Ficha técnica: Proyecto 4 de Análisis de Datos

Título del proyecto: Flight Delay and Cancellation

Objetivo:

Analizar y predecir los retrasos de vuelos mediante técnicas de análisis de datos como el riesgo relativo, y regresión lineal, para identificar rutas, aeropuertos, y aerolíneas con alta frecuencia de demora, entender las principales causas de estos retrasos, y mejorar la toma de decisiones proactivas en la gestión de vuelos.

Detalles del Objetivo:

1. Calcular el Riesgo Relativo de Retrasos por Rutas, Aeropuertos y Aerolíneas.
2. Identificar las Principales Causas de las Demoras.
3. Predecir el Tiempo de Retraso utilizando Regresión Lineal.

Equipo:
Trabajo en dupla.

Herramientas y Tecnologías:

- Google BigQuery
- Google Colab
- Google Slides
- Google Looker Studio
- Tableau

Lenguajes:

- SQL
- Python

Insumos:

- Conjunto de datos aquí

Diccionario de datos:

Tabla 1: DOT_CODE_DICTIONARY

- **Code:** Identificador numérico del U.S. Department of Transportation (DOT) para aerolíneas
- **Description:** descripción de la aerolínea

Tabla 2: AIRLINE_CODE_DICTIONARY

- **Code:** Código de operador único para agencias operadoras de aeronaves
- **Description:** descripción de la agencia operadora de aeronaves

Tabla 3: flights_202301

- **FL_DATE:** Fecha de vuelo (yyyymmdd)
- **AIRLINE_CODE:** Código de operador único. Cuando varios operadores han utilizado el mismo código, se utiliza un sufijo numérico para usuarios anteriores, por ejemplo, PA, PA(1), PA(2).
- **DOT_CODE:** Un número de identificación asignado por el DOT de EE. UU. para identificar una aerolínea (transportista) única. Una aerolínea (transportista) única se define como aquella que posee y reporta bajo el mismo certificado DOT independientemente de su código, nombre o compañía/corporación holding.
- **FL_NUMBER:** Número de vuelo
- **ORIGIN:** Aeropuerto de origen
- **ORIGIN_CITY:** Aeropuerto de origen, nombre de la ciudad
- **DEST:** Aeropuerto de destino
- **DEST_CITY:** Aeropuerto de destino, nombre de la ciudad
- **CRSDEPTIME:** Hora de salida registrada CRS (Sistema de control de reservas) (hora local: hhmm)
- **DEP_TIME:** Hora de salida real (hora local: hhmm)
- **DEP_DELAY:** Diferencia en minutos entre la hora de salida prevista y la real. Las salidas anticipadas arrojan cifras negativas.
- **ARR_DELAY:** Diferencia en minutos entre la hora de llegada prevista y la real. Las llegadas anticipadas arrojan cifras negativas.
- **TAXI_OUT:** Tiempo de taxi en la salida en minutos (taxi es el proceso de mover un avión mientras se encuentra en la pista)
- **WHEELS_OFF:** hora exacta de despegue (hora local: hhmm)
- **WHEELS_ON:** hora exacta de aterrizaje (hora local: hhmm)
- **TAXI_IN:** tiempo de taxi en la llegada en minutos
- **CRS_ARR_TIME:** Hora de llegada registrada en CRS (hora local: hhmm)
- **ARR_TIME**: Hora de llegada real (hora local: hhmm)
- **CANCELLED:** Indicador de vuelo cancelado (1=Sí)
- **CANCELLATION_CODE:** Especifica el motivo de la cancelación

- **DIVERTED:** Indicador de vuelo desviado (1=Sí)
- **CRSELAPSEDTIME:** Tiempo total de vuelo transcurrido en minutos registrado en CRS
- **ELAPSED_TIME:** Tiempo total de vuelo transcurrido en minutos real
- **AIR_TIME:** Tiempo de vuelo en el aire en minutos
- **DISTANCE:** Distancia entre aeropuertos (millas)
- **DELAY_DUE_CARRIER:** Retraso del operador en minutos
- **DELAY_DUE_WEATHER**: Retraso meteorológico en minutos
- **DELAY_DUE_NAS**: Retraso del Sistema Aéreo Nacional en Minutos
- **DELAY_DUE_SECURITY:** Retraso de seguridad en minutos
- **DELAYDUELATE_AIRCRAFT**: Retraso de aeronaves tardías en minutos.

Procesamiento y análisis:

3. Procesar y preparar la base de datos

1. **Conectar/importar datos a otras herramientas**

Se creó el proyecto4-flightdelay y el conjunto de datos Dataset en BigQuery.

- Tablas importadas: *DOT_CODE_DICTIONARY, AIRLINE_CODE_DICTIONARY y flights_202301.*

2. **Identificar y manejar valores nulos**

Se identifican valores nulos a través de comandos SQL COUNTIF, IS NULL, AS.

- **DOT_CODE_DICTIONARY:** 4 valores nulos.
    - Se separó la columna Description, en Name y Description.
    - Se eliminaron los siguientes códigos porque no hay información en la columna descripción, 22114, 22115, 22116, 22117.

- **AIRLINE_CODE_DICTIONARY:** 0 valores nulos.
  Se cambió el encabezado de la tabla de string_field_0 a AIRLINE_CODE, y de string_field_1 a NAME_AIRLINE, generando una vista de la tabla view_airline_code

- **flights_202301:**
    - DEP_TIME: 9,978 vuelos cancelados.
    - TAXI_OUT: 10,197 vuelos cancelados algunos tienen DEP_TIME y DEP_DELAY.

    - WHEELS_OFF: 10,197 son vuelos cancelados algunos tienen DEP_TIME y DEP_DELAY.
    - WHEELS_ON: 10,517 son vuelos cancelados.
    - TAXI_IN: 10,517 son vuelos cancelados.
    - ARR_TIME: 10,517 son vuelos cancelados.
    - ARR_DELAY: 11,640 son vuelos cancelados o desviados.
        - Vuelos cancelados=10,295
        - Vuelos desviados= 1,345
    - CRS_ELAPSED_TIME: 1 vuelo cancelado
    - ELAPSED_TIME: 11,640 vuelo cancelado o desviado
    - AIR_TIME: 11,640 vuelo cancelado o desviado
    - DELAY_DUE_CARRIER: 422,124 son los vuelos que llegan a tiempo.
    - DELAY_DUE_WEATHER: 422,124 son los vuelos que llegan a tiempo.
    - DELAY_DUE_NAS: 422,124 son los vuelos que llegan a tiempo.
    - DELAY_DUE_SECURITY: 422,124 son los vuelos que llegan a tiempo.
    - DELAY_DUELATE_AIRCRAFT: 422,124 son los vuelos que llegan a tiempo.
- Se IMPUTARON los valores nulos con el valor de 0 ya que son datos que no se obtuvieron debido a que los vuelos fueron cancelados o desviados, esto nos permite mantener la consistencia de la naturaleza de los datos.
- Para las variables DELAY_DUE_CARRIER, DELAY_DUE_WEATHER, DELAY_DUE_NAS, DELAY_DUE_SECURITY, DELAY_DUELATE_AIRCRAFT, se IMPUTARON colocando el valor de 0, ya que el número de datos nulos corresponde a los vuelos que llegaron a tiempo a su destino. Los valores nulos no indican falta de datos sino la ausencia de un evento, por ejemplo, retraso por clima, retrasó por la aerolínea.

3. Identificar y manejar valores duplicados

Se identifican valores duplicados a través de comandos SQL COUNT, GROUP BY, HAVING.

- **DOT_CODE_DICTIONARY:** no hay valores duplicados.
- **AIRLINE_CODE_DICTIONARY:** no hay valores duplicados.
- **flights_202301:** no hay valores duplicados.

4. Identificar y manejar datos fuera del alcance del análisis

Se manejan variables que no son útiles para el análisis a través de comandos SQL SELECT EXCEPT.

