UNIVERSIDAD NACIONAL DE SAN AGUSTÍN DE AREQUIPA

ESCUELA PROFESIONAL DE CIENCIA DE LA COMPUTACIÓN

CIENCIA DE DATOS

PIPELINE DE CIENCIA DE DATOS

PRESENTADO POR:

John Edson Sanchez Chilo

AREQUIPA – PERÚ
2025

Pipeline de Ciencia de Datos

Fuente de Datos

En este trabajo se analizarán dos fuentes de datos, ambas pertenecientes a Nueva Gales del Sur (NSW) en Australia. Los datos fueron obtenidos directamente de su página web https://data.nsw.gov.au/

Notificaciones de casos de COVID-19 en NSW, durante un período de aproximadamente dos años, desde el 25 de enero del 2020 al 07 de febrero de 2022.

	notification_date	postcode	lhd_2010_code	lhd_2010_name	lga_code19	lga_name19
0	2020-01-25	2134.0	X700	Sydney	11300	Burwood (A)
1	2020-01-25	2121.0	X760	Northern Sydney	16260	Parramatta (C)
2	2020-01-25	2071.0	X760	Northern Sydney	14500	Ku-ring-gai (A)
3	2020-01-27	2033.0	X720	South Eastern Sydney	16550	Randwick (C)
4	2020-03-01	2077.0	X760	Northern Sydney	14000	Hornsby (A)

Columna	notification_date	postcode	lga_code19	lhd_2010_code
Descripción	Representa la fecha en la que se encontro un caso de COVID-19	Codigo postal de cada Local Health District	Código del Local Goverment Area donde se detecto el caso	Código de 2010 que identifica el Local Health District
Tipo	Fecha	Número	Número	Texto
Formato/Medida	AA-MM-DD	Entero	Entero	Código Alfanumérico
Tipo de dato	Cuantitativo Discreto	Cualitativo Nominal	Cualitativo Nominal	Cualitativo Nominal
Minimo	25/01/2020	2000	10050	X700
Maximo	7/02/2022	2990	18710	X999
Valores repetidos	Contiene una gran cantidad de fechas repetidas, especialmente en los ultimos meses	Aunque es numérico, no tiene valor matemático directo. Es un código geográfico.	Contiene una gran cantidad de fechas repetidas, especialmente en los ultimos meses	Contiene una gran cantidad de fechas repetidas, especialmente en los ultimos meses
Valores unicos	No existen, ya que todos se repiten	No existen, ya que todos se repiten	No existen, ya que todos se repiten	No existen, ya que todos se repiten
Rango de valores	25/01/2020 a 7/02/2022	Son 764 valores distintos	Son 128 valores distintos	Son 15 valores distintos

Datos del censo de NSW, datos sociodemográficos y económicos a nivel de Local Government Areas (LGAs), extraídos del censo de 2016, con la intención de analizar características comunitarias y posiblemente correlacionarlas con los patrones de contagio.

	LGA_code	LGA_Name	LGA_Name_abbr	MedianAge	MedianMortgage	MedianPersonIncome	MedianRent	MedianFamilyIncome
0	LGA10050	Albury(C)	Albury	39	1421	642	231	1532
1	LGA10130	ArmidaleRegional(A)	Armidale.R.	36	1393	561	250	1465
2	LGA10250	Ballina(A)	Ballina	48	1733	601	340	1426
3	LGA10300	Balranald(A)	Balranald	41	950	624	150	1438
4	LGA10470	BathurstRegional(A)	Bathurst.R.	37	1670	646	280	1632

