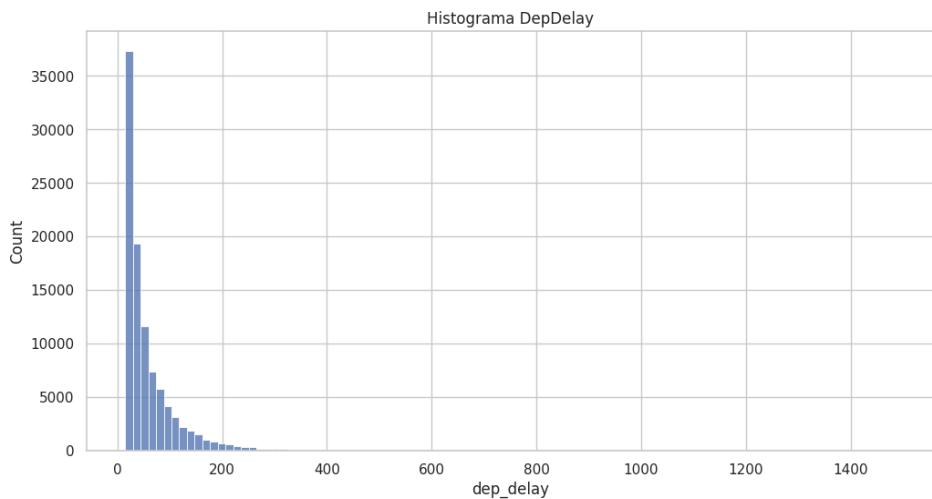


Gráficas exploratorias

1. Histograma DepDelay

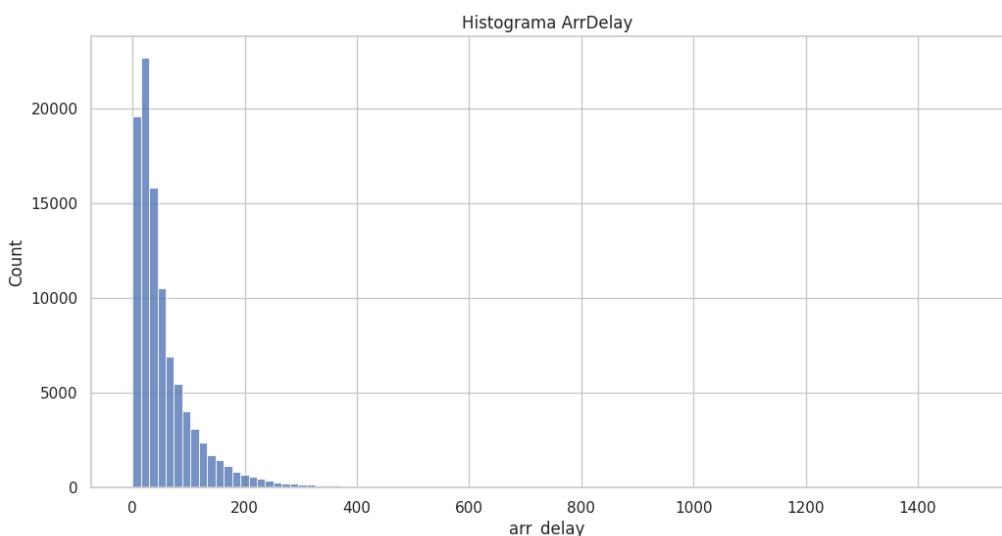
Caption: Distribución de los retrasos de salida (DepDelay); muestra la cola de retrasos y la presencia de outliers.



