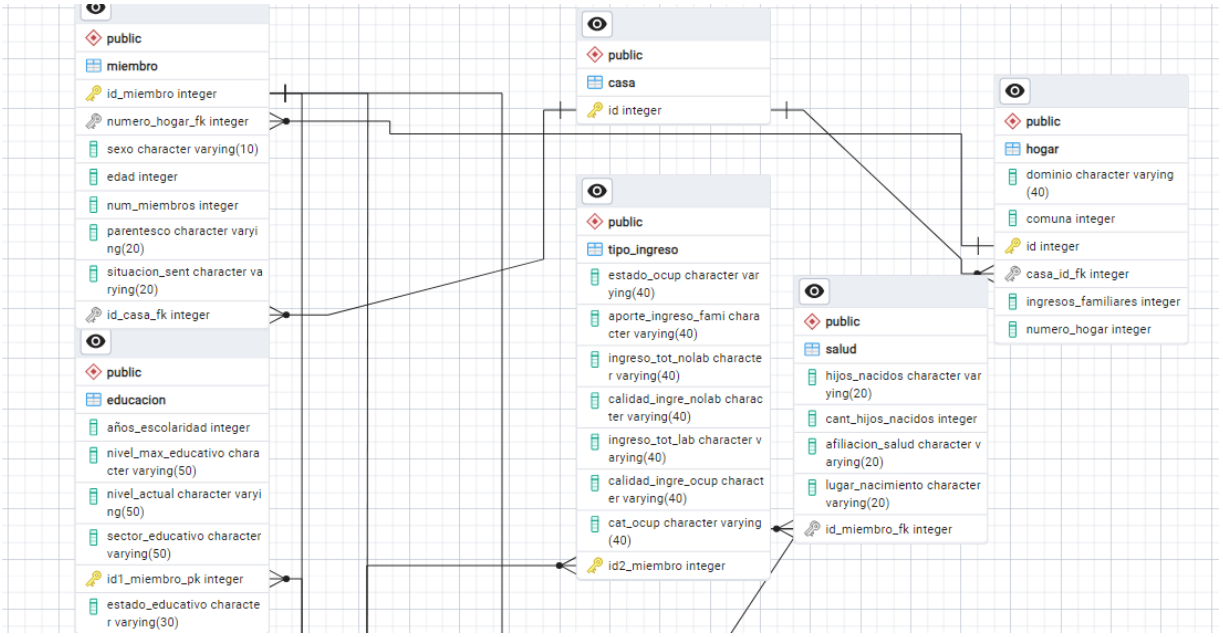
- **estado_educativo**: Indica el estado del nivel educativo actual de un individuo, como asiste o no asiste etc. Es categórica.
- **sector_educativo**: Se refiere al tipo de institución educativa en la que está inscrito un individuo, como pública, privada, etc. Es un string.
- **nivel_actual**: El nivel educativo actual de un individuo, que podría ser grado, curso, año, etc. Es un string.
- **nivel_max_educativo**: El nivel educativo máximo alcanzado por un individuo hasta el momento. Es categórica.
- **cantidad_hijos_nac_vivos:** El número de hijos que nacieron con vida de cada miembro y es un entero.
- **hijos_nacidos_vivos**: Describe si tiene hijos o no y es un string.
- **afiliacion_salud**: Describe el tipo de afiliación que tiene cada miembro y es un string.
- **Lugar_nacimiento**: describe la locación de donde nació un miembro y es un string.

- Relaciones:
  - Reside: esta relación nos indica cuantos hogares hay por casa
  - Vivir: Describe cuantas personas hay por hogar
  - Cotiza: indica los ingresos de cada miembro
  - Estudia: indica el nivel académico de cada miembro
  - Tener: indica si cada miembro tiene plan de salud

