AERONAUTICA CIVIL DE COLOMBIA MARZO DEL 2025

Cientifico De Datos : Nicolas Felipe Mogollon

INTRODUCCION

La aviación civil en Colombia desempeña un rol estratégico en la conectividad nacional e internacional, impulsando el crecimiento económico y el desarrollo turístico. En este marco, la Aeronáutica Civil de Colombia requiere de análisis rigurosos para optimizar sus operaciones, garantizar la eficiencia en la asignación de recursos y anticipar tendencias futuras. Este proyecto, enfocado en los vuelos registrados durante marzo de 2025, busca no solo caracterizar el comportamiento operativo del mes, sino también validar proyecciones y explorar relaciones causales que impacten en la toma de decisiones.

OBJETIVO

El enfoque principal esta en realizar un análisis estadístico integral (descriptivo e inferencial) de los datos de vuelos registrados en Colombia durante marzo de 2025 para caracterizar variables clave (*pasajeros, carga + correo, ciudad origen/destino*), validar proyecciones operativas y determinar relaciones significativas entre variables, utilizando técnicas estadísticas que contribuyan a la toma de decisiones estratégicas en la aeronáutica civil.

OBJETIVO ESPECIFICO

Parte Descriptiva

- Cleaning de la bases de datos .

- Generar tablas de frecuencia para las variables enfocadas al estudio

- Construir histogramas que visualicen la distribución de las variables de estudio.

- Calcular medidas de tendencia central y dispersión (media, varianza, desviación típica, coeficiente de variación) para las variables cuantitativas.

- Aplicar pruebas de normalidad para validar la parte inferencial (Anova)

- Elaborar conclusiones estadísticas que contribuyan a la investigación y toma de decisiones.

Parte Inferencial

- Validar con un nivel de confianza del 99% si el promedio de pasajeros por vuelo proyectado (800) es estadísticamente plausible.

- Determinar si la ciudad de destino influye significativamente en el número de pasajeros por vuelo.

- Evaluar si la ciudad de origen determina el número de pasajeros por vuelo.

- Verificar, tras eliminar los valores de "0 pasajeros", si el nuevo promedio proyectado de 900 pasajeros/vuelo es viable (99% confianza).

- Sintetizar conclusiones inferenciales para la toma de decisiones operativas.

MAPA CONCEPTUAL SOBRE LOS TEMAS TRABAJADOS

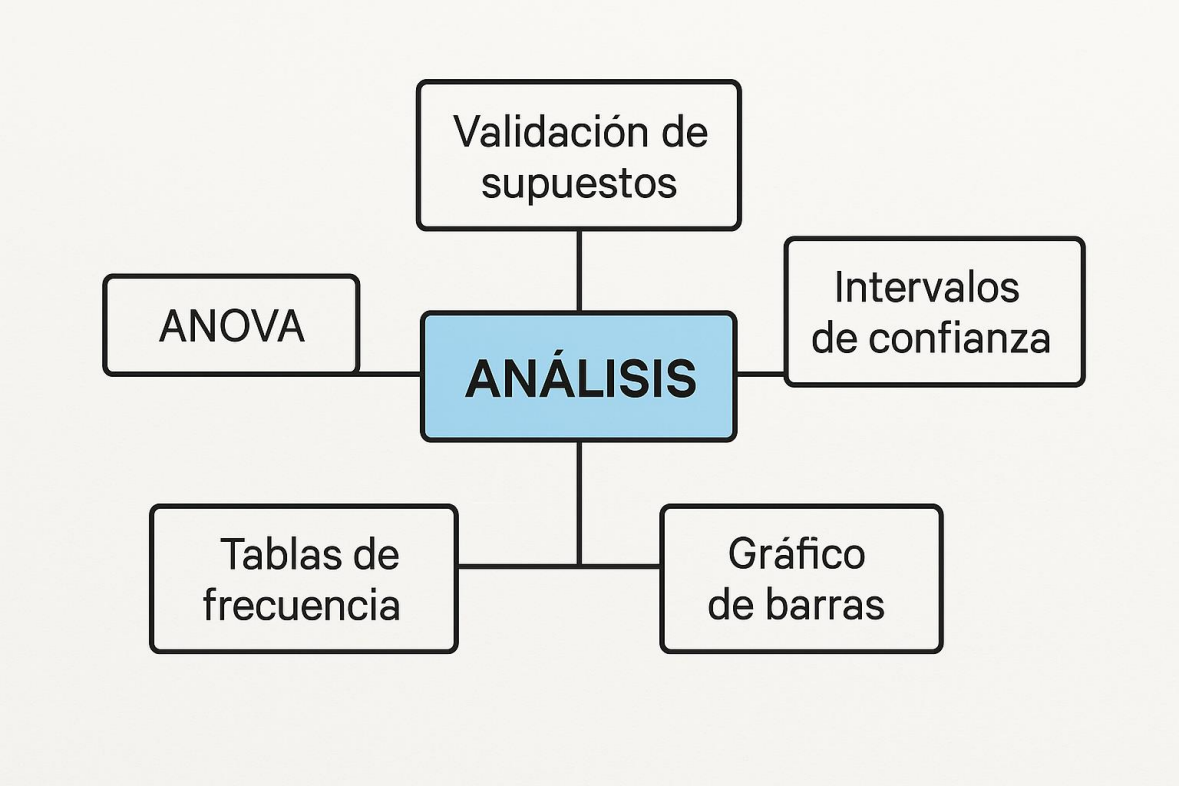


TABLA DE CONTENIDO

ANALISIS

Cleaning de el dataset

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7212 entries, 0 to 7211
Data columns (total 21 columns):
 #   Column                            Non-Null Count  Dtype
---  ------                            --------------  -----
 0   Sigla Empresa                     7212 non-null   object
 1   Nombre                            7212 non-null   object
 2   Fecha                             7212 non-null   datetime64[ns
 3   Año                               7212 non-null   int64
 4   Número de Mes                     7212 non-null   int64
 5   Origen                            7212 non-null   object
 6   Nombre.1                          7212 non-null   object
 7   Ciudad Origen                     7212 non-null   object
 8   Pais Origen                       7212 non-null   object
 9   Destino                           7212 non-null   object
 10  Nombre.2                          7212 non-null   object
 11  Ciudad Destino                    7212 non-null   object
 12  Pais Destino                      7212 non-null   object
 13  Tráfico (N/I)                     7212 non-null   object
 14  Tipo Vuelo                        7212 non-null   object
 15  Pasajeros                         7212 non-null   int64
 16  Carga + Correo (Kg)               7212 non-null   float64
 17  Tráfico (N/I) 2                   7212 non-null   object
 18  Tipo Vuelo 2                      7212 non-null   object
 19  Tipo Vuelo (Regular - No Regular)  7212 non-null   object
 20  Mes                               7212 non-null   object
dtypes: datetime64[ns](1), float64(1), int64(3), object(16)
memory usage: 1.2+ MB
```