- Se excluyen del análisis las variables DELAY_DUE_CARRIER, DELAY_DUE_WEATHER, DELAY_DUE_NAS, DELAY_DUE_SECURITY,

DELAY_DUELATE_AIRCRAFT, porque tienen muchos datos nulos, en su lugar se utilizan las variables en las que se imputaron los datos.

5. **Identificar y manejar datos discrepantes en variables categóricas**

- **DOT_CODE_DICTIONARY:** Separamos la columna Description, en Name y Description, utilizando las fórmulas =LEFT(A1, FIND("/", A1) - 1) y =RIGHT(A1, LEN(A1) - FIND("/", A1)) en google sheets.
- **AIRLINE_CODE_DICTIONARY:** Se cambió el encabezado de la tabla de string_field_0 a AIRLINE_CODE, y de string_field_1 a NAME_AIRLINE, generando una vista de la tabla view_airline_code, con los comandos WHERE y AS.
- Con los comandos REGEXP_CONTAINS, WHEN, ELSE, CASE, END, se comprobó la presencia o ausencia de caracteres especiales en las variables categóricas, y en caso de tenerlos que fueran adecuados para su definición.

6. **Identificar y manejar datos discrepantes en variables numéricas (OUTLIERS)**

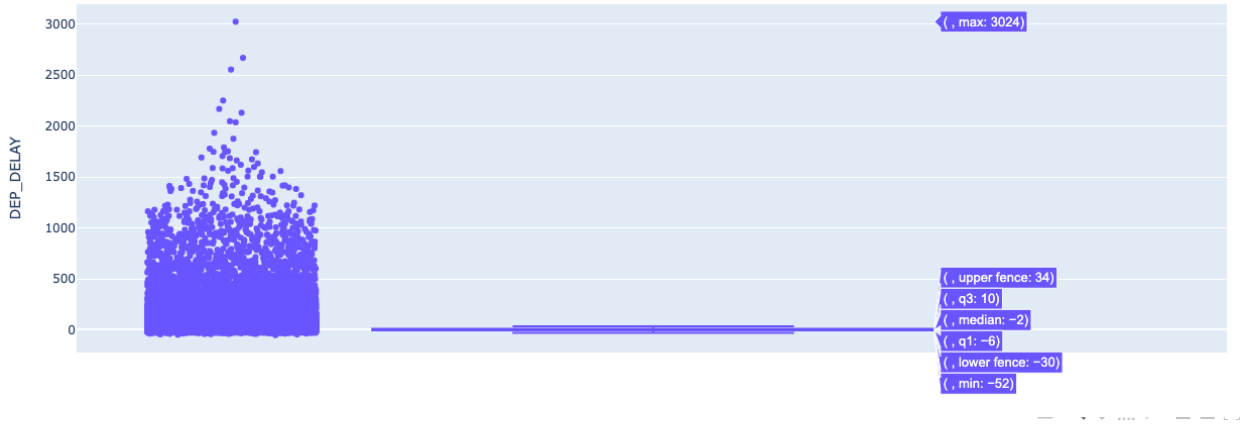
Con los comandos WITH, APPROX_QUANTILES, CASE, WHEN, ELSE, WHERE, se identifican los datos outliers de la tabla flights_202301. Se utilizó la metodología de rango intercuartil.

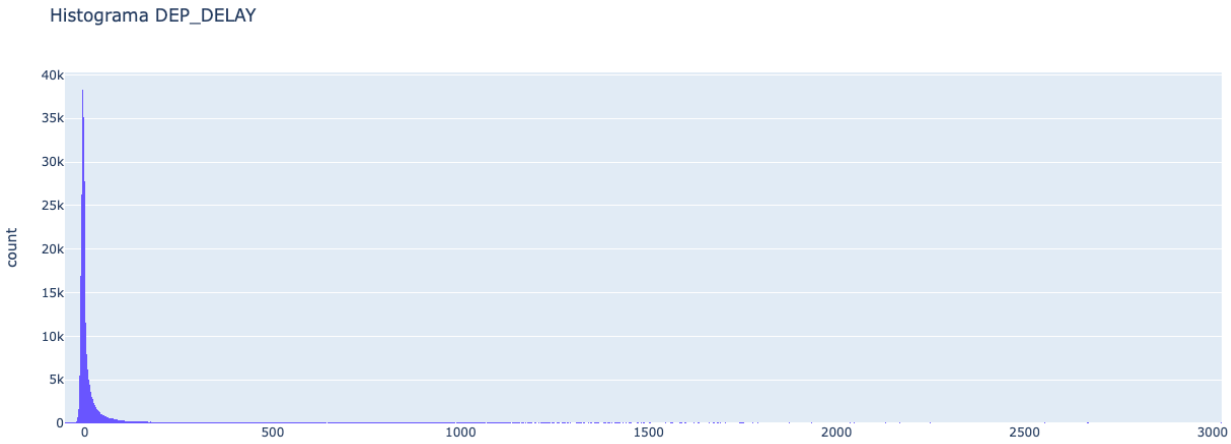
Variable	Outliers
DEP_DELAY	69,651
ARR_DELAY	47,759
DELAY_DUE_WEATHER	6,507
DELAY_DUE_NAS	8,437
DELAY_DUE_SECURITY	626
DELAY_DUE_LATE_AIRCRAFT	12,750
DELAY_DUE_CARRIER	12,005
TAXI_OUT	35,651
TAXI_IN	36,666
AIR_TIME	27,334
DISTANCE	32,257

- Se realizaron box plots e histogramas en google colab usando python para visualizar mejor los resultados encontrados.

DEP_DELAY

Row	CRS_DEP_TIME	DEP_TIME	DEP_DELAY	TAXI_OUT	WHEELS_OFF	WHEELS_ON	TAXI_IN	CRS_ARR_TIME
1	645	909	3024	71	1020	1130	6	833
2	1120	749	2669	29	818	1007	20	1340
3	1400	834	2554	11	845	1620	16	2159
4	930	2300	2250	21	2321	546	3	1640
5	700	1908	2168	*null*	*null*	*null*	*null*	1204
6	945	2115	2130	12	2127	2300	15	1237
7	800	1807	2047	34	1841	2007	9	955
8	905	1901	2036	27	1928	2355	5	1413
9	1347	2200	1933	18	2218	2349	5	1547
10	630	1345	1875	30	1415	1650	5	938
11	1635	2225	1790	16	2241	30	5	1900
12	635	1213	1778	12	1225	1440	5	949
13	1034	1546	1752	10	1556	2018	7	1510
14	1629	2136	1747	13	2149	2248	5	1804
15	1221	1725	1744	*null*	*null*	*null*	*null*	1624





ARR_DELAY

Row		DELAY_DUE_CARRIE	DELAY_DUE_WEATHI	DELAY_DUE_NAS	DELAY_DUE_SECURI	DELAY_DUE_LATE_A
1	07	3024	0	39	0	0
2	28	2669	0	18	0	0
3	54	2554	0	3	0	0
4	16	2229	0	0	0	0
5	44	2078	0	0	0	0
6	90	2047	0	14	0	0
7	35	0	0	0	0	2027
8	90	1807	0	0	0	120
9	14	1875	0	2	0	0
10	71	1775	0	0	0	0
11	73	0	0	3	0	1752
12	34	1740	0	0	0	0
13	73	1736	0	0	0	0
14	61	1729	0	0	0	0
15	29	1171	0	130	0	415

DELAY_DUE_WEATHER

Row	TANCE	DELAY_DUE_CARRIE	DELAY_DUE_WEATHI	DELAY_DUE_NAS	DELAY_DUE_SECURI	DELAY_DUE_LATE_A
1	311	0	1653	0	0	0
2	1187	0	1496	0	0	0
3	1371	0	1482	0	0	0
4	1334	0	1418	0	0	0
5	1334	0	1415	52	0	0
6	1415	0	1397	91	0	0
7	1162	0	1369	0	0	0
8	1415	0	1364	0	0	0
9	551	0	1217	0	0	0
10	1386	0	1163	66	0	0
11	1299	0	1154	12	0	0
12	659	0	1132	43	0	41
13	511	0	1125	103	0	0
14	468	0	1107	0	0	10
15	236	0	1100	0	0	0

