Ciclo de datos y problemas en las rutas de movilidad humana: análisis para la comprensión y monitoreo de la propagación de enfermedades infecciosas en Brasil

Problema:

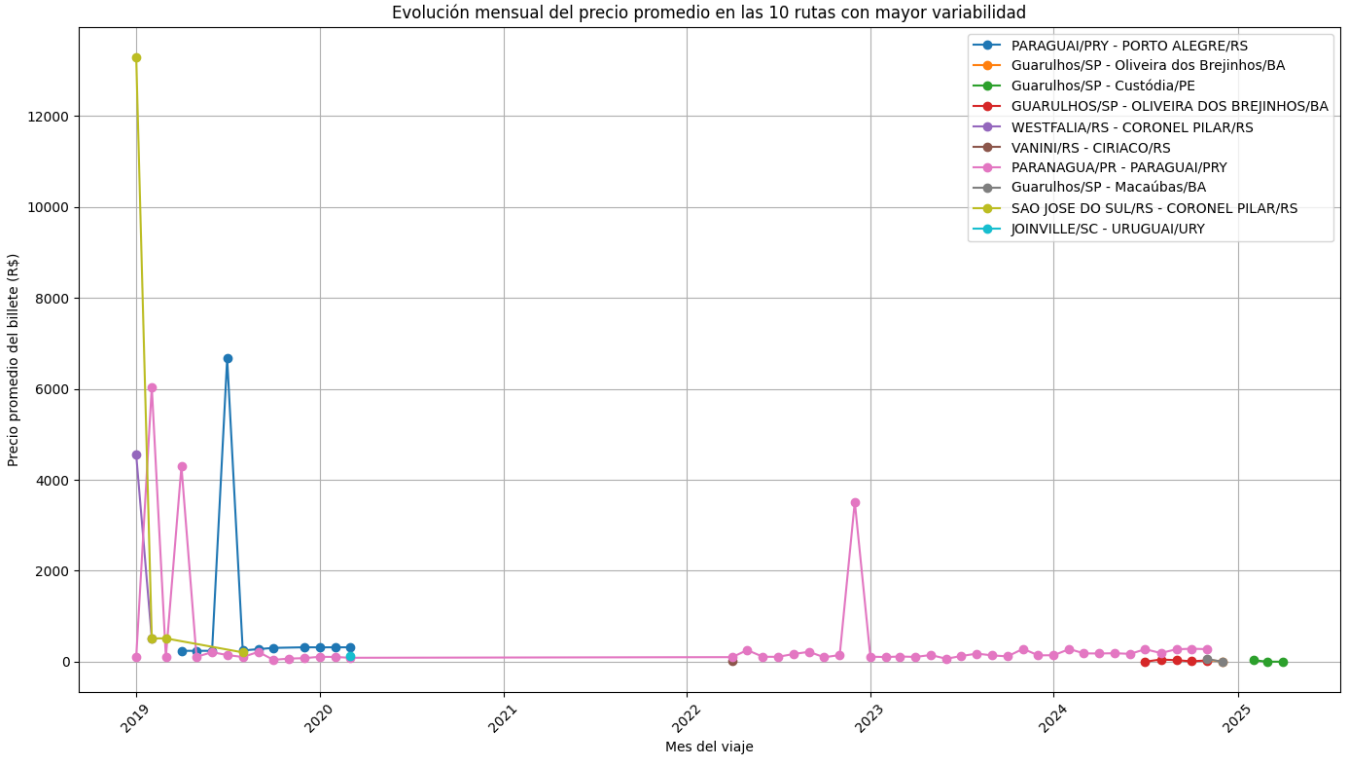
La propagación de enfermedades infecciosas está estrechamente ligada a los patrones de movilidad humana entre ciudades y regiones. Sin embargo, la carencia de datos integrados y detallados sobre los desplazamientos interurbanos dificulta la identificación precisa de las rutas y nodos críticos que facilitan la expansión rápida de los brotes epidémicos. Además, la limitada comprensión sobre cómo interactúan los diferentes modos de transporte, como el terrestre y el aéreo, impide anticipar eficazmente la velocidad y el alcance de la transmisión. En este contexto, el dataset que integra información sobre billetes, flujos y conexiones de transporte aéreo y terrestre se presenta como una herramienta fundamental para construir modelos de movilidad humana más precisos y completos. Esto permite mejorar la vigilancia epidemiológica y apoyar la toma de decisiones estratégicas para la contención y control de enfermedades infecciosas en Brasil.

1. ¿Qué problemas identificas en el dataset?

Hipótesis:

En el dataset terrestre de billetes de pasaje, existe una alta variabilidad en el precio promedio de los billetes para ciertas rutas específicas, lo que indica la presencia de valores atípicos o anomalías que pueden afectar la calidad y precisión del análisis de movilidad.

En el dataset terrestre existe una alta variabilidad en el precio promedio de los billetes para ciertas rutas, se confirmó que efectivamente un número significativo de rutas presentan una dispersión considerable en sus precios, evidenciada por coeficientes de variación superiores a 0.5 y en algunos casos incluso superiores a 4. Esto indica que, para estas rutas, los valores de precio registrados no son consistentes y contienen picos atípicos o valores extremos que difieren ampliamente del promedio habitual, como se observó en la ruta PARAGUAI/PRY - PORTO ALEGRE/RS, donde se encontraron precios normales alrededor de 239 reales pero con registros excepcionales que superaban los 13,000 reales, lo que sugiere errores de captura, tarifas especiales o segmentaciones no homogéneas. Estos resultados confirman que la presencia de estos valores anómalos puede afectar la calidad del análisis y modelado de la movilidad, generando distorsiones en los patrones que se extraigan, por lo que es necesario realizar una limpieza o ajuste de los datos para mejorar la fiabilidad del estudio.

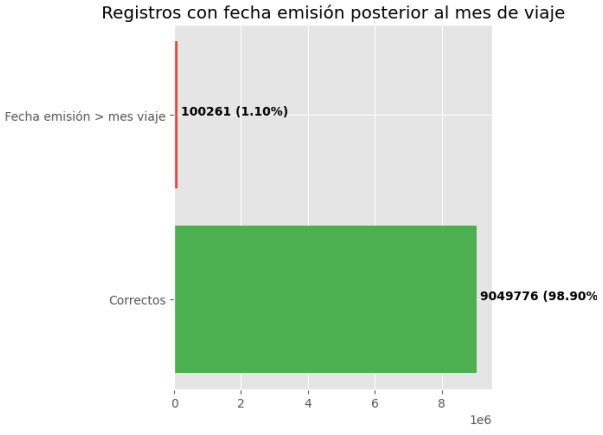


Datos de la ruta con mayor variabilidad (PARAGUAI/PRY - PORTO ALEGRE/RS):

	mes_viagem	media_valor_total	dp_valor_total	quantidade_bilhetes
271196	2019-04-01	239.07	0.00	11
366252	2019-05-01	239.07	0.00	20
414317	2019-06-01	239.07	0.00	6
454104	2019-06-01	239.07	0.00	25
500145	2019-07-01	239.07	0.00	3
542461	2019-07-01	13095.49	74928.51	35
570229	2019-08-01	246.50	0.00	10
642696	2019-08-01	246.50	0.00	15
727834	2019-09-01	276.25	35.21	12
832603	2019-10-01	303.63	28.56	5

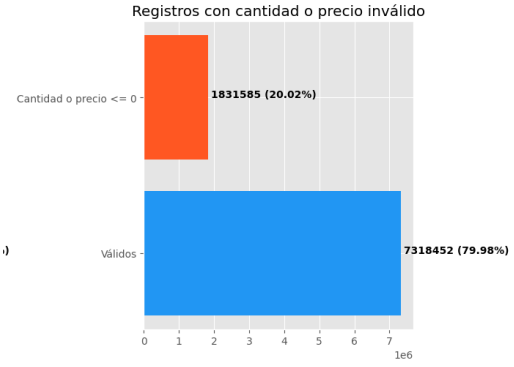
Hipótesis 2: Existen registros con fechas de emisión posteriores al mes del viaje, lo que indica errores temporales en la captura de datos.