```
df.duplicated().sum()
```

```
0
```

```
df=df[['Pasajeros','Carga + Correo (Kg)','Ciudad Origen','Ciudad Destino']]
```

La base de datos esta limpia no tiene inconsistencias en el tipo de datos según variable o valores faltantes como tampoco tiene registros duplicados, la base de datos tiene 7212 registros y 20 variables de las cuales para el estudio se hara un enfoque a las variables: Pasajeros, Carga + Correo (Kg), Ciudad Origen, Ciudad Destino

Dataset final

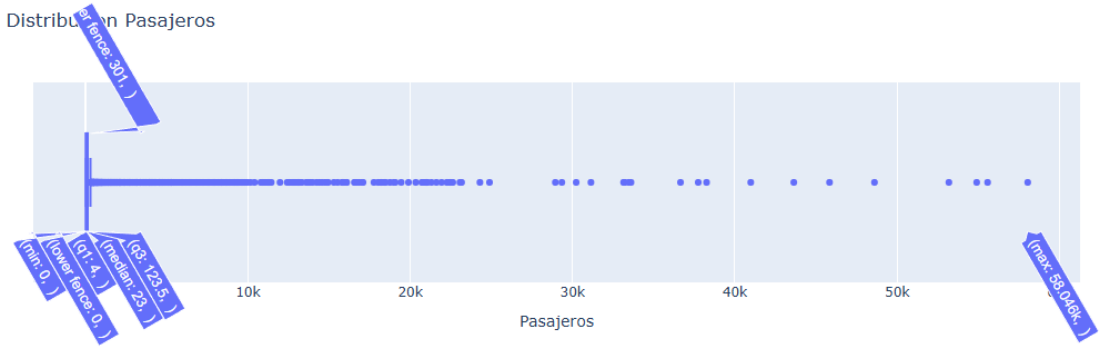
	Pasajeros	Carga + Correo (Kg)	Ciudad Origen	Ciudad Destino
0	0	5.0	LA CHORRERA	VILLAVICENCIO
1	1	5.0	MIRAFLORES - GUAVIARE	LETICIA
2	26	115.0	PUERTO CARREÑO	VILLAVICENCIO
3	3	5.0	LETICIA	LA CHORRERA
4	1	23.0	VILLAVICENCIO	MIRAFLORES - GUAVIARE
...	...	...	...	...
7207	219	0.0	TIGUANA	BOGOTÁ, D.C.
7208	945	0.0	BOGOTÁ, D.C.	VALENCIA
7209	1034	0.0	VALENCIA	BOGOTÁ, D.C.
7210	3828	0.0	SANTIAGO DE CALI	MADRID
7211	1250	9819.0	MADRID	SANTIAGO DE CALI

7212 rows × 4 columns

Exploration Data analysis (EDA)

En esta etapa esta enfocada a entender las variables de estudio todo con el enfoque de Elaborar conclusiones que contribuyan a la investigación inferencial y toma de decisiones.

Pasajeros



Aquí como podemos ver en la distribucion de pasajeros hay muchisimos outliers apartir de 301 pasajeros por ruta de vuelo serian un oulier pero despues de hacer un diagnostico se concluyo que no se puede eliminar estos valores ya que son muy representativos de la variable en cuestion a esta conclusion se llego cuando se filtro el data frame para entender mejor su variable “pasajeros”:

Nombre	Fecha	Año	Número de Mes	Origen	Nombre.1	Ciudad Origen	Pais Origen	Destino	...	Ciudad Destino	Pais Destino	Tráfico (N/I)	Tipo Vuelo	Pasajeros
AVIANCA	2025-03-01	2025	3	BOG	BOGOTA - EL DORADO	BOGOTÁ, D.C.	COLOMBIA	ADZ	...	SAN ANDRES - ISLA	COLOMBIA	N	R	13160
AVIANCA	2025-03-01	2025	3	BOG	BOGOTA - EL DORADO	BOGOTÁ, D.C.	COLOMBIA	AEP	...	BUENOS AIRES	ARGENTINA	I	R	2088

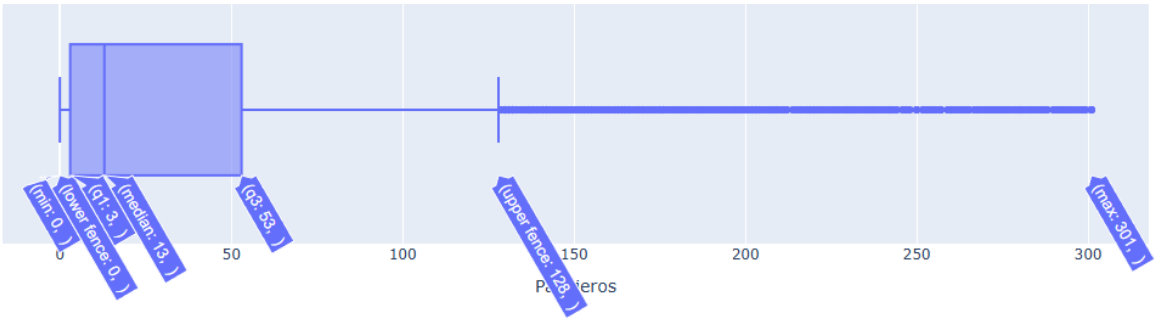
Como podemos ver tomamos de la aerolinea avianca los vuelos de bogota -san andres y bogota-buenos aires que son vuelos muy concurrentes en un mes vemos que el total de pasajeros transportados en cada ruta es de 13160 y 2068 lo cual nos hace entender que cada registro representa la suma total de ese trayecto en un mes osea en este caso cuantos pasajeros se tuvo en total en el mes de marzo en por ejemplo de los vuelos de la ruta de bogota – san andres, en si estos valores grandes en pasajeros que serian nuestros outliers representaria rutas muy concurridas como podria ser de bogota a buenos aires con la aerolinea avianca con 2068 pasajeros entonces no serian valores que se podrian eliminar o no tener en cuenta ,como conclusion se decidio dividir esta variable en dos la primera con vuelos <= 301 y la segunda con los vuelos con mas de 301 pasajeros para poder tener una graficacion y una tabla de frecuencias mas representativay legible de la variable.

Entrando en detalle de los 7212 registros del dataset cuando dividimos la variable quedaria asi:

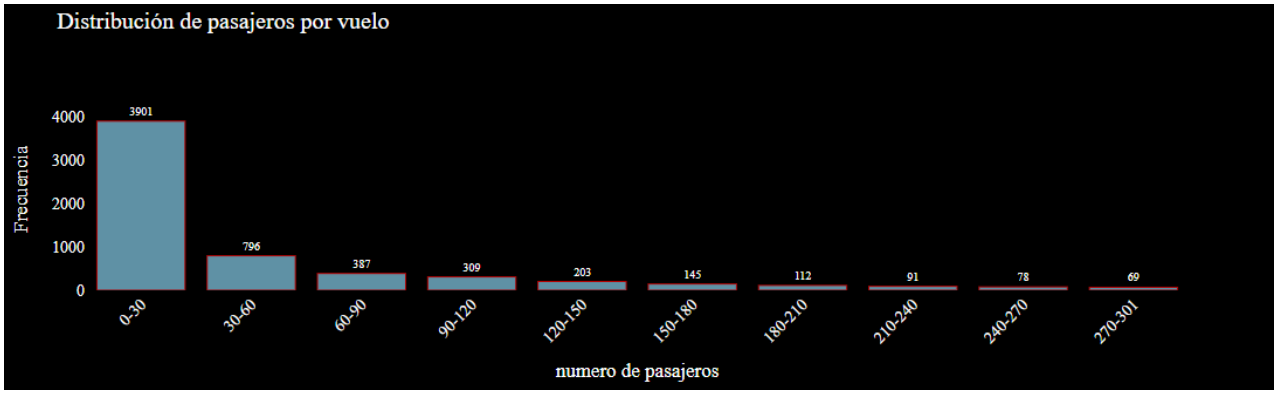
