Proyecto Integrador de Análisis de Datos (CRISP-DM)

PROYECTO POR:

LUIS CULAJAY
KAREN MORALES
BRANDON PORTILLO
JHONATAN SIMÓN

Grupo 3

PROYECTO INTEGRADOR DE ANÁLISIS DE DATOS (CRISP-DM)

- Business Understanding
- Data Understanding
- Data Preparation
- Modeling (Modelo en Estrella)
- Evaluation
- Deployment (Dashboard)

Grupo 3

Karen Morales
Data Administrator

Luis Culajay
Data Engineer

Jhonatan Simon
BI Analyst

Brandon Portillo
Data Analyst

Business Understanding

- KPI’s Iniciales
- Canvas Data Product
- Requerimientos analiticos
- Preguntas de negocio

KPI’s Iniciales

1. **Tipo de distribución por retrasos**
2. **Vuelos Programados vs real (llegada)**
3. **Vuelos Programados vs real (salida)**
4. **Ranking de aerolineas puntuales**
5. **Porcentaje de cancelación**

Preguntas de negocio

¿Qué aerolíneas y aeropuertos presentan mejor o peor puntualidad?
¿Cuáles son las causas de demora más relevantes y cómo se distribuyen por aerolínea/aeropuerto?
¿Qué rutas son más vulnerables a retrasos, cancelaciones o desvíos?
¿Qué tan diferentes son los tiempos programados vs reales y qué factores influyen en esa brecha?
¿Qué aeropuertos con mayor concurrencia muestran congestión operacional?
¿Las rutas largas tienen mejor o peor puntualidad que las rutas cortas?

CDP

Canvas Data Product

Problema

- Se desconoce qué aerolíneas presentan mejor o peor desempeño en puntualidad.
- No se sabe qué causas de demora son más frecuentes o relevantes.
- No está claro cuáles aeropuertos y rutas concentran mayor tráfico y riesgo operativo.
- Las cancelaciones y desvíos impactan la eficiencia del sistema y requieren identificación.

Datos

Se tiene un dataset de Bureau of Transportation Statistics (BTS) que contiene horarios, retrasos, causas de demora, aerolíneas, aeropuertos, tiempos operacionales y cancelaciones.

Hipótesis

- Algunas aerolíneas son más puntuales que otras.
- NAS y LateAircraftDelay serán las causas dominantes.
- Aeropuertos de alta concurrencia tendrán más retrasos.
- Algunas rutas presentan más cancelaciones y desvíos

Solución

La solución será un dashboard analítico que permita visualizar patrones de puntualidad, causas de retraso, concurrencia y rutas con problemas, entre otros, siendo un consolidado de los problemas encontrados en los vuelos.

KPI's

- Porcentaje de vuelos puntuales
- Porcentaje de desvíos
- Porcentaje de cancelación
- Porcentaje de desvíos
- Porcentaje de vuelos con retraso

Actores

- Cliente general: Entidades reguladoras de vuelos/tráfico aéreo
- Stakeholders: Aerolíneas, aeropuertos, reguladores.
- Usuarios: Analistas e investigadores
- Impacto en: planificación, eficiencia operativa y experiencia del pasajero.

Acciones

Las acciones derivadas del dashboard permitirán identificar patrones, segmentar aerolíneas y aeropuertos y visualizar horarios o rutas vulnerables para realizar mejores planificaciones en el tráfico aéreo

Valores / Riesgos

El resultado aportará un gran valor al generar visibilidad sobre la puntualidad y causas de demora Los riesgos se encuentran en los límites que pueda dar el dataset debido a la falta de datos externos.

