UNIVERSIDAD NACIONAL DE SAN AGUSTÍN DE AREQUIPA

ESCUELA PROFESIONAL DE CIENCIA DE LA COMPUTACIÓN

CIENCIA DE DATOS

INFORME DEL ANALISIS EXPLORATORIO DE DATOS

PRESENTADO POR:

John Edson Sanchez Chilo

AREQUIPA – PERÚ

2024

Análisis Exploratorio de Datos sobre Casos Covid-19 en NSW, y datos sociodemográficos por LGA

Plan de Análisis

1. **Motivación y Contexto**
   La enfermedad por coronavirus 2019 (COVID-19), causada por el virus SARS-CoV-2, emergió a finales de 2019 en Wuhan, China, y rápidamente se convirtió en una pandemia global que afectó todos los aspectos de la vida humana. Su rápida propagación y la falta de preparación para afrontar la pandemia por parte de los gobiernos y las autoridades pusieron a prueba los sistemas de salud, economía y la sociedad en su conjunto, afectando especialmente a ciertos sectores de la población.

   La pandemia, evidenció las profundas desigualdades estructurales presentes en muchas regiones. Las poblaciones que ya enfrentaban condiciones sociales y económicas precarias —como acceso limitado a servicios de salud, empleo, mayor densidad poblacional, o mayor número de personas de edad avanzada— fueron particularmente vulnerables a una mayor propagación del virus. En muchos casos, la ausencia de información clara sobre qué factores sociodemográficos influían con mayor peso en la dinámica de contagio dificultó la toma de decisiones efectivas por parte de las autoridades. Esta falta de conocimiento impidió aplicar medidas focalizadas que protegieran a los sectores más expuestos, limitando así la capacidad de respuesta y aumentando el impacto del virus en comunidades ya desfavorecidas.

2. **Preguntas de Hipótesis**
   - ¿Las zonas geográficas con mayor densidad poblacional tienden a tener un mayor número de contagios y mayor tasa de velocidad de propagación?
   - ¿El índice de contagio es mayor en familias con bajos índices económicos y mayor número de miembros por familia?
   - ¿Los casos de contagio están más concentrados en ciertos grupos DE EDADES (ej. adultos jóvenes vs. adultos mayores)?
   - ¿La forma de movilidad influye en gran medida sobre los contagios, siendo los buses los puntos de inflexión?
   - ¿Los LGAs con más hogares compartidos o menor cantidad de dormitorios presentan más contagios?

3. **Objetivos del Análisis**
   La revisión de donde es que se recolectaron los datos es esencial para saber el contexto de los mismos, entender a detalle sus valores y lo que representan, y entender a detalle los datos para saber cómo realizar la manipulación de los mismos. Por lo que, para investigar las Hipótesis propuestas, se tuvieron en consideración algunos pasos que seguiremos en el informe.

3.1. Análisis de Valores

- Revisar los valores usando en los conjuntos de datos, sus tipos, y analizar sus valores a partir de medidas estadísticas
- Identificar y examinar los valores que se desvían significativamente de la tendencia general de los datos.

3.2. Evaluación de la calidad de los datos

- Revisar la consistencia de los datos, verificando si hay errores de entrada o inconsistencias en la estructura de los registros
- Examinar la integridad de los datos en términos de valores faltantes o inconsistencias temporales.
- Considerar la precisión de los instrumentos de medición y la metodología de recolección de datos para evaluar la fiabilidad de los registros.

3.3. Análisis de correlación de los datos

- Calcular coeficientes de correlación para explorar las relaciones entre las variables de interés, como la concentración de contaminantes atmosféricos, la humedad, la temperatura y la velocidad del viento
- Realizar análisis gráficos, como diagramas de dispersión, para visualizar las relaciones entre las variables y determinar la fuerza y dirección de la correlación.

3.4. Interpretación de resultados

- Evaluar los hallazgos obtenidos en cada paso del análisis en función de las Hipótesis planteadas.
- Identificar patrones significativos, relaciones causales o tendencias emergentes que respalden o refuten las Hipótesis formuladas.

2. 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/

El objetivo principal de las autoridades y los entes sanitarios era recolectar información de la transmisión del virus a nivel comunitario en conjunción con factores sociodemográficos, para entender mejor los patrones de propagación, riesgos asociados y las intervenciones gubernamentales. La idea era explorar cómo los datos epidemiológicos, geo-localización, intervenciones y características sociales interactúan en la dinámica de la pandemia.

1. **Datos de casos de COVID-19 en NSW**

El archivo contiene notificaciones de casos de COVID-19 en el estado de Nueva Gales del Sur en Australia, durante un período de aproximadamente dos años, desde el 25 de enero del 2020 al 07 de febrero de 2022. Cada registro representa una notificación sobre un nuevo caso en una fecha y localidad específica de NSW, ingresando el Local Health District (LHD) como región y adicionalmente el Local Goverment Area (LGA) como subdivisión.