Pasajeros <= 301 total de registros (6091)

Pasajeros > 301 total de registros (1021)

Pasajeros <= 301 total de registros (6091)



	numero de pasajeros	Frecuencia	Frecuencia relativa
0	0-30	3901	64.045
1	30-60	796	13.068
2	60-90	387	6.354
3	90-120	309	5.073
4	120-150	203	3.333
5	150-180	145	2.381
6	180-210	112	1.839
7	210-240	91	1.494
8	240-270	78	1.281
9	270-301	69	1.133



Si analizamos la distribución, el 75% de las rutas que transportaron 310 pasajeros o menos en marzo se concentran en el rango de 0 a 53 pasajeros, lo que indica una marcada tendencia hacia valores bajos. Además, el 64.045% de los datos se encuentra únicamente en el intervalo de 0 a 30 pasajeros, reforzando la idea de una alta concentración en los tramos inferiores.

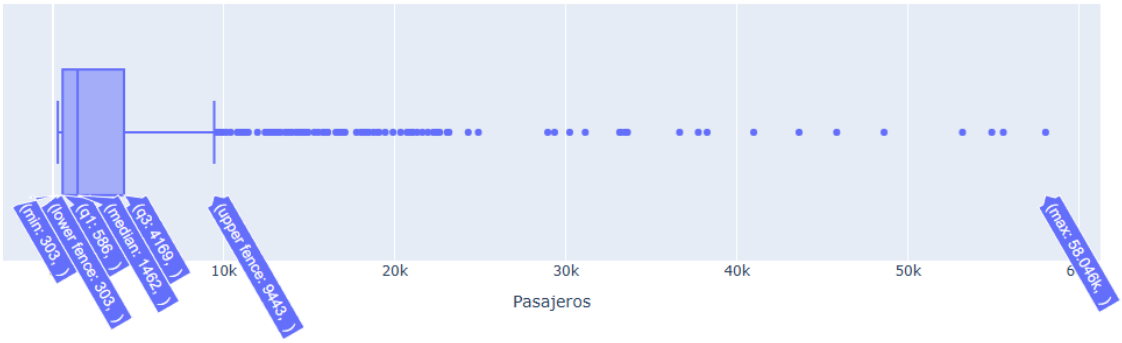
Por otro lado, al observar la tabla de frecuencias, los valores entre 60 y 301 pasajeros representan solo el 22.888% del total dentro de esta misma categoría, lo que sugiere que la mayoría de rutas movilizaron pocos pasajeros, y que los vuelos con más de 60 pasajeros son significativamente menos frecuentes.

Es importante tener en cuenta que esta variable de “pasajeros ≤ 301” es una subdivisión de la variable principal de pasajeros y representa aproximadamente el 84% del total. Si tomamos el intervalo de 0 a 120 pasajeros, que abarca el 88.54% de

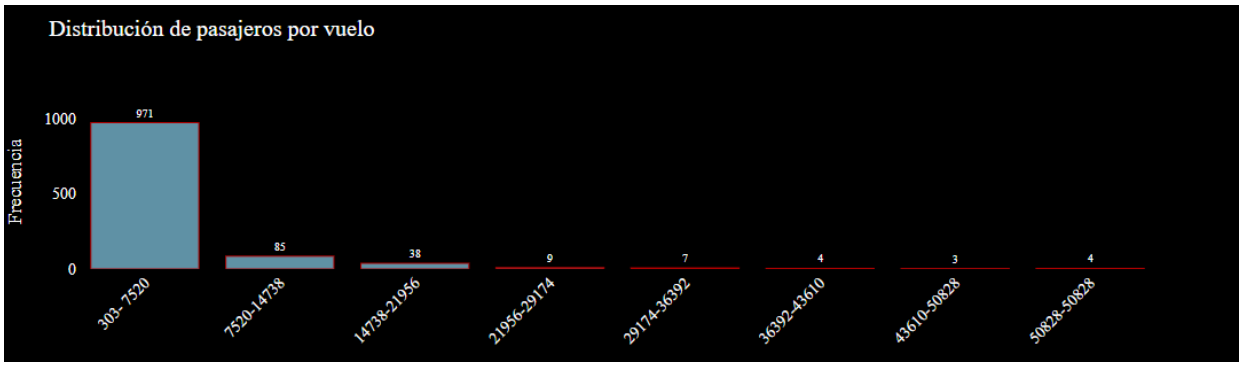
los datos dentro de esta subcategoría, podemos decir que dicho rango representa aproximadamente el 74% del total de la variable principal.

En resumen, incluso considerando el conjunto total de datos (sin dividir entre "pasajeros ≤ 301" y "pasajeros > 301"), se observa una clara concentración en los tramos de 0 a 120 pasajeros, lo cual refuerza la conclusión de que la mayoría de las rutas movilizaron un número reducido de pasajeros en el periodo de marzo de 2025.

Pasajeros > 301 (rutas mas concurrentes de vuelo) total de registros (1021)



	numero de pasajeros	Frecuencia	Frecuencia relativa
0	303- 7520	971	86.619
1	7520-14738	85	7.583
2	14738-21956	38	3.390
3	21956-29174	9	0.803
4	29174-36392	7	0.624
5	36392-43610	4	0.357
6	43610-50828	3	0.268
7	50828-50828	4	0.357

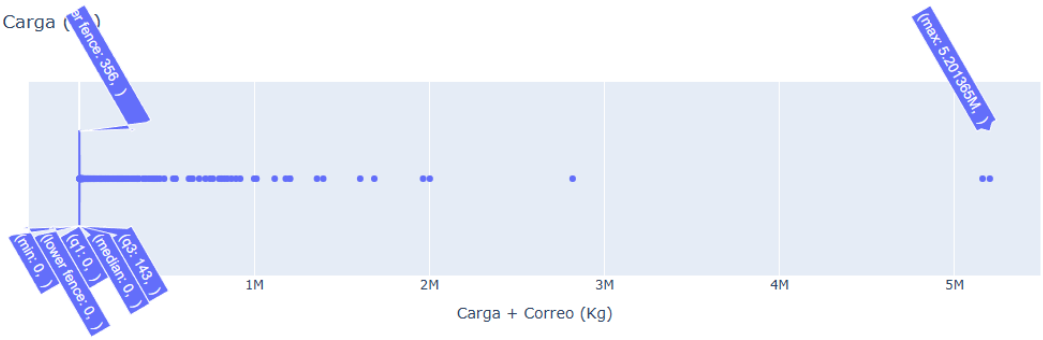


La subcategoría “pasajeros > 301” hace referencia a las rutas de vuelo con mayor concurrencia, generalmente asociadas a aerolíneas comerciales. Dentro de esta categoría, el 86% de los casos se encuentra en el rango de 303 a 7,520 pasajeros, mientras que el restante 14% se concentra en valores entre 7,520 y 50,828 pasajeros.

Si comparamos estos rangos con la variable total de pasajeros, se puede estimar que el intervalo de 303 a 7,520 pasajeros representa aproximadamente el 12% del total, mientras que el tramo de 7,520 a 50,828 pasajeros equivale a solo el 1.9% del total.

De acuerdo con el gráfico de distribución, a partir de 9,443 pasajeros los valores comienzan a ser considerados outliers dentro de esta subcategoría, lo que refuerza la idea de que los vuelos con volúmenes extremadamente altos de pasajeros son casos excepcionales ,pero como vimos anteriormente al filtrar el data frame tambien son muy significativos .