DELAY_DUE_NAS

Row		DELAY_DUE_CARRIE	DELAY_DUE_WEATHI	DELAY_DUE_NAS	DELAY_DUE_SECURI	DELAY_DUE_LATE_A	FL
1	567	0	0	1343	0	0	
2	715	0	0	1079	0	0	
3	486	0	0	1060	0	8	
4	526	0	0	1005	0	0	
5	299	0	0	968	0	0	
6	754	0	0	931	0	0	
7	754	0	0	926	0	0	
8	299	0	0	894	0	0	
9	419	0	0	867	0	0	
10	733	0	0	844	0	0	
11	502	0	0	806	0	0	
12	581	17	0	795	0	0	
13	419	0	0	775	0	221	
14	175	0	0	770	0	0	
15	271	68	0	764	0	0	

DELAY_DUE_SECURITY

Row	DELAY_DUE_CARRIE	DELAY_DUE_WEATHI	DELAY_DUE_NAS	DELAY_DUE_SECURI	DELAY_DUE_LATE_A	FL_YEAR	FL_MONTH	FL_DAY
1	0	0	0	234	0	2023	1	11
2	0	0	0	197	0	2023	1	11
3	0	0	0	192	0	2023	1	11
4	0	0	0	189	0	2023	1	24
5	0	0	0	168	0	2023	1	28
6	0	0	0	168	0	2023	1	24
7	0	0	0	155	0	2023	1	11
8	0	0	9	139	0	2023	1	19
9	0	0	0	132	0	2023	1	11
10	0	0	23	132	30	2023	1	24
11	0	0	12	122	38	2023	1	24
12	0	0	0	121	0	2023	1	12
13	0	0	13	107	0	2023	1	26
14	0	0	0	106	0	2023	1	11
15	0	0	0	106	0	2023	1	11

DELAY_DUE_LATE_AIRCRAFT

Row	DISTANCE	DELAY_DUE_CARRIE	DELAY_DUE_WEATHI	DELAY_DUE_NAS	DELAY_DUE_SECURI	DELAY_DUE_LATE_A	FL
1	1235	0	0	0	0	2027	
2	1573	0	0	3	0	1752	
3	583	1	0	0	0	1560	
4	312	0	0	0	0	1486	
5	1187	64	0	22	0	1439	
6	940	0	0	0	0	1397	
7	651	0	0	0	0	1363	
8	769	0	0	0	0	1309	
9	1345	3	0	0	0	1216	
10	1242	72	0	0	0	1211	
11	1360	0	0	0	0	1206	
12	1107	9	0	0	0	1181	
13	152	0	0	0	0	1163	
14	2039	0	0	0	0	1152	
15	837	0	0	0	0	1146	

DELAY_DUE_CARRIER

Row	AIR_TIME	DISTANCE	DELAY_DUE_CARRIE	DELAY_DUE_WEATHI	DELAY_DUE_NAS	DELAY_DUE_SECURI	DELAY_DUE_LATE_A	FL_YEAR
1	70	507	3024	0	39	0	0	2023
2	49	328	2669	0	18	0	0	2023
3	275	2454	2554	0	3	0	0	2023
4	205	1916	2229	0	0	0	0	2023
5	213	1744	2078	0	0	0	0	2023
6	86	590	2047	0	14	0	0	2023
7	95	814	1875	0	2	0	0	2023
8	91	590	1807	0	0	0	120	2023
9	229	1671	1775	0	0	0	0	2023
10	54	234	1740	0	0	0	0	2023
11	195	1573	1736	0	0	0	0	2023
12	119	761	1729	0	0	0	0	2023
13	182	1264	1688	0	0	0	0	2023
14	51	312	1684	0	7	0	0	2023
15	72	569	1623	0	15	0	10	2023

7. Crear nuevas variables

- Con los comandos WHEN, CASE, ELSE se crearon las variables ***ESTATUS_VUELO*** para agregar las siguientes etiquetas de identificación “A TIEMPO”, “DEMORADO”, “CANCELADO”, “DESVIADO” a cada vuelo. Adicional se creó la variable ***ETIQUETA_NUM*** para asignar el valor de “1” a los vuelos demorados, “0” a los no demorados, “2” a los vuelos cancelados y “3” a los vuelos desviados.
- Con el comando IF se creó la variable ***CAUSAS_DEMORA***, que asigna una etiqueta de acuerdo al tipo de demora de los vuelos, etiqueta a los vuelos

desviados, etiqueta los vuelos a tiempo, y para los vuelos cancelados especifica el motivo.

- Con el comando FORMAT_DATE se creó la variable ***DAY_OF_WEEK***.

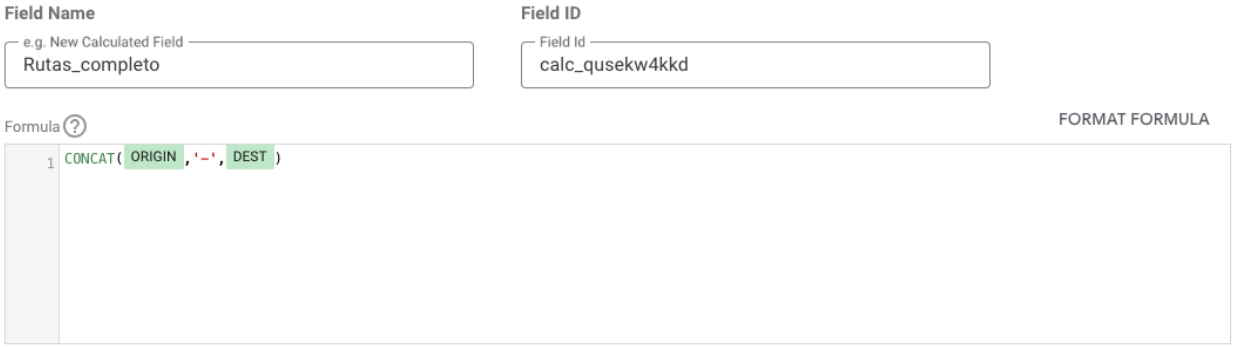
8. Unir tablas

- Con el comando JOIN se unieron las vistas view_consolidado_completo, view_airline_code y dot_code_dictionary.

4. Hacer un análisis exploratorio

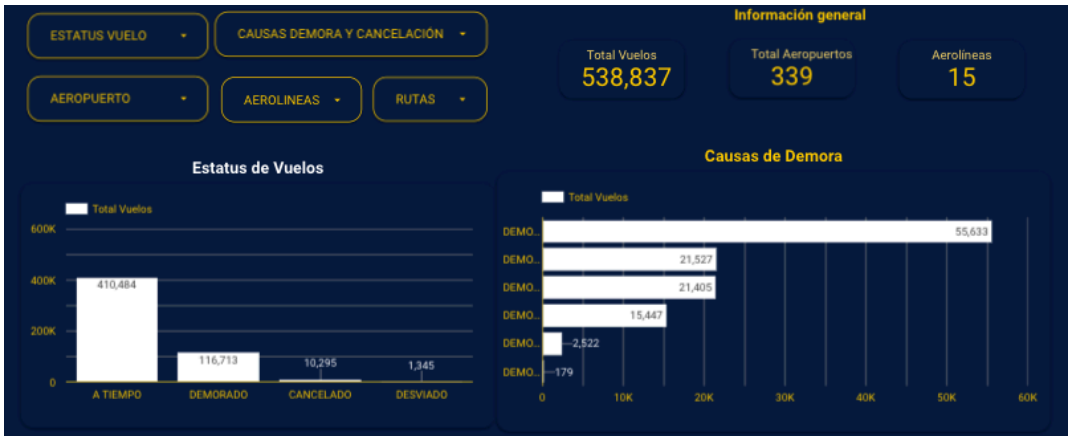
9. Agrupar datos según variables categóricas

- Se conectaron los datos a looker studio desde Bigquery.
- Se crearon campos calculados para la visualización de variable y elaboración de gráficos.