Rendimiento / Impacto

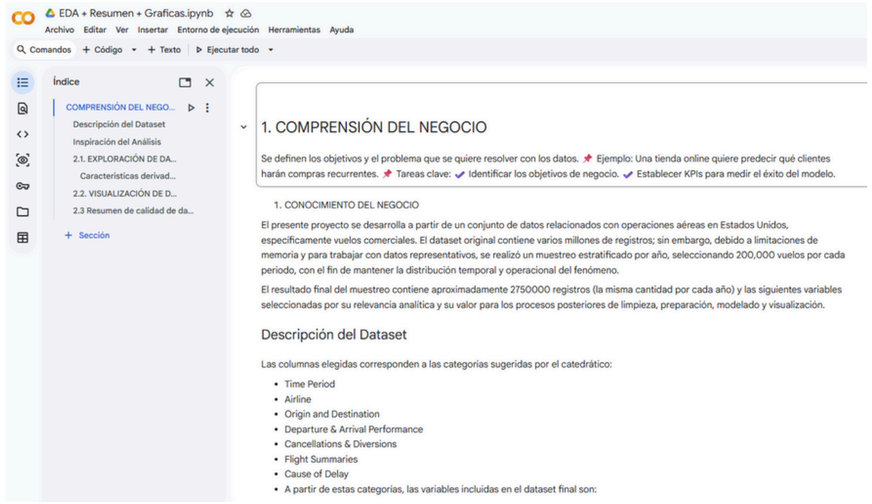
Capacidad de incrementar la tasa de vuelos puntuales, reducir el número de vuelos cancelados, entender cómo reducir la variabilidad entre tiempos programados y reales.

Data Understanding

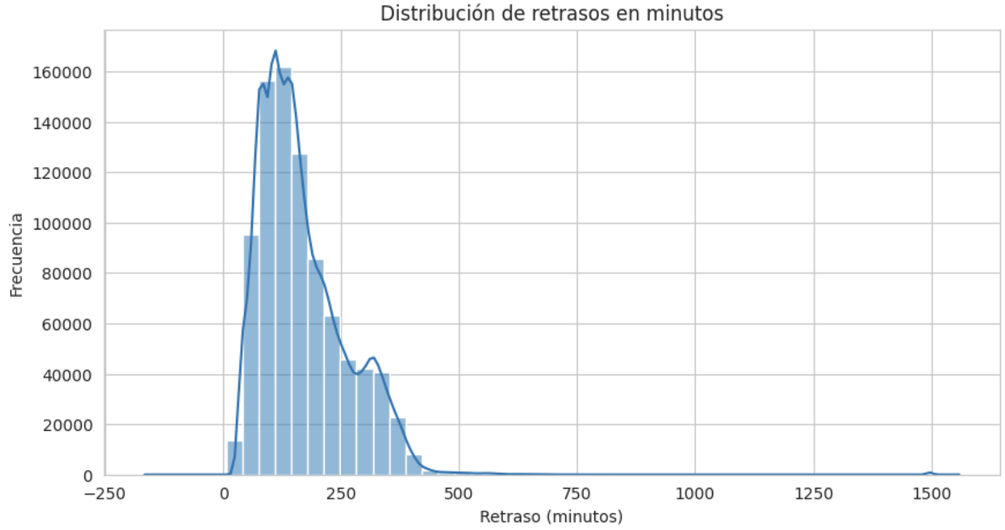
Descarga del dataset usando la interfaz BTS.
Columnas seleccionadas y justificación.
Exploración y análisis inicial (EDA).
Detección de valores faltantes y outliers.