Model**o Entidad Relación**

Modelo Relacional normalizado en tercera forma normal



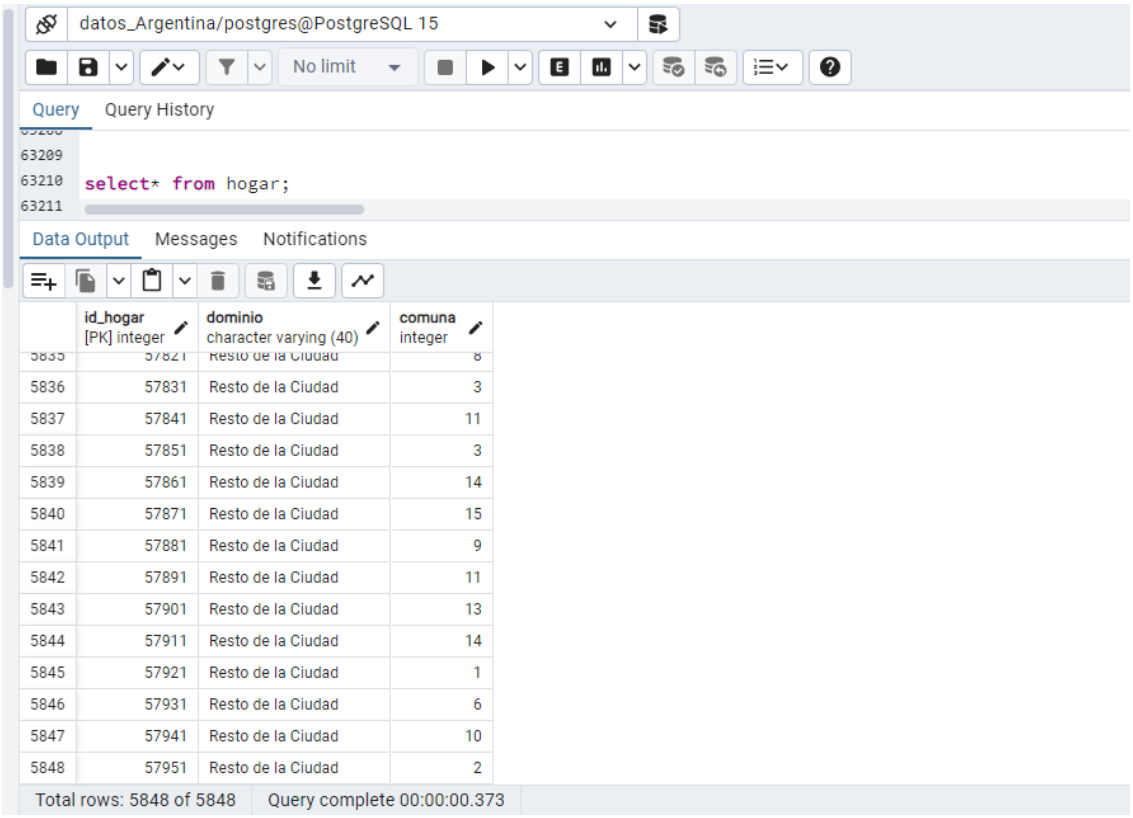


Enlace base de datos

https://docs.google.com/spreadsheets/d/1yvxhY7NS5FK6kLnpIqBCFv5bZn3UvFZzG7IOxGMUJy0/edit?usp=sharing