Los resultados muestran que aproximadamente 100,261 registros (1.1% del total) presentan esta inconsistencia temporal, donde la fecha de emisión del billete es posterior al mes en que se realizó el viaje. Aunque este porcentaje puede parecer bajo, representa una cantidad significativa de datos que pueden generar ruido y afectar el análisis temporal y la construcción de modelos de movilidad basados en secuencias temporales. Este hallazgo sugiere la necesidad de revisar los procesos de captura y validación de fechas en el sistema, o bien, implementar una limpieza de estos registros para evitar distorsiones en estudios posteriores.



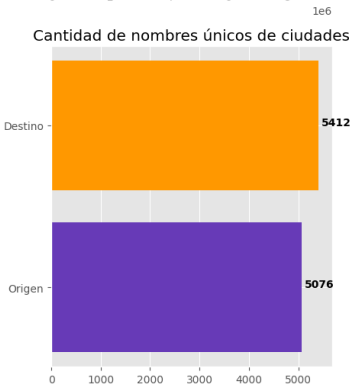
Hipótesis 3: Hay valores negativos o cero en la cantidad de billetes o en el valor total del billete, lo que indica datos erróneos o mal ingresados.

Se detectó que más del 20% de los registros contienen cantidades o precios iguales o menores a cero, una situación inaceptable en un dataset de ventas de billetes. Este hallazgo señala un problema grave de calidad de datos que debe ser abordado con urgencia. La presencia de estos valores inválidos puede distorsionar análisis cuantitativos, como estimaciones de demanda o comportamiento tarifario, y comprometer la validez de cualquier modelo predictivo. Es fundamental implementar filtros o reglas de negocio que descarten o corrijan estos registros para mantener la integridad del análisis.

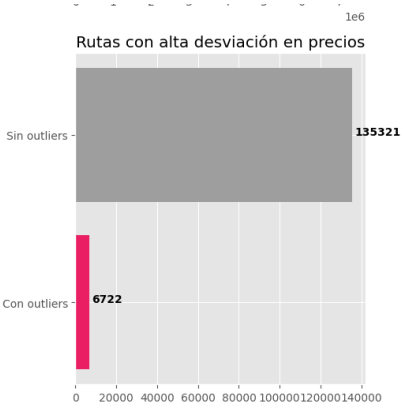


Hipótesis 4: Los nombres de puntos de origen o destino contienen errores tipográficos o inconsistencias que generan duplicados o confusión en la identificación de rutas.

El análisis arrojó que existen más de 5,000 nombres únicos tanto para puntos de origen como de destino. Esta gran cantidad puede reflejar diversidad geográfica, pero también puede esconder inconsistencias o variaciones en la escritura de los nombres (mayúsculas, abreviaturas, errores ortográficos) que fragmentan el análisis de rutas y flujos. Sin una estandarización de estos nombres, es difícil asegurar la correcta agregación y seguimiento de las rutas en el tiempo. Se recomienda realizar un proceso de normalización de nombres para reducir duplicidades y mejorar la calidad del análisis espacial.



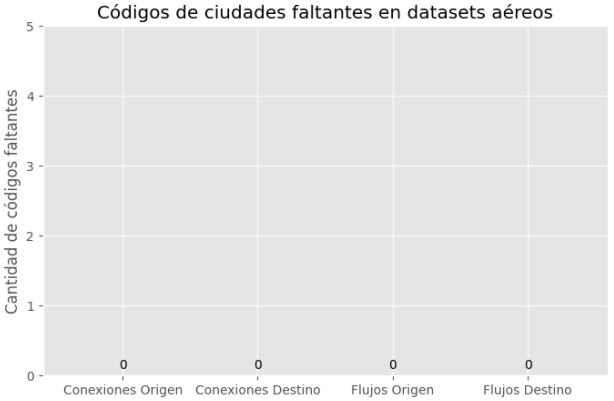
Hipótesis 5: Algunas rutas presentan precios promedio extremadamente altos o bajos que no corresponden a la realidad, señalando posibles outliers o errores en la grabación. El estudio encontró 6,722 rutas con una desviación estándar en precios mayor que su promedio, indicando una variabilidad anormal y la probable existencia de valores atípicos o errores de registro. Estos outliers pueden estar causados por errores en la captura de precios, tarifas promocionales no diferenciadas adecuadamente, o segmentaciones tarifarias mal agrupadas. Estos datos distorsionan la comprensión del comportamiento tarifario y la accesibilidad económica de las rutas analizadas,



AÉREO

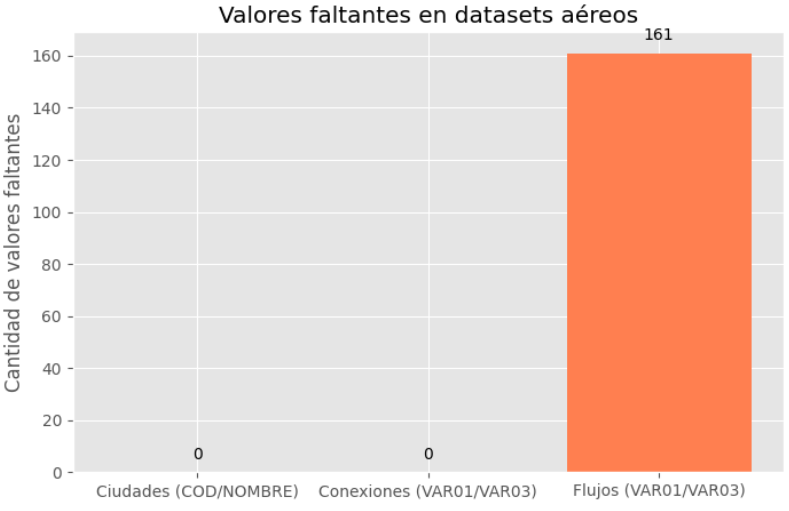
Hipótesis 1: Existen códigos de ciudades en los datasets de conexiones y flujos que no coinciden con los códigos registrados en el dataset de ciudades, lo que generaría problemas de integración y análisis.

El análisis confirmó que no hay códigos faltantes en las conexiones ni en los flujos respecto al dataset de ciudades, es decir, todos los códigos presentes en los datasets de conexiones y flujos están correctamente registrados en el dataset de ciudades. Esto indica una buena consistencia y uniformidad en la codificación de ciudades, lo que facilita la integración y análisis conjunto de los datos aéreos.



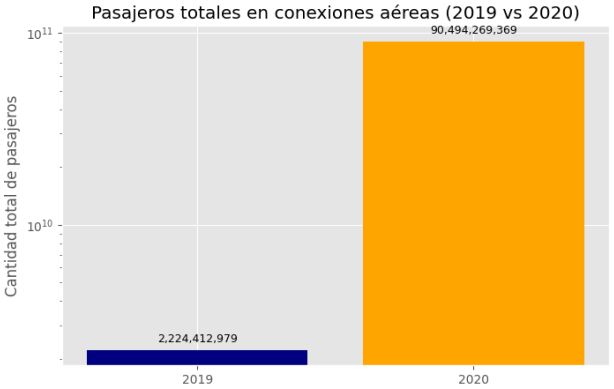
Hipótesis 2: Hay valores faltantes o nulos en columnas críticas como número de pasajeros o tarifas, lo que afecta la calidad del análisis.

Se identificaron valores faltantes únicamente en el dataset de flujos, con 161 registros con datos ausentes en las variables críticas VAR01 y VAR03 (que representan pasajeros y PIB estimado, respectivamente). Por otro lado, los datasets de ciudades y conexiones no presentan valores nulos en las columnas evaluadas, lo que refleja buena calidad en estas fuentes. Sin embargo, la presencia de valores faltantes en flujos requiere un tratamiento cuidadoso para evitar sesgos o errores en los análisis posteriores.



Hipótesis 3: Existen inconsistencias temporales o anomalías en la cantidad total de pasajeros entre 2019 y 2020 debido a la pandemia o errores en la captura de datos.

