Global Air Transportation Network

Philip Spaier and Emile Klerner

Source Data

Global Air Transportation Network

Valid for 2022
Source: Kaggle

- routes.csv (67661 rows)
- airports.csv (7697 rows)
- airplanes.csv (246 rows)
- airlines.csv (6162 rows)

Economic and Social Data

Valid for 2023
Source: World Bank DataBank

- country_gdp.csv (265 rows)
  - gdp
  - social and political stability
  - population size

Data Cleaning

Standardizing NULL Values

	index	Airline ID	Name	Alias	IATA	ICAO	Callsign
215	214	214	Air Berlin	\N	AB	BER	AIR BERLIN
216	215	215	Air Brousse	\N	<null>	ABT	AIR BROUSSE
217	216	216	Air Contractors	\N	AG	ABR	CONTRACT
218	217	217	Air Illinois	\N	<null>	AIL	AIR ILLINOIS

	Airline ID	Name	Alias	IATA	ICAO	Callsign
213	214	Air Berlin	<null>	AB	BER	AIR BERLIN
214	215	Air Brousse	<null>	<null>	ABT	AIR BROUSSE
215	216	Air Contractors	<null>	<null>	ABR	CONTRACT
216	217	Air Illinois	<null>	<null>	AIL	AIR ILLINOIS

Data Cleaning

Standardizing Boolean Values

	Source airport	Source airport ID	Destination airport	Destination airport ID	Codeshare	Stops	Equipment
5266	DFW	3670	MIA	3576	<null>	0	738 763 757
5267	DFW	3670	MKE	3717	<null>	0	M80 M83
5268	DFW	3670	MLI	4072	Y	0	ERD ER4
5269	DFW	3670	MLM	1821	Y	0	ER4

	Source airport	Source airport ID	Destination airport	Destination airport ID	Codeshare	Stops	Equipment
13921	HDS	811	JNB	813	1	0	DH4
13922	HDS	811	CPT	797	1	0	CR2
13923	GRU	2564	JNB	813	0	0	346 332 343
13924	GRJ	804	JNB	813	0	0	738 733

Data Cleaning

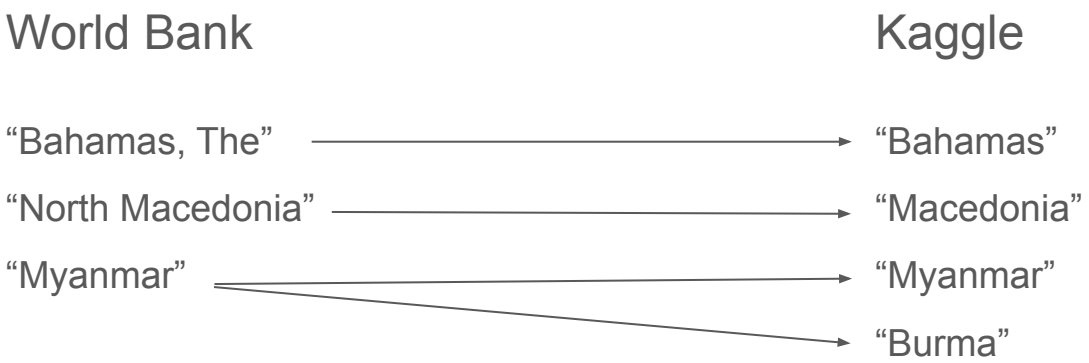
Additional Cleaning Steps:

- Removing rows where a primary key (PK) is empty
- Removing duplicate rows
- Removing all rows where a primary key (PK) is assigned to multiple unique rows (likely due to outdated data)
- Removing rows where almost all attributes are missing
- Removing empty strings

Matching Country Names

The Problem:

Country names are not the same in the Kaggle and WorldBank datasets

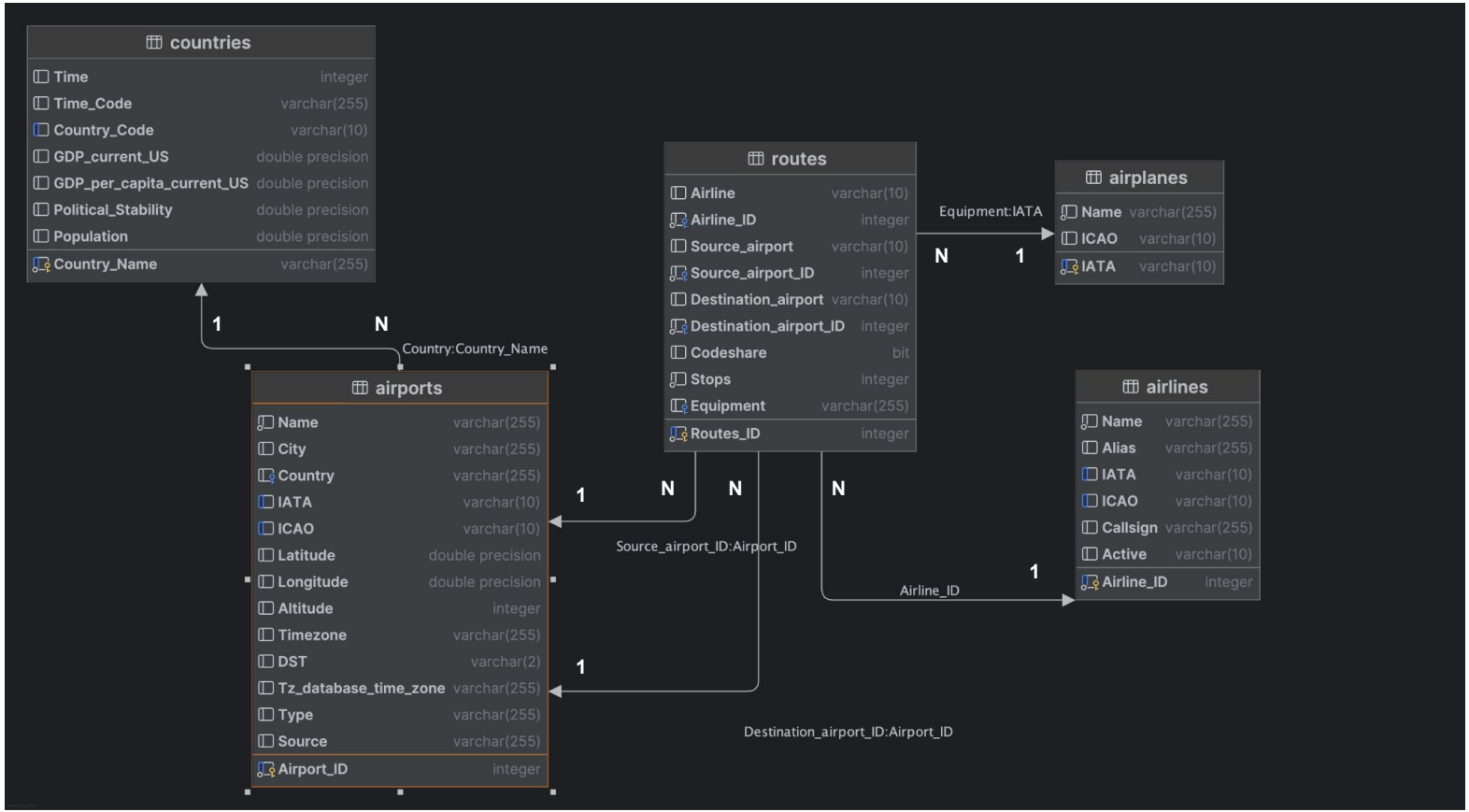


Matching Country Names

The Solution:

1. A fuzzy algorithm (rapidfuzz) matching country names with above 90% similarity and creating a lookup table
2. Manually matching countries that have not been correctly mapped
3. Removing entries from the WorldBank data relating to countries or entities not present in our airport data.

ER Diagram


countries
Time integer
Time_Code varchar(255)
Country_Code varchar(10)
GDP_current_US double precision
GDP_per_capita_current_US double precision
Political_Stability double precision
Population double precision
Country_Name varchar(255)
1
N
Country:Country_Name
airports
Name varchar(255)
City varchar(255)
Country varchar(255)
IATA varchar(10)
ICAO varchar(10)
Latitude double precision
Longitude double precision
Altitude integer
Timezone varchar(255)
DST varchar(2)
Tz_database_time_zone varchar(255)
Type varchar(255)
Source varchar(255)
Airport_ID integer
routes
Airline varchar(10)
Airline_ID integer
Source_airport varchar(10)
Source_airport_ID integer
Destination_airport varchar(10)
Destination_airport_ID integer
Codeshare bit
Stops integer
Equipment varchar(255)
Routes_ID integer
Equipment:IATA
N
1
airplanes
Name varchar(255)
ICAO varchar(10)
IATA varchar(10)
1
N
N
Source_airport_ID:Airport_ID
N
Airline_ID
1
airlines
Name varchar(255)
Alias varchar(255)
IATA varchar(10)
ICAO varchar(10)
Callsign varchar(255)
Active varchar(10)
Airline_ID integer
1
Destination_airport_ID:Airport_ID

