BIG DATA

Elías Ramos Calderón

New Technology School

Tokio.

New
Technology
School

Elías Ramos Calderón

RESPONSABLE DEL DESARROLLO DE LA APLICACIÓN

Fecha: 21, Julio, 2024

BIG DATA
PROYECTO FINAL
Elías Ramos Calderón

Tokio.

ÍNDICE:

Categorización de los Datos:

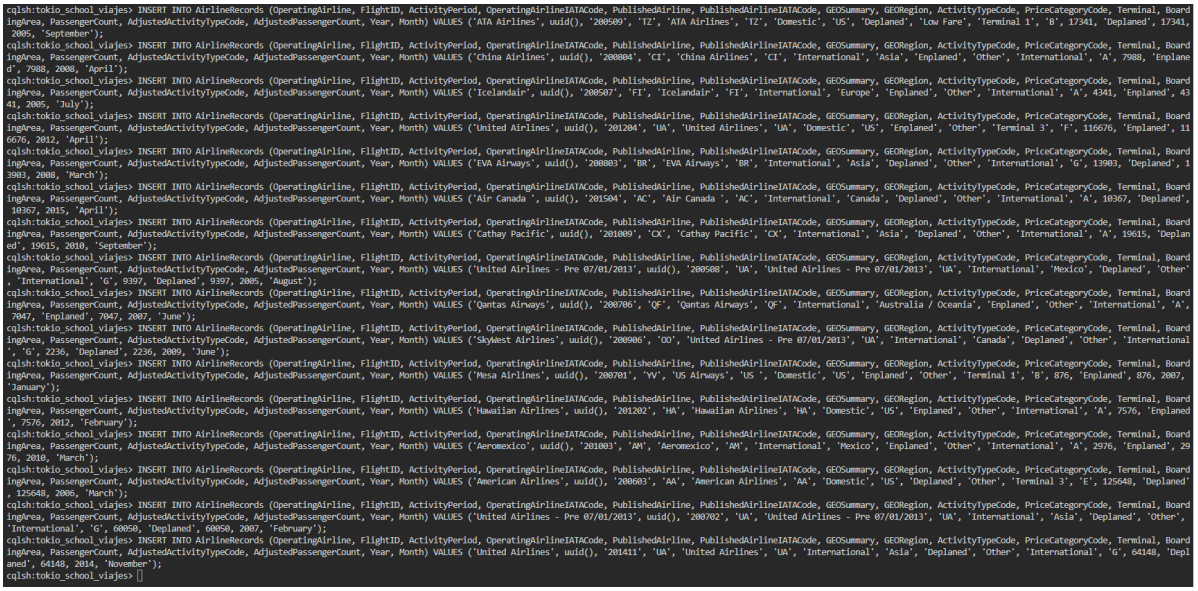
Nombre del Campo	Tipo de Dato
Activity Period	Numérico Discreto
Operating Airline	Categórico Nominal
Operating Airline IATA Code	Categórico Nominal
Published Airline	Categórico Nominal
Published Airline IATA Code	Categórico Nominal
GEO Summary	Categórico Binario
GEO Region	Categórico Nominal
Activity Type Code	Categórico Nominal
Price Category Code	Categórico Binario
Terminal	Categórico Nominal
Boarding Area	Categórico Nominal
Passenger Count	Numérico Discreto
Adjusted Activity Type Code	Categórico Nominal
Adjusted Passenger Count	Numérico Discreto
Year	Numérico Discreto
Month	Categórico Ordinal

Cassandra

```
cqlsh> CREATE KEYSPACE IF NOT EXISTS Tokio_School_Viajes WITH replication = {'class':'SimpleStrategy', 'replication_factor': 1};
cqlsh> DESCRIBE KEYSPACES;

tokio_school_viajes    system_schema  system_views  system_distributed
system_virtual_schema  system_auth    system        system_traces

cqlsh> USE Tokio_School_Viajes;
cqlsh:tokio_school_viajes> CREATE TABLE IF NOT EXISTS AirlineRecords (
                       ...     OperatingAirline TEXT,
                       ...     FlightID UUID,
                       ...     ActivityPeriod TEXT,
                       ...     OperatingAirlineIATACode TEXT,
                       ...     PublishedAirline TEXT,
                       ...     PublishedAirlineIATACode TEXT,
                       ...     GEOSummary TEXT,
                       ...     GEORegion TEXT,
                       ...     ActivityTypeCode TEXT,
                       ...     PriceCategoryCode TEXT,
                       ...     Terminal TEXT,
                       ...     BoardingArea TEXT,
                       ...     PassengerCount FLOAT,
                       ...     AdjustedActivityTypeCode TEXT,
                       ...     AdjustedPassengerCount INT,
                       ...     Year INT,
                       ...     Month TEXT,
                       ...     PRIMARY KEY (OperatingAirline, FlightID)
                       ... );
cqlsh:tokio_school_viajes> CREATE TABLE IF NOT EXISTS FlightBoarding (
                       ...     OperatingAirline TEXT,
                       ...     BoardingArea TEXT,
                       ...     FlightID UUID,
                       ...     ActivityPeriod TEXT,
                       ...     OperatingAirlineIATACode TEXT,
                       ...     PublishedAirline TEXT,
                       ...     PublishedAirlineIATACode TEXT,
                       ...     GEOSummary TEXT,
                       ...     GEORegion TEXT,
                       ...     ActivityTypeCode TEXT,
                       ...     PriceCategoryCode TEXT,
                       ...     Terminal TEXT,
                       ...     PassengerCount INT,
                       ...     AdjustedActivityTypeCode TEXT,
                       ...     AdjustedPassengerCount INT,
                       ...     Year INT,
                       ...     Month TEXT,
                       ...     PRIMARY KEY ((OperatingAirline, BoardingArea), FlightID)
                       ... )
                       ... ;
cqlsh:tokio_school_viajes>
```

Comenzamos creando la base de datos en "Cassandra", a la que nombraremos como Tokio_school_viajes. Se procederá a crear tantas tablas sean necesarias para la obtención de los diferentes datos. En mi caso, solo se necesito realizar dos, las cuales nombre como Airlinerecords y flghyboarding; estás fueron creadas utilizando la misma estructura, pero con diferentes claves primarias; con el fin de adquirir los datos necesarios.

Una vez creadas las tablas procedemos a insertar todos los registros en cada tabla correspondiente, como son los siguientes:

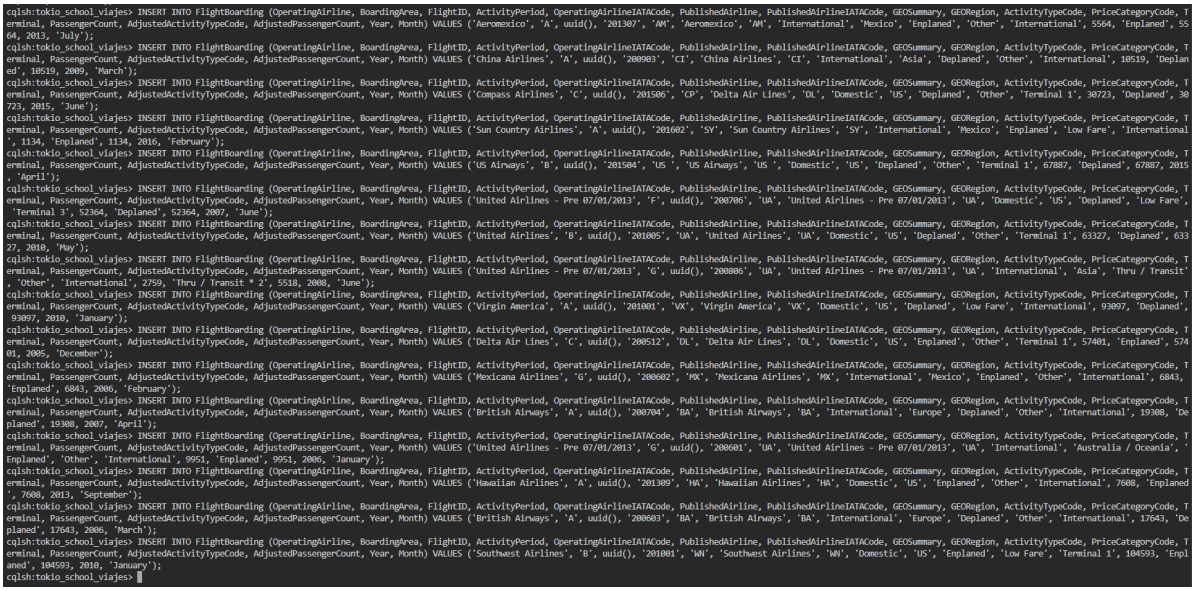


```
cqlsh:tokio_school_viajes> INSERT INTO AirlineRecords (OperatingAirline, FlightID, ActivityPeriod, OperatingAirlineIATACode, PublishedAirline, PublishedAirlineIATACode, GEOSummary, GEORegion, ActivityTypeCode, PriceCategoryCode, Terminal, Board
ingArea, PassengerCount, AdjustedActivityTypeCode, AdjustedPassengerCount, Year, Month) VALUES ('ATA Airlines', uuid(), '200509', 'TZ', 'ATA Airlines', 'TZ', 'Domestic', 'US', 'Deplaned', 'Low Fare', 'Terminal 1', 'B', 17341, 'Deplaned', 17341,
 2005, 'September');
cqlsh:tokio_school_viajes> INSERT INTO AirlineRecords (OperatingAirline, FlightID, ActivityPeriod, OperatingAirlineIATACode, PublishedAirline, PublishedAirlineIATACode, GEOSummary, GEORegion, ActivityTypeCode, PriceCategoryCode, Terminal, Board
ingArea, PassengerCount, AdjustedActivityTypeCode, AdjustedPassengerCount, Year, Month) VALUES ('China Airlines', uuid(), '200804', 'CI', 'China Airlines', 'CI', 'International', 'Asia', 'Enplaned', 'Other', 'International', 'A', 7988, 'Enplane
d', 7988, 2008, 'April');
cqlsh:tokio_school_viajes> INSERT INTO AirlineRecords (OperatingAirline, FlightID, ActivityPeriod, OperatingAirlineIATACode, PublishedAirline, PublishedAirlineIATACode, GEOSummary, GEORegion, ActivityTypeCode, PriceCategoryCode, Terminal, Board
ingArea, PassengerCount, AdjustedActivityTypeCode, AdjustedPassengerCount, Year, Month) VALUES ('Icelandair', uuid(), '200507', 'FI', 'Icelandair', 'FI', 'International', 'Europe', 'Enplaned', 'Other', 'International', 'A', 4341, 'Enplaned', 43
41, 2005, 'July');
cqlsh:tokio_school_viajes> INSERT INTO AirlineRecords (OperatingAirline, FlightID, ActivityPeriod, OperatingAirlineIATACode, PublishedAirline, PublishedAirlineIATACode, GEOSummary, GEORegion, ActivityTypeCode, PriceCategoryCode, Terminal, Board
ingArea, PassengerCount, AdjustedActivityTypeCode, AdjustedPassengerCount, Year, Month) VALUES ('United Airlines', uuid(), '201204', 'UA', 'United Airlines', 'UA', 'Domestic', 'US', 'Enplaned', 'Other', 'Terminal 3', 'F', 116676, 'Enplaned', 11
6676, 2012, 'April');
cqlsh:tokio_school_viajes> INSERT INTO AirlineRecords (OperatingAirline, FlightID, ActivityPeriod, OperatingAirlineIATACode, PublishedAirline, PublishedAirlineIATACode, GEOSummary, GEORegion, ActivityTypeCode, PriceCategoryCode, Terminal, Board
ingArea, PassengerCount, AdjustedActivityTypeCode, AdjustedPassengerCount, Year, Month) VALUES ('EVA Airways', uuid(), '200803', 'BR', 'EVA Airways', 'BR', 'International', 'Asia', 'Deplaned', 'Other', 'International', 'G', 13903, 'Deplaned', 1
3903, 2008, 'March');
cqlsh:tokio_school_viajes> INSERT INTO AirlineRecords (OperatingAirline, FlightID, ActivityPeriod, OperatingAirlineIATACode, PublishedAirline, PublishedAirlineIATACode, GEOSummary, GEORegion, ActivityTypeCode, PriceCategoryCode, Terminal, Board
ingArea, PassengerCount, AdjustedActivityTypeCode, AdjustedPassengerCount, Year, Month) VALUES ('Air Canada ', uuid(), '201504', 'AC', 'Air Canada ', 'AC', 'International', 'Canada', 'Deplaned', 'Other', 'International', 'A', 10367, 'Deplaned',
 10367, 2015, 'April');
cqlsh:tokio_school_viajes> INSERT INTO AirlineRecords (OperatingAirline, FlightID, ActivityPeriod, OperatingAirlineIATACode, PublishedAirline, PublishedAirlineIATACode, GEOSummary, GEORegion, ActivityTypeCode, PriceCategoryCode, Terminal, Board
ingArea, PassengerCount, AdjustedActivityTypeCode, AdjustedPassengerCount, Year, Month) VALUES ('Cathay Pacific', uuid(), '201009', 'CX', 'Cathay Pacific', 'CX', 'International', 'Asia', 'Deplaned', 'Other', 'International', 'A', 19615, 'Deplan
ed', 19615, 2010, 'September');
cqlsh:tokio_school_viajes> INSERT INTO AirlineRecords (OperatingAirline, FlightID, ActivityPeriod, OperatingAirlineIATACode, PublishedAirline, PublishedAirlineIATACode, GEOSummary, GEORegion, ActivityTypeCode, PriceCategoryCode, Terminal, Board
ingArea, PassengerCount, AdjustedActivityTypeCode, AdjustedPassengerCount, Year, Month) VALUES ('United Airlines - Pre 07/01/2013', uuid(), '200508', 'UA', 'United Airlines - Pre 07/01/2013', 'UA', 'International', 'Mexico', 'Deplaned', 'Other
', 'International', 'G', 9397, 'Deplaned', 9397, 2005, 'August');