Es importante aclarar que Australia se divide en estados, cada estado se subdivide en distritos de salud local (LHD), y cada distrito se divide en áreas de gobierno local (LGA). Los datos usados solo pertenecen al estado de Nueva Gales del Sur (NSW)

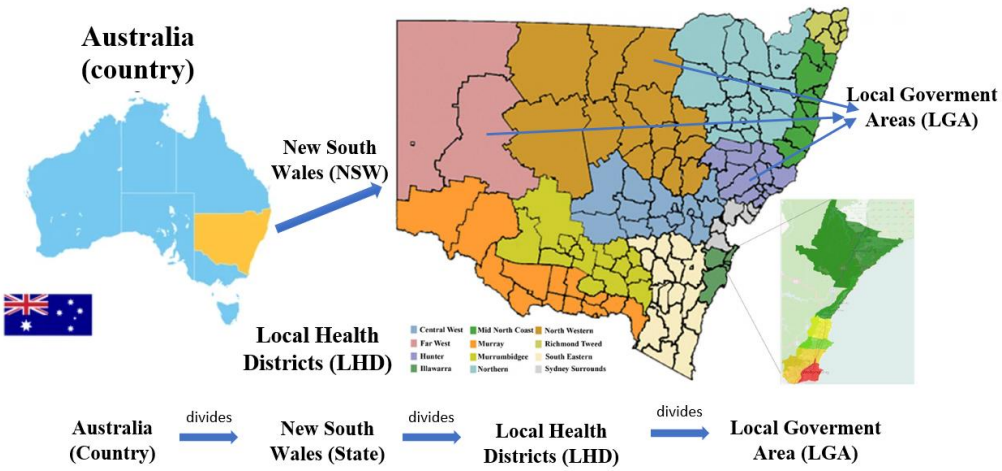


Figura 1. Mapa de subdivisiones de Australia

	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)

Figura 2. Cada registro se lee como, “Caso de Covid-19 el día 2020-01-25 en el distrito de Sydney (x700), con código postal 2134, en el área de Burwood con código 11300”.

Definición de columnas

notification_date: Fecha específica en la que se notificó sobre un caso de COVID-19

postcode: es el código postal que se utiliza para identificar el Local Health District

lhd_2010_code: Código utilizado para identificar al Local Health District (LHD)

lhd_2010_name: Nombre del Local Health District

lga_code19: Código de identificación del Local Goverment Area (LGA)

lga_name19: Nombre del LGA correspondiente

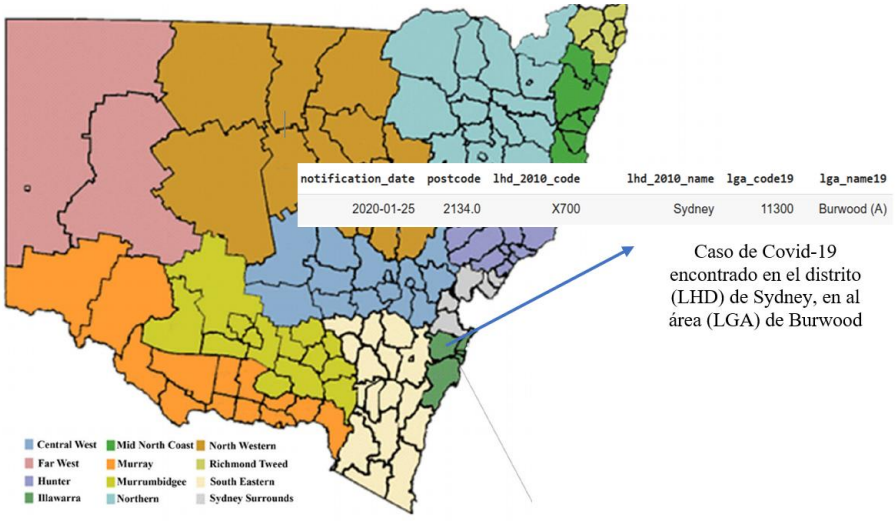


Figura 3. Representación de casos por Covid-19 por cada registro

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

Tabla 1. Notificaciones de Covid-19 en NSW

2. **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. La versión del censo de 2020 aún no estaba disponible en el momento de la investigación, por lo que se usaron los datos del censo de 2016.

	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

Figura 4. Tabla de datos sociodemográficos por LGA

De esta tabla se tiene un mayor número de columnas, por lo cual solo se colocará una descripción de los campos más resaltantes.

Columnas:

LGA_code: Código dúnico del Local Goverment Area

LGA_code, LGA_Name_abbr: Nombre del LGA

MedianAge: Edad media. Tiene valor numérico y se puede promediar.

MedianMortgage: Representa la media de hipoteca a pagar

MedianRent: Representa la media de renta a pagar

MedianPersonIncome: Media de ingreso por familia

OneMethodbyBus: N° de personas que tienen como Modo único de transporte al Bus

LowIncome%: Porcentaje de personas de bajos ingresos y su proporción.

Otras columnas más.