Carga + Correo(kg)

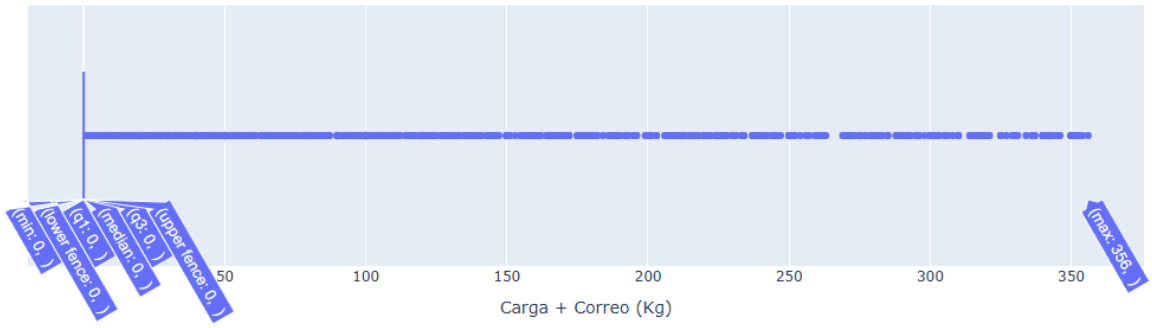


En esta variable al igual que la de pasajeros para poder interpretarla, leerla mejor con una tabla de frecuencias y visualizacion se decidio dividir la variable en dos la 1 variable seria <=356 en carga(kg) y la segunda variable seria >356 en carga (kg) “estos valores se escogieron basandose en el upper fence de la grafica de distribucion” , como ya miramos al filtrar el dataset para entender las variables, las cargas que llegan hasta 5 millones hacen referencia a la acumulacion en carga (kg) de alguna ruta de vuelo durante el mes de marzo

Carga (kg) <= 356 numero de registros (5755)

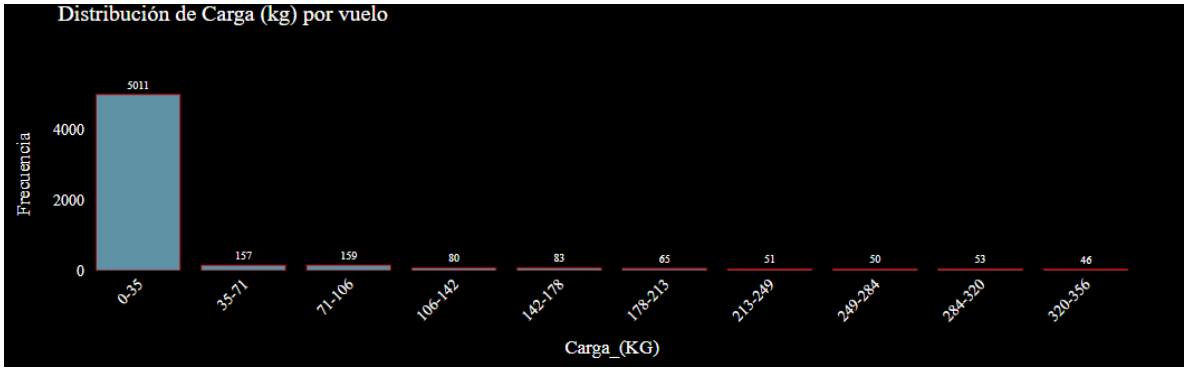
Carga (kg) > 356 numero de registros (1457)

Carga (kg) <= 356



)

	Carga_(KG)	Frecuencia	Frecuencia relativa
0	0-35	5011	87.072
1	35-71	157	2.728
2	71-106	159	2.763
3	106-142	80	1.390
4	142-178	83	1.442
5	178-213	65	1.129
6	213-249	51	0.886
7	249-284	50	0.869
8	284-320	53	0.921
9	320-356	46	0.799

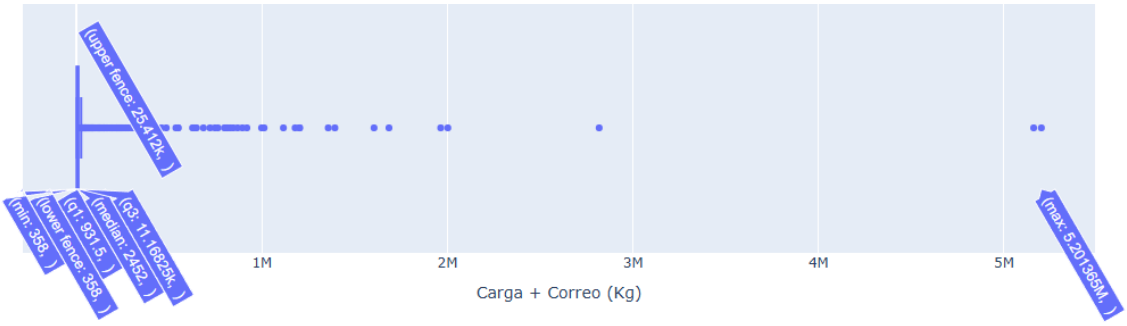


Como podemos observar, a pesar de dividir la variable en dos subcategorías, la distribución de la variable principal sigue presentando una gran cantidad de outliers. Esto indica que los valores de carga por ruta de vuelo varían considerablemente, reflejando un rango muy amplio.

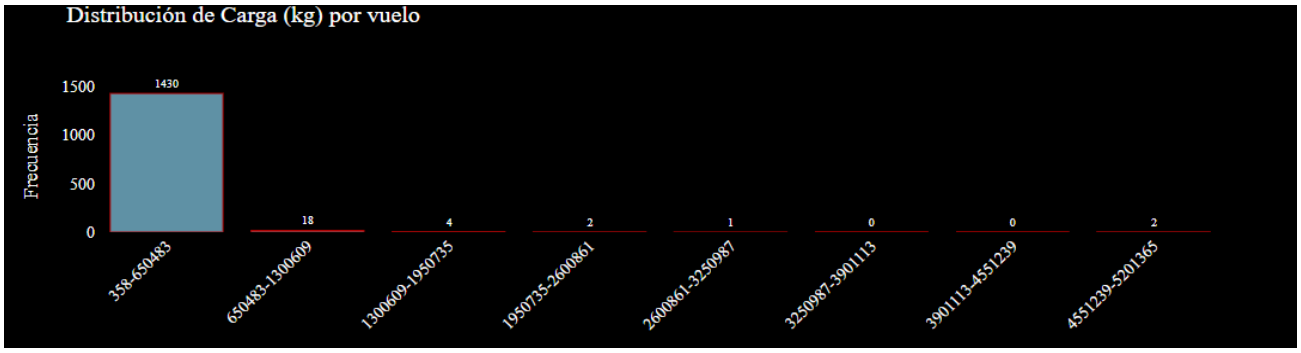
En la primera subcategoría (carga ≤ 356), el 75% de los datos corresponden a una carga de 0, lo que significa que a partir de cualquier valor mayor que cero ya podría considerarse un outlier dentro de esta subcategoría. Además, el rango de 0 a 71 unidades de carga abarca aproximadamente el 89% de los datos de esta subcategoría, lo cual representa alrededor del 70% de la variable principal.

Por otro lado, el intervalo de 71 a 356 unidades de carga representa solo el 11% de la subcategoría, lo que equivale aproximadamente al 8% del total de la variable principal. Esto confirma una fuerte concentración de valores bajos y una alta asimetría, con pocos casos que presentan valores significativamente más altos, considerados atípicos.

Carga (kg) > 356



	Carga_(KG)	Frecuencia	Frecuencia relativa
0	358-650483	1430	98.147
1	650483-1300609	18	1.235
2	1300609-1950735	4	0.275
3	1950735-2600861	2	0.137
4	2600861-3250987	1	0.069
5	3250987-3901113	0	0.000
6	3901113-4551239	0	0.000
7	4551239-5201365	2	0.137



En esta subcategoria que estarian todos los valores altos de carga (kg) el 98 % de los datos estaria entre 358 -650.483 que equivaldria de la variable total en 19 % el 75 % de los datos esta hasta 11.168 de carga que equivladria de la variable total a un 15 % y en el intervalo de 650.483 hasta 5.201.365 habria un total de 2 % que equivaldria de la variable total a un 0.4 % podemos ver que apartir de 25.412 kg se consideraria un outlier

Intervalos

Intervalo (kg)	Registros estimados	% del total general
0–71	5,168	71.65%
71–356	587	8.14%
356–650,483	1,430	19.83%
>650,483	27	0.37%
TOTAL	**7,212**	✅ **100.00%**

Ciudad de origen ------- Ciudad de destino