La distribución está fuertemente sesgada a la derecha: la mayoría de los vuelos presentan retrasos bajos, pero existe una cola larga de retrasos extremos (hasta >1,400 min). Esto confirma que la **mediana** es más representativa que la media y que unos pocos eventos extremos inflan las estadísticas agregadas

2. Histograma ArrDelay

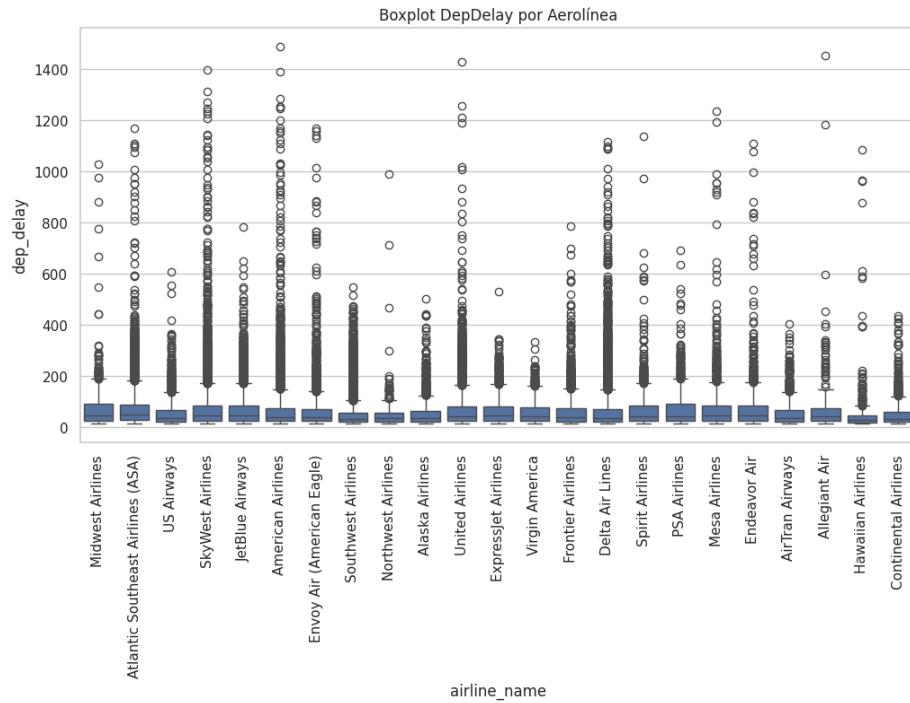
Caption: Distribución de los retrasos de llegada (ArrDelay); comparar forma y cola con DepDelay para ver transferencia de demora.



La distribución está fuertemente sesgada a la derecha, con la mayoría de vuelos con retrasos bajos y una cola de retrasos extremos (hasta >1,400 min). La alta correlación con dep_delay sugiere que los retrasos de salida se trasladan a la llegada en la mayoría de los casos.

3. Boxplot DepDelay por aerolínea

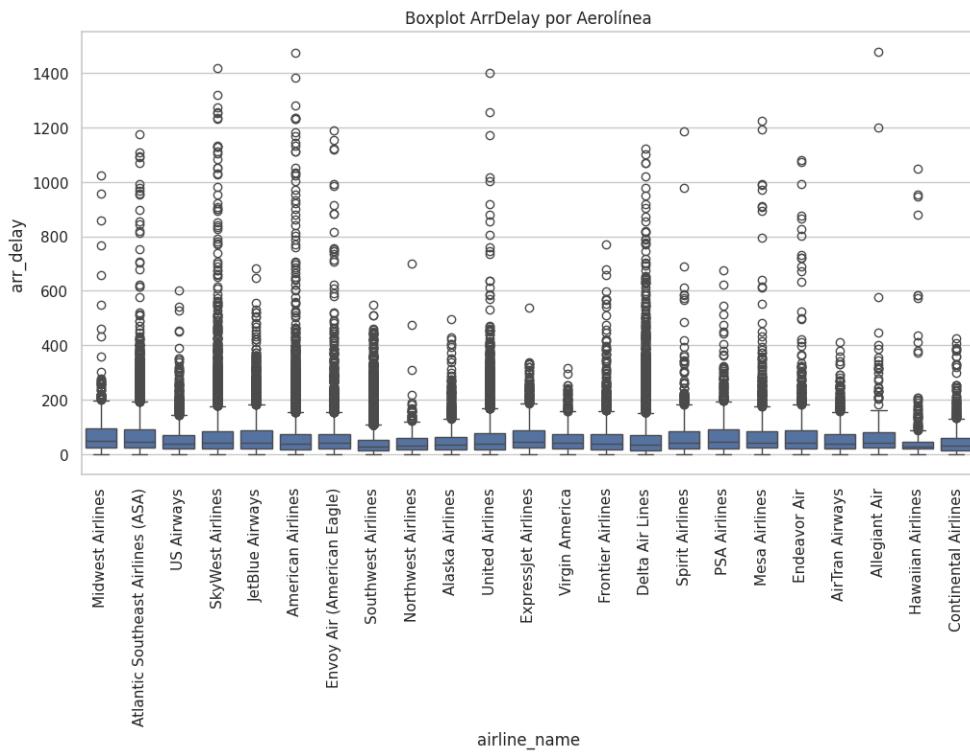
Caption: Mediana, IQR y outliers de DepDelay por aerolínea; útil para comparar severidad y variabilidad entre operadores.