Columna	LGA_code	LGA_code, LGA_Name_abbr	MedianAge	MedianMortgage	MedianRent	MedianPersonIncome	OneMethodbyBus	LowIncome%
Descripción	Código dúnico del Local Goverment Area	Nombre del LGA	Edad media. Tiene valor numérico y se puede promediar	Representa la media de hipoteca a pagar	Representa la media de renta a pagar	Media de ingreso por familia	N° de personas que tienen como Modo único de transporte al	Porcentaje de personas de bajos ingresos y su proporción
Tipo	Número	Texto	Número	Número Entero	Número Entero	Número Entero	Número Entero	Número Decimal
Formato/Medida	Entero	Nombre	Entero	Dólar Australiano	Dólar Australiano	Dólar Australiano	Entero	Porcentaje
Tipo de dato	Cualitativo Nominal	Cualitativo Nominal	Cuantitativo discreto	Cuantitativo Discreto	Cuantitativo Discreto	Cuantitativo Discreto	Cuantitativo Discreto	Cuantitativo Continuo
Minimo	10050	No tiene	32	433	90	439	0	0.184522
Maximo	18710	No tiene	54	3200	650	1386	1386	0.522246
Valores repetidos	Todos los valores son únicos	Todos los valores son únicos	40, 42, 36, 44, 45, …	1300, 1083, 1733, 1517, …	150, 200, 180, 250, …	600, 538, 439, 524, …	3, 8, 7, 10, …	No hay valores repetidos
Valores unicos	Todos los valores son únicos	Todos los valores son únicos	54, 50, 52, 53	1421, 1393, 1670, 950, …	245, 90, 105, 100, …	719, 634, 688, 620, …	54, 52, 127, …	0.365752, 0.412482, …
Rango de valores	Son 128 valores distintos	Son 128 valores distintos	32-54	433-3200	90-650	433-3200	433-3200	0.184522-0.522246

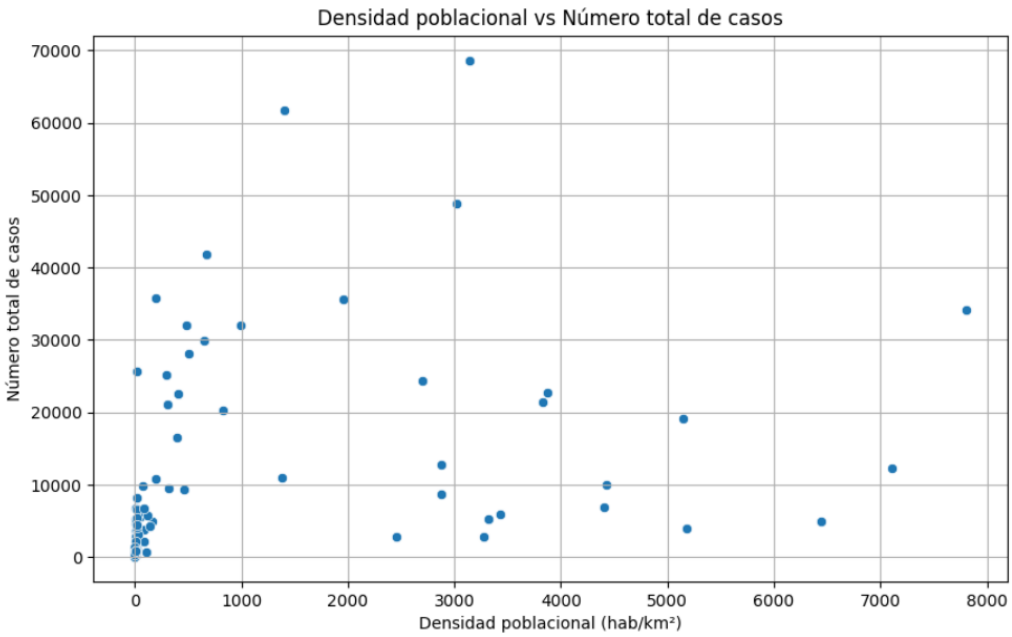
Hipótesis

1. **Las zonas geográficas con mayor densidad poblacional tienden a tener un mayor número de contagios y mayor tasa de velocidad de propagación**

La tabla de casos indica la ubicación geográfica en la cual fueron detectados, es decir el LGA, mientras que la segunda tabla contiene los datos para hallar la densidad poblacional a través de la cantidad de población y el área que ocupa el LGA.