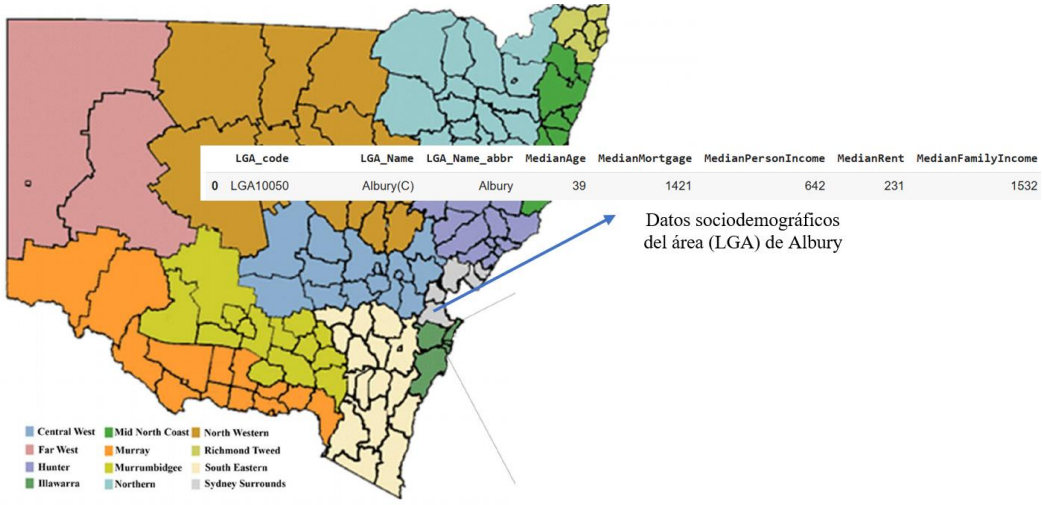


Figura 5. Representación de casos por Covid-19 por cada registro

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

Tabla 2. Datos sociodemográficos por LGA de NSW

Analisis Exploratorio de Datos

Análisis del comportamiento de los datos

Manipulación y transformación de datos del archivo
Se procedera con el análisis del archivo que contiene los datos a utilizar sobre los casos COVID por LGA y los datos sociodemográficos de cada LGA

- **Formato de los archivos:**

  Casos Covid-19 en NSW: El archivo se encuentra en formato CSV, es un formato ligero de intercambio de datos.

  Datos sociodemográficos por LGA: El archivo se encuentra en formato CSV, es un formato ligero de intercambio de datos, en el cual se separa la información por comas.

- **Encoding del archivo**
  Se revisa el encoding del archivo para saber en que formato se leerá posteriormente el archivo, en este caso el resultado del encoding es ascii, por lo cual enconding como UTF-8 también pueden ser usados como encoding para este archivo.

- **Tamaño del archivo**
  El archivo sobre casos tiene un peso de 61.76MB, con lo cual se puede decir que es un archivo de tamaño mediano, y por lo tanto se trabajará con el de manera directa, es decir no hay necesidad de trabajarlo por partes.
  El segundo archivo por otra parte, pesa solo 0.07MB, con lo cual se puede decir que es un archivo pequeño y se trabajará con el de forma directa
- **Transformación de datos**
  En los archivos CSV, se presentan dos tipos de datos a nivel granular, el primero son los casos de COVID por semana vistos en cada LGA, y el segundo son los datos particulares y sociodemográficos de cada LGA Entonces se procedera a manejar dos

tablas separadas que representaran nivel granular los datos sociodemográficos de cada LGA y los casos de cada LGA.

Preguntas:

Un registro es una entidad, describa que representa un registro
En la primera tabla cada registro representa una notificacion de un caso de COVID-19 en una localidad específica.
En la segunda tabla cada registro representa un LGA y sus diferentes datos sociodemográficos

¿Cuantos registros hay?
En la primera tabla de casos existen un total de 973412 registros.
En la segunda tabla de LGA's existen un total de 129 registros, esto debido a que aunque en NSW solo se considera 128 LGA's, NSW también contiene zonas no incorporadas que aunque estan dentro de su juridiscción no pertenecen a ningun LGA.

¿Son demasiado pocos?
No, son una cantidad adecuada para poder hacer el analisis de los datos de forma correcta

¿Son muchos y no tenemos Capacidad (CPU+RAM) suficiente para procesarlo?
No, en el caso de la primera tabla de casos, esta ocupa un total de 316.9MB de memoria RAM por lo cual esta dentro de las capacidades de Google Colab, y en el caso de la segunda tabla esta ocupa solo un 0.14MB de RAM, lo cual es bastante bajo, y por lo tanto se puede trabajar con ambas tablas sin complicaciones.

¿Hay datos duplicados?
Si en la primera tabla de casos, existen bastantes registros repetidos, lo cual es algo normal en esta tabla, ya que cada día en la misma localidad se puede detectar más de un caso de COVID-19.
No en la segunda tabla, ya que cada registro representa un LGA diferente, e incluso haciendo una busqueda de repeticiones por código de LGA o nombre de LGA, no se encuentran repeticiones, lo cual es correcto.

2. Tipo de Datos

Se procederá a revisar cada una de las columnas de ambas tablas para identificar sus tipo de datos, asi mismo también se hara un análisis para encontrar valores máximos, minimos, repetidos, nulo, etc.

¿Cuales son los tipos de datos de cada columnas?
Tabla de Casos de COVID-19:

- notification_date: Fecha
- postcode: Número entero positivo
- lhd_2010_code: texto
- lhd_2010_name: texto

- lga_code19: Número entero positivo
- lga_name19: texto

Tabla de LGA's

- LGA_code: Número entero positivo
- LGA_Name: texto
- MedianAge: Número entero positivo
- MedianMortgage: Número entero positivo

¿Entre qué rangos están los datos de cada columna?, valores únicos, min, max?

Tabla de casos de COVID-19 en NSW

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

Tabla de datos sociodemográficos por LGA

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

¿Todos los datos están en su formato adecuado?

Si todos los datos excepto por la columna "post_code" de la tabla de casos, esto debido a que en el caso de valores nulos de los que no se sabe su localidad, en lugar de colocar un valor vacío (que representa nulo), escribe literalmente el "None"

¿Los datos tienen diferentes unidades de medida?
Si, cada uno de los datos especialmente en la data sociodemografica tiene distintas unidades de medida, por ejemplo en el caso de ingresos esta en dolares australianos, en el caso de conteos de personas esta en un valor entero de personas, en el caso de algunos valores particulares, como habitaciones por persona y demas, estan en su respectiva medida.