Categoría	Nombre del campo	Nombre en Dataset	Justificacion
Time Period	Year	YEAR	Necesario para los filtros a implementar en el data product (dashboard) lo cual permitirá analizar y visualizar temporalidad de los vuelos
	Month	MONTH	
	DayofMonth	DAY_OF_MONTH	
	DayOfWeek	DAY_OF_WEEK	
Airline	Reporting_Airline	OP_UNIQUE_CARRIER	Campo clave para medir y comparar la tasa de puntualidad, eficiencia operativa y cantidad de vuelos entre aerolíneas. Permite responder qué compañías presentan mejor desempeño general.
Origin	OriginAirportID	ORIGIN_AIRPORT_ID	Identificador único del aeropuerto de salida. Permite establecer relaciones entre aeropuertos y rutas dentro del modelo en estrella y comparar desempeño por ubicación.
	OrigenCityMarketID	ORIGIN_CITY_MARKET_ID	Agrupa aeropuertos por ciudad, lo que facilita análisis a nivel de ciudad.
	Origin	ORIGIN	Código IATA del aeropuerto de salida. Facilita la identificación geográfica y visualización de rutas en dashboards o mapas.
	OriginCityName	ORIGIN_CITY_NAME	Proporciona el nombre de la ciudad asociada al aeropuerto, lo que permite contextualizar los resultados geográficamente.
	OriginState	ORIGIN_STATE_ABR	Codigo del estado. Facilita el analisis por region y comparaciones de desempeño de rutas o concurrencia de uso
	OriginStateName	ORIGIN_STATE_NM	Proporciona el nombre del estado asociada al aeropuerto, lo que permite contextualizar los resultados geográficamente.
Destination	DestAirportID	DEST_AIRPORT_ID	Identificador único del aeropuerto de llegada. Permite analizar eficiencia en el destino, frecuencia de vuelos y rutas con mayor congestión o demoras.
	DestCityMarketID	DEST_CITY_MARKET_ID	Relaciona vuelos a nivel de mercado de ciudad destino, útil para analizar rutas metropolitanas y volumen de tráfico aéreo interurbano.
	Dest	DEST	Código IATA del aeropuerto destino, fundamental para mapear rutas aéreas y analizar la conexión entre origen–destino.
	DestCityName	DEST_CITY_NAME	Permite visualizar y agrupar resultados por ciudad destino, útil para dashboards geográficos y reportes de movilidad.
	DestState	DEST_STATE_ABR	Codigo del estado. Facilita el analisis por region y comparaciones de desempeño de rutas o concurrencia de uso
	DestStateName	DEST_STATE_NM	Proporciona el nombre del estado asociada al aeropuerto, lo que permite contextualizar los resultados geográficamente.

Resumen



Gráficas exploratorias



MAPA CONCEPTUAL

TIME PERIOD

YEAR
Indica el año de operación del vuelo

MONTH
Indica el mes de operación del vuelo

DAYOFMONTH
Indica el mes de operación del vuelo

DAYOFWEEK
Indica el mes de operación del vuelo

ORIGIN

ORIGINAIRPORTID
Código del aeropuerto de salida

ORIGENCITYMARKETID
Código de la ciudad a la que pertenece el aeropuerto de salida

ORIGINSTATE
Código del estado donde se ubica el aeropuerto de origen.

ORIGIN
Código IATA de tres letras del aeropuerto de origen

ORIGINCITYNAME
Nombre de la ciudad donde se localiza el aeropuerto de origen.

ORIGINSTATENAME
Nombre del estado donde se encuentra el aeropuerto de origen.

ARRIVAL PERFORMANCE

CRSARRTIME
Hora programada de llegada del vuelo

ARRDELAY
Retraso de llegada de vuelo

ARRTIME
Hora real de llegada del vuelo

AIRLINE

REPORTING_AIRLINE
Código de la aerolínea que reporta la información

DESTINATION

DESTAIRPORTID
Código del aeropuerto de destino

DESTCITYMARKETID
Código de la ciudad a la que pertenece el aeropuerto de destino

DEST
Código IATA de tres letras del aeropuerto de destino

DESTCITYNAME
Nombre de la ciudad donde se localiza el aeropuerto de destino.

DESTSTATE
Código del estado donde se ubica el aeropuerto de destino.

DESTSTATENAME
Nombre del estado donde se encuentra el aeropuerto de destino.