```
df['Ciudad Origen'].nunique()
462
```

```
df['Ciudad Destino'].nunique()
471
```

Las variables ciudad_origen y ciudad de destino presentan hasta 471 valores únicos, lo que representa un desafío para la visualización y análisis debido a:

- Tablas de frecuencias: La gran cantidad de categorías generaría una tabla excesivamente larga y de difícil interpretación.
- Visualizaciones: En gráficos tradicionales (como barras), la alta cardinalidad produciría superposición de etiquetas y pérdida de legibilidad.

Solución implementada:

Para optimizar el análisis, se ha decidido:

1. Seleccionar las 20 ciudades con mayor frecuencia.
2. Agrupar el resto en una categoría denominada "Otros".

Este enfoque permitirá:

- Visualizar claramente las ciudades más representativas
- Mantener la información del resto de registros sin saturar los gráficos

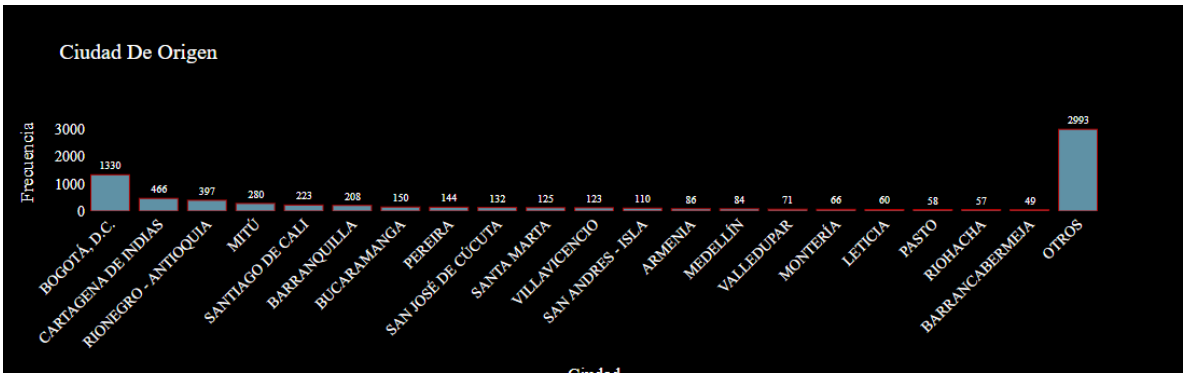
- Facilitar la identificación de patrones y tendencias principales

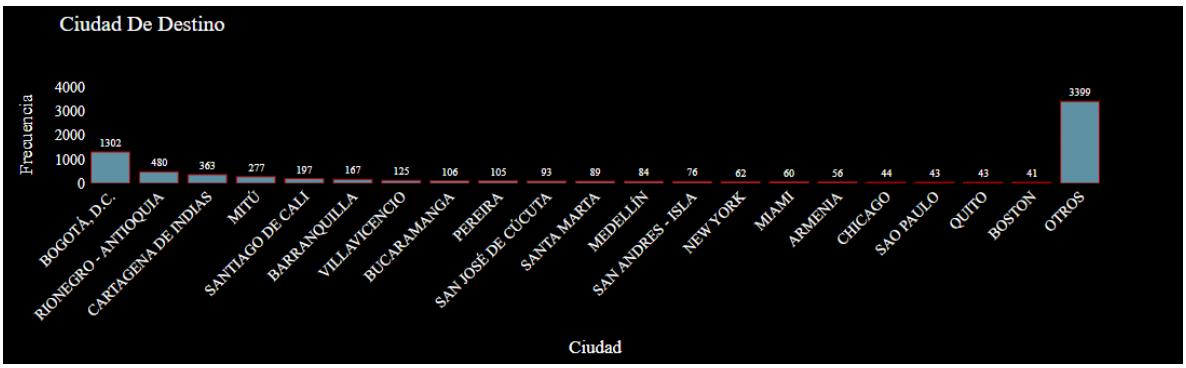
Ciudad de Origen

	Ciudad	Frecuencia	Porcentage
0	BOGOTÁ, D.C.	1330	18.441486
1	CARTAGENA DE INDIAS	466	6.461453
2	RIONEGRO - ANTIOQUIA	397	5.504714
3	MITÚ	280	3.882418
4	SANTIAGO DE CALI	223	3.092069
5	BARRANQUILLA	208	2.884082
6	BUCARAMANGA	150	2.079867
7	PEREIRA	144	1.996672
8	SAN JOSÉ DE CÚCUTA	132	1.830283
9	SANTA MARTA	125	1.733222
10	VILLAVICENCIO	123	1.705491
11	SAN ANDRES - ISLA	110	1.525236
12	ARMENIA	86	1.192457
13	MEDELLÍN	84	1.164725
14	VALLEDUPAR	71	0.984470
15	MONTERÍA	66	0.915141
16	LETICIA	60	0.831947
17	PASTO	58	0.804215
18	RIOHACHA	57	0.790349
19	BARRANCABERMEJA	49	0.679423
20	OTROS	2993	41.500277

Ciudad de Destino

	Ciudad	Frecuencia	Porcentage
0	BOGOTÁ, D.C.	1302	18.053245
1	RIONEGRO - ANTIOQUIA	480	6.655574
2	CARTAGENA DE INDIAS	363	5.033278
3	MITÚ	277	3.840821
4	SANTIAGO DE CALI	197	2.731559
5	BARRANQUILLA	167	2.315585
6	VILLAVICENCIO	125	1.733222
7	BUCARAMANGA	106	1.469773
8	PEREIRA	105	1.455907
9	SAN JOSÉ DE CÚCUTA	93	1.289517
10	SANTA MARTA	89	1.234054
11	MEDELLÍN	84	1.164725
12	SAN ANDRES - ISLA	76	1.053799
13	NEW YORK	62	0.859678
14	MIAMI	60	0.831947
15	ARMENIA	56	0.776484
16	CHICAGO	44	0.610094
17	SAO PAULO	43	0.596229
18	QUITO	43	0.596229
19	BOSTON	41	0.568497
20	OTROS	3399	47.129784





Bogotá se posiciona como el principal centro de operaciones aéreas tanto en vuelos de origen como de destino durante el mes de marzo de 2025. En la variable "Ciudad de Origen", concentra 1.330 vuelos, lo que representa el 18.44 % del total, muy por encima de Cartagena (6.46 %) y Rionegro - Antioquia (5.50 %). En cuanto a la "Ciudad de Destino", Bogotá también lidera con 1.302 vuelos (18.05 %), seguida nuevamente por Rionegro (6.66 %) y Cartagena (5.03 %).

Entre los destinos internacionales más frecuentes estan ciudades como Nueva York, Miami, Chicago, São Paulo, Quito y Boston, que en conjunto representan aproximadamente el 4 % del total de vuelos. Aunque esta proporción es menor, refleja los principales enlaces internacionales del país.

Además, las 20 ciudades más frecuentes en la variable "Ciudad de Origen" representan el 59 % del total de registros, mientras que en la variable "Ciudad de Destino" este valor es del 53 %. Esto indica una alta concentración de la actividad aérea en pocos puntos, lo cual es característico de un sistema dominado por hubs. Sin embargo, también se observa una “cola larga” de rutas secundarias: muchas ciudades con baja frecuencia de vuelos que, en conjunto, representan casi la mitad del flujo aéreo total. Esta distribución sugiere que, aunque hay pocos centros con alto tráfico, existe una dispersión considerable hacia ciudades menos concurridas.

Variable	Top 20 (%)	“Otros” (%)
Origen	59 %	41 %
Destino	53 %	47 %

TENDENCIA CENTRAL ---- DISPERSION

(PASAJEROS, CARGA (KG))

"Aqui se tuvo en cuenta la variable completa sin division "

Pasajeros