¿Cuáles son los datos categóricos, ¿hay necesidad de convertirlos en numéricos? Ninguna de las dos tablas cuenta con datos categóricos, por lo cual no hay necesidad de convertir nada.

3. Granularidad

¿Qué representa un registro? Describe qué representa cada fila
En la primera tabla, cada registro hace referencia a una notificación de un caso de COVID en una localidad(LGA) y fecha específica.
En la segunda tabla cada registro representa un LGA y sus factores sociodemográficos.

Si es una data etiquetada, ¿como interpretas la información de las clases?
Ninguna de las dos tablas es etiquetada

¿Hay niveles de granularidad de los datos? Por ejemplo, datos a nivel país, región, ciudad. Años, meses, días, horas, minutos, etc
Si, en la primera tabla de casos, se sabe que todos pertenecen al país de Autralia en le región de Nueva Gales del Sur, y la data específica que LHD (Local Health District) e internamente des este especiica que LGA(Local Goverment Area), aparte de también mencionar la fecha del caso de COVID-19.

4. Limpieza

¿Están todas las filas completas o tenemos campos con valores nulos?
En el caso de la primera tabla, si hay valores NULOS, hay algunos registros, de las cuales solo tenemos la fecha en la que se dió, mas no tenemos el resto de valores que determinan la localidad.
Se hizo una revisión del porcentaje de valores nulos, y se halló que era de 1.54%, por lo cual no es un cantidad significativa o que afecte a los datos, así que se tomó la decisión de eliminarlos del dataset.
En el caso de la segunda tabla, no había ningún valor NULO, por lo cual no había necesidad de ello.

5. Transformaciones

Para poder usar los datos de casos por Covid-19, se tuvo que armar resúmenes acerca de la cantidad de casos totales por LGA, los casos por persona por LGA, los casos por cada 100000 personas por cada LGA, ello con el fin de tener un buen entendimiento de como afectaron a los casos por cada LGA.

La segunda transformación es sobre los casos por LGA en ciertos intervalos de tiempo, en este caso por meses, de manera que se tenga los casos y los casos acumulados por LGA.

```
lga_code19,year_month,MonthlyCases,CumulativeCases
11300,2020-01,1,1
11300,2020-02,0,1
11300,2020-03,5,6
```

Figura. Transformación de casos por Covid-19 a tabla de evolución de casos por LGA en el tiempo

6. Análisis de medidas de tendencia

¿Siguen alguna distribución?
Al aplicar describe() y analizar la columna de fechas en la tabla de casos COVID, se observa una distribución temporal clara. Inicialmente, hay pocos casos registrados y a medida que pasan los días, la cantidad de casos aumenta, lo cual es característico de un brote epidémico. Esto muestra una distribución acumulativa a lo largo del tiempo.

En la tabla socioeconómica (socio_df), al aplicar describe() en columnas como MedianPersonIncome y AverageHouseSize, se observa que los valores varían ampliamente, lo que sugiere que la distribución no es uniforme. Estos campos podrían seguir distribuciones sesgadas o tener outliers, lo que se confirmaría mejor con histogramas.

Medidas de tendencia central: media aritmética, geométrica, armónica, mediana, moda, desviación estándar.

Tabla 1: Casos COVID
Esta tabla tiene principalmente columnas de tipo texto, pero algunas como postcode pueden ser numéricas. Sin embargo:

- notification_date es una columna de tipo fecha, por lo cual solo se le puede aplicar la moda
- postcode es un código categórico, no una cantidad con sentido aritmético → solo se puede aplicar la moda (para saber cuál es el más frecuente).
- Otras columnas como lga_name19, lhd_2010_code, etc., son también categóricas → moda únicamente.
- No hay columnas numéricas reales como cantidad de casos, edades o ingresos.

Tabla 2: Datos Socioeconómicos

Esta sí tiene campos numéricos reales, como ingresos, tamaño de casa, dormitorios, etc., por lo que se pueden aplicar todas las medidas de tendencia. Sin embargo ya que son bastantes campos, solo se ha elegido algunos como campos representativos.

Selección de columnas representativas por categoría Demográficos (edades, población, sexo):

MedianAge, Male, Female, Population, TotalMale(allages), TotalFemale(allages)

Ingresos (individuales, familiares, por hogar):

MedianPersonIncome, MedianFamilyIncome, MedianHouseholdIncome, (Se omiten los rangos semanales, ya que son muchos y difíciles de analizar individualmente sin más contexto)

Vivienda:

MedianMortgage, MedianRent, AverageBedroom, AverageHouseSize, Área (km², sirve para ver densidad poblacional si se cruza con población)

Otros relevantes:

LowIncomePersons, LowIncome%, LonePerson, LonePerson%, PercentofPublicTransportation

Correlación y covarianza: permite entender la relación entre dos variables aleatorias

Tabla 1 de Casos

- No hay fuertes correlaciones útiles entre las variables numéricas excepto la esperada entre year y month.
- Podrías considerar que la mayoría de estas variables son independientes entre sí desde el punto de vista lineal.

Tabla 2 sobre datos sociodemográficos
MedianAge tiene correlación negativa con muchas variables: lo cual sugiere que zonas con población más joven tienden a:

- Tener mayores ingresos familiares.
- Tener hogares más grandes (más personas por hogar).
- Tener más población total.