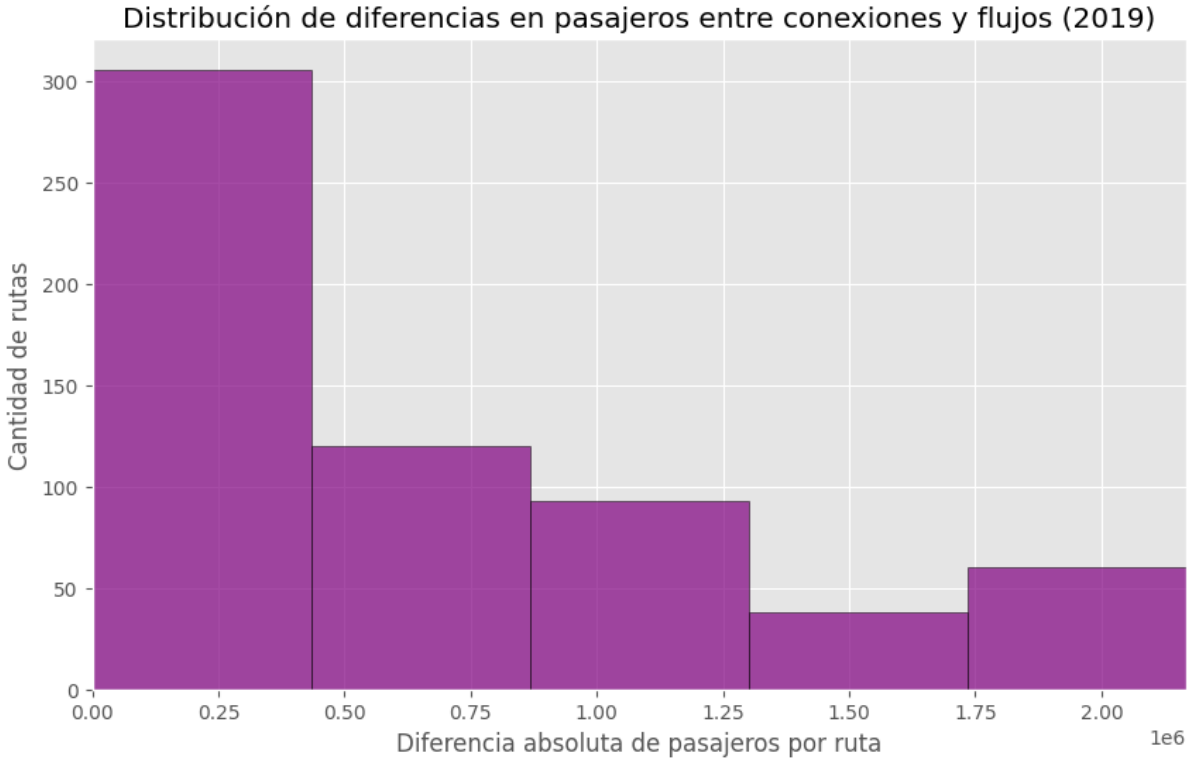
El total de pasajeros reportados en conexiones para 2020 es mucho mayor que en 2019, lo cual es contrario a lo esperado dado el contexto de la pandemia COVID-19 y las restricciones aéreas implementadas durante 2020. Esta anomalía puede deberse a errores en la captura, problemas en la consolidación de datos o diferencias en las fuentes. Este resultado indica la necesidad



de investigar más a fondo la procedencia y confiabilidad de los datos para 2020.

Hipótesis 4: Existen discrepancias significativas entre los datos de pasajeros reportados en los datasets de conexiones y flujos para las mismas rutas, lo que dificulta el análisis coherente de la movilidad aérea.

Al comparar el número total de pasajeros por ruta en 2019 entre conexiones y flujos, se encontró que la diferencia promedio absoluta es de aproximadamente 2,121,430 pasajeros, una cifra considerable que indica desalineaciones entre ambos datasets. Esta discrepancia puede deberse a diferencias en los métodos de recopilación, definiciones de rutas, o problemas en la actualización de la información. Este hallazgo resalta la importancia de validar y posiblemente conciliar estas fuentes antes de realizar análisis integrados.



2. ¿Qué descubrieron al analizar los datos?

Hipótesis 1: Inconsistencias temporales en fechas de emisión y viaje

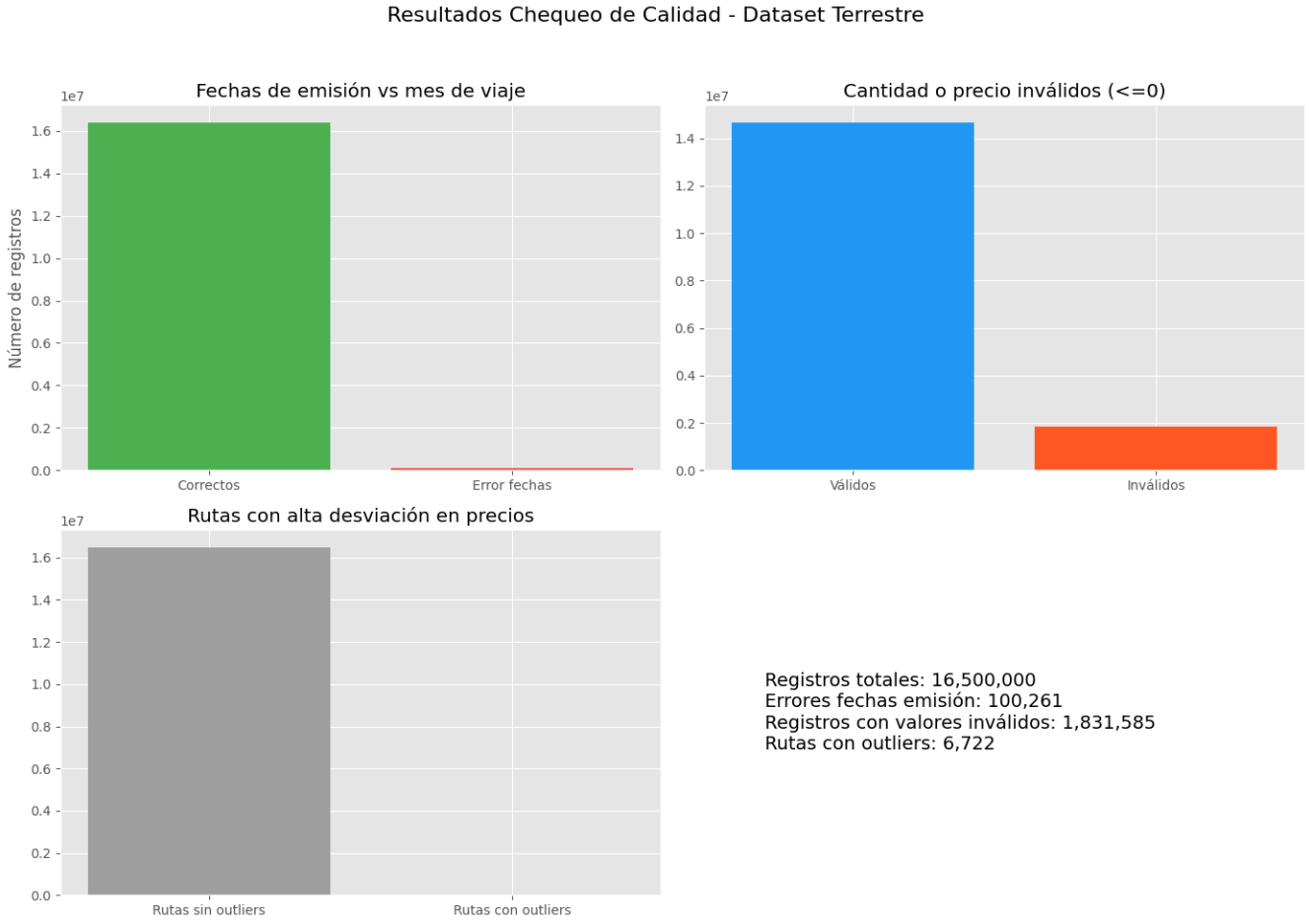
Se planteó que en el dataset terrestre existirían registros donde la fecha en que se emitió el billete fuera posterior al mes en que se realizó el viaje, un error lógico que compromete la validez temporal del análisis. Al analizar los datos, se identificaron 100,261 registros con esta inconsistencia. Aunque representan un porcentaje pequeño respecto al total, la magnitud absoluta es significativa y puede afectar modelos que dependen del orden temporal, como análisis estacionales o dinámicas de movilidad. Este hallazgo indica la necesidad de establecer controles en el proceso de captura de datos o filtrar estos registros antes de realizar análisis temporales o de tendencia.