```
Marzo/2025
--------------------------------
La media de pasajeros en ruta-vuelo es de : 628.07
La desviación estándar de los pasajeros es de : 2920
La varianza de los pasajeros fue de : 8527362
El coeficiente de variación fue de : 464.9%
```

El análisis de los pasajeros por ruta de vuelo en marzo de 2025 muestra una media de 628 pasajeros, valor que parece moderado pero que oculta una dispersión extrema en los datos. Esta irregularidad se confirma con un coeficiente de variación (CV) del 464.9%, muy superior al 100%, lo que indica que la variabilidad es 4.65 veces mayor que el promedio (casi cinco veces). Además, la varianza de 8,527,362 calculada

como el cuadrado de la desviación estándar (2,920²) refuerza esta interpretación, ya que un valor tan elevado revela que los datos están extremadamente alejados de la media.

La desviación estándar de 2,920 pasajeros señala que, en promedio, las rutas se desvían ±2,920 pasajeros respecto a la media. Esta enorme brecha sugiere una distribución bimodal: mientras algunas rutas, como la de la aerolinea avianca Bogotá-San Andrés, que llego a acumular hasta 13,160 pasajeros por ruta de vuelo en el mes de marzo ,otras registraron cero pasajeros. Esta dualidad se debe a:

1. *Rutas de carga--Rutas comerciales: Vuelos dedicados exclusivamente al transporte de mercancías, donde los pasajeros son irrelevantes VS rutas de vuelos muy concurridas "turisiticas "en las que se lleva un arto flujo de pasajeros al mes .*

2. *Baja demanda / Alta demanda estructural: Rutas con poca atracción comercial o geográfica VS rutas que por su naturaleza son muy demandadas*

CARGA (KG)