Questions

1) Which airlines are inactive?
2) Show all airports located in either Germany or Austria. Include name, city and country
3) Show all routes with one stop, with the name of the airline, departure and arrival airports (IATA code)
4) Which airline have more than 80 routes?
5) Which countries have the highest percentage of domestic routes?
6) Which aircraft types (Equipment codes) appear on the largest number of distinct routes?
7) What is the average number of unique destination countries reachable from each country
8) What are the top 5 countries where the ratio of total outgoing routes (from all airports in the country) to the country's GDP per Capita is the highest?
9) Which airports have the largest disparity between the number of outgoing and incoming routes?
10) What are the top 10 cities globally, based on the total number of airports?

Question 1 - SQL + Basic Relational Algebra

Which airlines are inactive? Provide ID and Name

```
query_question01 = """
SELECT
    "Airline_ID",
    "Name"
FROM airlines
WHERE "Active" = 'N';
"""
q1_df = pd.read_sql(query_question01, engine)
q1_df
```

SQL Query - Question 01

	Airline_ID	Name
0	2	135 Airways
1	4	2 Sqn No 1 Elementary Flying Training School
2	5	213 Flight Unit
3	6	223 Flight Unit State Airline
4	7	224th Flight Unit
...	...	...
4901	20963	Atlantic Air Cargo
4902	21056	Dummy
4903	21181	Air Andaman (2Y)
4904	21240	TDA Toa Domestic Airlines
4905	21251	Lynx Aviation (L3/SSX)

Result - Question 01

$$\pi_{Airline\_ID,\ Name}\left(\sigma_{Active = 'N'}(airlines)\right)$$

Relational Algebra - Question 01

Question 2 - SQL + Basic Relational Algebra

Show all airports located in either Germany or Austria. Include name, city and country

```
query_question02 = """
SELECT
    "Name",
    "City",
    "Country"
FROM airports
WHERE "Country" = 'Germany'
UNION
SELECT
    "Name",
    "City",
    "Country"
FROM airports
WHERE "Country" = 'Austria'
"""

q2_df = pd.read_sql(query_question02, engine)
q2_df
```

SQL Query - Question 02

	Name	City	Country
0	Hamburg Airport	Hamburg	Germany
1	Geilenkirchen Air Base	Geilenkirchen	Germany
2	Vilshofen Airport	Vilshofen	Germany
3	St. Johann In Tirol Airport	St. Johann in Tirol	Austria
4	Wipperfürth-Neye Airport	Wipperfuerth	Germany
...	...	...	...
264	Hohn Air Base	Hohn	Germany
265	Rügen Airport	Ruegen	Germany
266	Torgau-Beilrode Airport	Gransee	Germany
267	Aalen-Heidenheim/Elchingen Airport	Aalen-heidenheim	Germany
268	Rothenburg/Görlitz Airport	Rothenburg/ol	Germany

Result - Question 02

$$\pi_{Name, City, Country}(\sigma_{Country = \text{"Germany"}}(airports)) \cup \pi_{Name, City, Country}(\sigma_{Country = \text{"Austria"}}(airports))$$

Relational Algebra - Question 02

Question 3 - SQL + Extended Relational Algebra

Show all routes with one stop, with the name of the airline, departure and arrival airports (IATA code)

```sql
query_question03 = """
SELECT
    a."Name",
    r."Source_airport",
    r."Destination_airport"
FROM routes r
JOIN airlines a ON r."Airline_ID" = a."Airline_ID"
WHERE r."Stops" = 1
"""

q3_df = pd.read_sql(query_question03, engine)
q3_df
```

SQL Query - Question 03

	Name	Source_airport	Destination_airport
0	Canadian North	YRT	YEK
1	Air Canada	ABJ	BRU
2	Air Canada	YVR	YBL
3	Cubana de Aviación	FCO	HAV
4	AirTran Airways	HOU	SAT
5	AirTran Airways	MCO	ORF
6	Scandinavian Airlines System	ARN	GEV