10. Visualizar las variables categóricas

- Se realizaron gráficos de barras para la visualización de variables y exploración de datos en looker studio.
- Se crearon listas drop-down list para filtrar y explorar la información y se agregaron score cards con datos relevantes.



11. Aplicar medidas de tendencia central

- Se crearon tablas en looker studio con las medidas de tendencia central y de dispersión (media, promedio, rango, desviación estándar) de la variable ARR_DELAY para explorar los datos considerando todos los vuelos y los vuelos demorados.

Todos los vuelos

	Total Vuelos	AVG	Median ARR_D...	MIN ARR_DELAY	MAX ARR_DELAY	STD ARR_DELAY
1.	538,837	7.78	-5	-80	3,063	57.4

1 - 1 / 1

Todos los vuelos

	Estatus Vuelo	Total Vuelos	AVG	Median ARR...	MIN ARR_DE...	MAX ARR_D...	STD ARR_DE...
1.	DEMORADO	116,713	69.44	41	15	3,063	97.41
2.	A TIEMPO	410,484	-9.76	-10	-80	14	12.03
3.	CANCELADO	10,295	null	null	null	null	null
4.	DESVIADO	1,345	null	null	null	null	null

1 - 4 / 4

Vuelos Demorados

	Total Vuelos	AVG	Median ARR_D...	MIN ARR_DELAY	MAX ARR_DELAY	STD ARR_DELAY
1.	116,713	69.44	41	15	3,063	97.41

1 - 1 / 1

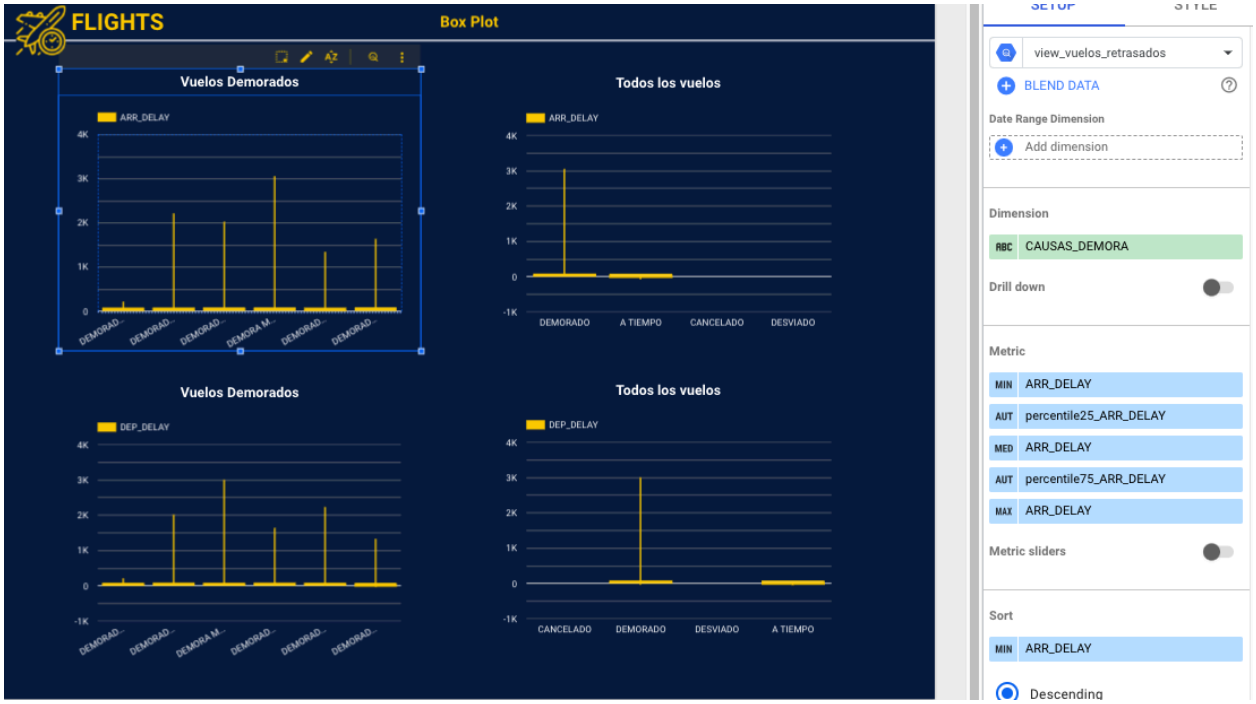
Vuelos Demorados

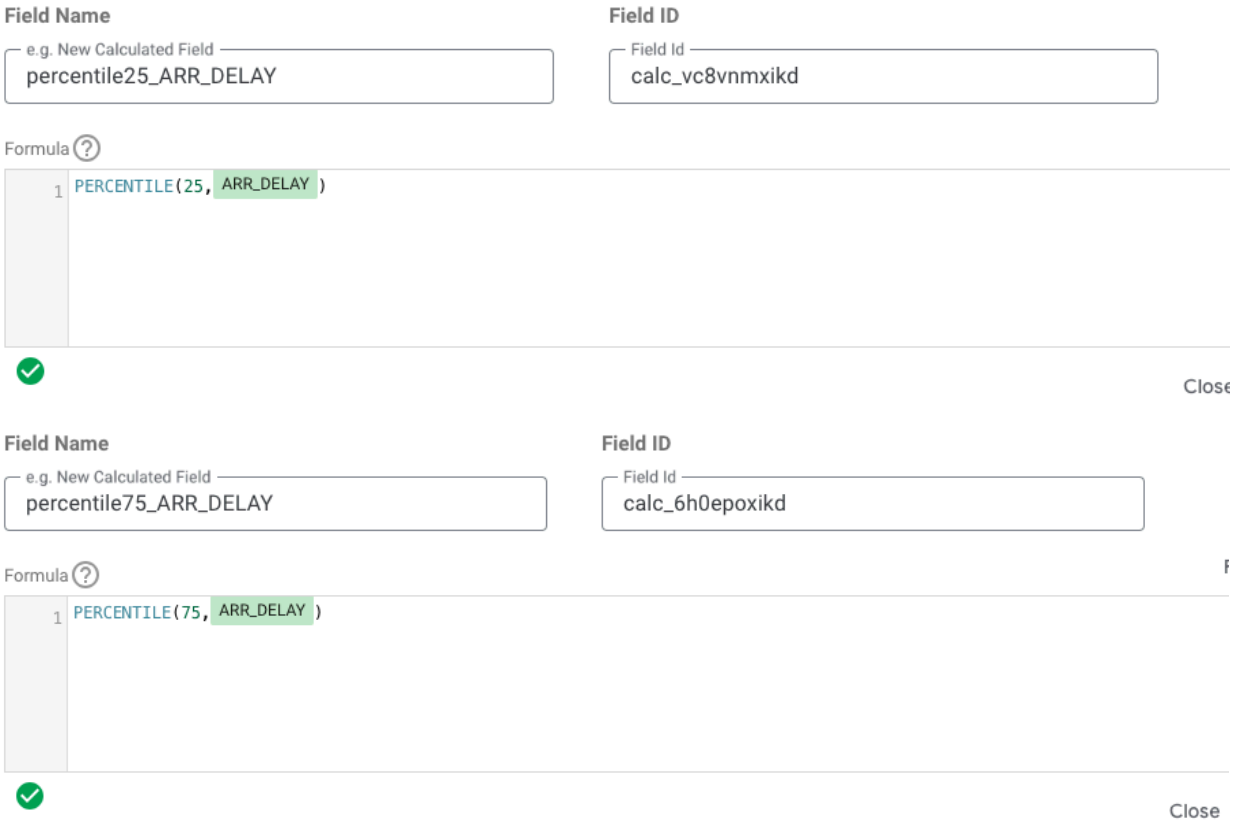
	Causas Demora	Total Vuelos	AVG	Median ARR_DELAY	MIN ARR_DELAY	MAX ARR_DELAY	STD ARR_DELAY
1.	DEMORA MULTIFACTOR	55,633	74.94	47	15	3,063	99.93
2.	DEMORADO POR OPER...	21,527	71.5	37	15	2,229	116.13
3.	DEMORADO POR NAS	21,405	46.04	27	15	1,343	53.58
4.	DEMORADO AERONAV...	15,447	72.88	47	15	2,027	86.2
5.	DEMORADO POR CLIMA	2,522	109.84	56	15	1,653	163.25
6.	DEMORADO POR SEGU...	179	46.01	31	15	234	37.94

