Análisis de 2 variables relacionadas

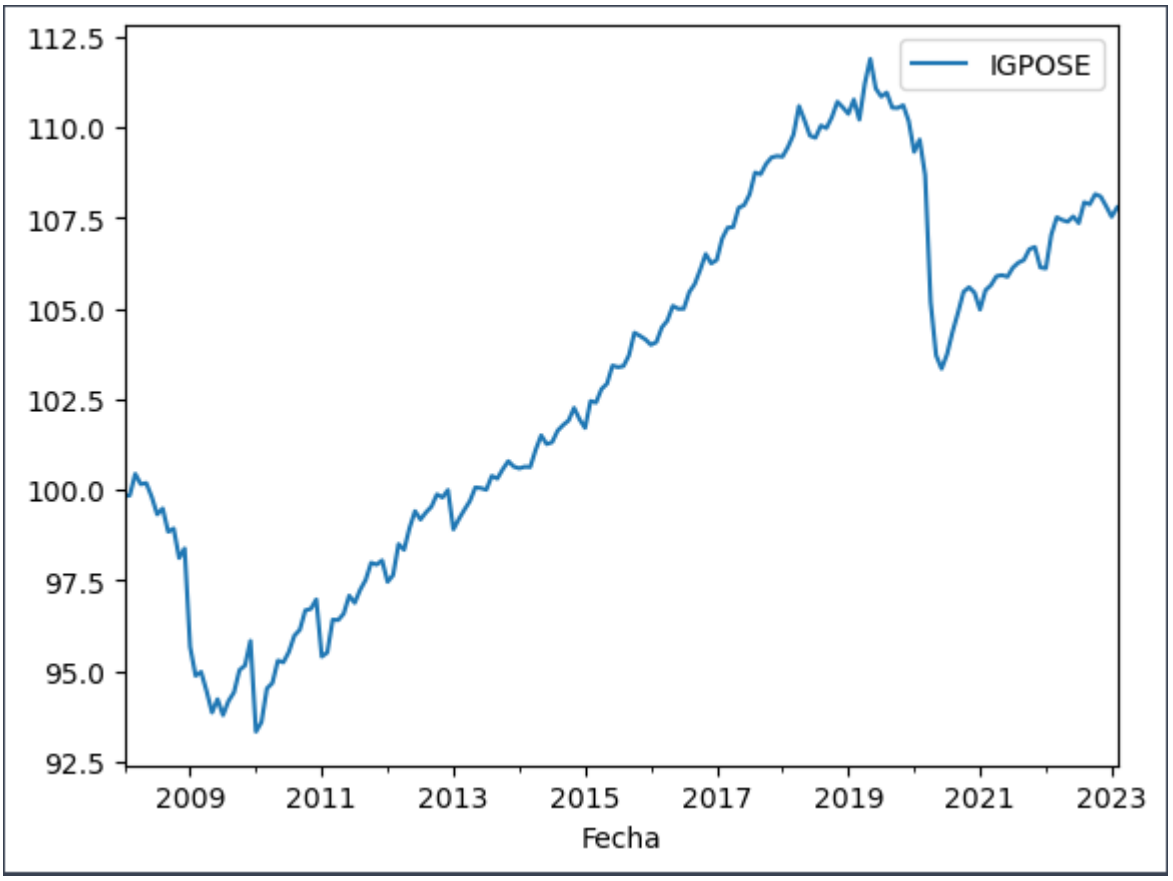
Mitsiu Alejandro Carreño Sarabia - E23S-18014

En el presente análisis se obtuvieron datos de las siguientes fuentes:

Base de datos de población:
https://www.inegi.org.mx/contenidos/masiva/indicadores/programas/ccpv/2020/cpv_00_csv.zip

Base de datos de ocupación:
https://www.inegi.org.mx/temas/personalo/#Tabulados

Lo primero que se analizó fue cuáles datos existen y que granularidad tenían, es decir, verificar que representa cada columna, en el primer archivo que se descargó de ocupación, los datos estaban en la unidad “Índice Global de Personal Ocupado de los Sectores Económicos” (IGPOSE) por lo que la tendencia iba del 95 al 115



pero los datos poblacionales estaban en número de personas, por lo que la comparación de ambas fuentes no revelaba datos correctos.

Una de las maneras en que se podía solucionar este problema era buscando cómo se obtiene el IGPOSE y aplicar una transformación a la población para que estuviera en la misma unidad y poder comparar.