- Densidad poblacional = Population / Area
- Número de contagios por LGA = agrupar por LGA desde tu dataset de COVID-19 (notification_date, lga_code19, etc.)
- Velocidad de propagación = evolución de casos en el tiempo por LGA (requiere notification_date bien estructurada y múltiples fechas).

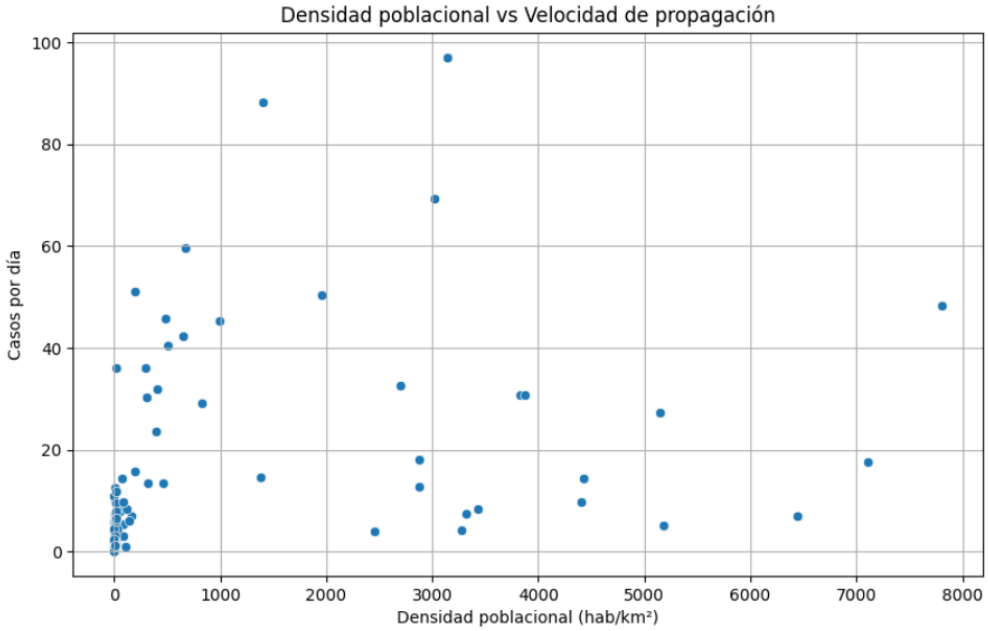
El primer caso, plantea evaluar a ver si mayor densidad poblacional → mayor número total de casos.



El gráfico nos muestra cierta correlación entre la cantidad de casos por LGA y la densidad poblacional que esta presenta, de la cual se puede afirmar dos cosas

- En su mayoría, los LGA de baja densidad poblacional, también presentan una baja cantidad de casos de Covid-19
- En el resto de casos, la variación por casos contra los LGA con mayor densidad poblacional, tiene una baja correlación, ya que tiene casos de LGAs con alta densidad pero poca variación de casos, y lugares con poca densidad y alta variación de casos.

El segundo caso es analizar la densidad poblacional contra la velocidad de propagación, para ello se debe primero Agrupar por LGA y fecha, luego calcular el número de días entre el primer y último caso, y por último dividir el número total de casos por ese número de días.