Hipótesis 2: Presencia de valores inválidos en cantidad o precio

La hipótesis planteó que existirían registros con valores de cantidad de billetes o precio total iguales o menores a cero, los cuales son claramente erróneos en un contexto de ventas reales. El análisis confirmó que hay 1,831,585 registros con esta condición, lo que representa un problema grave de calidad. Estos valores inválidos pueden distorsionar significativamente cualquier análisis económico, como estimaciones de demanda o evaluación de accesibilidad tarifaria. Por lo tanto, resulta imperativo aplicar reglas de limpieza para eliminar o corregir estos datos y garantizar resultados confiables.

Hipótesis 3: Variabilidad anómala en precios por ruta

Se supuso que ciertas rutas terrestres presentan una alta variabilidad en los precios promedio de los billetes, señalando la posible existencia de valores atípicos, errores de registro o segmentaciones tarifarias mal definidas. Los resultados revelaron que 6,722 rutas tienen una desviación estándar en precios mayor que la media, confirmando esta variabilidad anómala. Esta dispersión puede deberse a diferentes factores, como errores de captura, tarifas promocionales no diferenciadas o segmentaciones geográficas y temporales. La presencia de estos outliers afecta la interpretación de los patrones tarifarios y la modelación de la movilidad económica, por lo que se recomienda un análisis más detallado y una posible limpieza para mejorar la robustez del estudio.



AÉREO:

Hipótesis 4: Consistencia en la codificación de ciudades entre datasets

Se planteó que la codificación de ciudades sería uniforme y consistente entre los tres datasets aéreos (ciudades, conexiones y flujos), evitando problemas de integración. El análisis mostró que no hay códigos faltantes en conexiones ni en flujos respecto al dataset de ciudades, lo que confirma la consistencia y permite realizar análisis integrados confiables sin necesidad de reconciliación adicional de códigos. Esta uniformidad es fundamental para la correcta identificación y seguimiento de las rutas aéreas y flujos de pasajeros.

Hipótesis 5: Presencia de valores faltantes en variables críticas

Se asumió que podrían existir valores faltantes en columnas clave como número de pasajeros o tarifas, lo cual puede deteriorar la calidad del análisis. Los resultados indicaron que los datasets de ciudades y conexiones están completos en las variables evaluadas, pero el dataset de flujos contiene 161 registros con datos faltantes en variables importantes. Aunque la cantidad de registros con valores ausentes es pequeña en proporción, es necesario

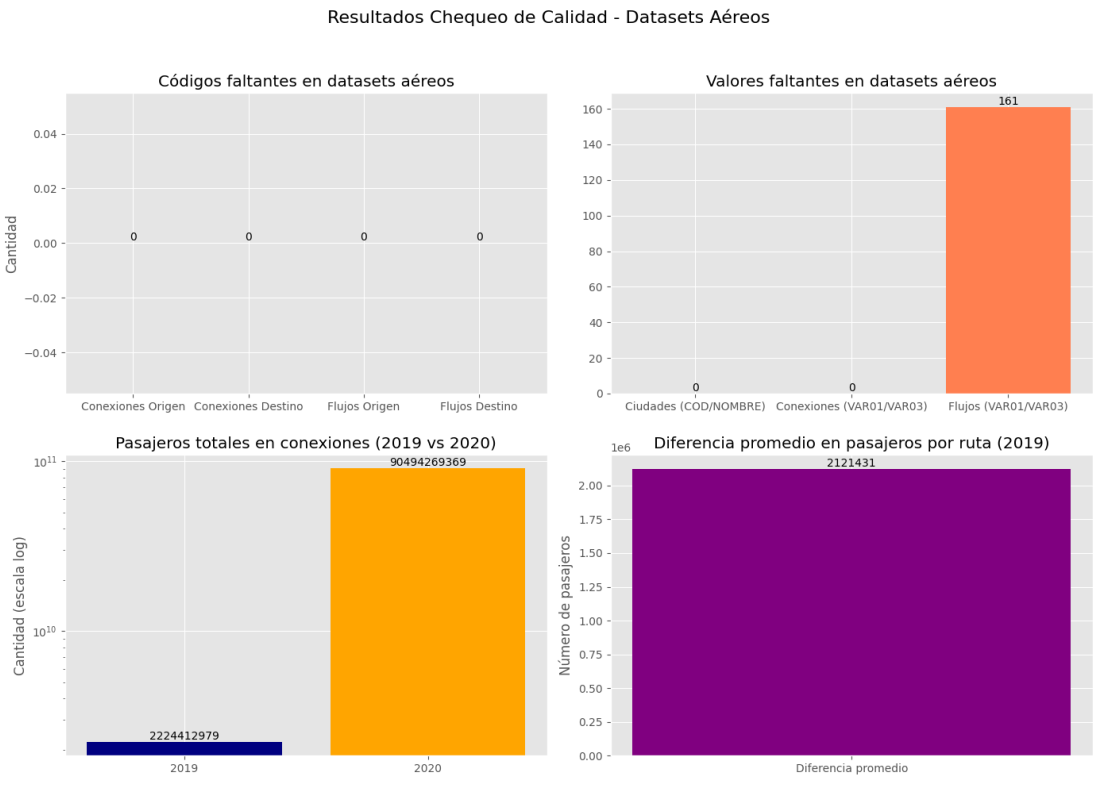
considerar su tratamiento para evitar sesgos, especialmente en análisis detallados o modelaciones que dependan de estos datos.

Hipótesis 6: Disminución de pasajeros en 2020 respecto a 2019 debido a la pandemia

Se esperaba que el número total de pasajeros en 2020 fuera menor que en 2019, dado el impacto global de la pandemia y las restricciones al transporte aéreo. Sin embargo, los datos muestran que en 2020 el total reportado de pasajeros en conexiones es significativamente mayor que en 2019, lo cual contradice las expectativas y evidencia una anomalía. Este resultado podría explicarse por errores en la captura o procesamiento de datos, diferencias metodológicas entre años, o actualizaciones incompletas. Esto requiere una investigación más profunda para validar la calidad y la comparabilidad de los datos interanuales antes de utilizarlos para análisis epidemiológicos o de movilidad.