DEPARTURE PERFORMANCE

CRSDEPTIME
Hora programada de salida del vuelo

DEPDELAY
Retraso de salida de vuelo

DEPTIME
Hora real de salida del vuelo

CANCELLATIONS & DIVERSIONS

CANCELLED
Identifica si el vuelo fue cancelado

DIVERTED
Identifica si el vuelo fue desviado

CANCELLATION CODE
Código que identifica la causa de la cancelación

```sql
-- Creacion de dimension para dias festivos
CREATE TABLE dim_dias_festivos (
    festivo_id INT IDENTITY(1,1) PRIMARY KEY,
    fecha DATE NOT NULL UNIQUE,
        mes INT NOT NULL,
    dia INT NOT NULL,
    descripcion VARCHAR(100) NOT NULL
);

INSERT INTO dim_dias_festivos (fecha, mes, dia, descripcion)
VALUES
('2024-01-01', 1, 1, 'Año Nuevo'),
('2024-05-01', 5, 1, 'Día del Trabajo'),
('2024-06-30', 6, 30, 'Día del Ejército'),
('2024-09-15', 9, 15, 'Día de la Independencia'),
('2024-10-20', 10, 20, 'Revolución'),
('2024-11-01', 11, 1, 'Día de Todos los Santos'),
('2024-12-24', 12, 24, 'Nochebuena'),
('2024-12-25', 12, 25'Navidad'),
('2024-12-31', 12, 31, 'Año Viejo');

-- Adicion de nuevas columnas a la tabla de hechos
ALTER TABLE fact_flights ADD
    retraso_total INT NULL,
    semana INT NULL,
    trimestre INT NULL,
    dia_festivo INT NULL;

ALTER TABLE fact_flights
ADD CONSTRAINT FK_fact_festivo
FOREIGN KEY (dia_festivo)
REFERENCES dim_dias_festivos(festivo_id);

-- Creacion de valores derivados
UPDATE fact_flights
SET retraso_total = ISNULL(retraso_salida, 0) + ISNULL(retraso_llegada, 0);

UPDATE f
SET semana = DATEPART(WEEK, d.full_date)
FROM fact_flights f
JOIN dim_date d ON f.fecha_id = d.date_id;

UPDATE f
SET trimestre = DATEPART(QUARTER, d.full_date)
FROM fact_flights f
JOIN dim_date d ON f.fecha_id = d.date_id;
```

Data preparation

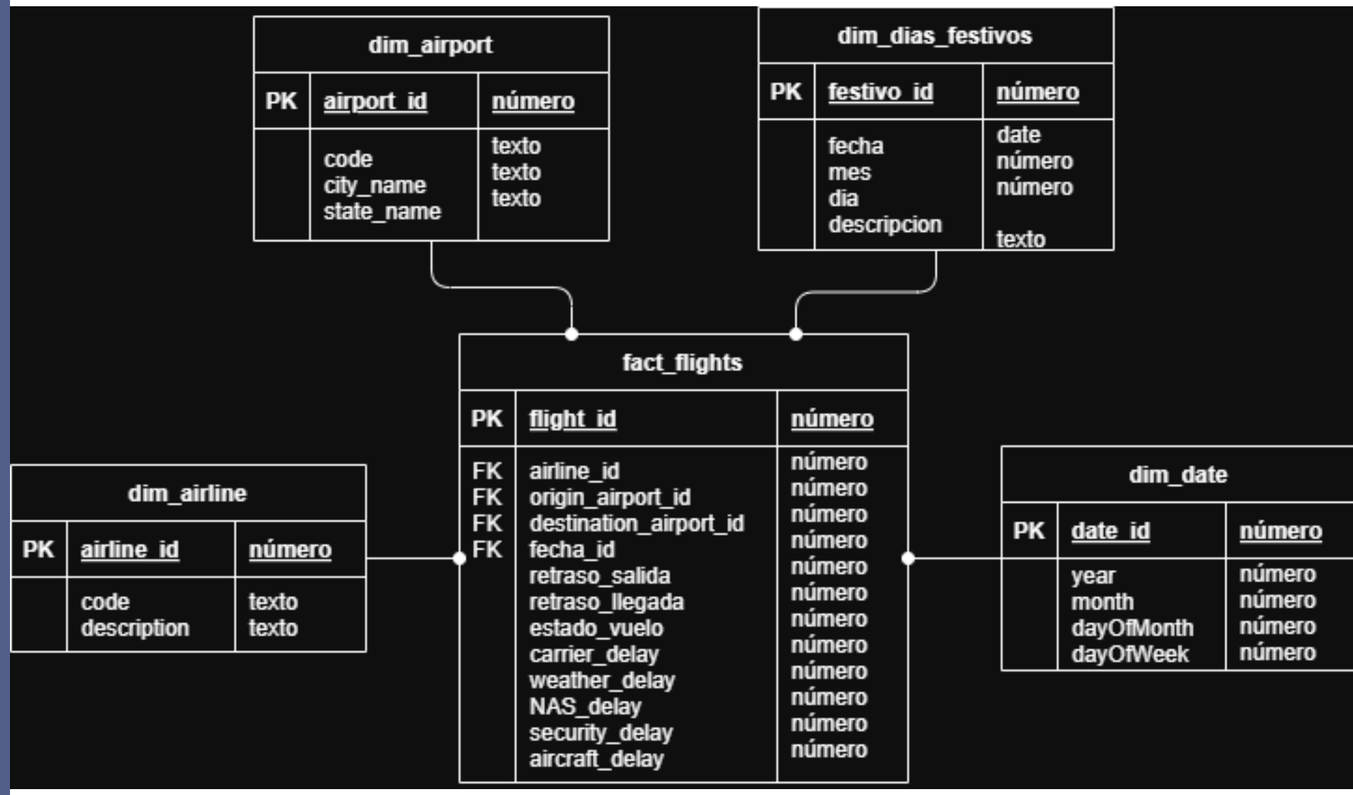
- Limpieza (duplicados, nulos, formatos de fecha).
- Estandarización de zonas horarias.
- Creación de columnas derivadas:
- retraso_total
- categoría de retraso
- semana, trimestre, día festivo
- Integración de tablas externas

```python
from datetime import datetime
import pandas as pd
import pyodbc
import warnings

# ============================================================
# === CONFIGURACIÓN
# ============================================================

warnings.filterwarnings("ignore", category=UserWarning)

airlines_csv = r"datos/aerolineas.csv"
flights_csv = r"datos/vuelos.csv"  # archivo de 10GB

chunk_size = 500_000  # medio millón de filas por chunk

connection_string = (
    "Driver={ODBC Driver 17 for SQL Server};"
    "Server=localhost;"
    "Database=DWBTS;"
    "Trusted_Connection=yes;"
)

conn = pyodbc.connect(connection_string)
cursor = conn.cursor()

# ============================================================
# === CARGAR DIMENSION AEROLÍNEAS (ya es pequeño, se queda igual)
# ============================================================
```