1 - 6 / 6

12. Visualizar distribución

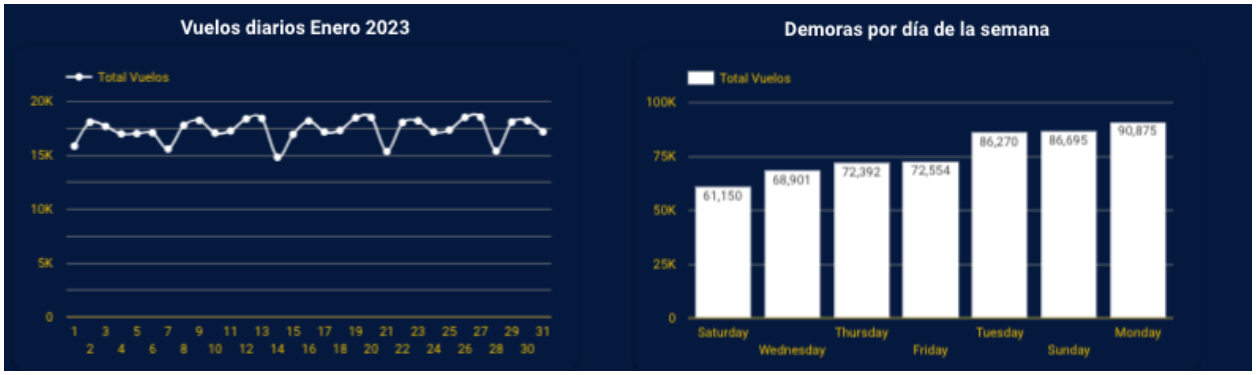
- Se crearon box plot para visualizar la distribución de las variables ARR_DELAY y DEP_DELAY por causa de demora y por estatus de vuelo.
- Se realizaron box plot e histogramas para las variables en google colab usando python.





13. Visualizar el comportamiento de los datos a lo largo del tiempo

- Se crearon gráficos de línea para observar el comportamiento de los datos a lo largo del mes y por día de la semana.



14. Calcular cuartiles, deciles o percentiles

- Con los comandos WITH, NTILE, COUNT, GROUP BY, MIN, MAX, JOIN, se calcularon los cuartiles de la variable ARR_DELAY para los vuelos demorados y para el consolidado con todos los vuelos, se contabilizó el número de vuelos por cuartil, el total de vuelos retrasados y se calculó el rango de cada cuartil.

Vuelos demorados

Row	cuartiles_delay	total_vuelos	total_vuelos_retrasad	min_delay	max_delay
1	1	29179	29179	15	24
2	2	29178	29178	24	41
3	3	29178	29178	41	79
4	4	29178	29178	79	3063

Todos los vuelos

Row	cuartiles_delay	total_vuelos	total_vuelos_retrasad	min_delay	max_delay
1	1	134710	0	-80	-16
2	2	134709	0	-16	-6
3	3	134709	67866	-6	10
4	4	134709	134709	10	3063

15. Calcular correlación entre variables

- Con el comando CORR se calculó la correlación entre las variables en BigQuery.
- En google colab usando python se creó una matriz de correlación incluyendo todas las variables numéricas.

Todos los vuelos

Row	CORR_DEP_ARR	CORR_ARR_CARRIER	CORR_ARR_AIRCRAF	CORR_ARR_NAS	CORR_ARR_SECURIT	CORR_ARR_WEATHE
1	0.966316306952...	0.679793352230...	0.575573704093...	0.352662587497...	0.026894945339...	0.330184132003...

CORR_DEP_CARRIER	CORR_DEP_AIRCRAF	CORR_DEP_NAS	CORR_DEP_SECURIT	CORR_DEP_WEATHE
0.689052552957...	0.586619588032...	0.266382411572...	0.025460448995...	0.324211611451...

CORR_DEP_TAXI_OU	CORR_ARR_TAXI_OU	CORR_DEP_TAXI_IN	CORR_ARR_TAXI_IN	CORR_DEP_DISTANC	CORR_ARR_DISTANC
0.061429694833...	0.211083526901...	0.015914291711...	0.101097163539...	0.028338291951...	0.005373996805...

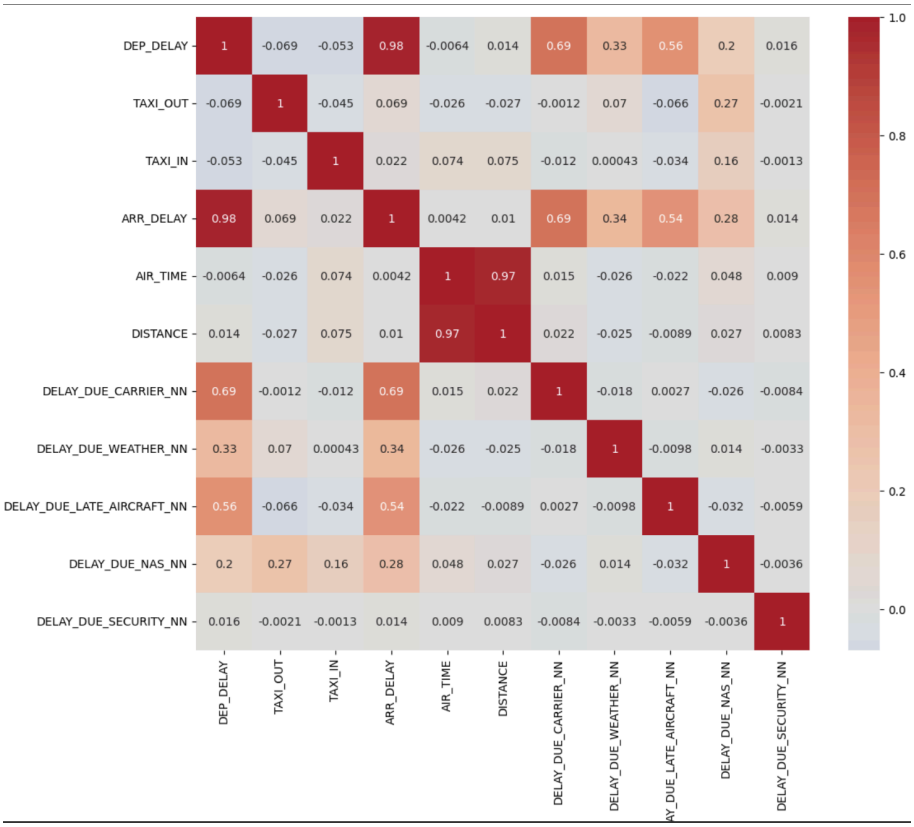
Vuelos demorados

Row	CORR_DEP_ARR	CORR_ARR_CARRIER	CORR_ARR_AIRCRAF	CORR_ARR_NAS	CORR_ARR_SECURIT	CORR_ARR_WEATHE
1	0.978421614242...	0.676766765798...	0.491530810149...	0.190173358088...	0.001742253960...	0.342733419836...

Row	CORR_DEP_CARRIER	CORR_DEP_AIRCRAF	CORR_DEP_NAS	CORR_DEP_SECURIT	CORR_DEP_WEATHE
1	0.675852015854...	0.514303806717...	0.108042372643...	0.004273547606...	0.324884537401...

CORR_DEP_TAXI_OU	CORR_ARR_TAXI_OU	CORR_DEP_TAXI_IN	CORR_ARR_TAXI_IN	CORR_DEP_DISTANC	CORR_ARR_DISTANC
-0.12392228175...	0.018944573071...	-0.07497092274...	0.002098646146...	0.004014667745...	0.002618237223...