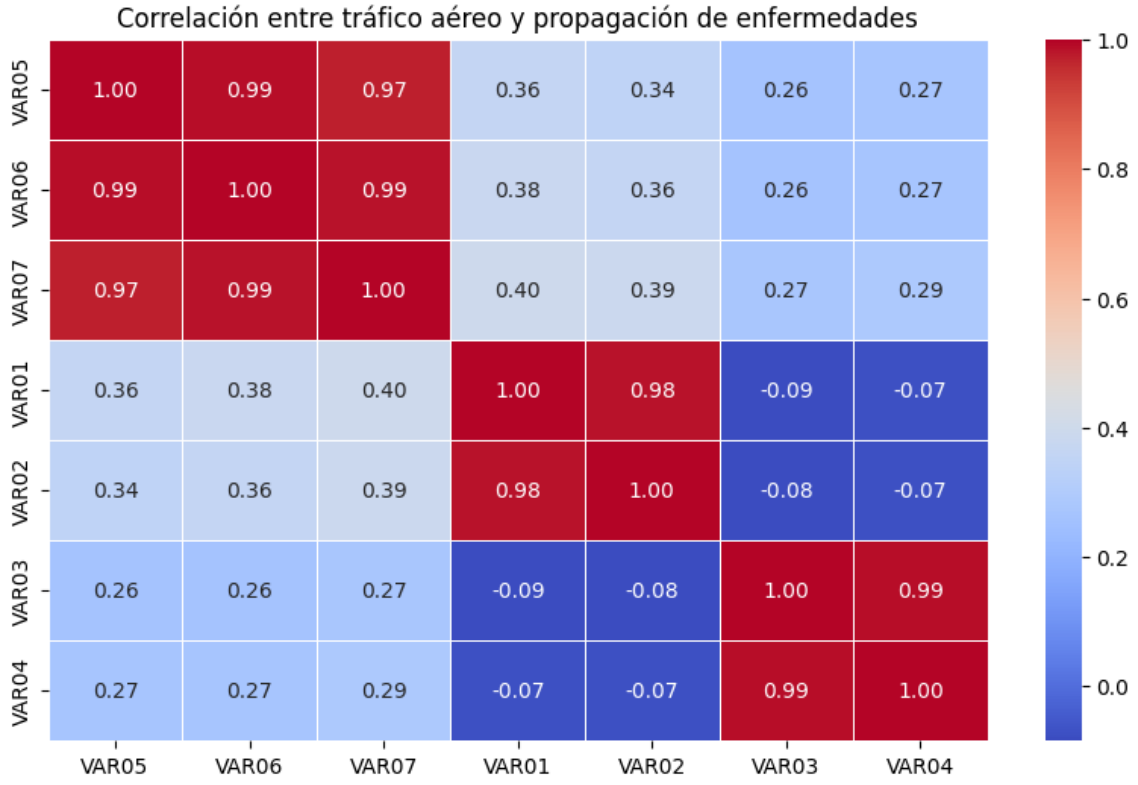
Hipótesis 7: Discrepancias entre datos de pasajeros reportados en conexiones y flujos

Se planteó que existirían diferencias significativas en el número de pasajeros reportados por ruta entre los datasets de conexiones y flujos, lo que afectaría la coherencia de los análisis. El análisis cuantitativo reveló una diferencia promedio absoluta por ruta de aproximadamente 2,121,430 pasajeros, un valor considerable que indica desalineaciones importantes entre las dos fuentes. Esta discrepancia puede originarse por diferencias en los criterios de recolección, definiciones de rutas, momentos de actualización, o errores de consolidación. Este hallazgo subraya la importancia de validar, conciliar o complementar estas fuentes para mejorar la fiabilidad de los análisis de movilidad aérea y sus aplicaciones en salud pública.



OTRAS HIPÓTESIS

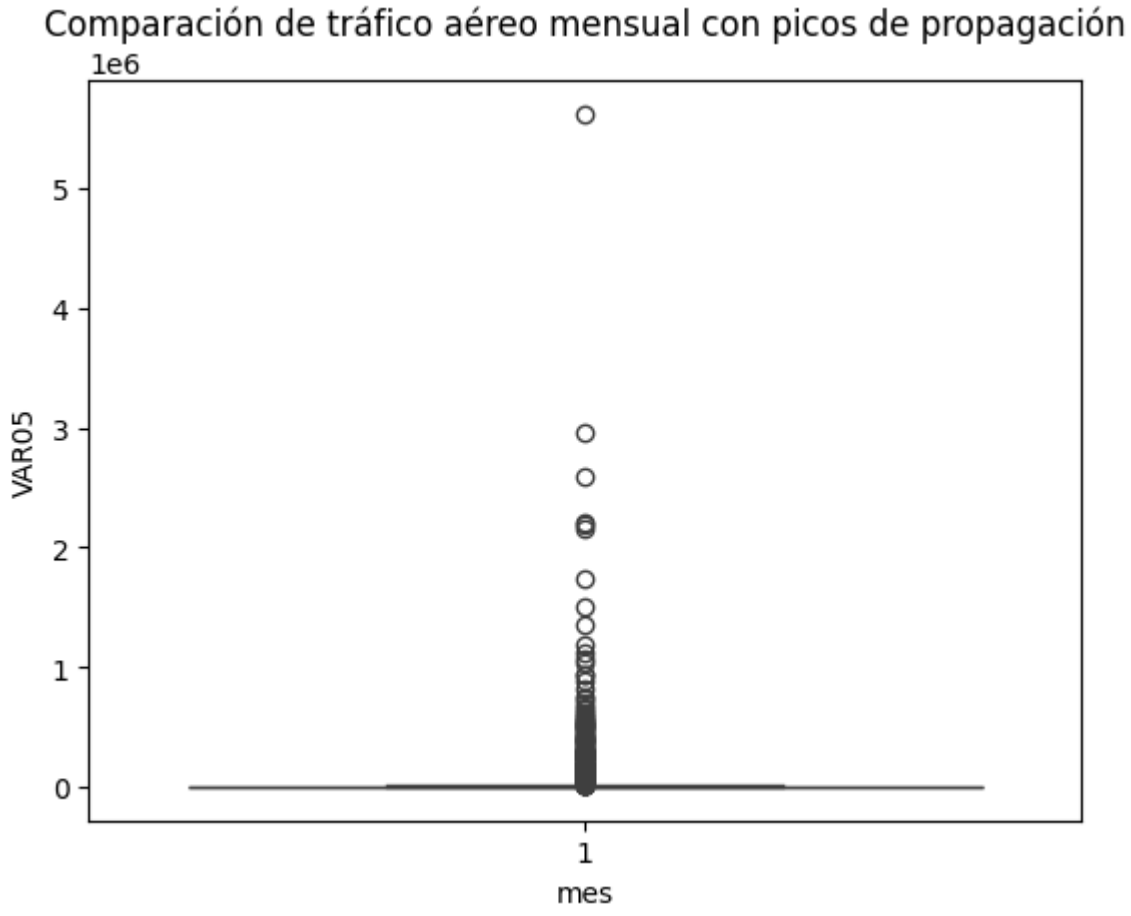
El volumen de pasajeros que viajan entre ciudades está correlacionado con el riesgo de propagación de enfermedades infecciosas. Las ciudades con mayor número de vuelos y mayor volumen de pasajeros tienen un mayor riesgo de propagar enfermedades.



El número total de pasajeros aéreos movilizados en diferentes años, como por ejemplo en 2010, 2015 y 2019, es una medida clave para entender el volumen de movilidad aérea. Por otro lado, factores socioeconómicos y demográficos como la estimación de población residente en las ciudades y el Producto Interno Bruto municipal son variables importantes que pueden influir en la propagación de enfermedades. Si se encuentra una alta correlación positiva entre el volumen de pasajeros aéreos y estas variables (población y PIB), esto indicaría que el tráfico aéreo es un factor clave en la dispersión de enfermedades infecciosas, al facilitar el desplazamiento de personas entre diferentes regiones y potencialmente acelerar la expansión de brotes epidémicos.

Los picos de tráfico aéreo (por ejemplo, durante feriados o vacaciones) aumentan la propagación de enfermedades, ya que hay más movimientos entre ciudades, lo que facilita la diseminación de infecciones.