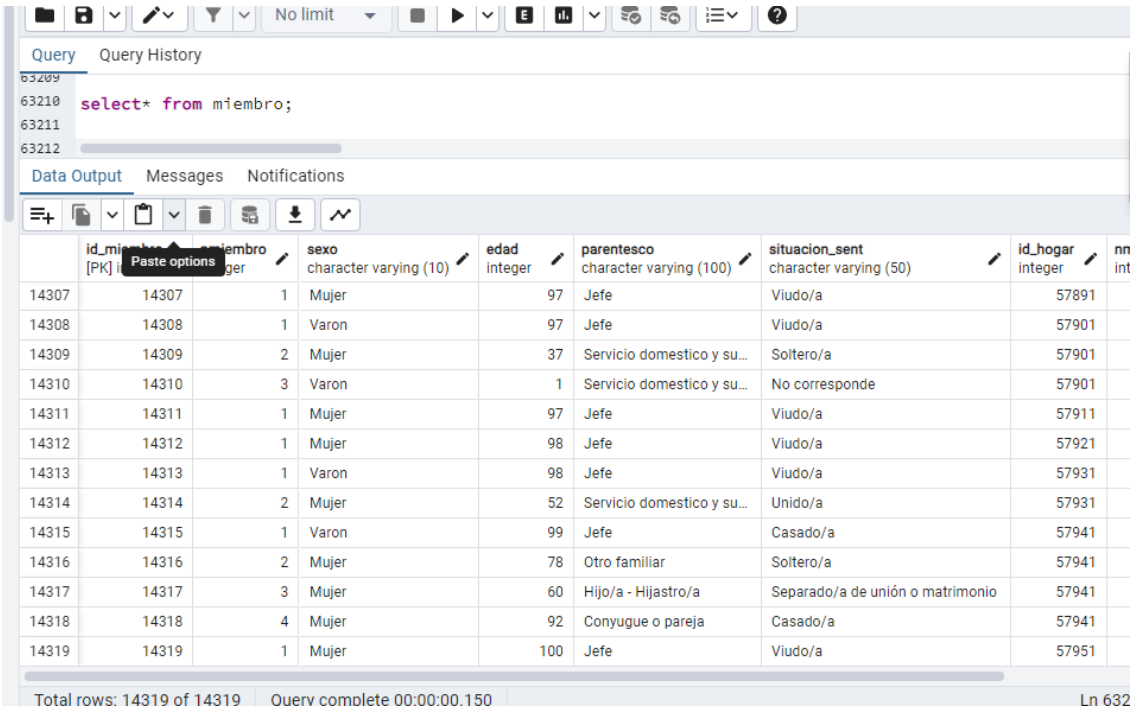
Descripción del proceso de carga

Se llevó a cabo una carga máxima de datos mediante la función 'INSERT INTO' en cinco tablas diferentes del sistema de bases de datos. Antes de iniciar el proceso, se verificó la integridad de los datos para asegurar que cumplieran con las restricciones de cada tabla, incluyendo claves primarias y restricciones de valor único. La operación se ejecutó dentro de transacciones para garantizar la consistencia de los datos, con un riguroso control de errores para manejar cualquier anomalía durante la inserción. Se mantuvo un registro detallado de la carga para auditar el proceso y cualquier error encontrado.



	id_hogar [PK] integer	dominio character varying (40)	comuna integer
5836	57831	Resto de la Ciudad	3
5837	57841	Resto de la Ciudad	11
5838	57851	Resto de la Ciudad	3
5839	57861	Resto de la Ciudad	14
5840	57871	Resto de la Ciudad	15
5841	57881	Resto de la Ciudad	9
5842	57891	Resto de la Ciudad	11
5843	57901	Resto de la Ciudad	13
5844	57911	Resto de la Ciudad	14
5845	57921	Resto de la Ciudad	1
5846	57931	Resto de la Ciudad	6
5847	57941	Resto de la Ciudad	10
5848	57951	Resto de la Ciudad	2

Total rows: 5848 of 5848 Query complete 00:00:00.373



	id_miembro [PK]	nro_miembro integer	sexo character varying (10)	edad integer	parentesco character varying (100)	situacion_sent character varying (50)	id_hogar integer	nn int
14307	14307	1	Mujer	97	Jefe	Viudo/a	57891	
14308	14308	1	Varon	97	Jefe	Viudo/a	57901	
14309	14309	2	Mujer	37	Servicio domestico y su...	Soltero/a	57901	
14310	14310	3	Varon	1	Servicio domestico y su...	No corresponde	57901	
14311	14311	1	Mujer	97	Jefe	Viudo/a	57911	
14312	14312	1	Mujer	98	Jefe	Viudo/a	57921	
14313	14313	1	Varon	98	Jefe	Viudo/a	57931	
14314	14314	2	Mujer	52	Servicio domestico y su...	Unido/a	57931	
14315	14315	1	Varon	99	Jefe	Casado/a	57941	
14316	14316	2	Mujer	78	Otro familiar	Soltero/a	57941	
14317	14317	3	Mujer	60	Hijo/a - Hijastro/a	Separado/a de unión o matrimonio	57941	
14318	14318	4	Mujer	92	Conyugue o pareja	Casado/a	57941	
14319	14319	1	Mujer	100	Jefe	Viudo/a	57951	