cqlsh:tokio_school_viajes> INSERT INTO AirlineRecords (OperatingAirline, FlightID, ActivityPeriod, OperatingAirlineIATACode, PublishedAirline, PublishedAirlineIATACode, GEOSummary, GEORegion, ActivityTypeCode, PriceCategoryCode, Terminal, Board
ingArea, PassengerCount, AdjustedActivityTypeCode, AdjustedPassengerCount, Year, Month) VALUES ('Qantas Airways', uuid(), '200706', 'QF', 'Qantas Airways', 'QF', 'International', 'Australia / Oceania', 'Enplaned', 'Other', 'International', 'A',
 7047, 'Enplaned', 7047, 2007, 'June');
cqlsh:tokio_school_viajes> INSERT INTO AirlineRecords (OperatingAirline, FlightID, ActivityPeriod, OperatingAirlineIATACode, PublishedAirline, PublishedAirlineIATACode, GEOSummary, GEORegion, ActivityTypeCode, PriceCategoryCode, Terminal, Board
ingArea, PassengerCount, AdjustedActivityTypeCode, AdjustedPassengerCount, Year, Month) VALUES ('SkyWest Airlines', uuid(), '200906', 'OO', 'United Airlines - Pre 07/01/2013', 'UA', 'International', 'Canada', 'Deplaned', 'Other', 'International
', 'G', 2236, 'Deplaned', 2236, 2009, 'June');
cqlsh:tokio_school_viajes> INSERT INTO AirlineRecords (OperatingAirline, FlightID, ActivityPeriod, OperatingAirlineIATACode, PublishedAirline, PublishedAirlineIATACode, GEOSummary, GEORegion, ActivityTypeCode, PriceCategoryCode, Terminal, Board
ingArea, PassengerCount, AdjustedActivityTypeCode, AdjustedPassengerCount, Year, Month) VALUES ('Mesa Airlines', uuid(), '200701', 'YV', 'US Airways', 'US ', 'Domestic', 'US', 'Enplaned', 'Other', 'Terminal 1', 'B', 876, 'Enplaned', 876, 2007,
'January');
cqlsh:tokio_school_viajes> INSERT INTO AirlineRecords (OperatingAirline, FlightID, ActivityPeriod, OperatingAirlineIATACode, PublishedAirline, PublishedAirlineIATACode, GEOSummary, GEORegion, ActivityTypeCode, PriceCategoryCode, Terminal, Board
ingArea, PassengerCount, AdjustedActivityTypeCode, AdjustedPassengerCount, Year, Month) VALUES ('Hawaiian Airlines', uuid(), '201202', 'HA', 'Hawaiian Airlines', 'HA', 'Domestic', 'US', 'Enplaned', 'Other', 'International', 'A', 7576, 'Enplaned
', 7576, 2012, 'February');
cqlsh:tokio_school_viajes> INSERT INTO AirlineRecords (OperatingAirline, FlightID, ActivityPeriod, OperatingAirlineIATACode, PublishedAirline, PublishedAirlineIATACode, GEOSummary, GEORegion, ActivityTypeCode, PriceCategoryCode, Terminal, Board
ingArea, PassengerCount, AdjustedActivityTypeCode, AdjustedPassengerCount, Year, Month) VALUES ('Aeromexico', uuid(), '201003', 'AM', 'Aeromexico', 'AM', 'International', 'Mexico', 'Enplaned', 'Other', 'International', 'A', 2976, 'Enplaned', 29
76, 2010, 'March');
cqlsh:tokio_school_viajes> INSERT INTO AirlineRecords (OperatingAirline, FlightID, ActivityPeriod, OperatingAirlineIATACode, PublishedAirline, PublishedAirlineIATACode, GEOSummary, GEORegion, ActivityTypeCode, PriceCategoryCode, Terminal, Board
ingArea, PassengerCount, AdjustedActivityTypeCode, AdjustedPassengerCount, Year, Month) VALUES ('American Airlines', uuid(), '200603', 'AA', 'American Airlines', 'AA', 'Domestic', 'US', 'Deplaned', 'Other', 'Terminal 3', 'E', 125648, 'Deplaned'
, 125648, 2006, 'March');
cqlsh:tokio_school_viajes> INSERT INTO AirlineRecords (OperatingAirline, FlightID, ActivityPeriod, OperatingAirlineIATACode, PublishedAirline, PublishedAirlineIATACode, GEOSummary, GEORegion, ActivityTypeCode, PriceCategoryCode, Terminal, Board
ingArea, PassengerCount, AdjustedActivityTypeCode, AdjustedPassengerCount, Year, Month) VALUES ('United Airlines - Pre 07/01/2013', uuid(), '200702', 'UA', 'United Airlines - Pre 07/01/2013', 'UA', 'International', 'Asia', 'Deplaned', 'Other',
'International', 'G', 60050, 'Deplaned', 60050, 2007, 'February');
cqlsh:tokio_school_viajes> INSERT INTO AirlineRecords (OperatingAirline, FlightID, ActivityPeriod, OperatingAirlineIATACode, PublishedAirline, PublishedAirlineIATACode, GEOSummary, GEORegion, ActivityTypeCode, PriceCategoryCode, Terminal, Board
ingArea, PassengerCount, AdjustedActivityTypeCode, AdjustedPassengerCount, Year, Month) VALUES ('United Airlines', uuid(), '201411', 'UA', 'United Airlines', 'UA', 'International', 'Asia', 'Deplaned', 'Other', 'International', 'G', 64148, 'Depl
aned', 64148, 2014, 'November');
cqlsh:tokio_school_viajes>
```