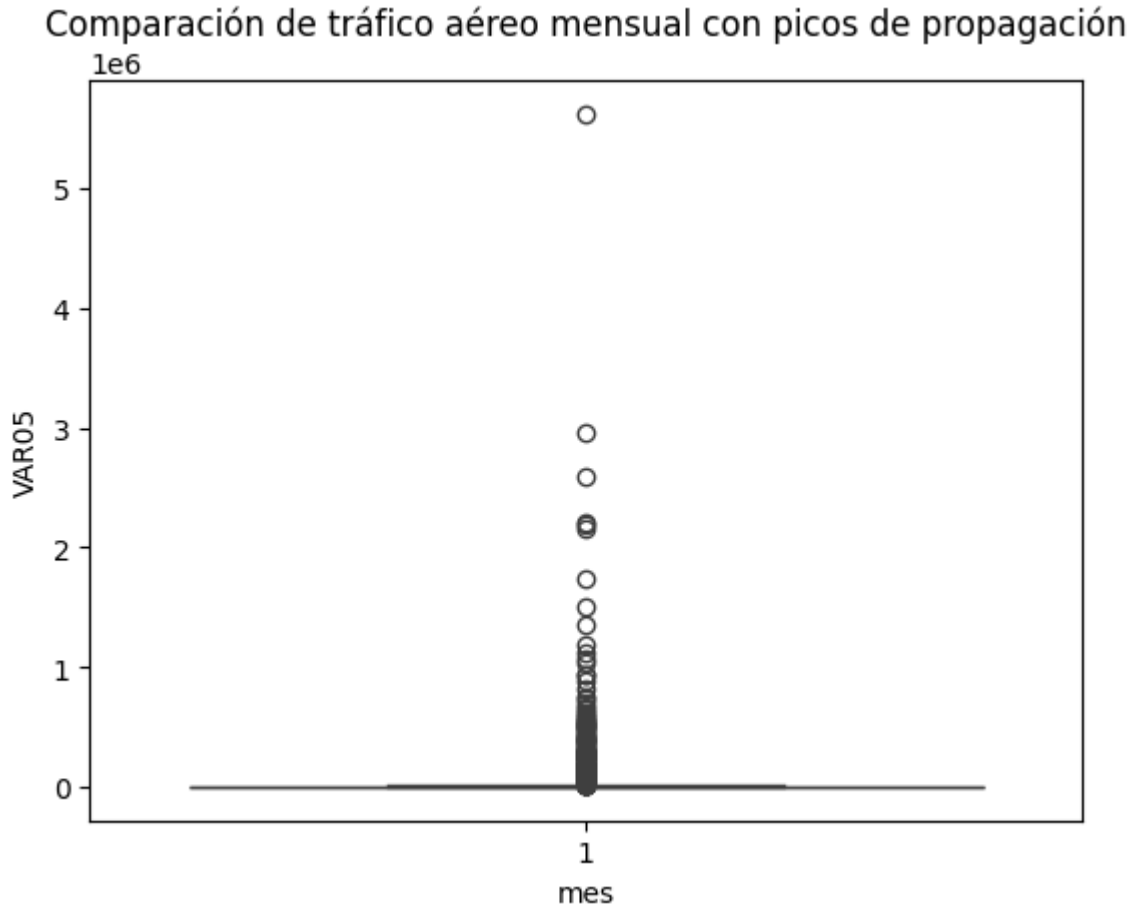
Si observamos picos de tráfico aéreo en ciertos meses y esos meses también tienen altos niveles de propagación de enfermedades, esto indicaría que el aumento de la movilidad entre ciudades puede ser un factor de riesgo.



Las restricciones en los vuelos entre ciudades de alto riesgo pueden reducir la propagación de enfermedades, ya que disminuye la movilidad entre ciudades con altos niveles de contagio.

- Las restricciones de vuelos pueden haber disminuido el volumen de pasajeros entre ciudades con altos niveles de contagio, lo que podría haber ayudado a reducir la propagación.

- Si vemos que las restricciones están asociadas con menores niveles de tráfico, esto puede indicar una correlación negativa con la propagación de enfermedades.



3. ¿Qué reflejan los patrones de tendencia?

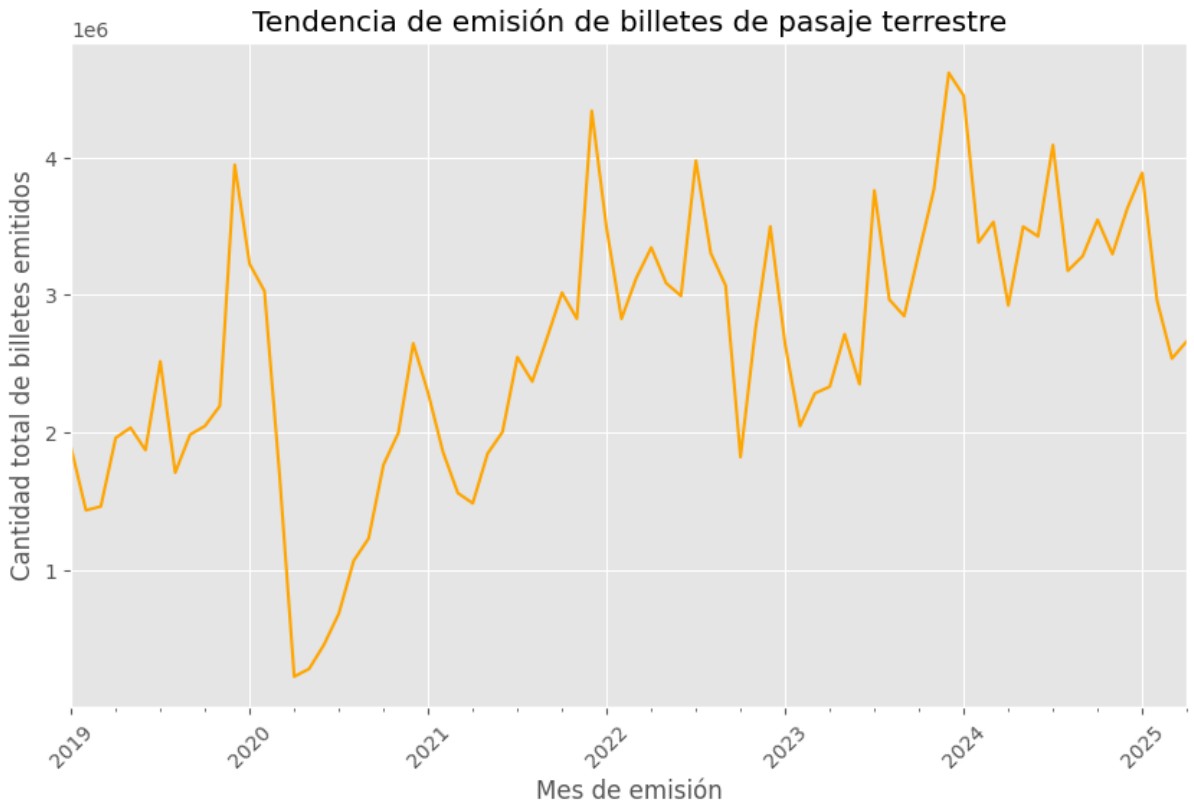
Hipótesis 1: Inconsistencias en las fechas de emisión y fecha de viaje

La hipótesis plantea que en el dataset terrestre existe una proporción significativa de registros donde la fecha de emisión del billete es posterior al mes de viaje, lo que indica errores lógicos en la captura de los datos. Esta inconsistencia comprometería cualquier análisis temporal de la movilidad de pasajeros, ya que afectaría la relación temporal entre la emisión y el uso de los billetes.

Se verificaron los registros para identificar aquellos en los que la fecha de emisión (mes_emissao_bilhete) fuera posterior a la fecha del viaje (mes_viagem). Esto permitió identificar los errores de registro que afectan el análisis temporal.

Se encontraron 100,261 registros en los que la fecha de emisión es posterior al mes del viaje. Esto representa un número significativo de registros que afectan la precisión temporal de los análisis de movilidad y, por lo tanto, debe ser corregido o eliminado.