AverageHouseSize vs Population: 0.55

- Lugares con más población tienden a tener casas con más personas en promedio.

Population vs Area: -0.31

- Zonas más pobladas suelen tener menor área → podrían ser zonas urbanas densas.

Análisis de outliers

¿Cuáles son los Outliers? (unos pocos datos aislados que difieren drásticamente del resto y "contaminan" ó desvían las distribuciones)

Tabla 1 de Casos de COVID-19
year: Tiene un 21.97% de valores outlier, y todos los valores son 2020 y 2021. Esto sugiere

que el resto de la tabla contiene datos de 2020, los cuales a pesar de ser unos pocos meses, al presentar mayor cantida de casos, contiene casi el 80% de datos restantes.

Tabla 2 de Datos Sociodemográficos

- Variables económicas (como MedianPersonIncome, MedianRent, MedianFamilyIncome) presentan entre 2% y 10% de valores atípicos. Esto puede indicar:

  Ciudades con ingresos o rentas muy altos o bajos.

  Posibles errores de captura si los valores son inesperadamente extremos.
- AverageBedroom: 44.19% de valores outliers, principalmente con valores como 0.7, 0.9, 1.1, 1.2, etc. Posiblemente estas son zonas con unidades pequeñas (tipo estudio), lo cual es inusual comparado con el resto.
- Demográficos (Male, Female, Population, edades): Muchos valores considerados outliers, alrededor del 10–15%, lo cual indica:

  Municipios con poblaciones atípicamente grandes o pequeñas.

  Podría tener sentido revisar si estas localidades tienen una demografía muy distinta al promedio.

Exploración y Visualización

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

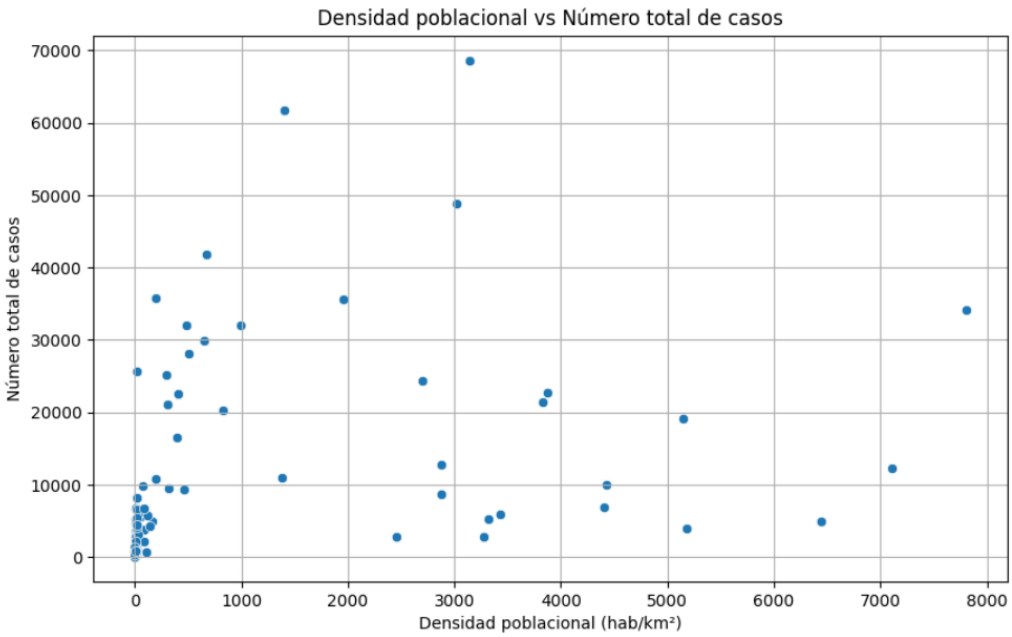
La tabla de casos indica la ubicación geográfica en la cual fueron detectados, es decir el LGA, mientras que la segunda tabla contiene los datos para hallar la densidad poblacional a través de la cantidad de población y el área que ocupa el LGA.

- 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.