- AirlineRecords

INSERT INTO AirlineRecords (OperatingAirline, FlightID, ActivityPeriod, OperatingAirlineIATACode, PublishedAirline, PublishedAirlineIATACode, GEOSummary, GEORegion, ActivityTypeCode, PriceCategoryCode, Terminal, BoardingArea, PassengerCount, AdjustedActivityTypeCode, AdjustedPassengerCount, Year, Month)

VALUES ('United Airlines - Pre 07/01/2013', uuid(), '200606', 'UA', 'United Airlines - Pre 07/01/2013', 'UA', 'International', 'Mexico', 'Deplaned', 'Other', 'International', 'G', 11085, 'Deplaned', 11085, 2006, 'June');



```
cqlsh:tokio_school_viajes> INSERT INTO FlightBoarding (OperatingAirline, BoardingArea, FlightID, ActivityPeriod, OperatingAirlineIATACode, PublishedAirline, PublishedAirlineIATACode, GEOSummary, GEORegion, ActivityTypeCode, PriceCategoryCode, T
erminal, PassengerCount, AdjustedActivityTypeCode, AdjustedPassengerCount, Year, Month) VALUES ('Aeromexico', 'A', uuid(), '201307', 'AM', 'Aeromexico', 'AM', 'International', 'Mexico', 'Enplaned', 'Other', 'International', 5564, 'Enplaned', 55
64, 2013, 'July');
cqlsh:tokio_school_viajes> INSERT INTO FlightBoarding (OperatingAirline, BoardingArea, FlightID, ActivityPeriod, OperatingAirlineIATACode, PublishedAirline, PublishedAirlineIATACode, GEOSummary, GEORegion, ActivityTypeCode, PriceCategoryCode, T
erminal, PassengerCount, AdjustedActivityTypeCode, AdjustedPassengerCount, Year, Month) VALUES ('China Airlines', 'A', uuid(), '200903', 'CI', 'China Airlines', 'CI', 'International', 'Asia', 'Deplaned', 'Other', 'International', 10519, 'Deplan
ed', 10519, 2009, 'March');
cqlsh:tokio_school_viajes> INSERT INTO FlightBoarding (OperatingAirline, BoardingArea, FlightID, ActivityPeriod, OperatingAirlineIATACode, PublishedAirline, PublishedAirlineIATACode, GEOSummary, GEORegion, ActivityTypeCode, PriceCategoryCode, T
erminal, PassengerCount, AdjustedActivityTypeCode, AdjustedPassengerCount, Year, Month) VALUES ('Compass Airlines', 'C', uuid(), '201506', 'CP', 'Delta Air Lines', 'DL', 'Domestic', 'US', 'Deplaned', 'Other', 'Terminal 1', 30723, 'Deplaned', 30
723, 2015, 'June');
cqlsh:tokio_school_viajes> INSERT INTO FlightBoarding (OperatingAirline, BoardingArea, FlightID, ActivityPeriod, OperatingAirlineIATACode, PublishedAirline, PublishedAirlineIATACode, GEOSummary, GEORegion, ActivityTypeCode, PriceCategoryCode, T
erminal, PassengerCount, AdjustedActivityTypeCode, AdjustedPassengerCount, Year, Month) VALUES ('Sun Country Airlines', 'A', uuid(), '201602', 'SY', 'Sun Country Airlines', 'SY', 'International', 'Mexico', 'Enplaned', 'Low Fare', 'International
', 1134, 'Enplaned', 1134, 2016, 'February');
cqlsh:tokio_school_viajes> INSERT INTO FlightBoarding (OperatingAirline, BoardingArea, FlightID, ActivityPeriod, OperatingAirlineIATACode, PublishedAirline, PublishedAirlineIATACode, GEOSummary, GEORegion, ActivityTypeCode, PriceCategoryCode, T
erminal, PassengerCount, AdjustedActivityTypeCode, AdjustedPassengerCount, Year, Month) VALUES ('US Airways', 'B', uuid(), '201504', 'US ', 'US Airways', 'US ', 'Domestic', 'US', 'Deplaned', 'Other', 'Terminal 1', 67887, 'Deplaned', 67887, 2015
, 'April');
cqlsh:tokio_school_viajes> INSERT INTO FlightBoarding (OperatingAirline, BoardingArea, FlightID, ActivityPeriod, OperatingAirlineIATACode, PublishedAirline, PublishedAirlineIATACode, GEOSummary, GEORegion, ActivityTypeCode, PriceCategoryCode, T
erminal, PassengerCount, AdjustedActivityTypeCode, AdjustedPassengerCount, Year, Month) VALUES ('United Airlines - Pre 07/01/2013', 'F', uuid(), '200706', 'UA', 'United Airlines - Pre 07/01/2013', 'UA', 'Domestic', 'US', 'Deplaned', 'Low Fare',
 'Terminal 3', 52364, 'Deplaned', 52364, 2007, 'June');
cqlsh:tokio_school_viajes> INSERT INTO FlightBoarding (OperatingAirline, BoardingArea, FlightID, ActivityPeriod, OperatingAirlineIATACode, PublishedAirline, PublishedAirlineIATACode, GEOSummary, GEORegion, ActivityTypeCode, PriceCategoryCode, T
erminal, PassengerCount, AdjustedActivityTypeCode, AdjustedPassengerCount, Year, Month) VALUES ('United Airlines', 'B', uuid(), '201005', 'UA', 'United Airlines', 'UA', 'Domestic', 'US', 'Deplaned', 'Other', 'Terminal 1', 63327, 'Deplaned', 633
27, 2010, 'May');
cqlsh:tokio_school_viajes> INSERT INTO FlightBoarding (OperatingAirline, BoardingArea, FlightID, ActivityPeriod, OperatingAirlineIATACode, PublishedAirline, PublishedAirlineIATACode, GEOSummary, GEORegion, ActivityTypeCode, PriceCategoryCode, T
erminal, PassengerCount, AdjustedActivityTypeCode, AdjustedPassengerCount, Year, Month) VALUES ('United Airlines - Pre 07/01/2013', 'G', uuid(), '200806', 'UA', 'United Airlines - Pre 07/01/2013', 'UA', 'International', 'Asia', 'Thru / Transit'
, 'Other', 'International', 2759, 'Thru / Transit * 2', 5518, 2008, 'June');
cqlsh:tokio_school_viajes> INSERT INTO FlightBoarding (OperatingAirline, BoardingArea, FlightID, ActivityPeriod, OperatingAirlineIATACode, PublishedAirline, PublishedAirlineIATACode, GEOSummary, GEORegion, ActivityTypeCode, PriceCategoryCode, T
erminal, PassengerCount, AdjustedActivityTypeCode, AdjustedPassengerCount, Year, Month) VALUES ('Virgin America', 'A', uuid(), '201001', 'VX', 'Virgin America', 'VX', 'Domestic', 'US', 'Deplaned', 'Low Fare', 'International', 93097, 'Deplaned',
 93097, 2010, 'January');
cqlsh:tokio_school_viajes> INSERT INTO FlightBoarding (OperatingAirline, BoardingArea, FlightID, ActivityPeriod, OperatingAirlineIATACode, PublishedAirline, PublishedAirlineIATACode, GEOSummary, GEORegion, ActivityTypeCode, PriceCategoryCode, T
erminal, PassengerCount, AdjustedActivityTypeCode, AdjustedPassengerCount, Year, Month) VALUES ('Delta Air Lines', 'C', uuid(), '200512', 'DL', 'Delta Air Lines', 'DL', 'Domestic', 'US', 'Enplaned', 'Other', 'Terminal 1', 57401, 'Enplaned', 574
01, 2005, 'December');
cqlsh:tokio_school_viajes> INSERT INTO FlightBoarding (OperatingAirline, BoardingArea, FlightID, ActivityPeriod, OperatingAirlineIATACode, PublishedAirline, PublishedAirlineIATACode, GEOSummary, GEORegion, ActivityTypeCode, PriceCategoryCode, T
erminal, PassengerCount, AdjustedActivityTypeCode, AdjustedPassengerCount, Year, Month) VALUES ('Mexicana Airlines', 'G', uuid(), '200602', 'MX', 'Mexicana Airlines', 'MX', 'International', 'Mexico', 'Enplaned', 'Other', 'International', 6843,
 'Enplaned', 6843, 2006, 'February');
cqlsh:tokio_school_viajes> INSERT INTO FlightBoarding (OperatingAirline, BoardingArea, FlightID, ActivityPeriod, OperatingAirlineIATACode, PublishedAirline, PublishedAirlineIATACode, GEOSummary, GEORegion, ActivityTypeCode, PriceCategoryCode, T
erminal, PassengerCount, AdjustedActivityTypeCode, AdjustedPassengerCount, Year, Month) VALUES ('British Airways', 'A', uuid(), '200704', 'BA', 'British Airways', 'BA', 'International', 'Europe', 'Deplaned', 'Other', 'International', 19308, 'De
planed', 19308, 2007, 'April');
cqlsh:tokio_school_viajes> INSERT INTO FlightBoarding (OperatingAirline, BoardingArea, FlightID, ActivityPeriod, OperatingAirlineIATACode, PublishedAirline, PublishedAirlineIATACode, GEOSummary, GEORegion, ActivityTypeCode, PriceCategoryCode, T
erminal, PassengerCount, AdjustedActivityTypeCode, AdjustedPassengerCount, Year, Month) VALUES ('United Airlines - Pre 07/01/2013', 'G', uuid(), '200601', 'UA', 'United Airlines - Pre 07/01/2013', 'UA', 'International', 'Australia / Oceania', '
Enplaned', 'Other', 'International', 9951, 'Enplaned', 9951, 2006, 'January');
cqlsh:tokio_school_viajes> INSERT INTO FlightBoarding (OperatingAirline, BoardingArea, FlightID, ActivityPeriod, OperatingAirlineIATACode, PublishedAirline, PublishedAirlineIATACode, GEOSummary, GEORegion, ActivityTypeCode, PriceCategoryCode, T
erminal, PassengerCount, AdjustedActivityTypeCode, AdjustedPassengerCount, Year, Month) VALUES ('Hawaiian Airlines', 'A', uuid(), '201309', 'HA', 'Hawaiian Airlines', 'HA', 'Domestic', 'US', 'Enplaned', 'Other', 'International', 7608, 'Enplaned
', 7608, 2013, 'September');
cqlsh:tokio_school_viajes> INSERT INTO FlightBoarding (OperatingAirline, BoardingArea, FlightID, ActivityPeriod, OperatingAirlineIATACode, PublishedAirline, PublishedAirlineIATACode, GEOSummary, GEORegion, ActivityTypeCode, PriceCategoryCode, T
erminal, PassengerCount, AdjustedActivityTypeCode, AdjustedPassengerCount, Year, Month) VALUES ('British Airways', 'A', uuid(), '200603', 'BA', 'British Airways', 'BA', 'International', 'Europe', 'Deplaned', 'Other', 'International', 17643, 'De
planed', 17643, 2006, 'March');
cqlsh:tokio_school_viajes> INSERT INTO FlightBoarding (OperatingAirline, BoardingArea, FlightID, ActivityPeriod, OperatingAirlineIATACode, PublishedAirline, PublishedAirlineIATACode, GEOSummary, GEORegion, ActivityTypeCode, PriceCategoryCode, T
erminal, PassengerCount, AdjustedActivityTypeCode, AdjustedPassengerCount, Year, Month) VALUES ('Southwest Airlines', 'B', uuid(), '201001', 'WN', 'Southwest Airlines', 'WN', 'Domestic', 'US', 'Enplaned', 'Low Fare', 'Terminal 1', 104593, 'Enpl
aned', 104593, 2010, 'January');
cqlsh:tokio_school_viajes>
```

- FlghyBoarding

INSERT INTO FlightBoarding (OperatingAirline, BoardingArea, FlightID, ActivityPeriod, OperatingAirlineIATACode, PublishedAirline, PublishedAirlineIATACode, GEOSummary, GEORegion, ActivityTypeCode, PriceCategoryCode, Terminal, PassengerCount, AdjustedActivityTypeCode, AdjustedPassengerCount, Year, Month)