Las cajas muestran que algunas aerolíneas tienen medianas de retraso notablemente más altas y/o mayor dispersión; la presencia de muchos outliers indica eventos extremos que afectan la variabilidad. Esto confirma que la comparación por medianas es más robusta que por medias y que hay operadores con problemas sistemáticos o rutas con mayor variabilidad.

4. Boxplot ArrDelay por aerolínea

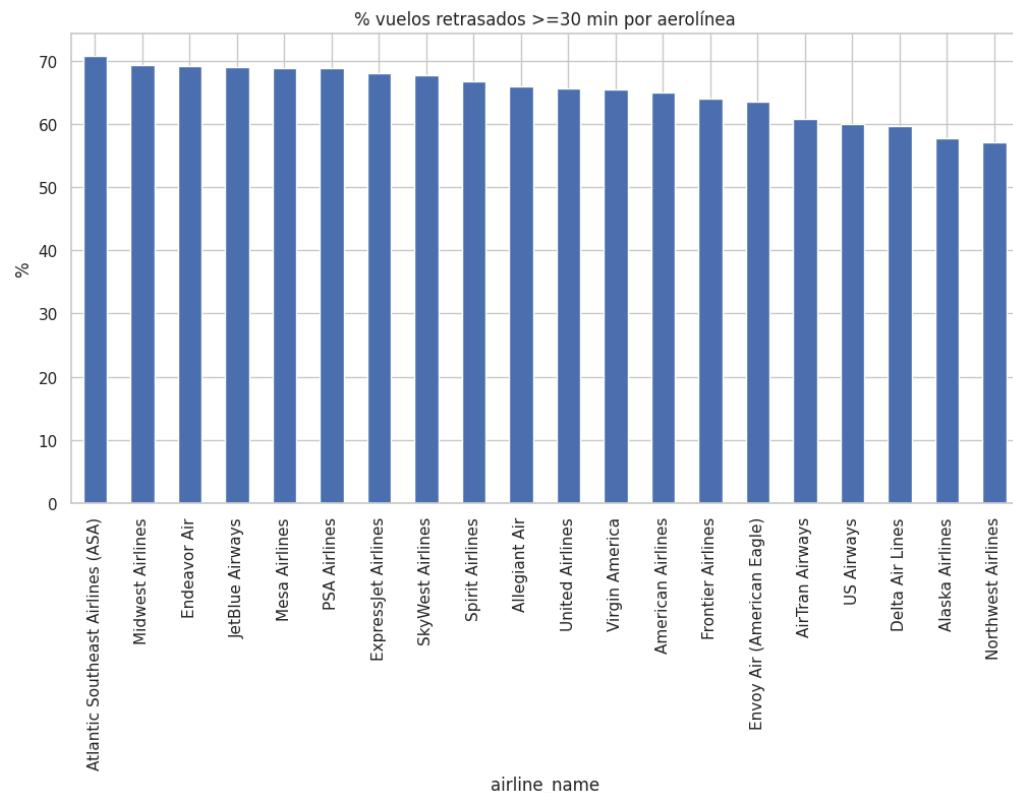
Caption: Mediana e IQR de ArrDelay por aerolínea; permite identificar aerolíneas con mayor dispersión en llegadas.