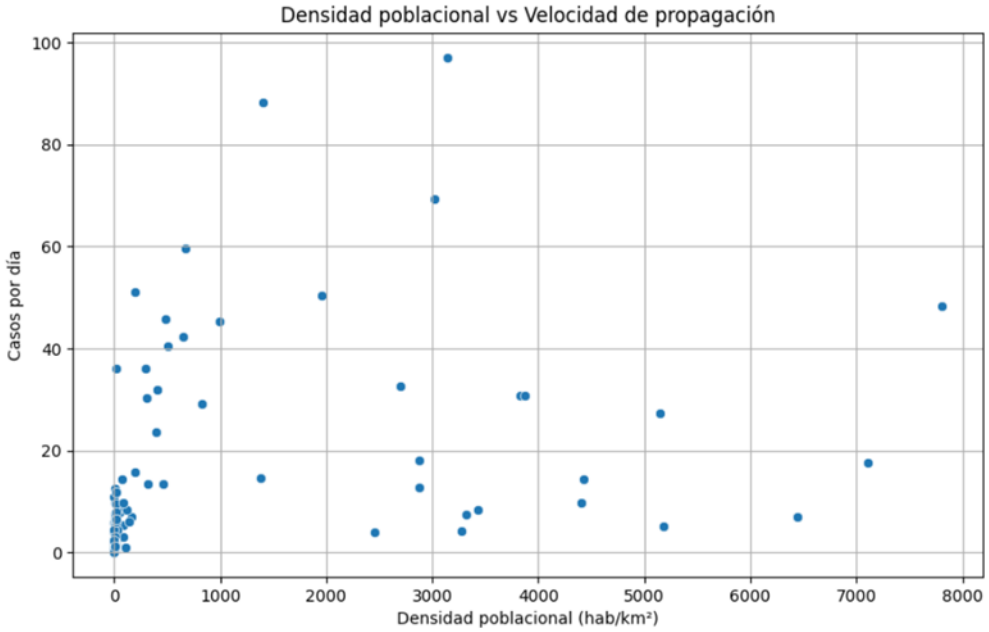


El gráfico nos muestra cierta correlación entre la cantidad de casos por LGA y la densidad poblacional que esta presenta, de la cual se puede afirmar dos cosas

- 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.

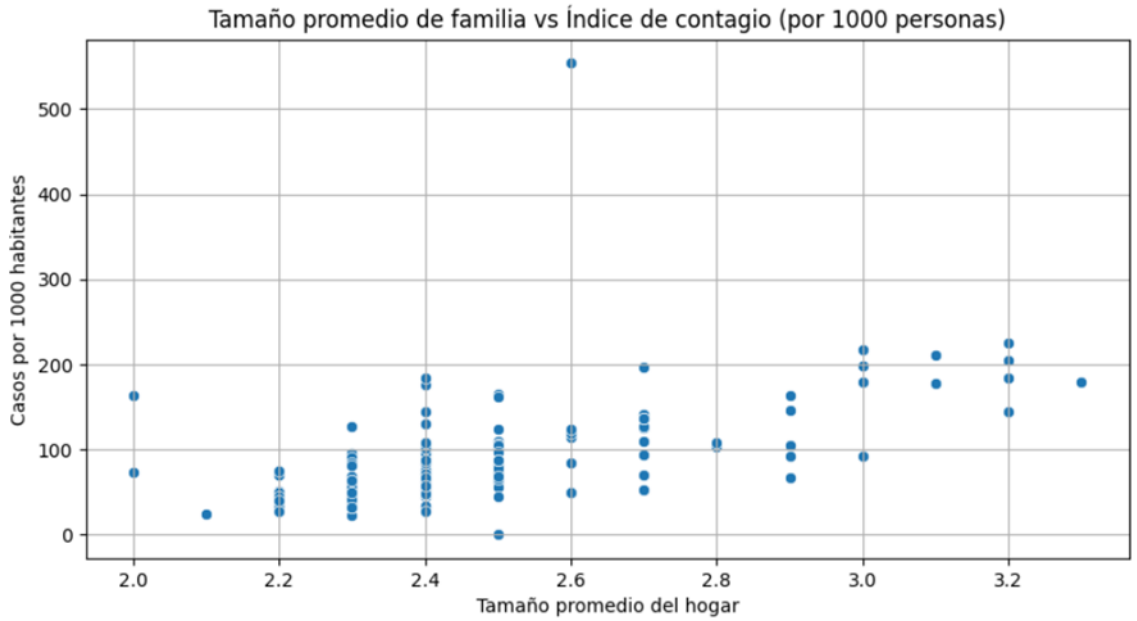
2. **El índice de contagio es mayor en familias con bajos índices económicos y mayor número de miembros por familia**

Para poder comprobar esta Hipótesis es necesario saber:

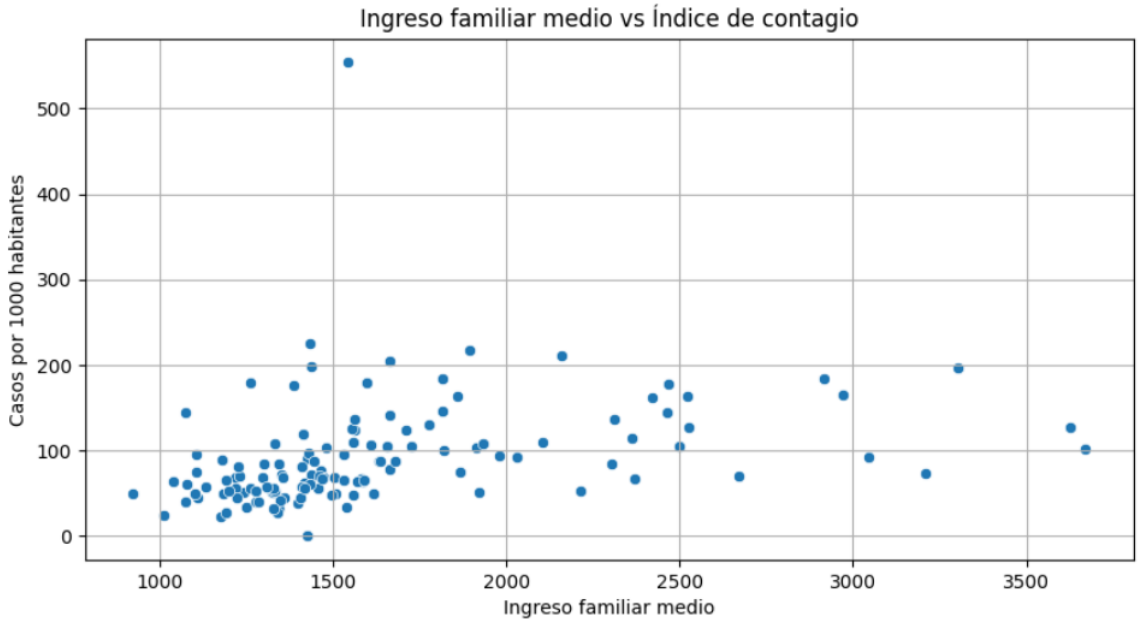
- Índice de contagio por zona → número de casos de COVID por población total (o por 1000 habitantes).