Desafortunadamente no encontré una fórmula de transformación directa. Por ello, decidí buscar otra base de datos de ocupación que estuviera en unidades de número de personas.

Cambiando la fuente de población ocupada, el otro detalle que noté es la temporalidad de los datos.

La información de ocupación estaba presentada de manera trimestral.

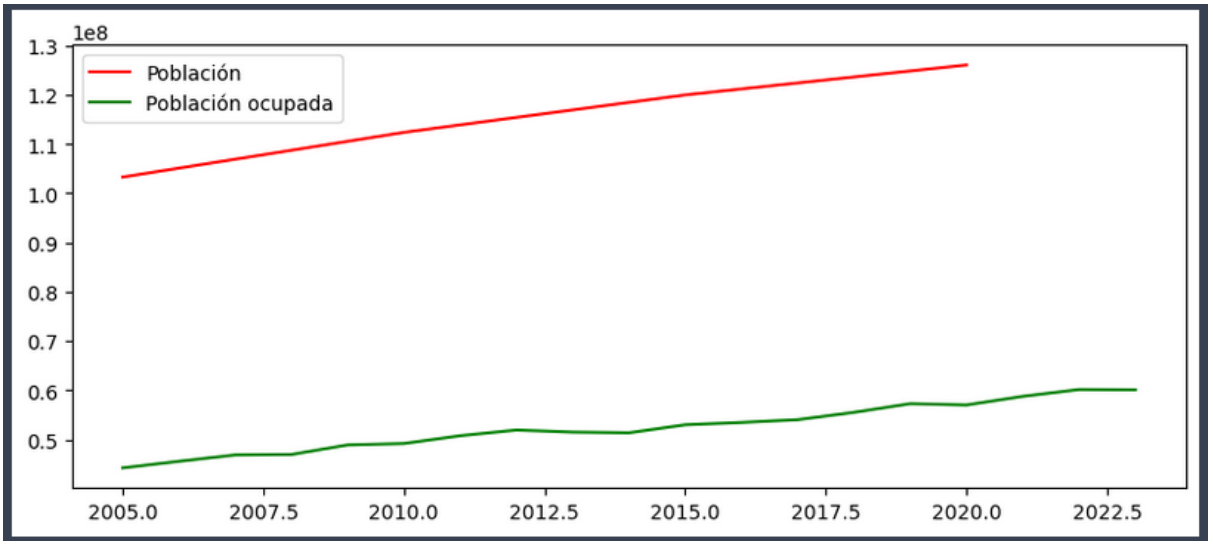
	year	tot
4	2005/01	43099847
5	2005/02	43180433
6	2005/03	44000204
7	2005/04	44245519
8	2006/01	44306012
...	...	...
72	2022/01	58085314
73	2022/02	59338419
74	2022/03	59480471
75	2022/04	60145456
76	2023/01	60089308

Por lo que se tuvo que preprocesar la información, primero, se hizo un tratamiento de valores para quitar el mes del formato original YYYY/mm una vez que se tenía el valor del año unicamente se realizó un agrupamiento por año y se tomó el valor máximo de los trimestres, con esto se obtuvo un único valor por año.

	year	tot
4	2005	43099847
5	2005	43180433
6	2005	44000204
7	2005	44245519
8	2006	44306012
...	...	...
72	2022	58085314
73	2022	59338419
74	2022	59480471
75	2022	60145456
76	2023	60089308

Para el dataset de población, se observó que existe una frecuencia de 5 años, por lo que los registros corresponden a 2005, 2010, 2015 y 2020, de igual manera es suficiente para estimar el valor en los años intermedios.

Una vez que se compararon ambas fuentes se obtuvo la siguiente gráfica.



En donde se puede apreciar que ambas cifras tienen una tendencia creciente, lo que concuerda con el crecimiento poblacional del país, otro aspecto a notar, es la diferencia marginal entre la población total y la ocupada y existen varios aspectos socioeconómicos que lo explican, los niños y adultos mayores no son empleados, los datos de ocupación únicamente cuentan a la población mayor de 15 años con ocupación formal, entonces queda fuera del conteo la población de trabajos informales.

Otra manera de realizar este análisis hubiese sido comparar la población ocupada, contra la población que está en el grupo de edad de actividad económica (15-55 años), ahí se esperaría que la cercanía entre líneas sea mayor.