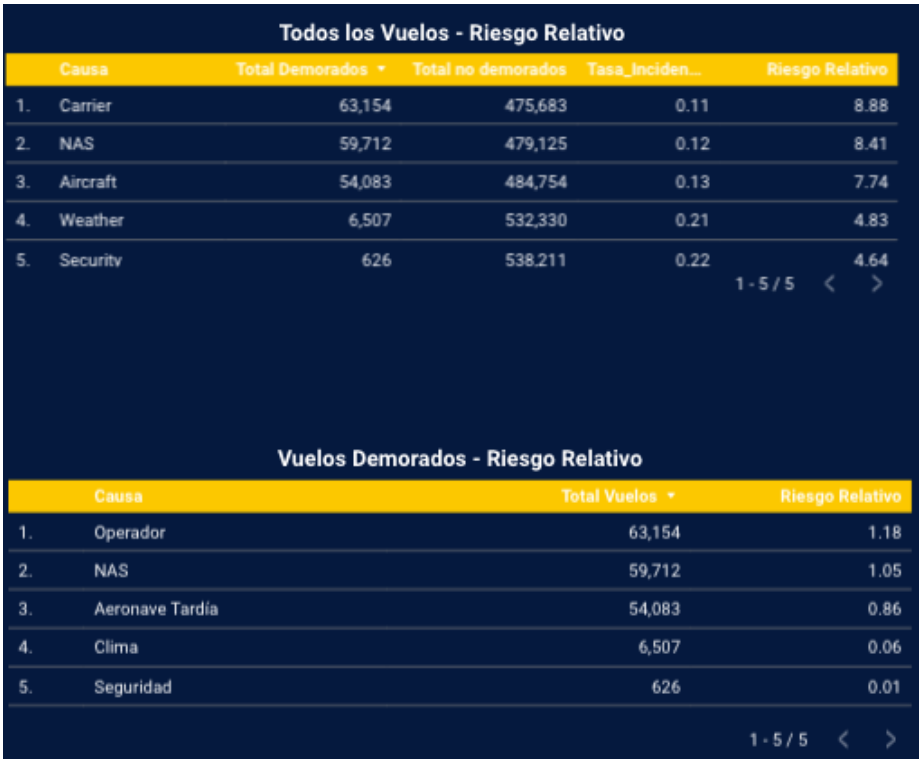
Matriz de Correlación



5. Aplicar técnica de análisis

16. Calcular riesgo relativo

- Con los comandos WITH, SUM, CASE, WHEN, THEN, ELSE, END, SAFE_DIVIDE, COUNT, GROUP BY, se calculó el riesgo relativo en BigQuery para las variables DELAY_DUE_CARRIER_NN, DELAY_DUE_SECURTY_NN, DELAY_DUE_WEATHER_NN, DELAY_DUE_LATE_AIRCRAFT_NN, DELAY_DUE_NAS_NN, adicional se hizo el cálculo por Aerolínea, Aeropuerto y Ruta, obteniendo una tabla con el total de vuelos expuestos, total de vuelos no expuestos, las tasas de incidencia de ambos grupos y el riesgo relativo.



Todos los Vuelos - Riesgo Relativo

	Causa	Total Demorados	Total no demorados	Tasa_Inciden...	Riesgo Relativo
1.	Carrier	63,154	475,683	0.11	8.88
2.	NAS	59,712	479,125	0.12	8.41
3.	Aircraft	54,083	484,754	0.13	7.74
4.	Weather	6,507	532,330	0.21	4.83
5.	Securitv	626	538,211	0.22	4.64

1 - 5 / 5

Vuelos Demorados - Riesgo Relativo

	Causa	Total Vuelos	Riesgo Relativo
1.	Operador	63,154	1.18
2.	NAS	59,712	1.05
3.	Aeronave Tardía	54,083	0.86
4.	Clima	6,507	0.06
5.	Seguridad	626	0.01

1 - 5 / 5

Aerolíneas - Riesgo Relativo

	Aerolínea	Descripción	Vuelos Demor...	Total Vuelos	Tasa incidencia	Riesgo Relativ...
1.	F9	Frontier Airline...	4,505	13,285	0.34	1.59
2.	NK	Spirit Air Lines	6,209	21,876	0.28	1.33
3.	G4	Allegiant Air	2,412	8,615	0.28	1.3
4.	B6	JetBlue Airways	6,084	23,249	0.26	1.22
5.	UA	United Air Line...	12,965	56,657	0.23	1.06

1 - 15 / 15

Aeropuertos- Riesgo Relativo

	Aeropuerto	Ciudad	Vuelos Demor...	Total Vuelos	Tasa incidencia	Riesgo Relativ...
1.	PPG	Pago Pago, TT	6	11	0.55	2.52
2.	CKB	Clarksburg/Fai...	5	10	0.5	2.31
3.	LNK	Lincoln, NE	12	28	0.43	1.98
4.	IAG	Niagara Falls, ...	12	28	0.43	1.98
5.	PSM	Portsmouth, NH	8	19	0.42	1.94

1 - 100 / 339

Rutas- Riesgo Relativo

	RUTA	Ciudad Orig...	Ciudad Dest...	Vuelos Dem...	Total Vuelos	Tasa incide...	Riesgo Rela...
1...	MCO - SJU	Orlando, FL	San Juan, PR	208	580	0.36	1.27
2...	ORD - ATL	Chicago, IL	Atlanta, GA	188	530	0.35	1.25
3...	LAS - SAN	Las Vegas, ...	San Diego, ...	192	559	0.34	1.21
4...	ORD - LGA	Chicago, IL	New York, NY	289	879	0.33	1.16
5...	LAX - SFO	Los Angeles...	San Francis...	283	900	0.31	1.1

1 - 15 / 15

17. Validar hipótesis

- Las aerolíneas con mayor número de vuelos demorados tienen mayor riesgo de que sus vuelos se retrasen.

Conclusión: Se refuta la hipótesis, la aerolínea con más cantidad de vuelos demorados es WN (Southwest Airlines Co.) con 21, 830, tiene un riesgo relativo de 0.87, por lo que hay menor riesgo de demora en comparación con el resto. La aerolínea que tiene el mayor riesgo relativo es F9 (Frontier Airlines Inc.) con un riesgo relativo de 1.59.

- Los vuelos que enfrentan condiciones meteorológicas adversas (como tormentas o niebla) tienen un riesgo relativo significativamente mayor de sufrir retrasos en comparación con aquellos que no enfrentan condiciones meteorológicas adversas.

Conclusión: Se refuta la hipótesis, los vuelos demorados por condiciones meteorológicas tiene un riesgo relativo menor a otras causas de demora, por ejemplo, retrasos por operador, retrasos por aeronave tardía y retrasos por NAS.

- Las demoras por causa del operador son más frecuentes en las aerolíneas con mayor riesgo relativo.

Conclusión: Se válida la hipótesis, la causa de demora con mayor riesgo relativo es la demora del operador.

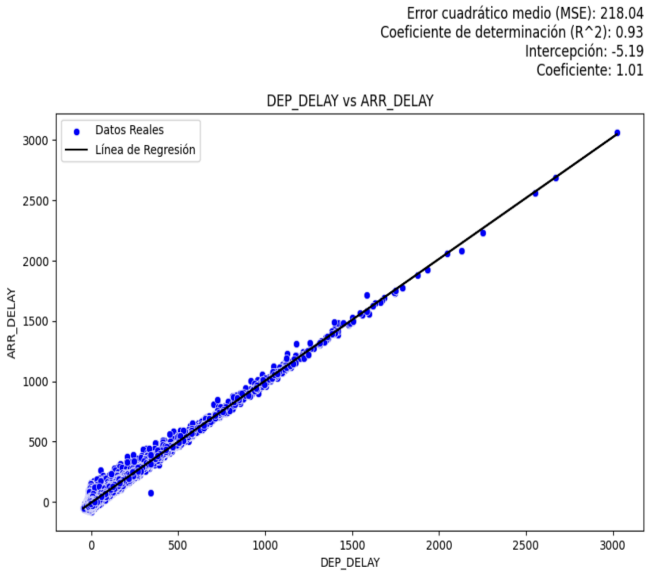
18. Regresión lineal

Se aplicó la técnica de regresión lineal para entender la relación entre el tiempo de retraso de llegada de un vuelo y las diversas causas de retraso representadas por las variables del análisis.