Modeling

```
--CREATE DATABASE DWBTS;

--USE DWBTS;

CREATE TABLE dim_airline (
    airline_id INT IDENTITY(1,1) PRIMARY KEY,
    code VARCHAR(10),
    description VARCHAR(200)
);

CREATE TABLE dim_date (
    date_id INT IDENTITY(1,1) PRIMARY KEY,
    year INT NOT NULL,
    month INT NOT NULL,
    day_of_month INT NOT NULL,
    day_of_week INT NOT NULL,
    full_date DATE NOT NULL
);

CREATE TABLE dim_airport (
    airport_id INT IDENTITY(1,1) PRIMARY KEY,
    code VARCHAR(10) NOT NULL,
    city_name VARCHAR(100),
    state_name VARCHAR(100)
);

CREATE TABLE fact_flights (
    flight_id BIGINT IDENTITY(1,1) PRIMARY KEY,
    airline_id INT NOT NULL,
    origin_airport_id INT NOT NULL,
    destination_airport_id INT NOT NULL,
    fecha_id INT NOT NULL,
    retraso_salida INT,
    retraso_llegada INT,
    cancelado BIT,
    desviado BIT,
    carrier_delay INT,
    weather_delay INT,
    NAS_delay INT,
```

Modelo de estrella



Evaluation

- Informe de evaluación
- Matriz de evaluación
- Lista de mejoras

Informe de evaluación

1. Proposito
2. Hallazgos clave
3. Conclusión
4. Calidad y complitud del modelo
5. Coherencia entre KPIs del Dashboard y Datos Reales
6. Comparación vs. Objetivos del Business Understanding
7. Validación Cruzada por Roles (Cross-Checking)
8. Conclusión y Próximos Pasos

Lista de mejoras

1. Implementar filtros para outliers en KPIs
2. Automatización completa del proceso ETL
3. Integración de datos meteorológicos externos
4. Incluir datos de tráfico aéreo y control de espacio aéreo
5. Desarrollo de alertas proactivas en el dashboard
6. Optimización del modelo estrella para consultas avanzadas
7. Incorporar análisis predictivo con Machine Learning
8. Mejorar la visualización de KPIs en el dashboard
9. Implementar control de calidad automatizado en cada carga
10. Ampliar el período de análisis con datos más recientes

