

# Proyecto Integrador de Análisis de Datos (CRISP-DM)

## Dataset TranStats (BTS)

**Curso:** Análisis de Datos

**Objetivo general:** Evaluar la aplicación efectiva del contenido del curso de Introducción al Análisis de Datos y otras estrategias según la metodología CRISP-DM mediante la construcción de una solución analítica completa usando datos reales de transporte aéreo de Estados Unidos.

### 1. Contexto General del Proyecto

El Bureau of Transportation Statistics (BTS) ofrece un conjunto muy amplio de datos sobre vuelos, aerolíneas, horarios, demoras, aeropuertos y características operacionales. Los estudiantes deberán utilizar:

**Dataset base:**

[https://www.transtats.bts.gov/DL\\_SelectFields.aspx?gnoyr\\_VQ=FGJ&QO\\_fu146\\_anzr=b0-gvzr](https://www.transtats.bts.gov/DL_SelectFields.aspx?gnoyr_VQ=FGJ&QO_fu146_anzr=b0-gvzr),

El grupo elegirá un subconjunto de datos comprendidos en 11 años dependiendo del siguiente esquema :

- Grupo 1 : 2009 a 2019
- Grupo 2 : 2010 a 2020
- Grupo 3 : 2011 a 2021
- Grupo 4 : 2012 a 2022
- Grupo 5 : 2013 a 2023
- Grupo 6 : 2014 a 2024

Considerar la siguientes categorías de información, elegir los campos que mayor provecho analítico tengan:

- Time Period
- Airline
- Origin and Destination
- Departure Performance and Arrival Performance
- Cancellations and Diversions
- Flight Summaries
- Cause of Delay
- Gate of Return Information at Origin Airport

La solución deberá cubrir el ciclo completo CRISP-DM.

## **2. Roles del Equipo (Máximo 5 estudiantes)**

Cada estudiante debe asumir **uno de los siguientes perfiles** (si hay 5 personas, puede haber un rol duplicado, excepto Data Administrator):

### **1. Data Administrator (obligatorio, solo uno)**

- Se encarga de depuración, diccionario de datos, metadatos, versionado, gobierno de datos.
- Lidera la documentación CRISP-DM.

### **2. BI Analyst**

- Diseña el modelo dimensional. Modelar un esquema estrella o constelación con información de cada campo utilizado y crear una dimensión con información de “Get Lookup Table”
- Define KPIs, métricas, mockups y wireframes del dashboard.
- Hace pruebas de validación con stakeholders (en este caso con compañeros).

### **3. Data Analyst**

- Realiza análisis exploratorio, estadístico y narrativo.
- Formula hallazgos, insights y conclusiones.
- Aplica análisis descriptivo y correlacional.

### **4. Data Engineering**

- Construye el modelo en estrella en SQL.
- Gestiona los contenedores en Docker para base de datos.
- Genera pipelines ETL (puede ser en Python, SQL scripts).

### 3. Fases del Proyecto Según CRISP-DM

A continuación se detallan los entregables por fase.

#### 3.1. Business Understanding

##### ✓ Objetivos analíticos del proyecto

El equipo debe escoger al menos **4 objetivos**, por ejemplo:

- Identificar los aeropuertos con mayor incidencia de retrasos.
- Determinar patrones estacionales en los vuelos.
- Calcular qué aerolíneas presentan mejor desempeño puntual.
- Analizar el impacto del clima o la geografía en los retrasos.

##### ✓ Entregables

1. Documento “Business Understanding” (máx. 2 páginas)
2. Canvas Data Product
3. KPI iniciales
4. Lista de requerimientos analíticos
5. Preguntas de negocio

#### 3.2. Data Understanding

##### Tareas requeridas

1. Descarga del dataset usando la interfaz BTS.
2. Documentar las columnas seleccionadas y justificación.
3. Exploración y análisis inicial (EDA).
4. Detección de valores faltantes y outliers.

Data Analyst

##### Entregables

1. Diccionario de datos forma
- ✗ 2. Resumen de calidad de dato
- ✗ 3. Gráficas exploratorias (mín. 10)
- ✓ 4. Mapa conceptual del dataset

### 3.3. Data Preparation

#### Actividades

1. Limpieza (duplicados, nulos, formatos de fecha).
2. Estandarización de zonas horarias.
3. Creación de columnas derivadas:
  - o retraso\_total
  - o categoría de retraso
  - o semana, trimestre, día festivo
4. Integración de tablas externas (opcional: lista de aeropuertos FAA).

#### Entregables

1. Carpeta con scripts SQL o Python
2. Dataset final listo para modelad
3. Resumen de transformaciones (ETL)

### 3.4. Modeling (Modelo en Estrella)

El equipo deberá construir un **modelo dimensional** con:

#### Tabla de hechos (fact\_flights)

- flight\_id
- airline\_id
- origin\_airport\_id
- destination\_airport\_id
- fecha\_id
- retraso\_salida
- retraso\_llegada
- distancia
- estado\_vuelo, otras a su discreción

#### Tablas de dimensiones

- dim\_airline
- dim\_airport

- ✓ dim\_fecha
- ✓ dim\_vuelo (opcional), otras a su discreción

#### **Entregables**

- ✓ Diagrama físico del modelo en estrella
- ✓ Script SQL para crear la base  
Archivo con carga de datos (INSERT o COPY)

### **3.5. Evaluation**

El equipo deberá validar:

1. Calidad y completitud del modelo
2. Coherencia entre KPIs del dashboard y datos reales
3. Comparación vs. objetivos del Business Understanding
4. Validación cruzada por roles (cross-checking)

#### **Entregables**

1. Informe de evaluación (máx. 3 páginas)
2. Matriz de indicadores validados
3. Lista de mejoras futuras

### **3.6. Deployment (Dashboard)**

#### **Dashboard en Looker Studio**

Debe incluir al menos:

1. KPI de puntualidad por aerolínea
2. Mapa geográfico por aeropuerto
3. Distribución de retrasos
4. Comparación programado vs. real
5. Ranking de aerolíneas/aeropuertos
6. Filtros por fecha, aerolínea, aeropuerto

Mejor!		Solo	Mal!
A	D	B	C
1	-2		
2	5		

#### **Entregables:**

1. Dashboard compartido en modo público
2. Repositorio GitHub del proyecto
3. Manual técnico de instalación

## 4. Evaluación Final del Proyecto

Criterio	Peso
Business Understanding	10%
Data Understanding	15%
Data Preparation	20%
Modelo en Estrella	20%
Dashboard	15%
Análisis final e insights	20%

## 5. Entregables Finales del Grupo

1. Documento formal CRISP-DM (PDF)
2. Modelo en estrella (diagrama + SQL)
3. Dataset limpio
4. Dashboard Looker Studio
5. Presentación final (10 - 15 min)

## Referencias (APA 7)

Bureau of Transportation Statistics. (2024). *TranStats Data Library*. U.S. Department of Transportation. <https://www.transtats.bts.gov/>

Han, J., Kamber, M., & Pei, J. (2022). *Data mining: Concepts and techniques* (4th ed.). Morgan Kaufmann.

Larose, D. T., & Larose, C. D. (2019). *Discovering knowledge in data: An introduction to data mining* (3rd ed.). Wiley.

Sharma, S. (2021). *Data engineering with Python*. Packt Publishing.

Kimball, R., & Ross, M. (2013). *The data warehouse toolkit: The definitive guide to dimensional modeling* (3rd ed.). Wiley.