Las distribuciones muestran diferencias claras en mediana y dispersión entre aerolíneas; varias presentan colas largas de outliers en llegadas, lo que indica eventos extremos que afectan la puntualidad. La fuerte correlación con `dep_delay` sugiere que muchos retrasos de llegada se originan en la salida, aunque hay casos donde la llegada empeora o mejora respecto a la salida.

5. % vuelos con DepDelay \geq 30 min por aerolínea (Top 20)

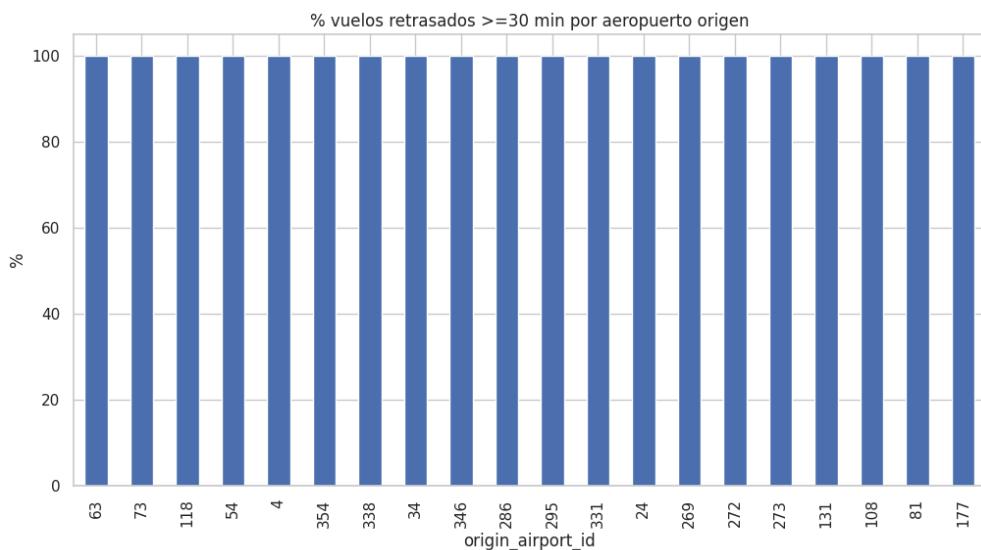
Caption: Frecuencia de retrasos relevantes (≥ 30 min) por aerolínea; orden descendente para priorizar acciones operativas.



Algunas aerolíneas presentan tasas de retraso significativamente más altas (p. ej. Atlantic Southeast Airlines (ASA) y Midwest Airlines en el top), mientras que otras muestran tasas mucho menores (p. ej. Northwest). Esto indica diferencias operativas o de red entre operadores; la variabilidad sugiere que no es un problema uniforme del sistema sino concentrado en ciertos carriers o rutas.

6. % vuelos con DepDelay \geq 30 min por aeropuerto origen (Top 20)

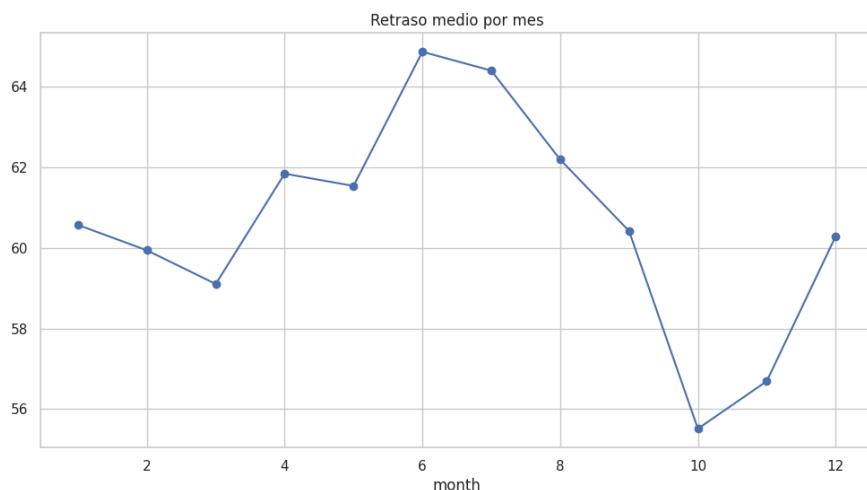
Caption: Incidencia de retrasos \geq 30 min por aeropuerto de origen; ayuda a localizar orígenes con mayor impacto relativo.



