

```
import pandas as pd
from pathlib import Path

# Load raw CSV
df = pd.read_csv(Path("/Users/gracepolito/Desktop/Master of Data Science/691 Applied Data Science/Airline Business Intelligence Database/data/bts_flights_2024.csv"))

# Normalize column names
df.columns = df.columns.str.lower().str.strip()

# Keep only needed fields
cols = [
    "year", "month", "carrier", "carrier_name", "airport", "airport_name",
    "arr_flights", "arr_del15", "arr_cancelled", "arr_diverted",
    "arr_delay", "carrier_delay", "weather_delay", "nas_delay", "security_delay",
    "late_aircraft_delay"
]
df = df[cols]

# Rename columns for DB consistency
df = df.rename(columns={
    "carrier": "airline_iata",
    "airport": "airport_iata",
    "arr_flights": "arrivals",
    "arr_del15": "arrivals_delayed_15min",
    "arr_delay": "total_arrival_delay_min"
})

# Fill NaNs with 0 for numeric delay values
num_cols = [c for c in df.columns if "delay" in c or "arrivals" in c]
df[num_cols] = df[num_cols].fillna(0)

# Add a unique key for loading
df["snapshot_id"] = (
    df["year"].astype(str) + "_" +
    df["month"].astype(str).str.zfill(2) + "_" +
    df["airline_iata"] + "_" + df["airport_iata"]
)

# Export cleaned version
df.to_csv("/Users/gracepolito/Desktop/Master of Data Science/691 Applied Data Science/Airline Business Intelligence Database/data/bts_cleaned.csv", index=False)
print("✅ Cleaned data saved to data/bts_cleaned.csv")
```