Total rows: 14319 of 14319 Query complete 00:00:00.150 Ln 632

```
63209
63210 select* from educacion;
63211
63212
```

Data Output | Messages | Notifications

	id1_miembro [PK] integer	años_escolaridad integer	nivel_max_educativo character varying (50)	nivel_actual character varying (50)	sector_educativo character varying (50)	estado_educati character varyi
14306	14306	7	Primario comun	No corresponde	No corresponde	No asiste pero
14307	14307	7	Primario comun	No corresponde	No corresponde	No asiste pero
14308	14308	17	Secundario/medio comun	No corresponde	No corresponde	No asiste pero
14309	14309	12	Primario especial	No corresponde	No corresponde	No asiste pero
14310	14310	0	No corresponde	No corresponde	No corresponde	Nunca asistio
14311	14311	0	No corresponde	No corresponde	No corresponde	Nunca asistio
14312	14312	3	Sala de 5	No corresponde	No corresponde	No asiste pero
14313	14313	17	Secundario/medio comun	No corresponde	No corresponde	No asiste pero
14314	14314	7	Primario comun	No corresponde	No corresponde	No asiste pero
14315	14315	5	Sala de 5	No corresponde	No corresponde	No asiste pero
14316	14316	9	EGB (1° a 9° año)	No corresponde	No corresponde	No asiste pero
14317	14317	12	Primario especial	No corresponde	No corresponde	No asiste pero
14318	14318	7	Primario comun	No corresponde	No corresponde	No asiste pero
14319	14319	15	Secundario/medio comun	No corresponde	No corresponde	No asiste pero

Total rows: 14319 of 14319 | Query complete 00:00:00.177

```
63209
63210 select* from tipo_ingreso;
63211
63212
```

Data Output | Messages | Notifications

	id2_miembro [PK] integer	estado_ocup character varying (80)	cat_ocup character varying (80)	ingreso_tot_nolab integer	calidad_ingre_nolab character varying (80)	ingr inte
14307	14307	Inactivo	No corresponde	22500	Tuvo ingresos pero no declara monto	
14308	14308	Inactivo	No corresponde	58000	Tuvo ingresos y declara monto	
14309	14309	Ocupado	Asalariado	2600	Tuvo ingresos y declara monto	
14310	14310	Inactivo	No corresponde	0	No corresponde	
14311	14311	Inactivo	No corresponde	16000	Tuvo ingresos y declara monto	
14312	14312	Inactivo	No corresponde	12250	Tuvo ingresos y declara monto	
14313	14313	Inactivo	No corresponde	80000	Tuvo ingresos y declara monto	
14314	14314	Ocupado	Asalariado	0	No tuvo ingresos	
14315	14315	Inactivo	No corresponde	24000	Tuvo ingresos pero no declara monto	
14316	14316	Inactivo	No corresponde	11000	Tuvo ingresos y declara monto	
14317	14317	Inactivo	No corresponde	11000	Tuvo ingresos y declara monto	
14318	14318	Inactivo	No corresponde	11000	Tuvo ingresos y declara monto	
14319	14319	Inactivo	No corresponde	82000	Tuvo ingresos y declara monto	

Total rows: 14319 of 14319 | Query complete 00:00:00.196

```
63209
63210 select* from salud;
63211
63212
```

Data Output | Messages | Notifications

	id3_miembro [PK] integer	hijos_nacidos character varying (5)	cant_hijos_nacidos integer	afiliacion_salud character varying (90)	lugar_nacimiento character varying (5
14306	14306	Si	1	Solo obra social	Pais no limitrofe
14307	14307	Si	2	Solo obra social	Pais no limitrofe
14308	14308	No	0	Otros	Pais no limitrofe
14309	14309	Si	1	Solo obra social	Otra provincia
14310	14310	No	0	Solo obra social	CABA
14311	14311	Si	2	Otros	CABA
14312	14312	Si	1	Solo obra social	Pais no limitrofe
14313	14313	No	0	Solo obra social	CABA
14314	14314	Si	3	Solo obra social	Pais limitrofe
14315	14315	No	0	Solo obra social	Pais no limitrofe
14316	14316	No	0	Solo obra social	Partido GBA
14317	14317	Si	2	Solo obra social	CABA
14318	14318	Si	1	Solo obra social	CABA
14319	14319	Si	1	Solo obra social	CABA