- **ARR_DELAY vs DEP_DELAY**

Correlación: r=0.98

Interpretación: se observa una **correlación positiva muy fuerte**, indica que, a medida que aumenta el retraso de salida, el retraso de llegada también tiende a aumentar casi en la misma proporción.

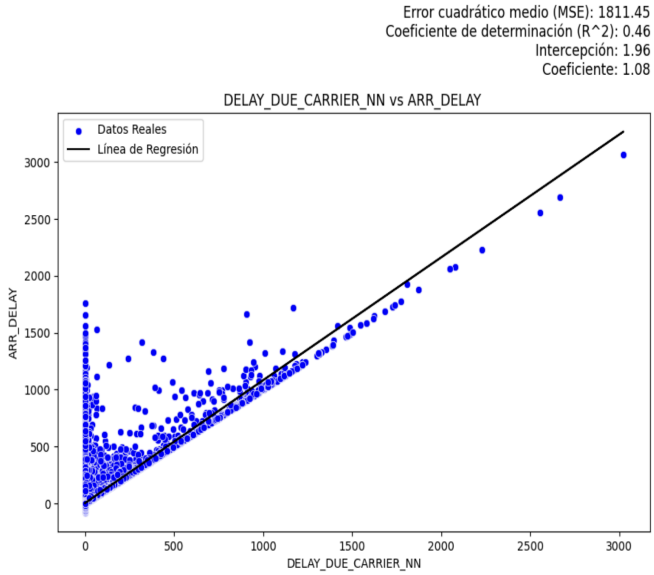


Coeficiente de determinación (R^2): 0.93, lo que indica que el 93% de la variabilidad en el retraso de llegada puede explicarse por el retraso de salida.

- **ARR_DELAY vs DELAY_DUE_CARRIER**

Correlación: r=0.69

> **Interpretación:** se observa una **correlación positiva moderada**, sugiere que las demoras atribuidas a la aerolínea (por razones como problemas operativos, mantenimiento, etc.) contribuyen significativamente al retraso en la llegada.

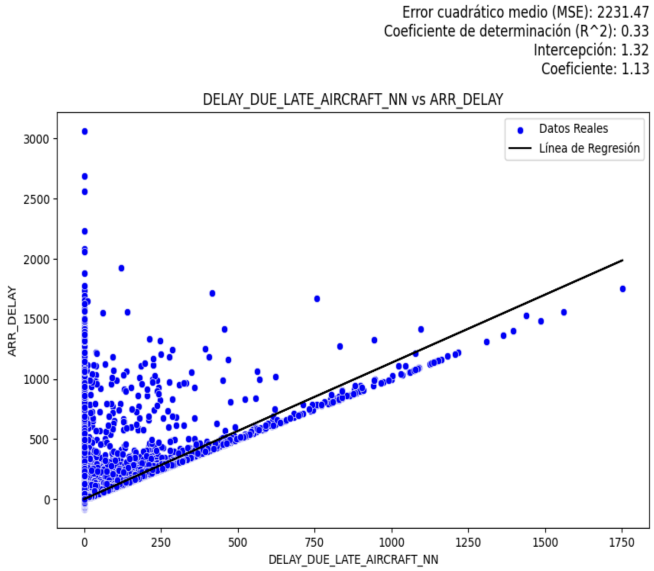


Coeficiente de determinación (R^2): 0.46, lo que indica que el 46% de la variabilidad en el retraso de llegada puede explicarse por los retrasos causados por la aerolínea. Esta es una relación moderada.

- **ARR_DELAY vs DELAY_DUE_LATE_AIRCRAFT**

Correlación: r=0.54

> **Interpretación:** Existe una **correlación positiva moderada**, la correlación indica que, aunque las llegadas tardías de vuelos anteriores influyen en el retraso actual, no es el único factor determinante.

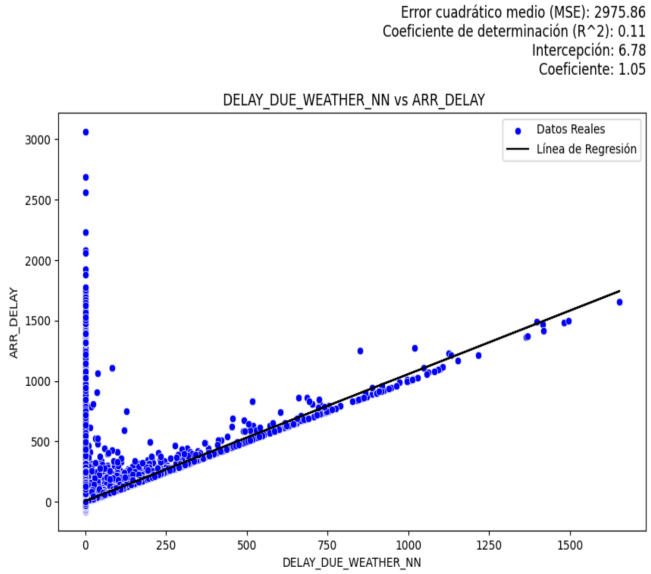


Coeficiente de determinación (R^2): 0.33, lo que indica que el 33% de la variabilidad en el retraso de llegada puede explicarse por los retrasos causados por la llegada tardía de la aeronave.

- **ARR_DELAY vs DELAY_DUE_WEATHER**

Correlación: r=0.34

> **Interpretación**: Existe una **correlación positiva débil**, sugiere que, si bien el mal tiempo contribuye a los retrasos de llegada, no es un factor tan influyente como las demoras operativas o de la aerolínea.

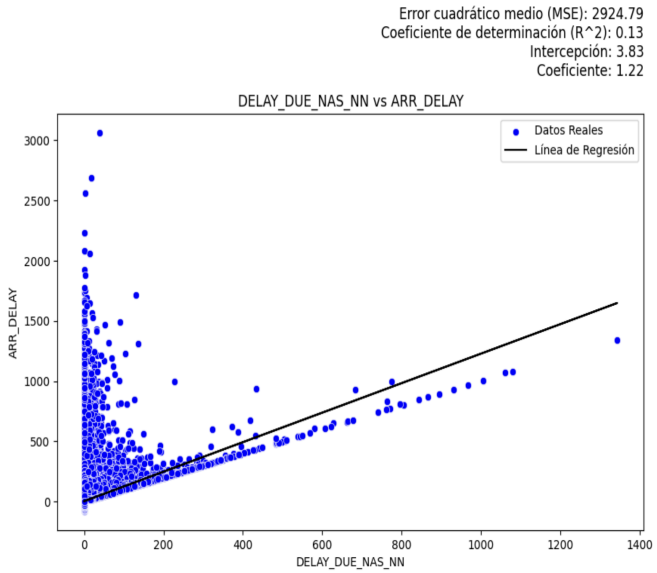


Coeficiente de determinación (R^2): 0.11, sugiere que solo el 11% de la variabilidad en el retraso de llegada puede explicarse por los retrasos debido al clima.

- **ARR_DELAY vs DELAY_DUE_NAS**

Correlación: r=0.28

Interpretación: Existe una **correlación positiva muy débil**, un coeficiente de 0.28 indica que las demoras relacionadas con la gestión del tráfico aéreo, la congestión, etc., tienen una influencia limitada en los retrasos de llegada.