```
Marzo/2025
--------------------------------
La media de Carga(kg) en ruta-vuelo es de : 10790
La desviación estándar de Carga(kg) es de : 118972
La varianza de Carga(kg) fue de : 14154429265
El coeficiente de variación fue de : 1102.6%
```

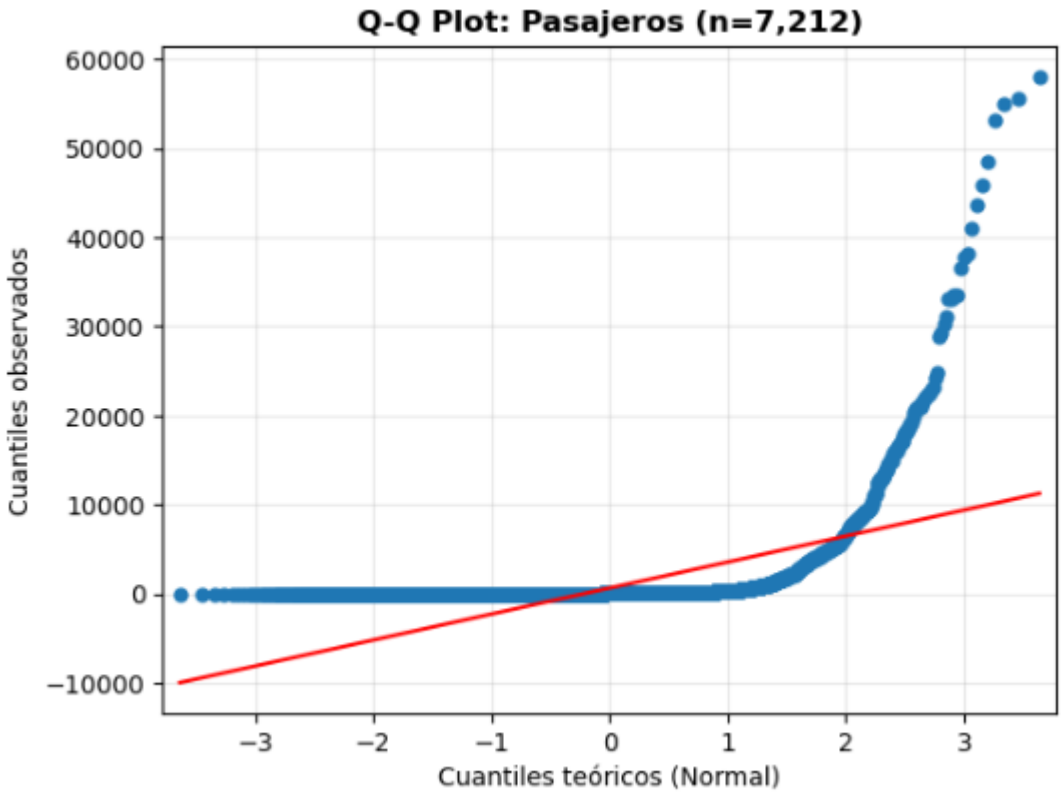
- La media de carga fue de 10.790 toneladas por ruta de vuelo pero tambien la desviacion estandar fue de 118.972 toneladas lo que indica que algunas rutas de vuelo transportan cargas enormes mientras que otras llevan muy poco o nada.esta variable a comparacion de la variable pasajeros tiene una dispersion/variabilidad relativa mas alta esto lo confirmamos con el coeficiente de variacion “CV” al tener 1102 % de variacion en comparacion al 464.9 % de la variable pasajeros la variable “Carga kg muestra más del doble de variabilidad relativa que la variable pasajeros”.la varianza de carga tambien fue muy elevada con 14.154.429.265,mirando esto podemos deducir que hay rutas de vuelos de carga pura en comparacion con rutas de vuelos sin casi nada o nula la carga esta afirmacion se confirma con el analisis que se hizo en los anteriores puntos mencionados en el que una carga de ruta de vuelo por el mes de marzo de 0-71 kg equivale al 71 % del total de los datos y apartir de ahí hasta llegar a 5.21365 M equivaldria al restante 29 %

Validacion De Supuestos (ANOVA)

Normalidad

Primero vamos a utilizar la grafica de residuos qqplot que nos sirve para comparar los quantiles de nuestra variable (linea azul)con los quantiles teoricos que llevaria una variable normal perfecta (linea roja)

Pasajeros



El Q-Q plot confirma que los datos de pasajeros no son normales. Los puntos azules (quantiles reales) se desvían drásticamente de la línea roja (quantiles teóricos de una normal), especialmente en los extremos. Esta divergencia es tan evidente que descarta cualquier posibilidad de normalidad esto se debe a la asimetria y los outliers anteriormente diagnosticados en esta variable.

El otro test de normalidad que se escogio fue el de anderson-darling al no tener un límite superior en total de registros: funciona bien en muestras muy grandes sin “romperse”

Estadistica A^2 > valor critico = Rechazamos H_0: los datos no siguen una distribución normal

```
Test de anderson-darling
------------------------
Estadístico A² = 2049.1027
Valor crítico al 5% = 0.7870
→ Rechazamos H₀: los datos no siguen una distribución normal (al 5%)
```

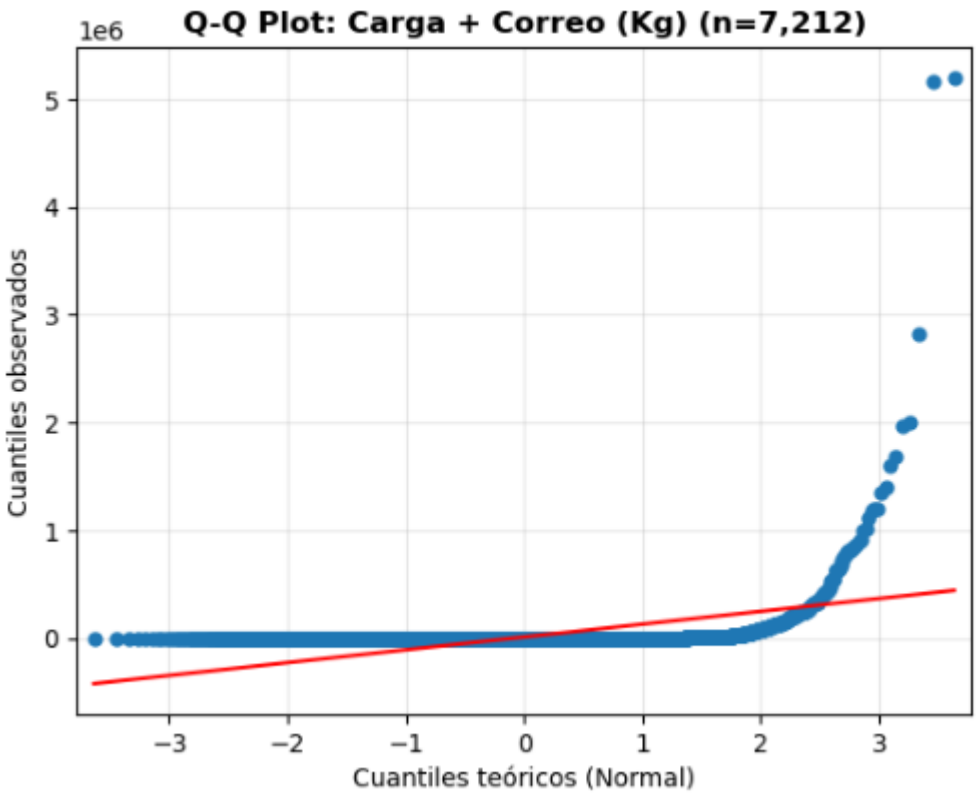
Aunque este test de normalidad es confiable y con la grafica qqplot verificamos que no hay normalidad este test anderson darling se vuelve cada vez mas sensible a cualquier desviacion minima según la cantidad de registros en mi caso como tengo 7212 registros y sabiendo que la variable Pasajeros tiene alta variabilidad terminamos de hacer el diagnostico de esta variable con el test de

normalidad Jarque–Bera este es un test que se centra en la asimetría (skewness) y curtosis (kurtosis) lo cual es ideal para la variable pasajeros al tener una distribucion tan dispareja(mucha variabilidad y outliers)

Valor p < nivel de significancia : Rechazamos H_0: los datos NO siguen una distribución normal

```
Test jarque bera
----------------
Valor-p: 0.0000
Significancia:  0.05
→ Rechazamos H₀: los datos NO siguen una distribución normal (p < 0.05)
```

Carga (KG)



Lo mismo ocurre con esta variable “Carga (kg)” al ser una variable con muchisima variabilidad e incluso mas del doble en variabilidad relativa con respecto a pasajeros la cantidad de outliers es grande en esta variable como lo vimos en los boxplots lo cual impide que esta variable lleve una distribucion normal

En este caso solo vamos a aplicar el test de normalidad jarque bera por ser ideal para una variable con tanta variabilidad, outliers sin dejar atrás la cantidad de registros que se tiene