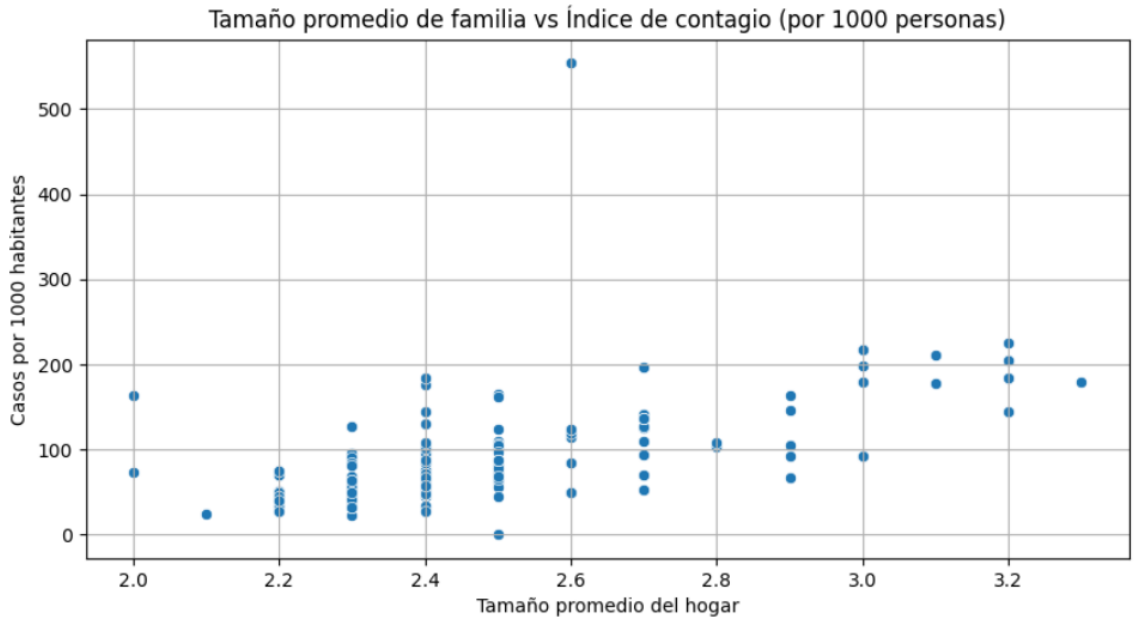
El gráfico nos muestra cierta correlación entre la densidad poblacional y la velocidad de propagación medida en la cantidad de casos por día, y si uno incluso lo compara con el anterior gráfico de densidad contra cantidad de casos, podremos ver que los gráficos son bastante similares, y por ello podemos concluir que:

- La correlación entre la densidad poblacional y la velocidad de propagación es baja, ya que hay varios casos de LGA con alta densidad poblacional pero baja velocidad de propagación y al revés, LGAs con aba densidad poblacional y alta velocidad de propagación.

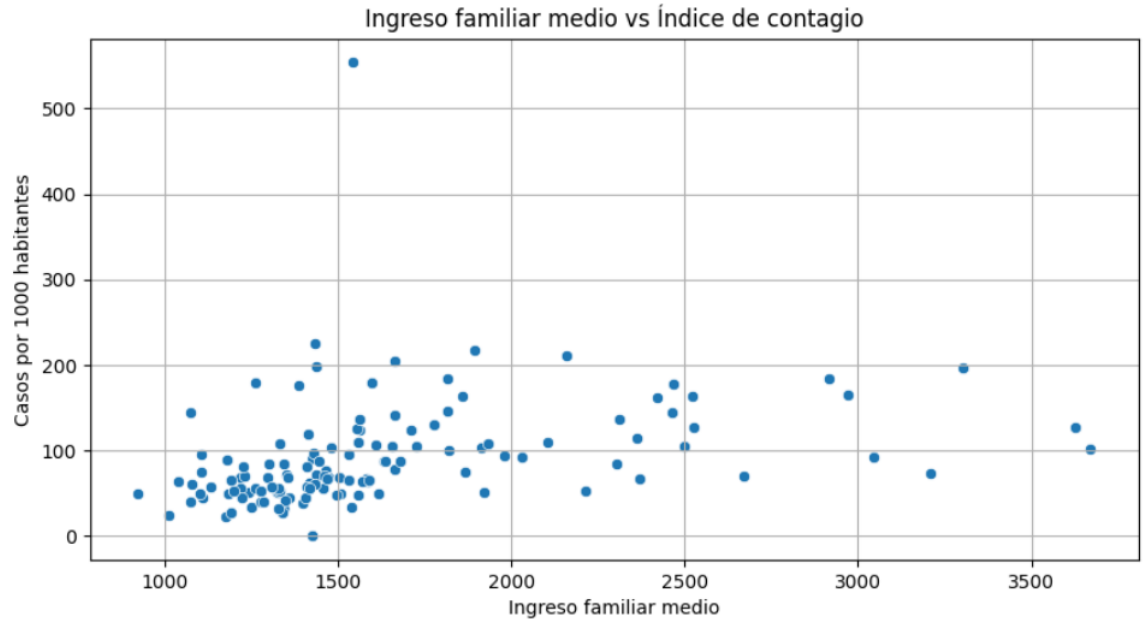
2. **El índice de contagio es mayor en familias con bajos índices económicos y mayor número de miembros por familia**