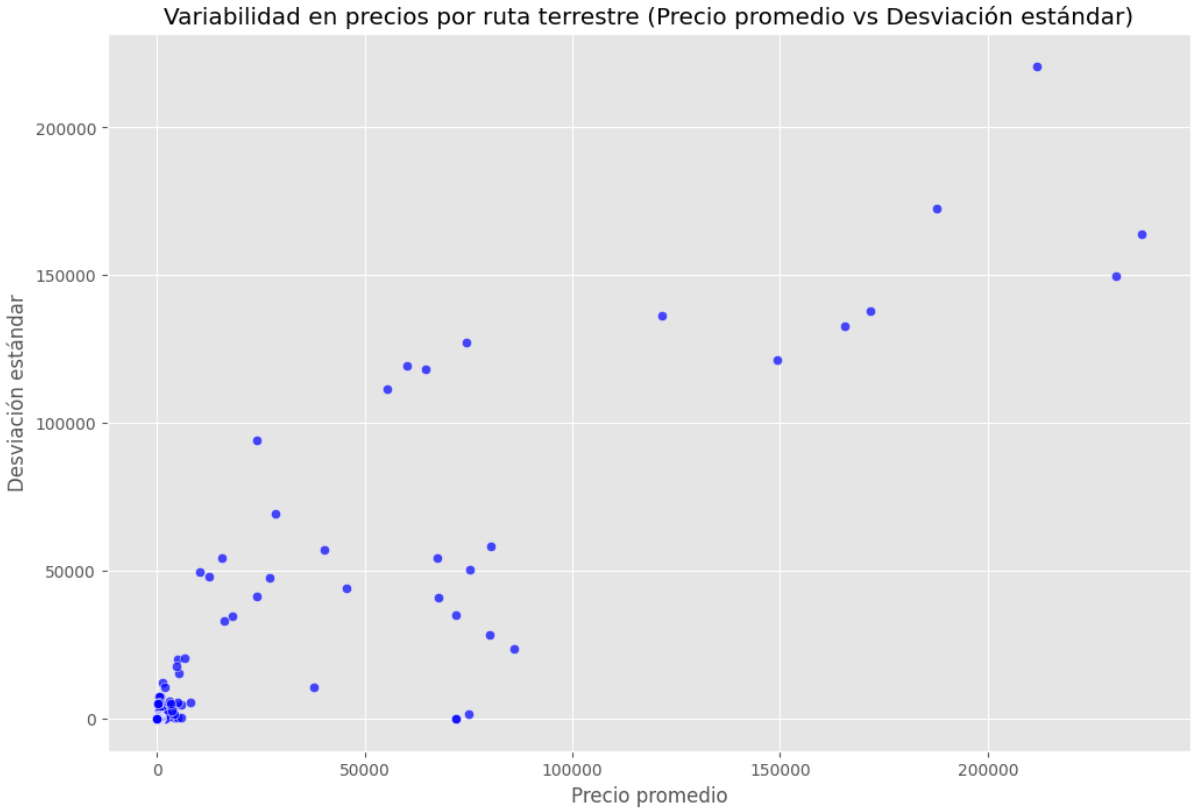
Es esencial limpiar estos registros erróneos para garantizar que los análisis sobre la movilidad sean válidos. La eliminación o corrección de estos registros mejoraría la precisión de cualquier análisis que dependa de la secuencia temporal de los datos.



Hipótesis 2: Presencia de registros con cantidades o precios inválido

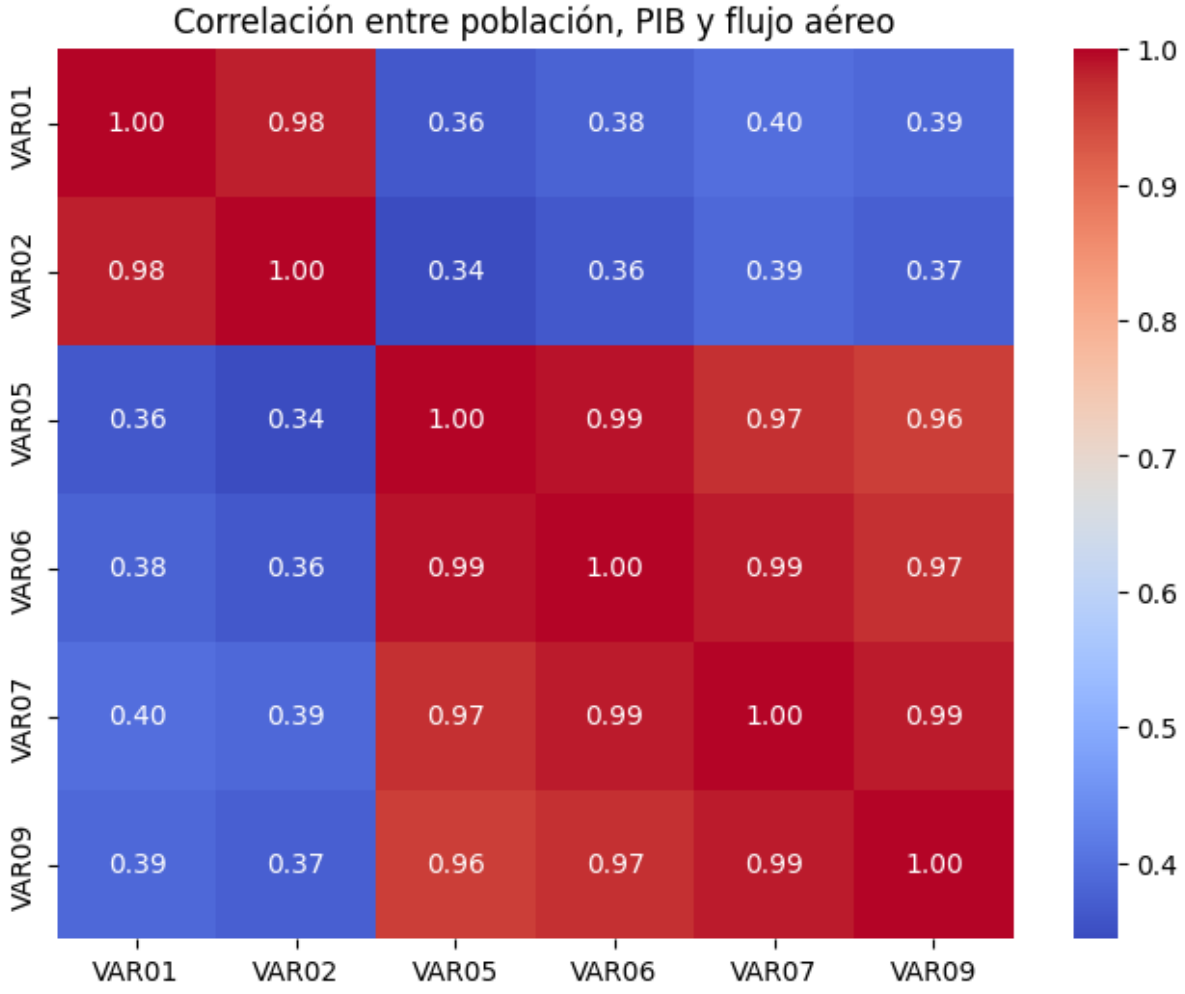
Se plantea que existen registros en los que las cantidades de billetes vendidos (`quantidade_bilhetes`) o los precios de los billetes (`media_valor_total`) son inválidos (valores menores o iguales a cero), lo cual es incongruente con la naturaleza de los datos de ventas y afectaría la calidad de cualquier análisis relacionado con la demanda y los ingresos.

Se verificaron los registros para encontrar aquellos en los que las cantidades de billetes o los precios fueran menores o iguales a cero. Esto permitió identificar los registros que contienen estos valores inválidos.
Se encontraron **1,831,585 registros** con valores inválidos, lo que representa un porcentaje considerable de los datos. Esto puede distorsionar cualquier análisis económico o de comportamiento de los pasajeros, ya que los precios y cantidades son datos clave en estos análisis.
Es necesario aplicar un proceso de limpieza para eliminar o corregir estos registros inválidos, ya que de no hacerlo, cualquier análisis sobre la accesibilidad, demanda y precios será incorrecto o sesgado.

