Resumen de calidad de dato

Hallazgos principales

- **Columnas de aeropuerto faltantes:** origin_code, origin_name, origin_city, origin_state, dest_code, dest_name, dest_city, dest_state están completamente vacías (100 000 filas).
- **Retrasos de llegada y total con nulos:** arr_delay y total_delay tienen 606 valores nulos.
- **Indicadores de retraso:** DepDel15, DepDel30, DepDel60 existen y no tienen nulos; DepDel15 no discrimina porque el mínimo de dep_delay es 15.
- **Outliers significativos:** 722 vuelos con dep_delay > 360 min y 711 con arr_delay > 360 min.
- **Integridad de claves y duplicados:** no hay duplicados (0) y las claves (airline_id, origin_airport_id, destination_airport_id, fecha_id) están completas.
- **Distance:** no hay distancias cero (Distance == 0: 0).

Se detectaron 606 nulos en arr_delay/total_delay y ~720 outliers en retrasos > 6 horas; las medias están afectadas por estos extremos.
Se crearán y reportarán DepDel30 y DepDel60; se priorizará el uso de medianas e IQR y se extraerán outliers para investigación operativa.

Acciones recomendadas (priorizadas)

1. **Rellenar datos de aeropuerto:** mapear origin_airport_id y destination_airport_id a dim_airport para recuperar airport_code y airport_name y reemplazar las columnas vacías.
2. **Revisar nulos en arr_delay/total_delay:** inspeccionar flight_status para las 606 filas nulas y decidir imputación o exclusión según caso (cancelled/diverted vs error ETL).
3. **Investigar outliers > 360 min:** extraer top outliers, revisar causas (flight_status, demoras por causa operativa o meteorológica) y documentar si deben excluirse de estadísticas agregadas.
4. **Usar métricas robustas:** reportar mediana y percentiles por aerolínea/aeropuerto; usar DepDel30 y DepDel60 para frecuencia informativa.
5. **Documentar transformaciones:** anotar en el diccionario las imputaciones, umbrales usados y criterios de exclusión.

KPIs sugeridos para reportes de calidad y operación

- % vuelos ≥ 30 min (pct_del30) y % vuelos ≥ 60 min (pct_del60) por aerolínea y aeropuerto.
- Mediana dep_delay por aerolínea y por aerolínea×origen.
- Número de outliers > 360 min por aerolínea.
- % filas con arr_delay nulo y su relación con flight_status.