- 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)



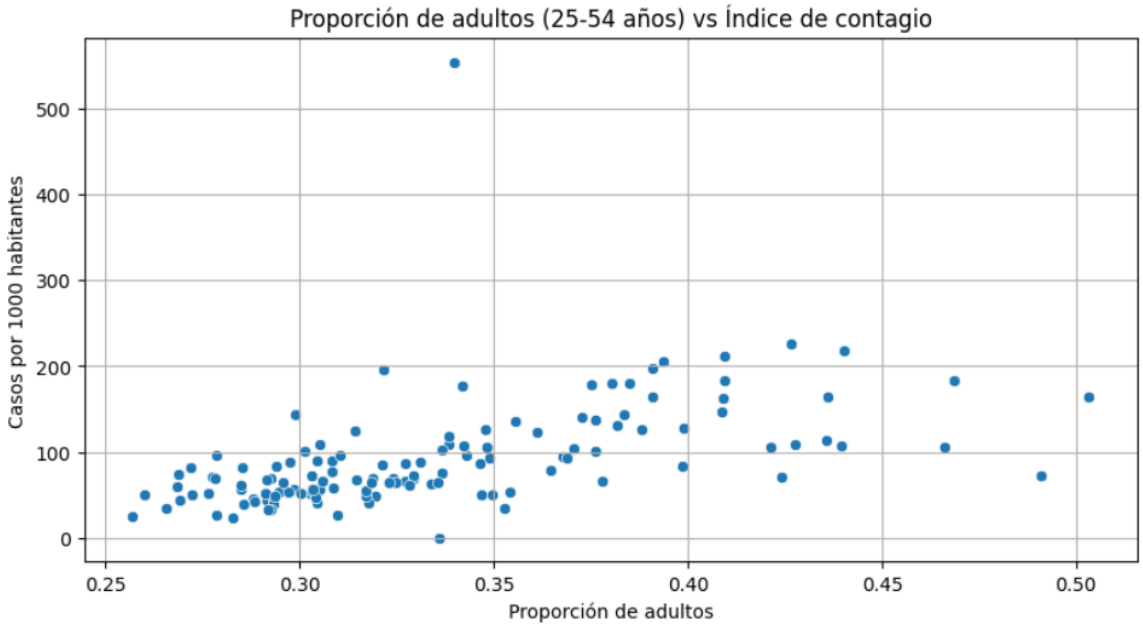
El segundo gráfico nos muestra la relación entre el ingreso familiar medio y la cantidad de casos por cada 1000 habitantes, lo cual nos da una gran sorpresa, viendo que el ingreso familiar medio no afecta ni tiene una alta correlación con la reducción de contagios.

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

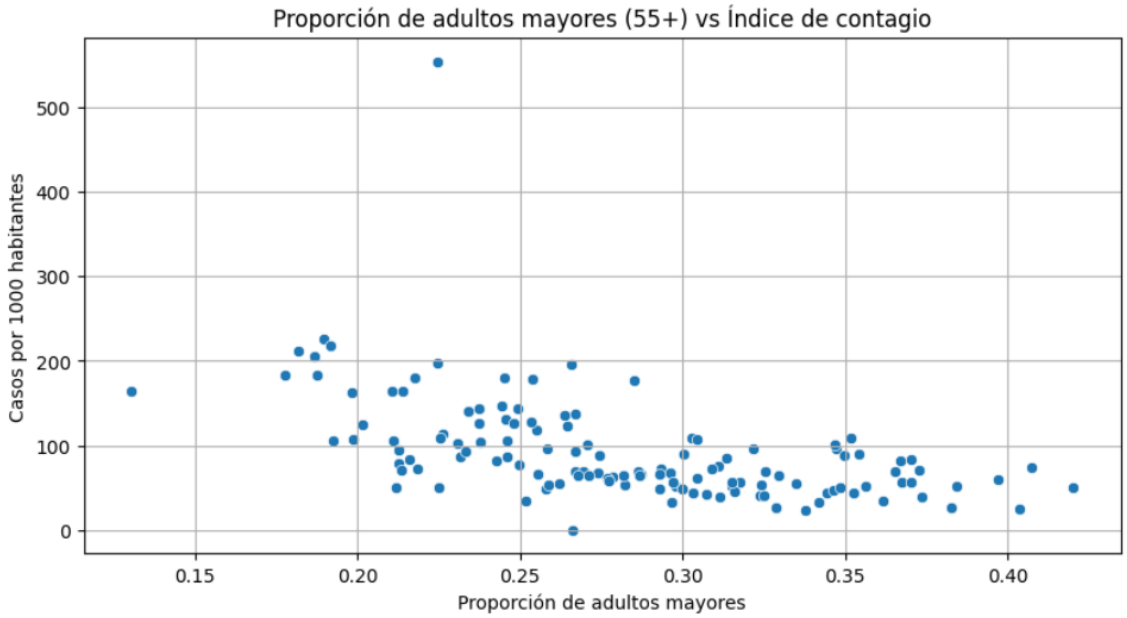
El objetivo de esta Hipótesis era analizar si LGAs con más población joven (o adulta mayor) tienen mayor tasa de contagios.