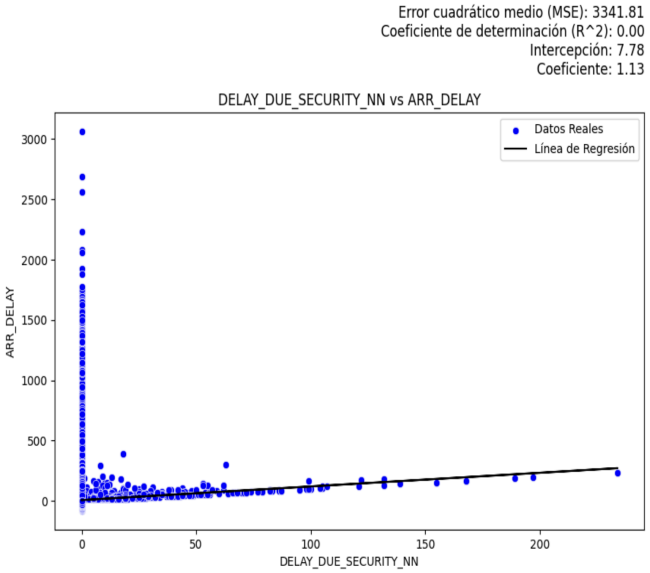


Coeficiente de determinación (R^2): 0.13, lo que sugiere que sólo el 13% de la variabilidad en el retraso de llegada puede explicarse por los retrasos debido al Sistema de Navegación Aérea.

- **ARR_DELAY vs DELAY_DUE_SECURITY**

Correlación: r=0.0014

> **Interpretación:** Existe una **correlación casi nula**, un coeficiente de 0.0014 indica que no hay prácticamente ninguna relación entre estas dos variables.



Coeficiente de determinación (R^2): 0.00, lo que sugiere que el 0% de la variabilidad en el retraso de llegada puede explicarse por los retrasos debidos a medidas de seguridad.

Resultados y Conclusiones:

- Existe una relación lineal muy fuerte entre DEP_DELAY y ARR_DELAY (retraso en la llegada), lo que significa que la mayoría del retraso en la llegada se debe directamente al retraso en la salida.
- Los retrasos causados por la aerolínea tienen una correlación moderada con los retrasos en la llegada. Aunque estos retrasos son significativos, hay otros factores que también influyen, como el clima, la congestión del tráfico aéreo, y problemas operativos.
- Existe una correlación moderada entre los retrasos debidos a la llegada tardía de una aeronave y los retrasos en la llegada. Este es un factor más relevante que el clima, pero no el más determinante.
- Las condiciones meteorológicas adversas influyen en los retrasos, pero la correlación es débil. Esto sugiere que, aunque el mal clima puede contribuir, no es un factor predominante.
- Hay una correlación débil entre los retrasos causados por problemas en el Sistema de Navegación Aérea (NAS) y los retrasos en la llegada. Aunque existe una relación, no es suficientemente fuerte como para ser un factor clave.
- Los retrasos debidos a problemas de seguridad no tienen un impacto significativo en los retrasos de llegada. La correlación es prácticamente inexistente.
- Frontier Airlines (F9): Esta aerolínea tiene el mayor riesgo relativo de 1.59. Esto sugiere que sus vuelos tienen un 59% más de probabilidades de experimentar retrasos en comparación con la tasa de referencia.
- Spirit Air Lines (NK) y Allegiant Air (G4): Ambas aerolíneas también presentan un riesgo relativo elevado, con valores de 1.33 y 1.30 respectivamente. Esto indica un riesgo significativamente alto de retrasos comparado con otras aerolíneas.
- Los pasajeros que viajan con Frontier, Spirit, y Allegiant pueden esperar un mayor riesgo de retrasos, y estas aerolíneas podrían beneficiarse de iniciativas para mejorar la puntualidad.
- Pago Pago, TT (PPG) y Clarksburg/Fairmont, WV (CKB) son aeropuertos que tienen los riesgos relativos más altos, de 2.52 y 2.31 respectivamente. Los vuelos desde estos aeropuertos tienen más del doble de probabilidad de retrasarse en comparación con otros.
- Lincoln, NE (LNK) y Niagara Falls, NY (IAG), tienen riesgo relativo alto, de alrededor de 1.98, lo que sugiere que estos aeropuertos tienen problemas similares con la puntualidad.

- Orlando, FL (MCO) a San Juan, PR (SJU), esta ruta tiene un riesgo relativo de 1.27, lo que indica que los vuelos en esta ruta son un 27% más propensos a experimentar retrasos en comparación con la media.
- Chicago, IL (ORD) a Atlanta, GA (ATL) y Las Vegas, NV (LAS) a San Diego, CA (SAN), ambas rutas tienen un riesgo relativo superior a 1.20, lo que sugiere una mayor probabilidad de retrasos.
- El análisis de riesgo relativo destaca las aerolíneas, aeropuertos, y rutas que tienen mayores probabilidades de retrasos. Estas áreas identificadas podrían ser objetivos prioritarios para mejoras operacionales o de infraestructura. Además, comunicar esta información a los pasajeros podría mejorar su experiencia, al permitirles tomar decisiones informadas sobre sus viajes.
- Estas conclusiones pueden ser utilizadas para proponer soluciones, como la revisión de procesos operativos en aerolíneas con altos riesgos relativos, o la implementación de mejoras en la infraestructura en aeropuertos problemáticos.

Recomendaciones:

Para Aerolíneas con Alto Riesgo Relativo:

- Mejorar la Gestión del Tiempo: Se podrían implementar programas más rigurosos de gestión del tiempo y planificación de vuelos. Esto incluye revisar y ajustar los horarios de salida y llegada para reducir la congestión en aeropuertos y mejorar la eficiencia operativa.
- Optimización de Recursos: Revisar la asignación de recursos, como el personal de tierra y el mantenimiento de aeronaves, para asegurarse de que las operaciones sean lo más eficientes posible y que cualquier problema se resuelva rápidamente.
- Comunicación Proactiva: Las aerolíneas con alto riesgo relativo deben comunicar proactivamente a los pasajeros sobre posibles retrasos y ofrecer opciones de reprogramación o compensaciones cuando sea necesario, mejorando la percepción del servicio.

Para Aeropuertos con Alto Riesgo Relativo:

- Mejoras en Infraestructura: Considerar inversiones en infraestructura aeroportuaria, especialmente en aquellos aeropuertos con recursos limitados, para mejorar la capacidad de manejo de vuelos y reducir la probabilidad de retrasos.
- Optimización de Procesos Operacionales: Revisar y optimizar los procesos operacionales, como la asignación de puertas, la gestión del tráfico aéreo y el manejo del equipaje, para minimizar los cuellos de botella que puedan estar causando retrasos.

- Coordinación con Aerolíneas: Trabajar en estrecha colaboración con las aerolíneas para mejorar la coordinación de los horarios de vuelo, especialmente en aeropuertos con capacidades limitadas.

Para Rutas con Alto Riesgo Relativo:

- Análisis de Factores Contribuyentes: Realizar un análisis detallado de las rutas con alto riesgo relativo para identificar factores específicos que contribuyan a los retrasos, como la congestión aérea, condiciones meteorológicas, o problemas en las operaciones en tierra.
- Revisar la frecuencia de vuelos: Considerar ajustar la frecuencia de vuelos en rutas problemáticas, o cambiar los horarios de salida y llegada para evitar picos de congestión.
- Planes de Contingencia: Desarrollar planes de contingencia específicos para estas rutas, que incluyan alternativas para los pasajeros en caso de retrasos significativos, como la reprogramación o desvío a aeropuertos alternativos.

Limitaciones/Próximos pasos:

- Realizar un análisis detallado de las cancelaciones.
- Crear un modelo predictivo para estimar la probabilidad de retraso.

Enlaces de interés:

Para la toma de decisiones se consultó la información de los siguientes enlaces.

https://aspm.faa.gov/aspmhelp/index/ASPM_Data_Download__Cancelled_Flights.html

https://en.wikipedia.org/wiki/2023_FAA_system_outage

https://www.marketwatch.com/guides/travel-insurance/trip-cancellation-compensation/#:~:text=How%20To%20Get%20Delayed%20or,airport%20to%20file%20a%20claim

https://www.transportation.gov/airconsumer/airline-cancellation-delay-dashboard

Dashboard:

https://lookerstudio.google.com/u/0/reporting/859bcf09-a71b-4a22-ad3d-3d68eccf0ae4/page/p_rk3uym2lkd/edit

https://public.tableau.com/app/profile/ysabel.mata5447/viz/DELAYED_FLIGHTS/Dashboard_vuelos?publish=yes