Para poder comprobar esta Hipótesis es necesario saber:

- Índice de contagio por zona → número de casos de COVID por población total (o por 1000 habitantes).
- Ingresos medios o nivel socioeconómico por zona → alguna medida como "household income", "socioeconomic index", "SEIFA".
- Tamaño promedio de las familias → usualmente como average household size o número de personas por vivienda.



El gráfico nos muestra la relación entre el tamaño promedio del hogar (cantidad de habitantes) y su relación con la cantidad de casos por cada 1000 habitantes, con lo cual se puede ver una tendencia a un mayor número de contagios cuando se tiene un mayor promedio de personas por hogar (tamaño del hogar)



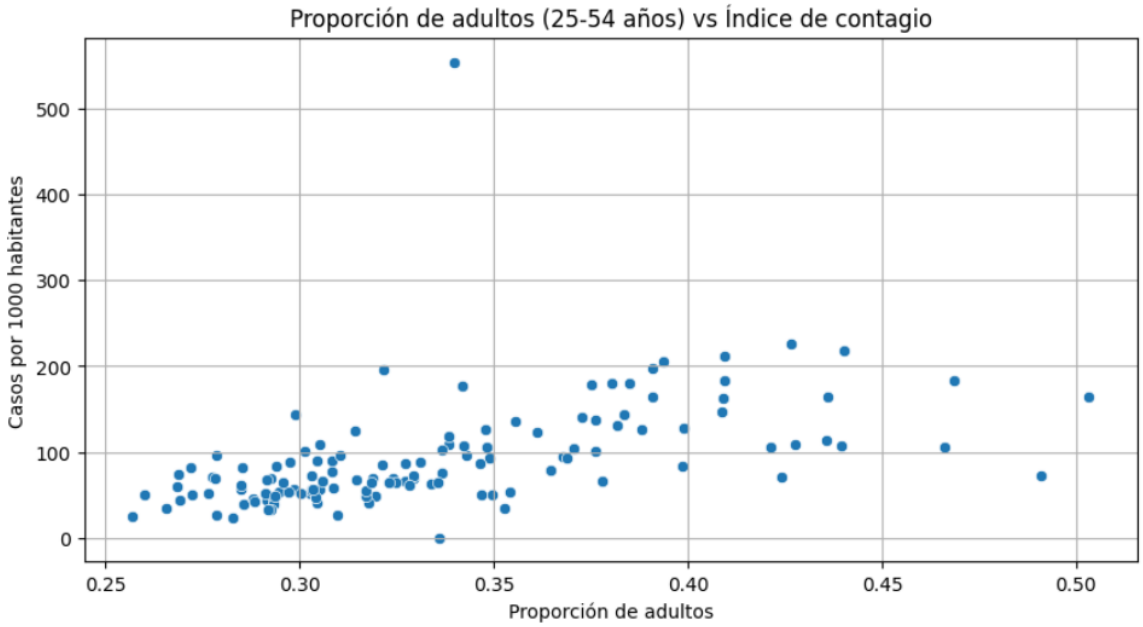
El segundo gráfico nos muestra la relación entre el ingreso familiar medio y la cantidad de casos por cada 1000 habitantes, lo cual nos da una gran sorpresa, viendo que el ingreso familiar medio no afecta ni tiene una alta correlación con la reducción de contagios.