VALUES ('Air France', 'A', uuid(), '201311', 'AF', 'Air France', 'AF', 'International', 'Europe', 'Deplaned', 'Other', 'International', 7170, 'Deplaned', 7170, 2013, 'November');

```
cqlsh:tokio_school_viajes> SELECT * FROM AirlineRecords WHERE OperatingAirline = 'Air China';

 operatingairline | flightid                             | activityperiod | activitytypecode | adjustedactivitytypecode | adjustedpassengercount | boardingarea | georegion | geosummary    | month     | operatingairlineiatacode | passengercount
 | pricecategorycode | publishedairline | publishedairlineiatacode | terminal | year
------------------+--------------------------------------+----------------+------------------+--------------------------+------------------------+--------------+-----------+---------------+-----------+--------------------------+----------------
      Air China | 002600dd-e57d-4ea2-8dbe-ab327e3abad8 |         201304 |        Deplaned |                Deplaned |                   8737 |            G |      Asia | International |     April |                       CA |           8737
|          Other |          Air China |                        CA | International | 2013
      Air China | 0099aee8-f825-48ae-abce-165ef0f6d084 |         201412 |        Enplaned |                Enplaned |                   7391 |            G |      Asia | International |  December |                       CA |           7391
|          Other |          Air China |                        CA | International | 2014
      Air China | 029358e4-b46c-4474-832b-bb0b3b9d3829 |         200605 |        Enplaned |                Enplaned |                   5680 |            G |      Asia | International |       May |                       CA |           5680
|          Other |          Air China |                        CA | International | 2006
      Air China | 030d9a72-4317-46a1-b13b-594dfa0b2ef7 |         200712 |        Deplaned |                Deplaned |                   6366 |            G |      Asia | International |  December |                       CA |           6366
|          Other |          Air China |                        CA | International | 2007
      Air China | 07beecdc-7cd7-4649-a739-a9c40df70b8a |         200511 |        Deplaned |                Deplaned |                   4550 |            G |      Asia | International |  November |                       CA |           4550
|          Other |          Air China |                        CA | International | 2005
      Air China | 0ef8fbae-8b08-43f8-88f6-47ff931123ba |         201003 |        Enplaned |                Enplaned |                   5277 |            G |      Asia | International |     March |                       CA |           5277
|          Other |          Air China |                        CA | International | 2010
      Air China | 127cab06-abf6-4fdc-9091-65efff896728 |         200702 |        Enplaned |                Enplaned |                   4424 |            G |      Asia | International |  February |                       CA |           4424
|          Other |          Air China |                        CA | International | 2007
      Air China | 1c47094f-2cba-4ad5-bc96-c99044c7c439 |         200604 |        Deplaned |                Deplaned |                   4728 |            G |      Asia | International |     April |                       CA |           4728
|          Other |          Air China |                        CA | International | 2006
      Air China | 1c72a7e9-751a-4b48-9911-4dfd4d8bb57d |         201205 |        Deplaned |                Deplaned |                   7353 |            G |      Asia | International |       May |                       CA |           7353
|          Other |          Air China |                        CA | International | 2012
      Air China | 1d29e774-4fc0-464f-aa78-8a0122b56aeb |         200911 |        Deplaned |                Deplaned |                   4849 |            G |      Asia | International |  November |                       CA |           4849
|          Other |          Air China |                        CA | International | 2009
      Air China | 1d3bcd69-7281-44e0-88c4-111e4c95a585 |         201501 |        Deplaned |                Deplaned |                   7177 |            G |      Asia | International |   January |                       CA |           7177
|          Other |          Air China |                        CA | International | 2015
      Air China | 1e4df035-7f70-489f-876e-9ce9d6d754c0 |         201012 |        Enplaned |                Enplaned |                   7194 |            G |      Asia | International |  December |                       CA |           7194
|          Other |          Air China |                        CA | International | 2010
      Air China | 1f322957-33c5-420b-8b94-607efcdec179 |         200707 |        Deplaned |                Deplaned |                   7874 |            G |      Asia | International |      July |                       CA |           7874
|          Other |          Air China |                        CA | International | 2007
      Air China | 22309719-1c0b-4fec-ad29-18223a46e07f |         201501 |        Enplaned |                Enplaned |                   5248 |            G |      Asia | International |   January |                       CA |           5248
|          Other |          Air China |                        CA | International | 2015
      Air China | 2a72b0a0-82b5-4575-a5f1-e845e6f7a123 |         201503 |        Deplaned |                Deplaned |                   6576 |            G |      Asia | International |     March |                       CA |           6576
|          Other |          Air China |                        CA | International | 2015
      Air China | 32a0ec0c-b174-422a-bf0a-3d1f5528f990 |         201504 |        Deplaned |                Deplaned |                   7465 |            G |      Asia | International |     April |                       CA |           7465
|          Other |          Air China |                        CA | International | 2015
      Air China | 391e446c-97fa-401d-adlc-bfbeae8e9118 |         201511 |        Enplaned |                Enplaned |                   8925 |            G |      Asia | International |  November |                       CA |           8925
|          Other |          Air China |                        CA | International | 2015
      Air China | 3b638e2d-12ab-4bcb-874f-4218fabf8c19 |         201105 |        Enplaned |                Enplaned |                   7497 |            G |      Asia | International |       May |                       CA |           7497
|          Other |          Air China |                        CA | International | 2011
      Air China | 3c7db96f-0432-4e87-ba40-274f7f6dba88 |         201309 |        Deplaned |                Deplaned |                   8581 |            G |      Asia | International | September |                       CA |           8581
|          Other |          Air China |                        CA | International | 2013
      Air China | 3de6550d-9845-4806-82f2-2f9dc20870ef |         201406 |        Deplaned |                Deplaned |                   8116 |            G |      Asia | International |      June |                       CA |           8116
|          Other |          Air China |                        CA | International | 2014
      Air China | 41b633ad-4c29-4889-a392-cbfe02a0ccf6 |         201102 |        Deplaned |                Deplaned |                   6053 |            G |      Asia | International |  February |                       CA |           6053
|          Other |          Air China |                        CA | International | 2011
      Air China | 42c9cf74-5387-4f3c-8264-4ed721690d38 |         200706 |        Enplaned |                Enplaned |                   7273 |            G |      Asia | International |      June |                       CA |           7273
```



```
      Air China | ea91f780-5806-4c28-aede-de554a4bae5e |         201509 |        Deplaned |                Deplaned |                   9430 |            G |      Asia | International | September |                       CA |           9430
|          Other |          Air China |                        CA | International | 2015
      Air China | efcc1cd9-7974-4b1d-9cd4-291430603529 |         201006 |        Deplaned |                Deplaned |                   6903 |            G |      Asia | International |      June |                       CA |           6903
|          Other |          Air China |                        CA | International | 2010
      Air China | f6d34cda-0322-4b00-a963-c1eff01495e6 |         201512 |        Deplaned |                Deplaned |                   9120 |            G |      Asia | International |  December |                       CA |           9120
|          Other |          Air China |                        CA | International | 2015
      Air China | fd7058ff-7f60-48ed-87df-4231887694b4 |         200904 |        Enplaned |                Enplaned |                   5275 |            G |      Asia | International |     April |                       CA |           5275
|          Other |          Air China |                        CA | International | 2009

(66 rows)
```

En estas primeras imágenes se observan los registros obtenidos de la aerolínea Air China de la tabla AirlineRecords.

SELECT * FROM AirlineRecords WHERE OperatingAirline = 'Air China';

```
cqlsh:tokio_school_viajes> SELECT * FROM FlightBoarding WHERE OperatingAirline = 'Air Berlin' AND BoardingArea = 'G';

 operatingairline | boardingarea | flightid                             | activityperiod | activitytypecode | adjustedactivitytypecode | adjustedpassengercount | georegion | geosummary    | month     | operatingairlineiatacode | passengercount
 | pricecategorycode | publishedairline | publishedairlineiatacode | terminal | year
------------------+--------------+--------------------------------------+----------------+------------------+--------------------------+------------------------+-----------+---------------+-----------+--------------------------+----------------
      Air Berlin |            G | 9e808902-83d2-421d-b1e7-be9a4242b339 |         201006 |        Enplaned |                Enplaned |                   2548 |    Europe | International |      June |                       AB |           2548
|          Other |      Air Berlin |                        AB | International | 2010
      Air Berlin |            G | b61d8a8c-6ce8-4ee1-b4bb-adee0a2096eb |         201009 |        Enplaned |                Enplaned |                   2343 |    Europe | International | September |                       AB |           2343
|          Other |      Air Berlin |                        AB | International | 2010
      Air Berlin |            G | bf272667-2b96-42c6-b419-9f57a63250a9 |         201010 |        Enplaned |                Enplaned |                   1689 |    Europe | International |   October |                       AB |           1689
|          Other |      Air Berlin |                        AB | International | 2010

(3 rows)
```

En esta imagen se observan los registros de la aerolínea Air Berlin y el área de embarque G de la tabla FlightBoarding.

SELECT * FROM FlightBoarding WHERE OperatingAirline = 'Air Berlin' AND BoardingArea = 'G';

PySpark

```
from google.colab import drive
drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

```
!pip install pyspark
```

```
Collecting pyspark
  Downloading pyspark-3.5.1.tar.gz (317.0 MB)
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 317.0/317.0 MB 2.3 MB/s eta 0:00:00
  Preparing metadata (setup.py) ... done
Requirement already satisfied: py4j==0.10.9.7 in /usr/local/lib/python3.10/dist-packages (from pyspark) (0.10.9.7)
Building wheels for collected packages: pyspark
  Building wheel for pyspark (setup.py) ... done
  Created wheel for pyspark: filename=pyspark-3.5.1-py2.py3-none-any.whl size=317488491 sha256=1aba4f5f5e38c79c6b58c00f4fa197bfe8d2bebeb21c1c13fc96d1801b27ec62
  Stored in directory: /root/.cache/pip/wheels/80/1d/60/2c256ed38dddce2fdd93be545214a63e02fbd8d74fb0b7f3a6
Successfully built pyspark
Installing collected packages: pyspark
Successfully installed pyspark-3.5.1
```