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.



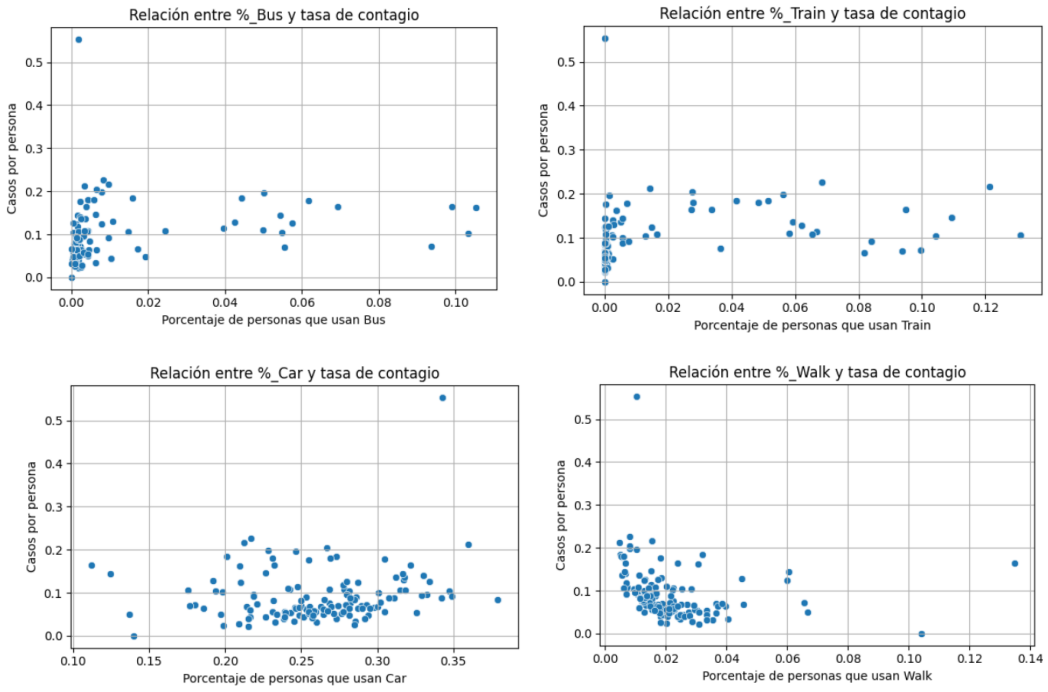
El segundo gráfico nos muestra la relación entre la cantidad de adultos mayores y la cantidad de casos Covid-19, observando que a mayor población de adultos mayores es menor la tasa de contagios, esto debido a que en su mayoría ya no pertenecen a la población económicamente activa, y, por tanto tienen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

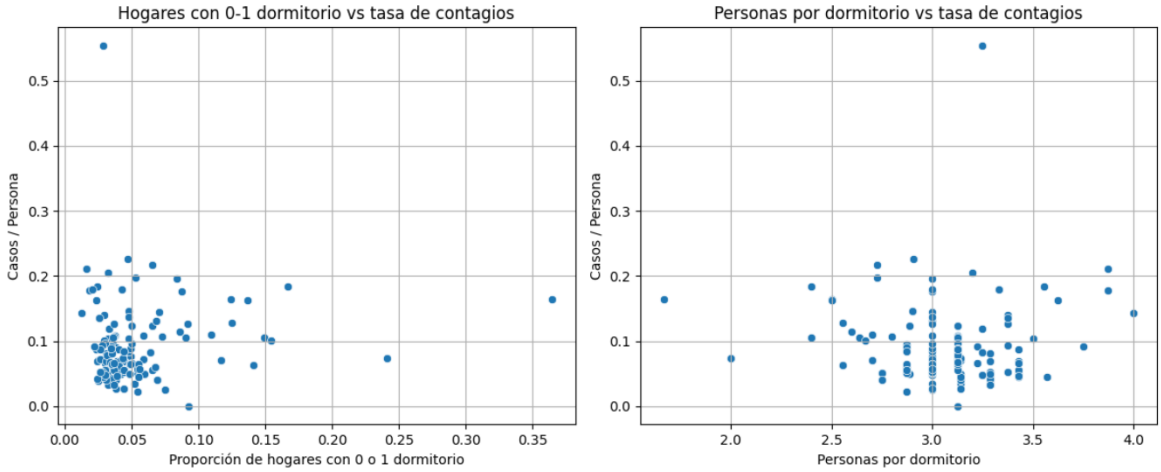
- Y en el caso de las personas que se movilizan a pie, podemos ver que tiene una correlación negativa, es decir que se presenta menor cantidad de casos de Covid-19 sobre LGAs con mayor porcentaje de personas que se movilizan a pie.

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.

Anexos

Enlace de Código de Data Wrangling: https://colab.research.google.com/drive/1yGfLV_b--PrjRN122SAGYTaLLXb7hlcT?usp=sharing

Enlace de código del Pipeline de Ciencia de Datos: https://colab.research.google.com/drive/1M3cNGs9Cc3rZ9oDIqoKOFkAMF_ywCVz_?usp=sharing

Enlace del dashboard: https://j0hn-sc.github.io/Dashboard_TCD/

Código fuente del dashboard: https://github.com/J0hn-SC/Dashboard_TCD