Result - Question 03

$$\Pi_{a.Name,\ r.Source\_airport,\ r.Destination\_airport}\left(\rho_r\left(\sigma_{Stops=1}(routes)\right) \bowtie_{r.Airline\_ID = a.Airline\_ID} \rho_a(airlines)\right)$$

Relational Algebra - Question 03

Question 4 - SQL + Extended Relational Algebra

Which airline have more than 80 routes?

```
SELECT
    a."Name",
    COUNT(*) AS Route_Count
FROM routes r
JOIN airlines a ON r."Airline_ID" = a."Airline_ID"
GROUP BY a."Name"
HAVING COUNT(*) > 80;
"""
q4_df = pd.read_sql(query_question04, engine)
q4_df
```

SQL Query - Question 04

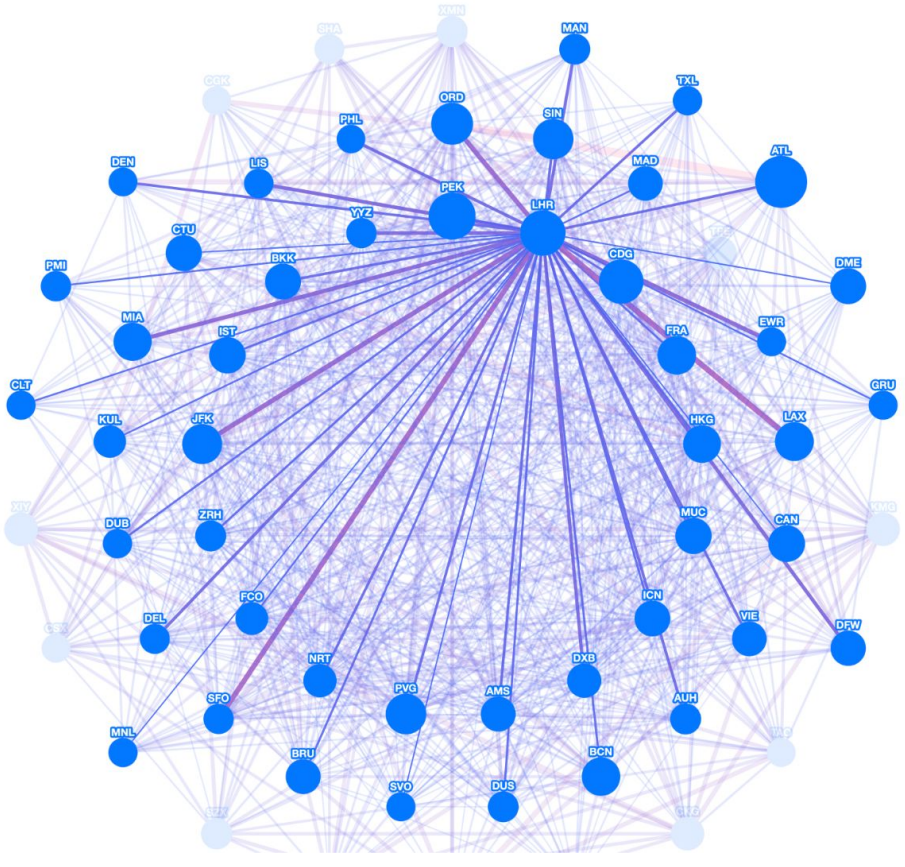
	Name	route_count
0	Air Bourbon	210
1	TransAsia Airways	92
2	Air India Limited	364
3	Meridiana	140
4	EVA Air	114
...	...	...
136	LOT Polish Airlines	114
137	Sriwijaya Air	106
138	LAN Airlines	285
139	Iberia Airlines	797
140	Philippine Airlines	144

Result - Question 04

$$\Pi_{Name, Route\_Count}(\sigma_{Route\_count > 80}(\gamma_{Name;\ COUNT(*) \rightarrow Route\_Count}(routes \bowtie_{routes.Airline\_ID = airlines.Airline\_ID} airlines)))$$

Relational Algebra - Question 04

Graphical Representation



Legend

Airports (Nodes)

Top **60 airports** selected by route count. **Larger circles = more connections.**

Routes (Edges)

Line thickness shows route frequency:

- Few flights
- Medium
- Many flights

Only routes with **≥2 connections** shown.

Values

Connections: Total routes in raw data.
Neighbors: Connected top airports (≥2 connections).

Interaction

Click airport to show only its routes. Click again to reset.

Layout

Concentric

Rearrange Airports

Reset View

Route Network Visualization
LHR

Connections: 985
Neighbors: 46