```
from pyspark.sql import SparkSession
from pyspark.sql.functions import avg
from pyspark.sql.functions import col
import matplotlib.pyplot as plt
import seaborn as sns

spark = SparkSession.builder.appName("Proyecto_Final").getOrCreate()
```

```
data_path = '/content/drive/MyDrive/Proyecto Final/Proyecto BigData/'

data = spark.read.options(inferSchema='True',delimiter=',', header=True).csv(data_path + 'Air_Traffic_Passenger_Statistics.csv')
data.take(2)
```

Lo primero, será conectar Google colab con Google Drive, para que nos permita posteriormente cargar el fichero csv. A continuación instalamos pyspark, la cual nos permitirá realizar las tareas necesarias sobre el fichero. Una vez instalado comenzaremos a importar las librerías; finalizamos creando sesión y cargando el fichero csv.

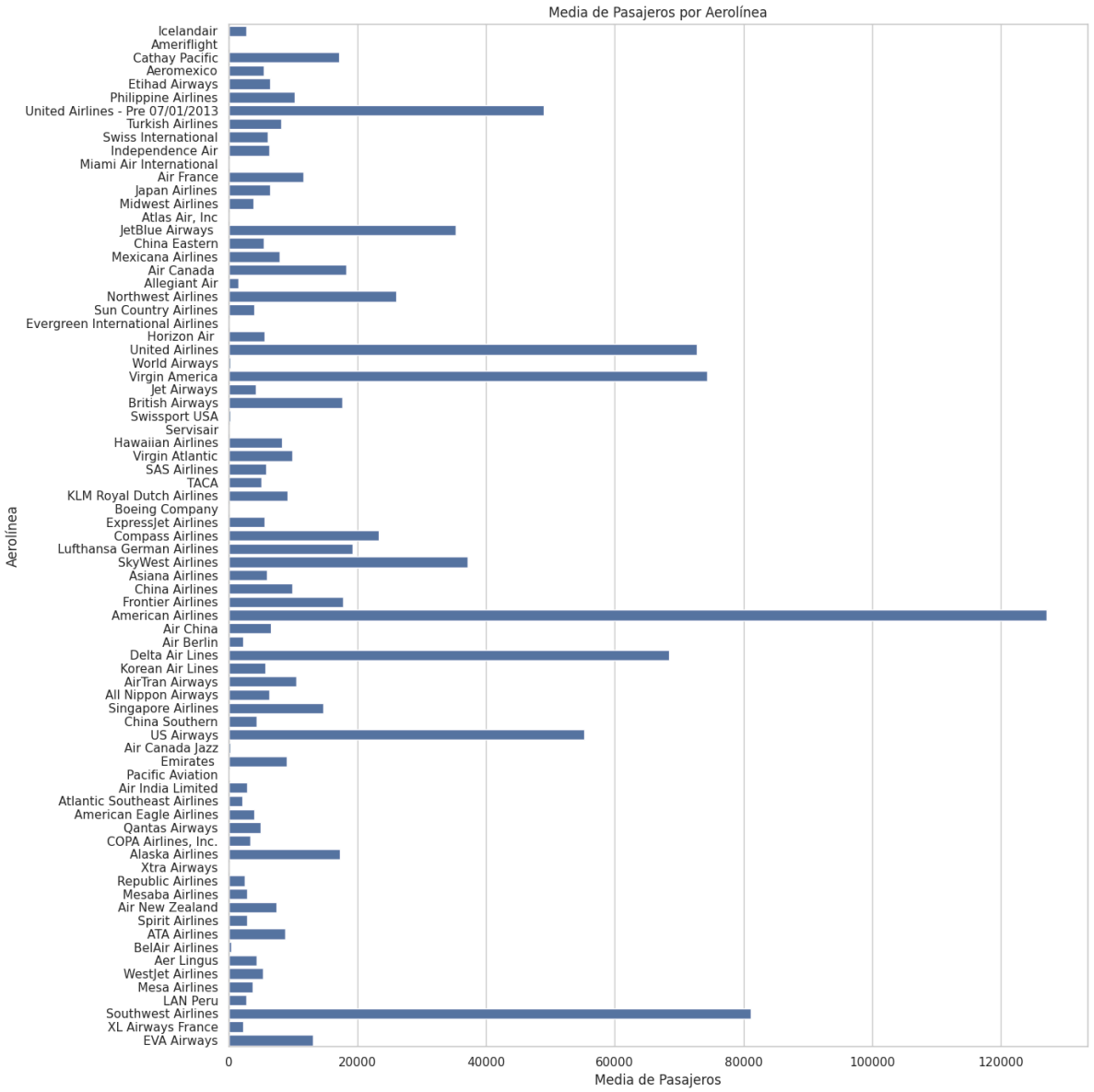
```python
# 1. ¿Cuántas compañías diferentes aparecen en el fichero?
unique_companies_count = data.select("Operating Airline").distinct().count()
print("Número de compañías diferentes:", unique_companies_count)

# Mostrar todas las compañías únicas
unique_companies = data.select("Operating Airline").distinct()
unique_companies.show(unique_companies.count(), truncate=False)
```

```
Número de compañías diferentes: 77
+--------------------------------+
|Operating Airline               |
+--------------------------------+
|Icelandair                      |
|Ameriflight                     |
|Cathay Pacific                  |
|Aeromexico                      |
|Etihad Airways                  |
|Philippine Airlines             |
|United Airlines - Pre 07/01/2013|
|Turkish Airlines                |
|Swiss International             |
|Independence Air                |
|Miami Air International         |
|Air France                      |
|Japan Airlines                  |
|Midwest Airlines                |
|Atlas Air, Inc                  |
|JetBlue Airways                 |
|China Eastern                   |
```

Sacaremos el total de las compañías que tenemos en el fichero, y mostraremos todas aquellas que sean únicas.

```python
# 2. ¿Cuántos pasajeros tienen de media los vuelos de cada compañía?
average_passengers_per_airline = data.groupBy("Operating Airline").agg(avg("Passenger Count").alias("Average Passenger Count"))
#average_passengers_per_airline.show(average_passengers_per_airline.count(), truncate=False)

# Convertir el DataFrame de PySpark a un DataFrame de Pandas
average_passengers_per_airline_pd = average_passengers_per_airline.toPandas()

# Configurar el estilo de los gráficos
sns.set(style="whitegrid")

# Crear el gráfico de barras
plt.figure(figsize=(14, 14))
ax = sns.barplot(x="Average Passenger Count", y="Operating Airline", data=average_passengers_per_airline_pd)

# Configurar el título y las etiquetas de los ejes
plt.title('Media de Pasajeros por Aerolínea')
plt.xlabel('Media de Pasajeros')
plt.ylabel('Aerolínea')

# Mostrar el gráfico
plt.tight_layout()
plt.show()
```

Lo que se está realizando aquí es sacar la media de pasajeros que tienen las compañías y mostrarlo en un gráfico.



Mediante el siguiente gráfico de barras, podemos observar la media de pasajeros de los vuelos de cada compañía.

```python
# 3. Eliminar registros duplicados por el campo “GEO Región”, manteniendo sólo aquel con mayor número de pasajeros
df_sorted = data.orderBy(col("GEO Region"), col("Passenger Count").desc())
df_deduplicated = df_sorted.dropDuplicates(["GEO Region"])

# Mostrar el DataFrame deduplicado
df_deduplicated.show(df_deduplicated.count(), truncate=False)
```

Activity Period	Operating Airline	Operating Airline IATA Code	Published Airline	Published Airline IATA Code	GEO Summary	GEO Region	Activity Type Code	Price Category
200708	United Airlines - Pre 07/01/2013	UA	United Airlines - Pre 07/01/2013	UA	International	Asia	Deplaned	Other
201501	Air New Zealand	NZ	Air New Zealand	NZ	International	Australia / Oceania	Enplaned	Other
200708	Air Canada	AC	Air Canada	AC	International	Canada	Deplaned	Other
201410	TACA	TA	TACA	TA	International	Central America	Deplaned	Other
201507	United Airlines	UA	United Airlines	UA	International	Europe	Deplaned	Other
201407	United Airlines	UA	United Airlines	UA	International	Mexico	Deplaned	Other
201507	Emirates	EK	Emirates	EK	International	Middle East	Deplaned	Other
201101	LAN Peru	LP	LAN Peru	LP	International	South America	Deplaned	Other
201308	United Airlines	UA	United Airlines	UA	Domestic	US	Deplaned	Other

Eliminaremos los registros duplicados, que hagan referencia a la GEO Región , dejando únicamente los que tengan mayor número de pasajeros.

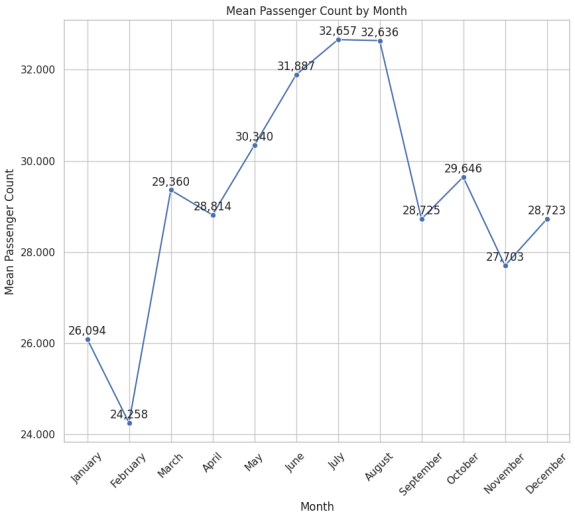
```python
# 4. Volcar los resultados a un archivo CSV
output_path_avg = "/content/drive/MyDrive/Proyecto Final/Proyecto BigData/average_passengers_per_airline.csv"
output_path_deduplicated = "/content/drive/MyDrive/Proyecto Final/Proyecto BigData/deduplicated_geo_region.csv"