La mayoría de los aeropuertos en el top muestran tasas muy altas de retrasos \geq 30 min, lo que sugiere que el problema está concentrado en ciertos orígenes. Dado que `origin_code` textual está vacío en el dataset, los valores se muestran por `origin_airport_id`; es necesario mapearlos a códigos/nombres para análisis operativo y comunicación.

7. Retraso medio por mes

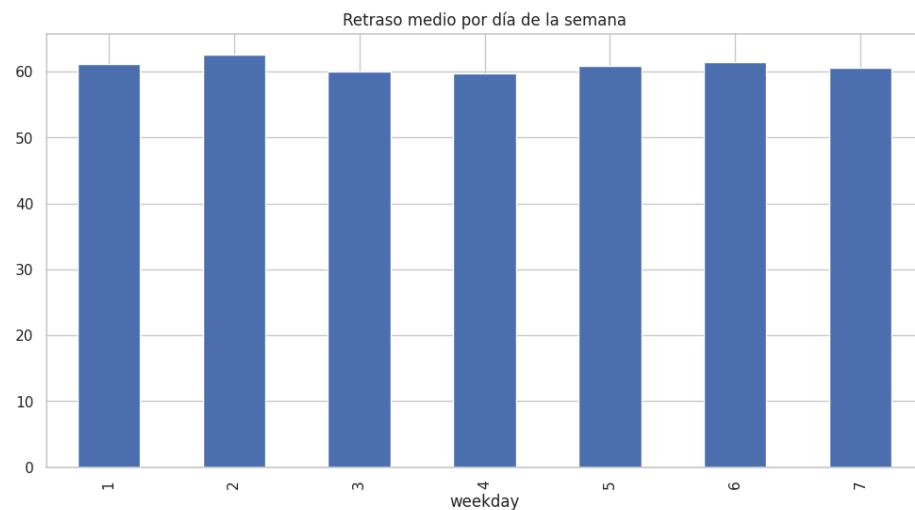
Caption: Evolución temporal de la media de DepDelay por mes; identifica estacionalidad y picos periódicos.



Se observan variaciones mensuales con un pico pronunciado en junio y una caída notable en octubre, lo que sugiere estacionalidad o eventos puntuales (picos de demanda, condiciones climáticas o cambios operativos) que afectan la puntualidad. Dado el sesgo por outliers, la media puede estar influida por eventos extremos.

8. Retraso medio por día de la semana

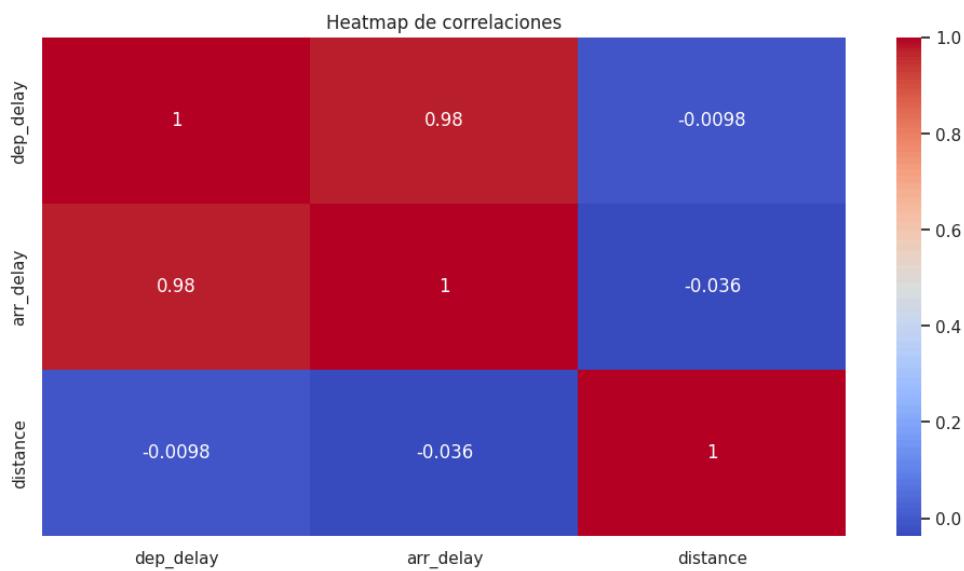
Caption: Promedio de DepDelay por weekday; revela patrones semanales y días críticos para operaciones.