```
Test jarque bera
----------------
Valor-p: 0.0000
Significancia:  0.05
→ Rechazamos H₀: los datos NO siguen una distribución normal (p < 0.05)
```

Valor p < nivel de significancia : Rechazamos H_0: los datos NO siguen una distribución normal

heterocedasticidad

p-value < 0.05 → Rechaza H_0 → Se concluye que hay diferencias significativas de varianza entre los grupos.

Aquí se aplica el test de levene para mirar si hay diferencias significativas de varianza entre los grupos y dependiendo los reultados se obtaria por usar un anova clasico o un anova welch

```
Levene test para Ciudad de origen  - p-value: 0.0000
Levene test para Ciudad de destino  - p-value: 0.0000
```

$$0.00 < 0.05$$

La prueba de hipotesis nos concluye que hay evidencia muy fuerte para rechazar la igualdad de varianzas → heterocedasticidad. ----- ANOVA WELCH

Conclusiones Parciales

el diagnóstico descriptivo nos deja claro la enorme variabilidad y el sesgo de la variable pasajeros que es la variable objetivo en las proyecciones estimadas por la aeronautica civil , esta variable no nos aportaria normalidad, Sin embargo, gracias al Teorema del Límite Central (TLC), podemos asumir normalidad aproximada en la distribución de las medias muestrales siempre que los grupos analizados tengan un tamaño mayor a 30. Para asegurar el cumplimiento de este criterio, se procedió a reagrupar las ciudades de origen y destino (variables consideradas en las proyecciones de la aeronautica civil) de modo que aquellas con una frecuencia menor a 30 observaciones fueran etiquetadas como "Otros". Esto garantiza que cada grupo muestral cumpla con los requisitos del TLC, tambien para las proyecciones se va a utilizar el ANOVA de Welch, el cual es más robusto frente a varianzas desiguales que es el caso de nuestras variables de estudio.

La ciudad de origen o destino determina el numero de pasajeros?

Ciudad de origen

```
welch_anova=pg.welch_anova(dv='Pasajeros',between='Ciudad Origen',data=df)
print(welch_anova.to_string(float_format='{:.4f}'.format)) # cambiando el numero de decimales
```

```
          Source  ddof1    ddof2       F  p-unc    np2
0  Ciudad Origen     31 726.7524 13.9143 0.0000 0.0205
```

Ciudad de destino

```
welch_anova_2=pg.welch_anova(dv='Pasajeros',between='Ciudad Destino',data=df)
print(welch_anova.to_string(float_format='{:.4f}'.format))
```

```
          Source  ddof1    ddof2       F  p-unc    np2
0  Ciudad Origen     31 726.7524 13.9143 0.0000 0.0205
```

Utilizando el anova welch miramos que la variabilidad que explica la variable ciudad de origen, ciudad de destino es la misma en la variable dependiente pasajeros con solo el 2 % (np2 --- 0.0205)

Pero si usamos el p-value (p-unc 0.0000) “ambas variables dieron el mismo p-value” para hacer una prueba de hipotesis y mirar si la variabilidad que explican estas variables son estadisticamente significativas :

0.0000 < 0.05 = se rechaza la hipotesis nula H0 entonces se diria que estas variables en cuestion tienen una gran significancia en la variacion de la variable dependiente Pasajeros

Aunque la magnitud del efecto es baja, la influencia de la Ciudad de Origen,Ciudad de Destino sobre el número de pasajeros es estadísticamente significativa.
Es decir, sí existe una relación entre la ciudad de origen, Ciudad de Destino y el número de pasajeros, pero esta no es muy fuerte en términos de variabilidad explicada (solo un 2 %) lo que implica que, aunque estas variables tienen un impacto real, su capacidad para explicar la variación en el número de pasajeros es limitada.

Proyecciones estimadas por la aeronautica civil

1. La aeronautica civil quiere primero saber si el promedio de numero de pasajeros para el proximo mes con un 99 % de confianza es de 800

```
st.t.interval(confidence=0.99,df=len(df)-1,loc=np.mean(df["Pasajeros"]),scale=st.sem(df["Pasajeros"]))
(539.4721915849736, 716.6631384205727)
```

Utilizando intervalos de confianza con t-studen los resultados que dieron en cuanto el promedio de pasajeros con un 99 % de confianza es de 539-716 pasajeros lo cual no cumple o no entra en el rango de lo proyectado por la aeronautica civil de 800 pasajeros entonces se podria esperar que para el siguiente mes se mueva el promedio de numero de pasajeros entre 539-716

2. La aeronautica civil quiere saber si se puede llegar a un promedio de 900 pasajeros el otro mes si no se tiene en cuenta los vuelos con numero de pasajeros 0

```
st.t.interval(confidence=0.99,df=len(df)-1,loc=np.mean(df["Pasajeros"]),scale=st.sem(df["Pasajeros"]))
```

```
(583.3113911546966, 774.4895680779173)
```

Eliminando todos los vuelos con numero de pasajeros 0 el nuevo promedio en intervalo con t-student es de 583-774 pasajeros con un 99 % de confianza lo cual no entra entre lo estipulado,proyectado de 900 pasajeros para el siguiente mes

CONCLUSIONES FINALES

La alta variabilidad presente en la variable cuantitativa *Pasajeros* ha impactado negativamente las proyecciones establecidas por la Aeronáutica Civil. Incluso después de excluir los vuelos con cero pasajeros, no se alcanza la proyección estimada de pasajeros promedio para el próximo mes.

Por otro lado, aunque las variables *Ciudad de Origen* y *Ciudad de Destino* resultaron ser estadísticamente significativas para explicar la cantidad de pasajeros por vuelo, su poder explicativo es bajo, ya que únicamente explican alrededor del 2% de la variabilidad de la variable Pasajeros.

Por lo tanto, se recomienda considerar la incorporación de otras variables adicionales que puedan tener una mayor capacidad explicativa sobre el número de pasajeros por vuelo y ajustar las proyecciones teniendo en cuenta el rango en promedio de la variable pasajeros "539-716".

BIBLIOGRAFIAS

1. Teorema del Límite Central (TLC)

Field, A. (2018). *Discovering statistics using IBM SPSS Statistics* (5th ed.). Sage.

- **Relevancia:** Capítulo 2 explica el TLC y su aplicación en pruebas paramétricas, incluyendo condiciones para su uso (e.g., tamaño muestral ≥ 30).
- **Enlace:** DOI: 10.4135/9781526445781

2. ANOVA de Welch

Delacre, M., Lakens, D., & Leys, C. (2017). Why psychologists should by default use Welch's *t*-test instead of Student's *t*-test. *International Review of Social Psychology*, *30*(1), 92-101.

- **Relevancia:** Aunque enfocado en pruebas *t*, justifica el uso de versiones robustas como Welch para diseños con varianzas desiguales, extendible a ANOVA.
- **Enlace:** DOI: 10.5334/irsp.82

3. Validación de supuestos en ANOVA

Tabachnick, B. G., & Fidell, L. S. (2019). *Using multivariate statistics* (7th ed.). Pearson.

- Relevancia: Capítulo 4 detalla cómo verificar normalidad, homocedasticidad e independencia en ANOVA, con alternativas no paramétricas.
- Enlace: ISBN: 978-0134790541