# Escribir los resultados en archivos CSV
average_passengers_per_airline.write.csv(output_path_avg, header=True, mode="overwrite")
df_deduplicated.write.csv(output_path_deduplicated, header=True, mode="overwrite")
```

Para terminar, volcaremos los datos obtenidos de la media y los obtenidos en el punto anterior.

Análisis de datos

Procederemos a la elaboración de un análisis descriptivo utilizando los datos obtenidos; con ellos, calcularemos la media y desviación estándar. Proseguiremos realizando una matriz de correlación, para observar la relación entre los datos, y finalizamos aplicando el algoritmo "modelo k-means", el cual nos ayudará a identificar patrones en el tráfico de pasajeros; la decisión de uso de este algoritmo, se debe a la sencillez y eficacia para la mayoría de los conjuntos de datos de clustering .

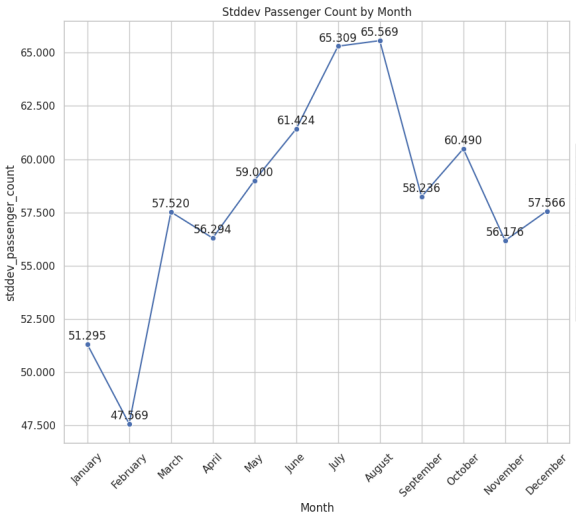
Passenger Count per Month



Month	mean_passenger_count	stddev_passenger_count	count_passenger_count
January	26094	51295	1268
February	24258	47569	1255
March	29360	57520	1253
April	28814	56294	1149
May	30340	59000	1172
June	31887	61424	1183
July	32657	65309	1303
August	32636	65569	1310
September	28725	58236	1297
October	29646	60490	1295
November	27703	56176	1263
December	28723	57566	1259

La media de pasajeros muestra el promedio de la cantidad de pasajeros mensuales a lo largo del año. Los datos revelan que los meses de verano (junio, julio y agosto) tienen las medias más altas, con valores superiores a 31,000 pasajeros. En junio, la media es de 31,887 pasajeros, alcanzando su punto máximo en julio y agosto con medias de 32,657 y 32,636, respectivamente. Esto indica un aumento significativo en la demanda de pasajeros durante estos meses.

En contraste, los meses de invierno, como enero y febrero, presentan las medias más bajas, con 26,094 y 24,258 pasajeros respectivamente. Estos valores reflejan una menor demanda en comparación con los meses de verano. Los meses de transición, como septiembre y octubre, tienen medias intermedias de aproximadamente 28,000 a 29,000 pasajeros, mostrando una disminución gradual en comparación con los picos estivales pero aún superiores a los meses más fríos.



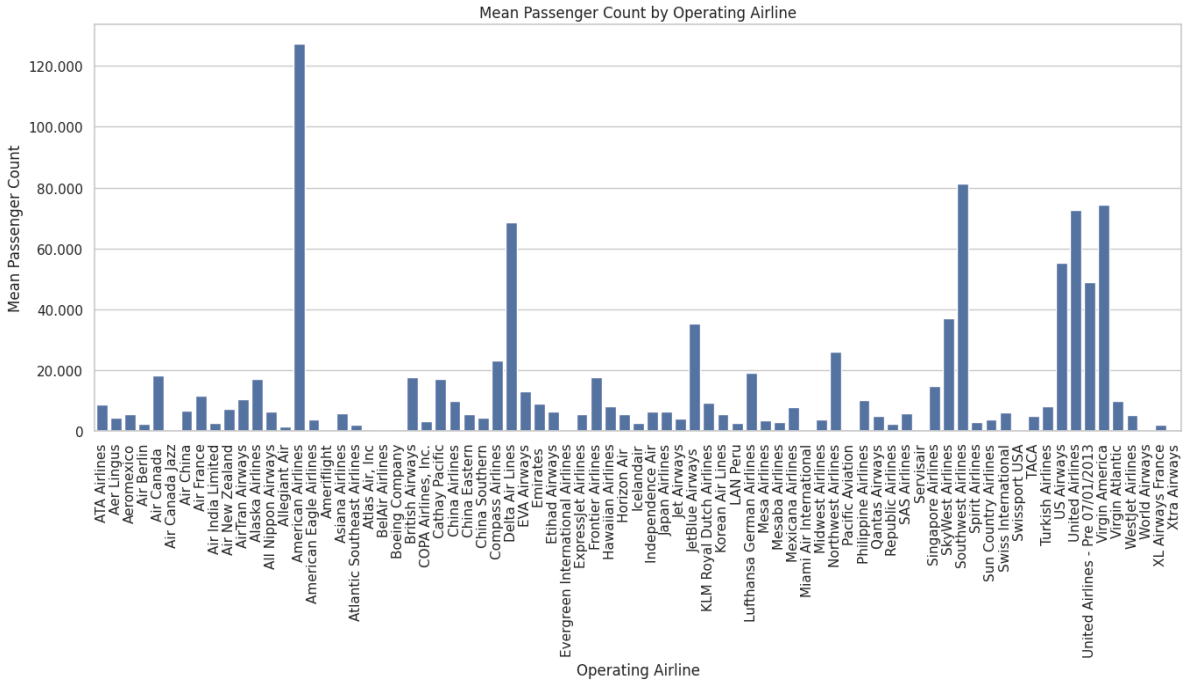
Month	mean_passenger_count	stddev_passenger_count	count_passenger_count
January	26094	51295	1268
February	24258	47569	1255
March	29360	57520	1253
April	28814	56294	1149
May	30340	59000	1172
June	31887	61424	1183
July	32657	65309	1303
August	32636	65569	1310
September	28725	58236	1297
October	29646	60490	1295
November	27703	56176	1263
December	28723	57566	1259

La desviación estándar indica la variabilidad del número de pasajeros respecto a la media mensual. Los datos muestran que la desviación estándar es significativamente mayor durante los meses de verano. En julio y agosto, las desviaciones estándar son de 65,309 y 65,569, respectivamente, lo que sugiere una alta variabilidad en el número de pasajeros durante estos meses, posiblemente debido a la estacionalidad y fluctuaciones en la demanda.

En los meses de invierno, como enero y febrero, la desviación estándar también es alta, con 51,295 y 47,569, pero no alcanza los niveles observados en los meses de verano. Esto indica que, aunque hay cierta variabilidad en la demanda durante estos meses más fríos, la fluctuación es menos pronunciada en comparación con el verano. La desviación estándar disminuye ligeramente en los meses intermedios como septiembre y octubre, con valores de 58,236 y 60,490, reflejando una variabilidad algo menor en la demanda durante la transición de la temporada estival a la invernal.

En resumen, la alta desviación estándar en los meses de verano refleja una mayor variabilidad en la demanda de pasajeros, mientras que en los meses de invierno, aunque la variabilidad sigue presente, es menos pronunciada en comparación con el verano.

Passenger Count per Operating Airlines

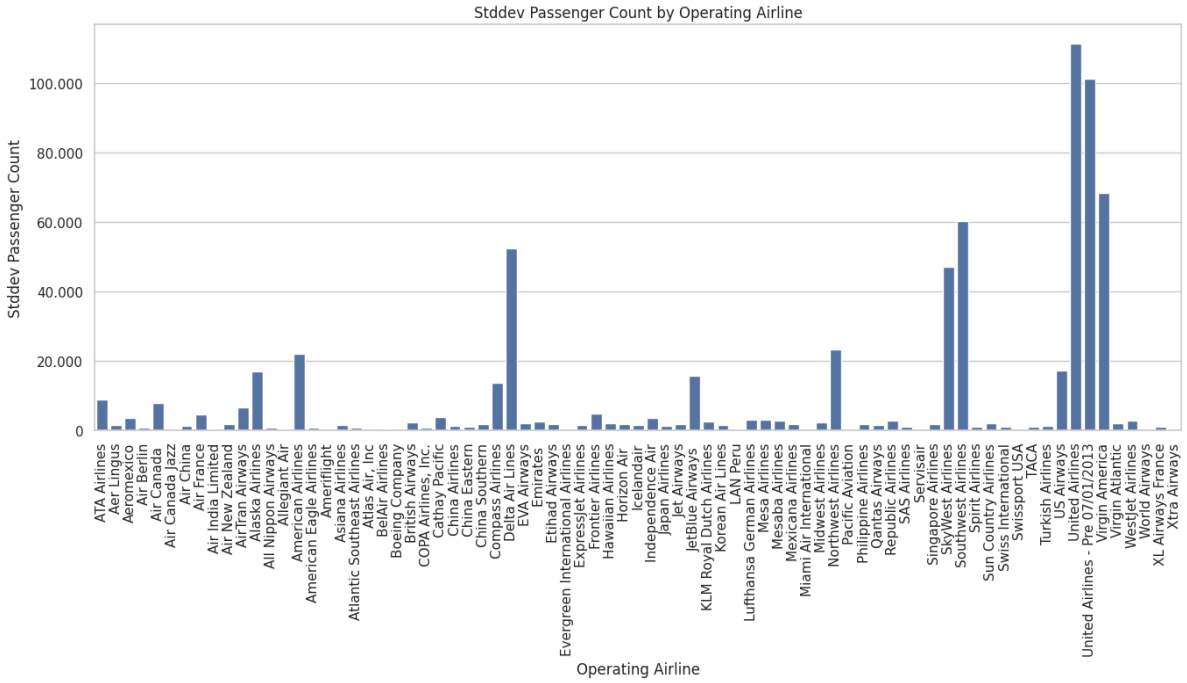


El análisis de la media de pasajeros para diferentes aerolíneas muestra una amplia variabilidad en el número promedio de pasajeros transportados por cada aerolínea. Las aerolíneas con las medias más altas son American Airlines y Southwest Airlines, con medias de 127,164 y 81,188 pasajeros, respectivamente. Esto sugiere que estas aerolíneas tienen una gran capacidad operativa y una alta demanda de sus servicios.