Total rows: 14319 of 14319 Query complete 00:00:00.151

Descripción de análisis identificados

1. El primer escenario es analizar los ingresos (ingreso_total) de las personas que terminaron sus estudios universitarios (nivel_max_educativo) y compararlo con las personas que no los han completado.
   Este escenario ayudaría a comprender y cuantificar los posibles beneficios de hacer un pregrado en la Argentina de 2019, respondiendo preguntas como ¿Qué tanto crece tu salario? ¿Vale la pena hacer un pregrado? ¿Estos beneficios cambian si la institución era pública o privada (sector_educativo)?
2. El segundo escenario es analizar la situación laboral (estado_ocupacional, cat_ocupacional, calidad_ingreso_lab) de la gente que vive con sus padres (num_miembro_padre y num_miembro_madre) vs la gente que no.
   Haciendo este análisis podríamos ver cómo vivir con los padres afecta la situación laboral y viceversa. ¿La gente que vive con sus padres gana menos? ¿La gente que vive sin sus padres tiene índices más bajos de desempleo? ¿Qué tipos de trabajo tienen las personas que viven con sus padres vs las que no? Son algunas de las preguntas que se pueden realizar.

3. Otro escenario es ver cuántas familias con cabeza de hogar mujer (edad, parentesco) superan el salario mínimo en argentina durante el 2019, comparándolo con las de cabeza de hogar hombre.

   Teniendo en cuenta el tamaño y variedad de la base de datos, analizar esto nos daría un vistazo en la desigualdad de género en argentina, respondiendo preguntas como ¿cómo afecta el salario ser madre cabeza de hogar? ¿Es más común estar en condiciones precarias siendo mujer?
4. Además, se pueden comparar los ingresos familiares (ingresos_familiares, ingresos_per_capita_familiar) y el salario (ingreso_total_lab) según el tipo de escolaridad que tiene una persona (sector_educativo), estudiando actualmente en una universidad pública o privada.

   Hacer esta proporción entre ingresos familiares y salario y analizarlo según el tipo de institución nos puede ayudar a entender quiénes necesitan trabajar mientras estudian y quiénes pueden depender de sus padres, respondiendo preguntas como, ¿los estudiantes de universidad privada necesitan trabajar más para sostenerse? ¿Los sostiene más la familia? ¿Cuántos estudian y trabajan y dónde estudian?
5. Otra situación por analizar es comparar los salarios (ingreso_total_lab) según los distintos lugares de nacimiento (Lugar_nacimiento) presentes en la encuesta. Con esto se conocería qué ciudades tienen mejores oportunidades laborales, sobre todo al compararse con el costo de vida de cada ciudad (aunque esta información no se encuentra en la base de datos), respondiendo preguntas como ¿Cuál es la ciudad con mayores salarios? ¿Qué ciudades son más igualitarias?
6. La sexta situación es la comparación del nivel de educación (nivel_max_educativo) según el tamaño de las familias (cantidad de personas que comparten el mismo id). Teniendo en cuenta el alto costo de estudiar (tanto en matrículas como en demorarse a entrar al mercado laboral) es importante analizar si el tamaño de las familias incluye en lo que se permite que gasten en educación, respondiendo preguntas como ¿Las familias más grandes tienen hijos menos educados?
7. La última situación para analizar es ver si las personas con ingresos per cápita superiores a la media (ingresos_per_capita_familiar) tienen salarios superiores a la media (ingreso_total_lab). Así se vería la influencia de la afluencia en el trabajo en la

Argentina de 2019, respondiendo preguntas como ¿Qué tanta influencia tiene venir de una familia acomodada a la hora de tener trabajo? ¿Entre la gente que logró un trabajo con altos ingresos cuántos vienen de familias acomodadas?

Descripción proceso creación módulos de consulta en Python

Enlace repositorio GIT

https://github.com/andrefm3435/ing_datozzz_best_group