3. **Los casos de contagio están más concentrados en ciertos grupos de edades (ej. adultos jóvenes vs. adultos mayores)**

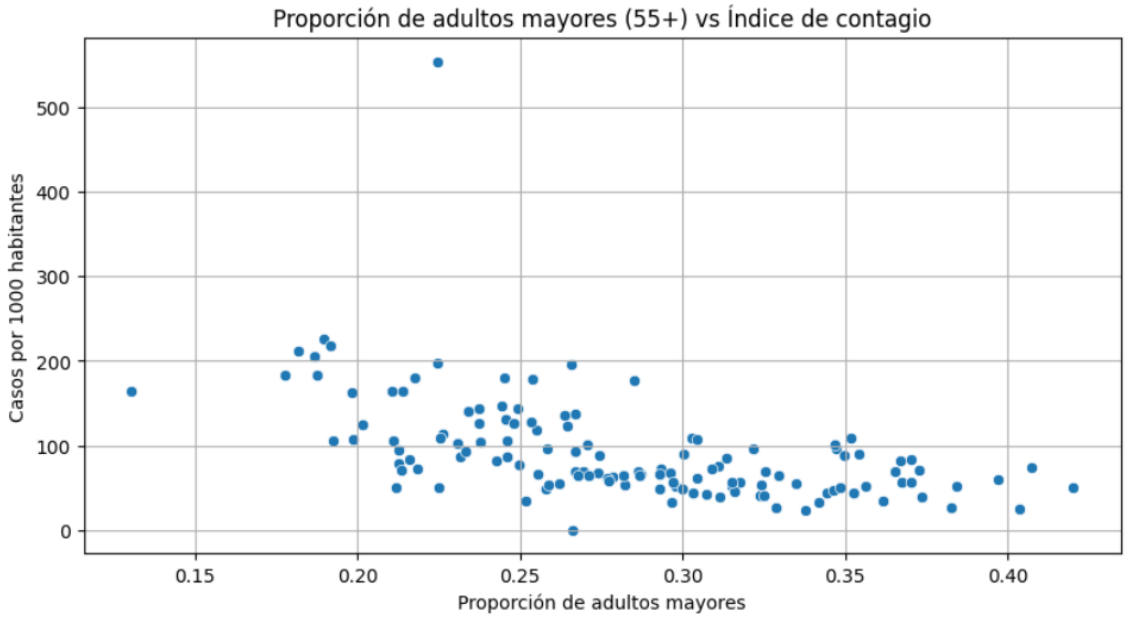
El objetivo de esta Hipótesis era analizar si LGAs con más población joven (o adulta mayor) tienen mayor tasa de contagios.

Entonces se procederá a agrupar por edades, donde:

- Adultos jóvenes y medios (25-34, 35-44, 45-54)
- Adultos mayores (55-64, 65-74, 75-84, 85+)



El primer gráfico, nos permite entender que LGAs con mayor cantidad de población adulta, tienden a tener una mayor cantidad de casos de Covid-19, ello en parte debido a que son la población económicamente activa, en comparación con los otros grupos de edades.