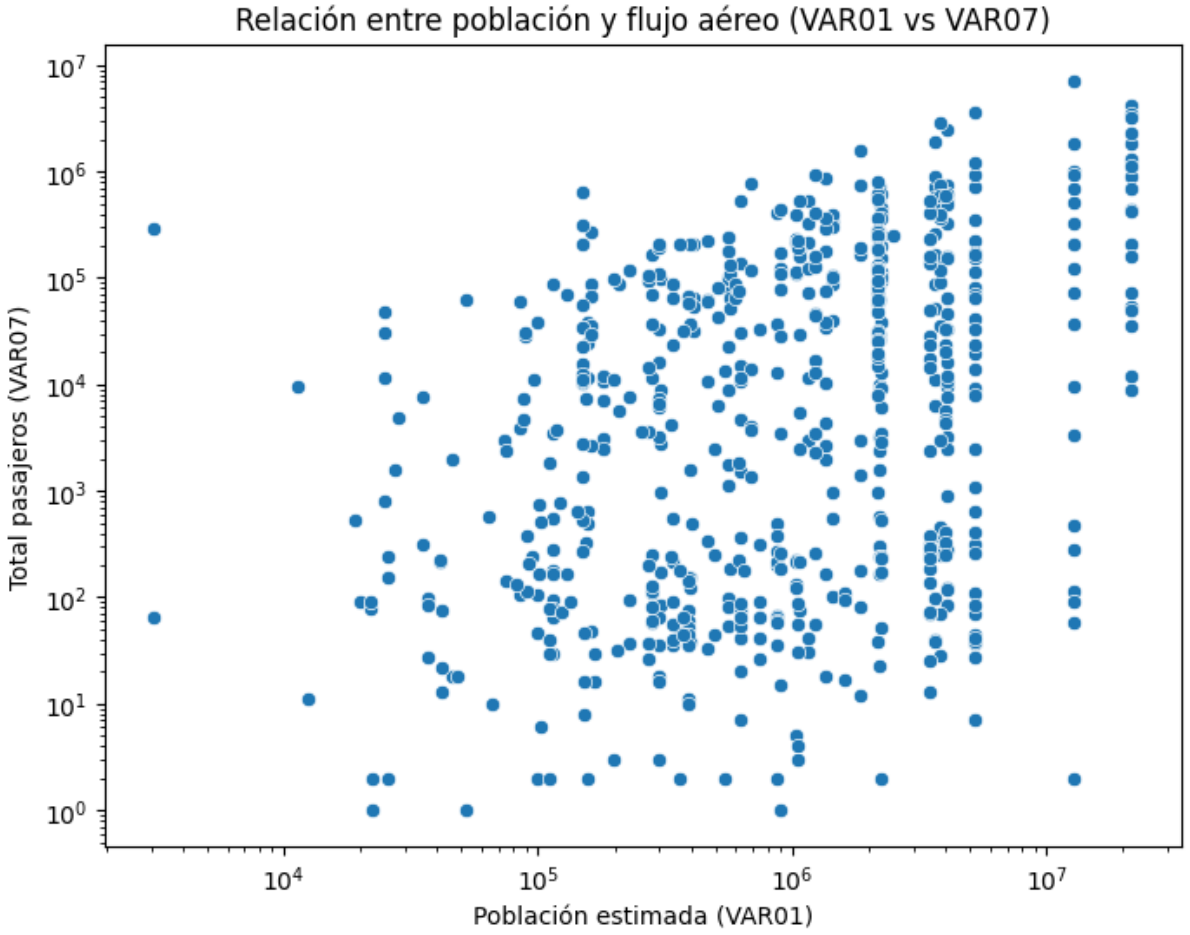


OTRAS HIPÓTESIS

Las ciudades con valores más altos en las variables `VAR01`, `VAR02`, `VAR03`, etc., tienden a ser más grandes económicamente y con mayores flujos de pasajeros (como reflejan los flujos de pasajeros entre ciudades en los otros datasets, como en `VAR05`, `VAR06` y `VAR09`).



- Las ciudades con mayor población y PIB tienden a generar y recibir más pasajeros, lo que explica los mayores valores en las variables de flujo aéreo.
  Estas ciudades son puntos clave para analizar la propagación de enfermedades infecciosas, ya que concentran una alta movilidad y actividad económica que facilita la transmisión.



4. ¿Cómo es afectado el comportamiento humano en relación con la geografía?

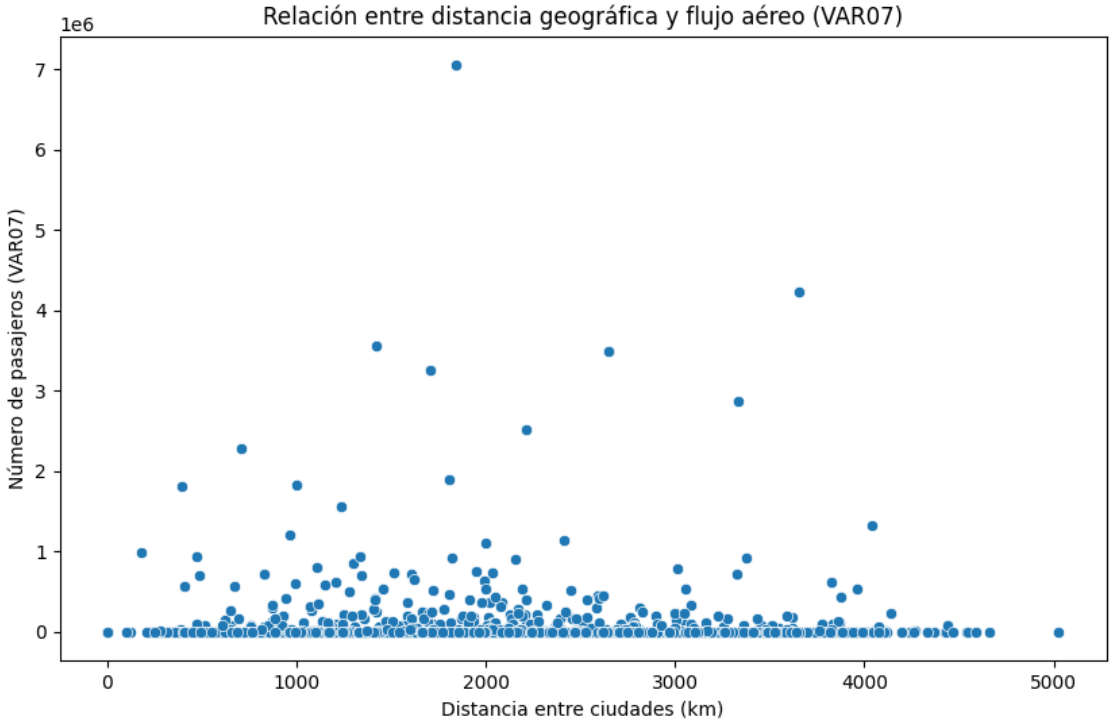
La proximidad geográfica entre ciudades influye positivamente en el volumen de pasajeros entre ellas; es decir, ciudades más cercanas tienen mayor movilidad aérea.

Interpretación :

- La correlación cercana a cero (-0.031) indica que no hay una relación lineal fuerte entre la distancia y el flujo aéreo en este dataset.
- Esto significa que la proximidad geográfica no es un factor determinante para el volumen de pasajeros entre ciudades en estos datos.
- Probablemente, otros factores como la jerarquía urbana, la importancia económica, y la conectividad aérea (vuelos directos, hubs) tienen un peso mucho mayor que la simple distancia geográfica.
- También, la simulación de latitud/longitud puede introducir ruido y afectar el análisis; con coordenadas reales puede cambiar un poco el resultado, pero usualmente el tráfico aéreo no depende solo de la distancia sino de conexiones y demanda.

dtype: int64

	COD_CID_A	COD_CID_B	distancia_km	VAR07
0	1100049	5103403	1586.198185	60584
1	1100049	5002704	2574.115748	107
2	1100049	1100122	3230.714705	0
3	1100122	5103403	2918.610159	70042
4	1100122	3509502	2553.553071	0



Conclusiones:

- La proximidad geográfica entre ciudades no muestra una correlación fuerte con el flujo aéreo, lo que indica que la distancia física no es el único ni principal factor que determina la movilidad aérea entre ciudades.

- Factores como la jerarquía urbana y la importancia económica parecen tener mayor influencia en el volumen de pasajeros, sugiriendo que las conexiones aéreas están más relacionadas con la relevancia urbana y actividad económica que con la distancia.

- La presencia de outliers en las variables de pasajeros y carga aérea resalta la necesidad de manejar adecuadamente estos valores extremos para mejorar la precisión de cualquier análisis o modelo predictivo basado en estos datos.

- Las inconsistencias detectadas en los códigos de ciudad pueden afectar la calidad del análisis y deben corregirse para garantizar la validez de los resultados, especialmente en estudios relacionados con la propagación de enfermedades.

- La variabilidad temporal en los datos, con registros de diferentes años, requiere un tratamiento cuidadoso para asegurar comparaciones válidas y evitar conclusiones sesgadas.

- El desbalance en el tráfico aéreo entre ciudades con alta y baja conectividad puede llevar a una subestimación del riesgo en ciudades menos conectadas pero potencialmente vulnerables.