Matriz de evaluación

Indicador	Descripción	Método de Validación	Resultado	Comentarios
% Puntualidad	Porcentaje de vuelos cuya hora de llegada **ARR_DELAY** y salida **DEP_DELAY** superan los 0 minutos, agrupados por aerolínea.	Consulta directa SQL sobre la tabla **fact_flights**.	Exitoso	Los resultados coinciden con los valores mostrados en el dashboard. Se observa que algunos aeropuertos presentan salidas antes del horario programado, lo que explica ligeras variaciones negativas en los retrasos.
Distribución de retrasos	Análisis de la distribución de minutos de demora por causa: **CarrierDelay, WeatherDelay, NASDelay, SecurityDelay y LateAircraftDelay.**	Consulta directa SQL agregada por tipo de causa.	Exitoso	Los resultados concuerdan con el dashboard. Las mayores demoras se atribuyen a factores operativos de la aerolínea y demoras por aeronaves provenientes de vuelos anteriores.
Comparación programado vs. real	Porcentaje de vuelos cuya hora de llegada **ARR_DELAY** y salida **DEP_DELAY** superan los 0 minutos, agrupados por aeropuerto.	Conteo y comparación directa de registros con **ARR_DELAY** > 0 y **DEP_DELAY** > 0.	Atípico	Se identificaron valores extremos con retrasos de varias horas, atribuibles a condiciones operativas o climatológicas.
Ranking de aerolíneas	Ordenamiento de aerolíneas según el volumen de vuelos registrados y su desempeño promedio en puntualidad.	Agregación por tipo de demora y conteo total de vuelos.	Exitoso	Los resultados reflejan correctamente el ranking, destacando a Southwest Airlines Co. como la aerolínea con mayor número de operaciones y posicionamiento principal en el análisis.
Total de cancelaciones y desvíos	Conteo de vuelos cancelados (**CANCELLED** = 1) y desviados (**DIVERTED** = 1) según los registros del modelo.	Conteo simple y validación de totales.	Exitoso	Los valores del dashboard coinciden con los obtenidos mediante consulta directa. No se identificaron inconsistencias entre las métricas agregadas y los datos fuente.