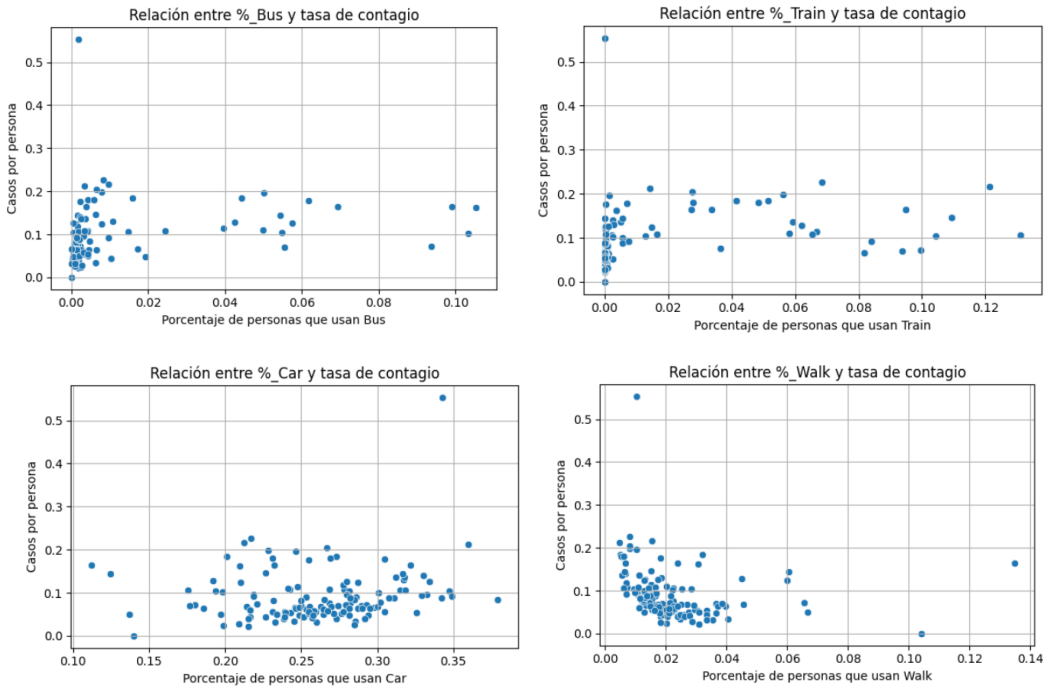
El segundo gráfico nos muestra la relación entre la cantidad de adultos mayores y la cantidad de casos Covid-19, observando que a mayor población de adultos mayores es menor la tasa de contagios, esto debido a que en su mayoría ya no pertenecen a la población económicamente activa, y, por tanto tienen menor riesgo de contagio al permanecer en sus casas.

También se puede corroborar ello, haciendo una revisión de la correlación que existe de los grupos de edad con la cantidad de casos, obteniendo:

- Correlación jóvenes: 0.432
- Correlación adultos: 0.504
- Correlación mayores: -0.524

4. **La forma de movilidad influye en gran medida sobre los contagios, siendo los buses los puntos de inflexión**

Dentro del dataset de datos sociodemográficos tenemos campos que indican los porcentajes de personas que toman ciertos medios de transporte (OneMethodbyBus, OneMethodbyTrain, OneMethodbyCarasDriver, etc), y son estos los que se usaran para verificar que tanta influencia tiene el medio de transporte predominante por sobre la cantidad de casos.



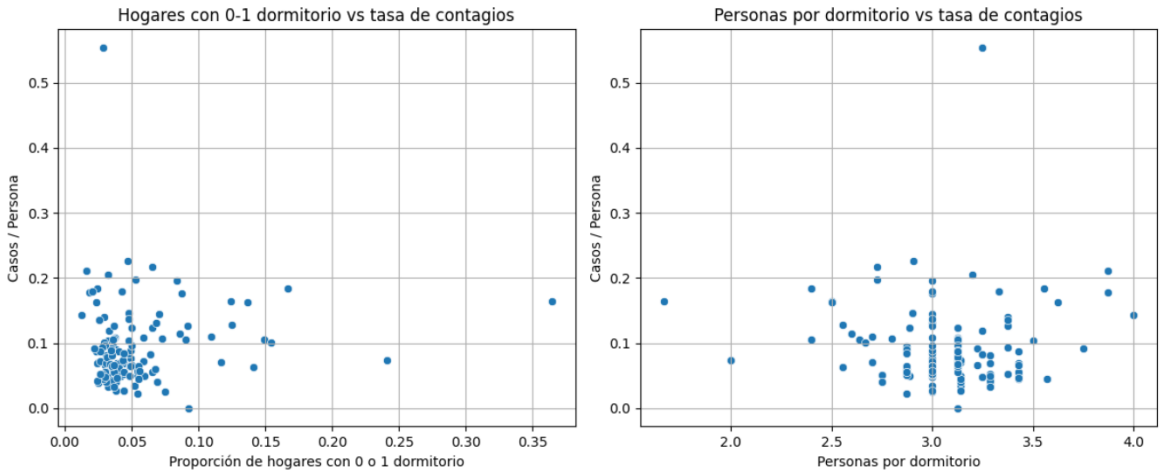
Los gráficos muestran el nivel de correlación entre el porcentaje de personas en un LGA que utilizan cierto medio de transporte