Las diferencias entre días son moderadas; se observa un pico en el día 2 y valores más bajos en los días 3 y 4, lo que sugiere un patrón semanal leve. Dado el sesgo y la presencia de outliers, la media puede ocultar variaciones internas por franjas horarias o rutas.

9. Heatmap de correlaciones entre variables

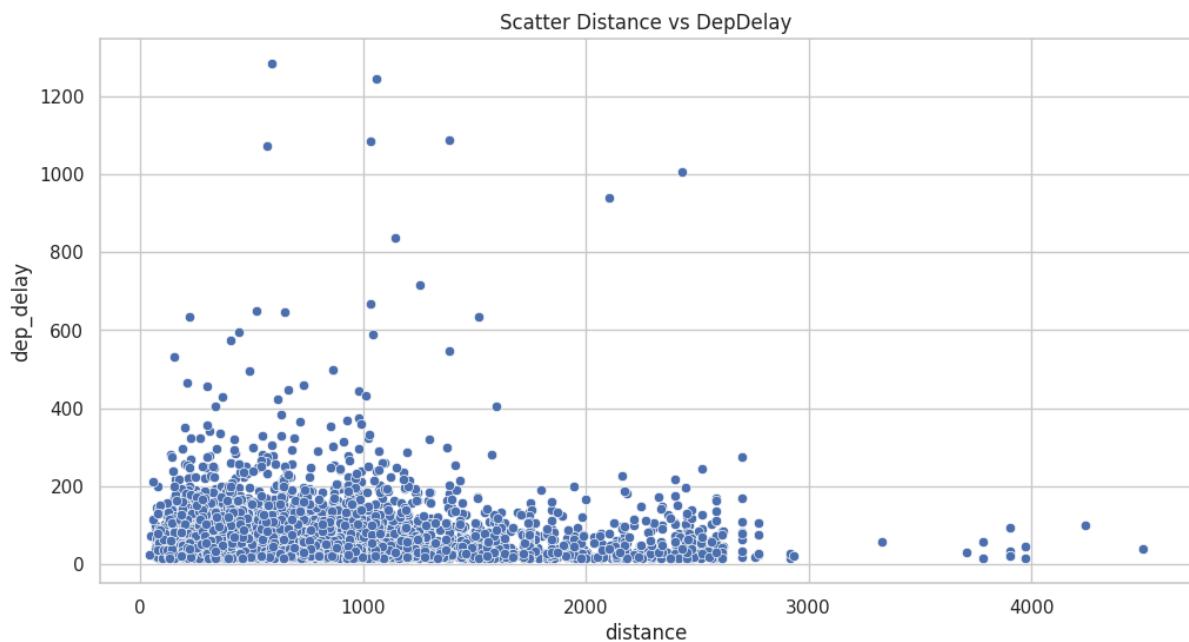
Caption: Matriz de correlaciones entre DepDelay, ArrDelay, Distance y otras variables numéricas disponibles; destaca relaciones fuertes y débiles.



Existe una **correlación muy alta** entre dep_delay y arr_delay (~0.98), lo que indica que los retrasos de salida se trasladan casi íntegramente a la llegada. La **relación con distance es prácticamente nula** (≈ -0.01 a -0.04), por lo que la longitud del vuelo no explica los retrasos en este conjunto de datos.

10. Scatter Distance vs DepDelay (muestra)

Caption: Relación entre distancia del vuelo y DepDelay en una muestra aleatoria; incluye tendencia para evaluar dependencia.



No se observa una relación lineal clara: la mayoría de los vuelos con retrasos bajos se concentran en distancias cortas y medias, y la distancia no explica la variabilidad de los retrasos salvo algunos outliers. Esto confirma el bajo coeficiente de correlación entre distance y dep_delay.