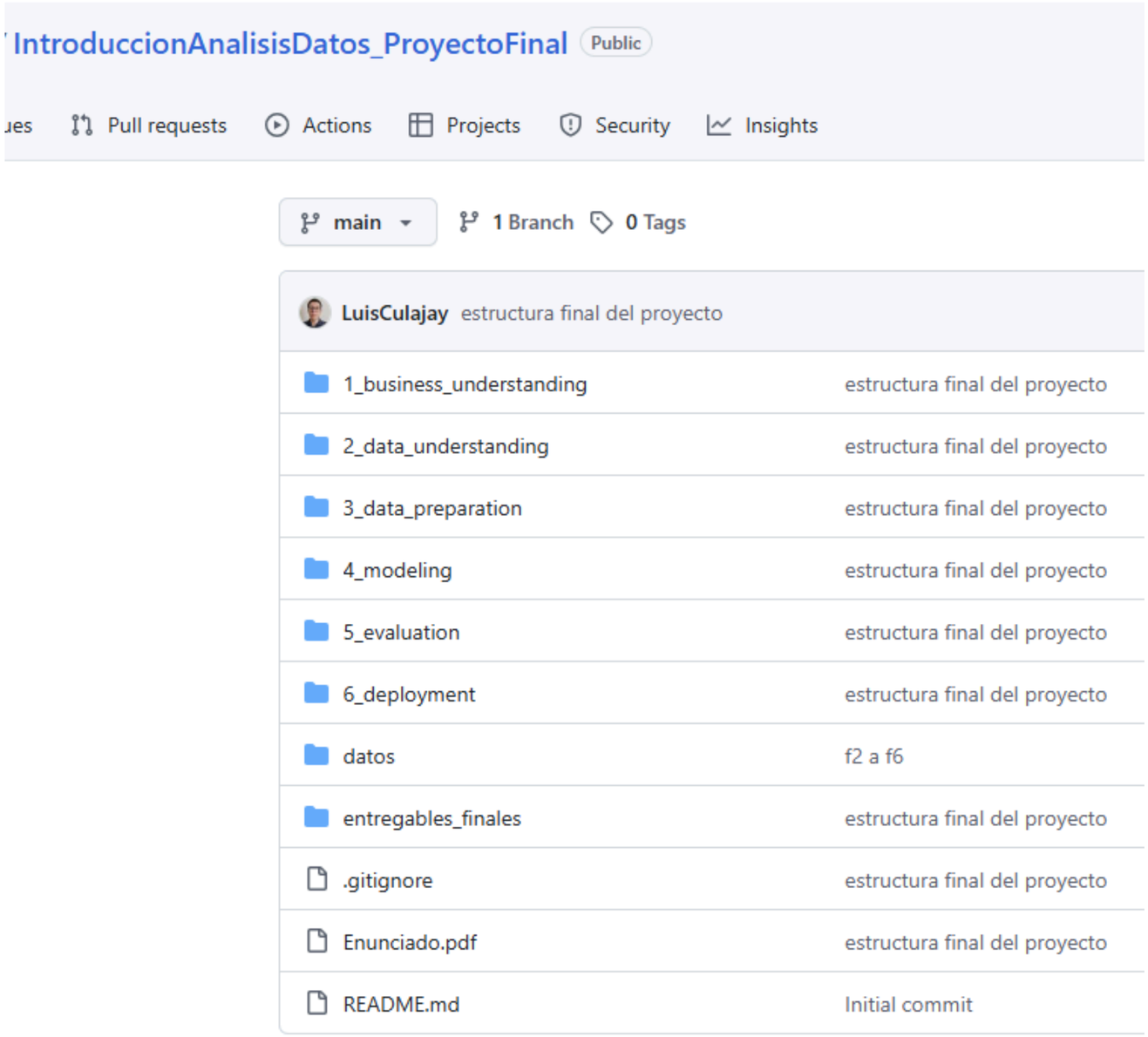
Deployment

- Manual Tecnico
- Repositorio GitHub
- Dashboard

MANUAL TÉCNICO

Elemento	Descripción
Nombre del Proyecto:	Análisis de Desempeño y Retrasos en Vuelos de EE.UU. (2011–2021)
Metodología aplicada:	CRISP-DM
Herramientas principales:	Python, SQL Server, Power BI, Docker
Equipo:	Data Administrator – Karen Morales Data Engineer – Luis Culajay BI Analyst – Jhonatan Simon Data Analyst – Brandon Portillo

GitHub



Fecha

1/1/2011

12.00 mill.
Total de vuelos

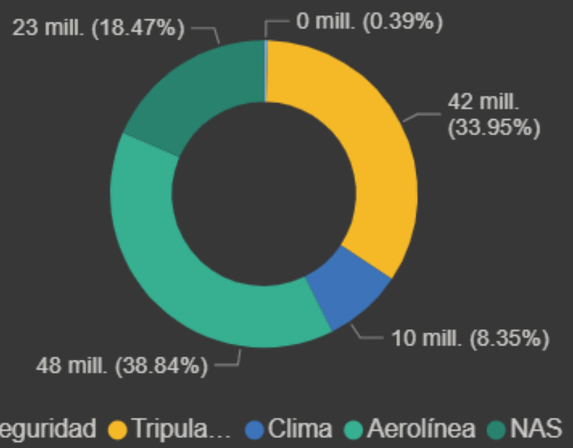
29 mil
Vuelos desviados

193 mil
Vuelos cancelados

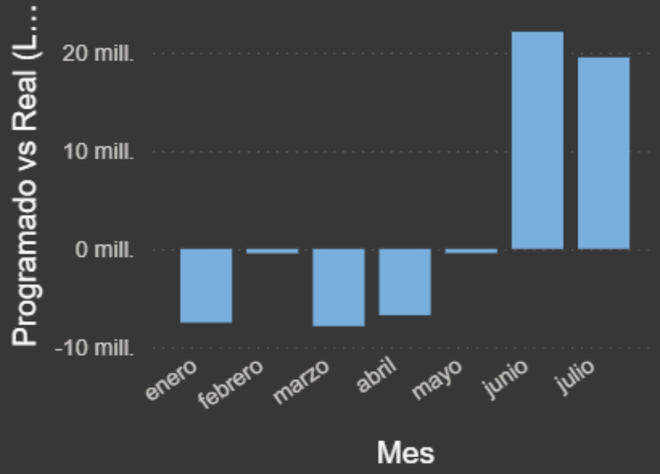
Aerolínea

- Altair Airlines Inc.
- Altus Airlines
- Amerford Airways Inc.
- America West Airlines…
- Americair Inc.
- American Air Transport
- American Airlines Inc.
- American Central Airli…
- American Flag Airline…
- American Flight Group