- En el caso de bus, se puede cierta correlación, aunque baja, sobre el número de casos por persona en contraparte con el porcentaje de personas que usan el bus
- En el caso de tren, sucede algo similar al casos de buses, presenta una baja correlación, a pesar de que ambos son puntos de agrupación de personas y por tanto puntos de infección, pero al parecer no lo son tanto.
- En el caso de el uso de automóvil, del cual él gráfico nos indica que la mayor cantidad de personas se movilizan de esta manera, tiene de igual forma una baja correlación con la cantidad de casos.
- Y en el caso de las personas que se movilizan a pie, podemos ver que tiene una correlación negativa, es decir que se presenta menor cantidad de casos de Covid-19 sobre LGAs con mayor porcentaje de personas que se movilizan a pie.

5. **Los LGAs con más hogares compartidos o menor cantidad de dormitorios presentan más contagios.**

La idea es que el hacinamiento o falta de espacio personal puede facilitar la propagación del COVID-19. Todo ello se puede ver gracias a los campos de:

- Cantidad de dormitorios por hogar (NoneBedroom, 1Bedroom, …, 5Bedrooms)
- Personas por hogar (AverageHouseSize)
- Personas por habitación (AverageHouseSize / AverageBedroom)



El primer gráfico nos muestra la correlación entre la proporción de hogares con 0 o 1 dormitorios, los cuales están orientados a personas que viven solas, de la cual se puede observar que no se tiene una tendencia para saber si la cantidad de casos varia según la el porcentaje de hogares de 1 dormitorio en un LGA

En el segundo gráfico por otra parte se puede ver una tendencia positiva a tener un mayor numero de casos por persona en hogares con mayor número de personas compartiendo un dormitorio.

Conclusiones

Los factores sociales, económicos y estructurales tienen una influencia clara en la propagación del COVID-19 a nivel local. La densidad poblacional, la edad, el hacinamiento en viviendas y el tipo de transporte son variables que, al combinarse, pueden agravar los contagios en comunidades vulnerables.

Observamos una correlación positiva entre la densidad de población y la tasa de contagios. Esto sugiere que en LGAs más densamente poblados hay una mayor exposición al virus, posiblemente por mayor interacción social y menor posibilidad de distanciamiento físico.

Aunque no tuvimos acceso directo a datos de PEA, sí analizamos por edades. Los grupos de edad más activos laboralmente (25-44 años) están más expuestos y presentan mayores tasas de contagio. Por el contrario, LGAs con mayor porcentaje de adultos mayores (65+) presentan menos casos en proporción, con lo cual se concluyó que los adultos jóvenes pueden tener un mayor riesgo de exposición, posiblemente por razones laborales o sociales.

Además, se encontró una correlación positiva moderada entre el uso del bus para ir al trabajo y la tasa de contagios. El uso del auto mostró una correlación negativa (o muy baja) con contagios. El uso del tren también se relaciona positivamente, pero no tanto como el bus. El uso de transporte público, especialmente el bus, podría haber facilitado mayores contagios por el contacto cercano entre personas. Esto refuerza indirectamente la idea de que la actividad económica y movilidad laboral aumentan la exposición al virus.

También respecto al hacinamiento, LGAs con mayor proporción de hogares con 0 o 1 dormitorio y más personas por dormitorio muestran tendencia a tener más contagios. Las correlaciones apoyan la hipótesis de que el hacinamiento aumenta el riesgo de transmisión intradomiciliaria. En pocas palabras, el acceso desigual a vivienda digna puede ser un factor estructural que influye en la propagación del virus.

Todos estos aspectos sociodemográficos tienen una mayor o menor influencia de la cantidad de casos totales, por cada 1000 habitantes o por personas, la tasa de contagios y la velocidad de propagación, es allí que se fundamente la importancia de analizarlos para saber como tomar decisiones que pueda solucionar estos problemas en tiempos de pandemia.