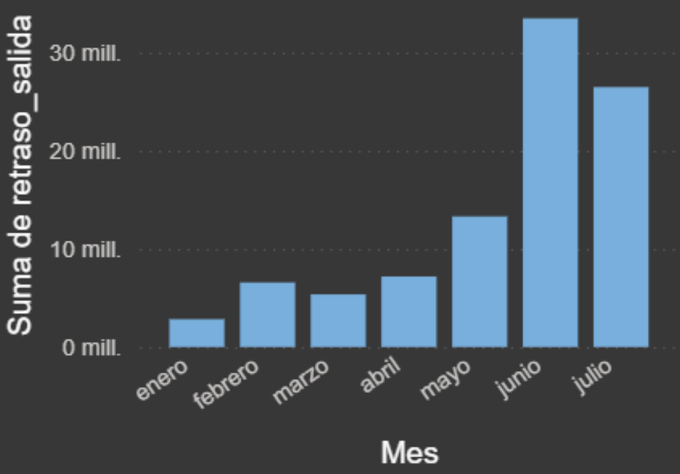
Distribución de retrasos



Programado vs Real (Llegada)



Programado vs Real (Salida)

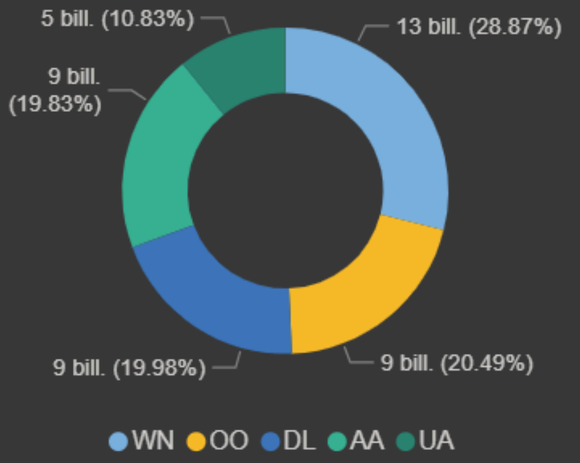


Aeropuerto

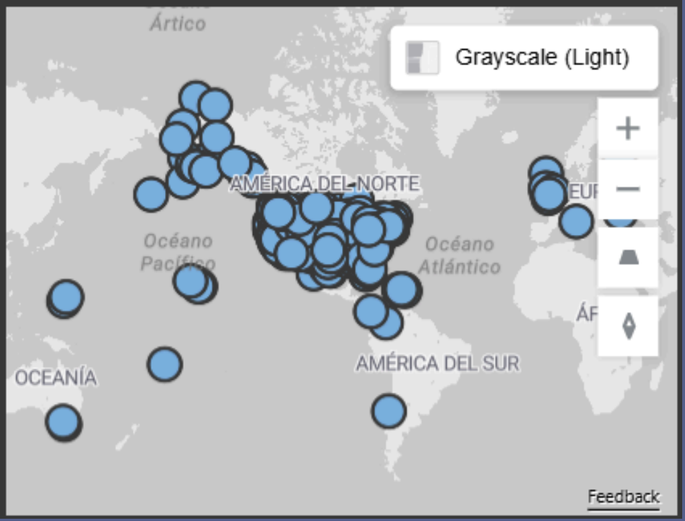
- Aberdeen
- Abilene
- Adak Island
- Aguadilla
- Akron
- Alamosa
- Albany
- Albuquerque
- Alexandria

Aerolinea	Descripcion	Puntualidad
UA	United Air Lines Inc.	7391228
NK	Spirit Air Lines	6242708
WN	Southwest Airlines Co.	30563424
OO	SkyWest Airlines Inc.	15300240
YX	Republic Airline	2397300
OH	PSA Airlines Inc.	2021464
YV	Mesa Airlines Inc.	6998192
B6	JetBlue Airways	10220716
QX	Horizon Air	277132
HA	Hawaiian Airlines Inc.	-28436
F9	Frontier Airlines Inc.	3354304
MQ	Envoy Air	4976516
9E	Endeavor Air Inc.	-2511368
Total		**113804532**

Ranking de aerolíneas



Mapa geográfico por aeropuerto



Muchas gracias

GRUPO 3