Por otro lado, aerolíneas más pequeñas o con operaciones limitadas como Evergreen International Airlines y Atlas Air, Inc. tienen medias extremadamente bajas, de 2 y 34 pasajeros, respectivamente. Esto indica operaciones muy específicas o vuelos muy limitados en comparación con las grandes aerolíneas.

Aerolíneas como JetBlue Airways y Virgin America también muestran medias significativas, con 35,261 y 74,405 pasajeros, respectivamente, lo que refleja su posición como aerolíneas de tamaño mediano con una demanda considerable.

La desviación estándar del conteo de pasajeros ofrece una visión sobre la variabilidad en el número de pasajeros transportados por cada aerolínea. United Airlines presenta una desviación estándar extremadamente alta, de 111,408, lo que indica una gran variabilidad en el número de pasajeros que transporta, probablemente debido a su gran número de vuelos y rutas variadas.



De manera similar, Southwest Airlines y Virgin America también muestran altas desviaciones estándar de 60,358 y 68,540, respectivamente, sugiriendo una variabilidad significativa en la cantidad de pasajeros por vuelo.

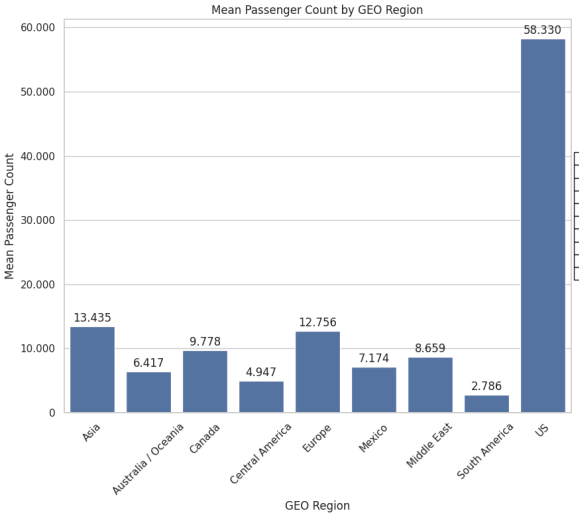
En contraste, aerolíneas como Air India Limited y Swissport USA presentan desviaciones estándar muy bajas, de 333 y 109, respectivamente, lo que indica una mayor consistencia en el número de pasajeros transportados.

En resumen, las aerolíneas más grandes tienden a mostrar una mayor variabilidad en el número de pasajeros, reflejada en desviaciones estándar más altas, mientras que las aerolíneas más pequeñas o con operaciones más específicas muestran una menor variabilidad. Esta información puede ser útil para entender la consistencia operativa y la capacidad de respuesta de cada aerolínea a la demanda de pasajeros.

En la siguiente imagen se podrían ver los datos de la media y la desviación estándar:

Operating Airline	mean_passenger_count	stddev_passenger_count	count_passenger_count
ATA Airlines	8745	8883	44
Aer Lingus	4407	1589	98
Aeromexico	5464	3719	180
Air Berlin	2321	753	36
Air Canada	18252	8036	366
Air Canada Jazz	294	124	14
Air China	6618	1485	259
Air France	11589	4566	258
Air India Limited	2834	333	8
Air New Zealand	7452	1885	259
AirTran Airways	10569	6649	226
Alaska Airlines	17252	16965	751
All Nippon Airways	6386	730	258
Allegiant Air	1517	296	16
American Airlines	127164	22044	272
American Eagle Airlines	4007	972	106
Ameriflight	5	3	22
Asiana Airlines	5903	1642	258
Atlantic Southeast Airlines	2177	806	22
Atlas Air, Inc	34	44	2
BelAir Airlines	415	307	22
Boeing Company	18	0	1
British Airways	17625	2490	258
COPA Airlines, Inc.	3418	925	14
Cathay Pacific	17121	4000	258
China Airlines	9858	1312	258
China Eastern	5498	1107	72
China Southern	4321	1792	32
Compass Airlines	23359	13643	88
Delta Air Lines	68498	52442	386
EVA Airways	13116	2180	258
Emirates	9071	2669	180
Etihad Airways	6476	1929	34
Evergreen International Airlines	2	0	2
ExpressJet Airlines	5632	1651	32
Frontier Airlines	17788	4894	260
Hawaiian Airlines	8282	2164	258
Horizon Air	5578	1843	216
Icelandair	2800	1720	20
Independence Air	6391	3566	10
Japan Airlines	6470	1481	259
Jet Airways	4280	1761	16
JetBlue Airways	35261	15782	222
KLM Royal Dutch Airlines	9222	2517	258
Korean Air Lines	5678	1492	258
LAN Peru	2786	397	90
Lufthansa German Airlines	19302	3158	258
Mesa Airlines	3711	3165	117
Mesaba Airlines	2865	2818	44
Mexicana Airlines	7994	1982	124
Miami Air International	107	57	16
Midwest Airlines	3883	2351	116
Northwest Airlines	26109	23299	240
Pacific Aviation	160	0	2
Philippine Airlines	10249	1855	258
Qantas Airways	4991	1515	134
Republic Airlines	2452	2936	24
SAS Airlines	5866	1152	72
Servisair	90	58	36
Singapore Airlines	14747	1969	258
SkyWest Airlines	37084	47114	963
Southwest Airlines	81188	60358	309
Spirit Airlines	2921	1139	24
Sun Country Airlines	3993	2238	250
Swiss International	6062	1010	139
Swissport USA	259	109	15
TACA	5066	1107	258
Turkish Airlines	8162	1466	24
US Airways	55318	17369	304
United Airlines	72732	111408	892
United Airlines - Pre 07/01/2013	48915	101345	2154
Virgin America	74405	68540	362
Virgin Atlantic	9847	2020	258
WestJet Airlines	5338	2858	103
World Airways	262	8	3
XL Airways France	2223	1146	31
Xtra Airways	73	0	2

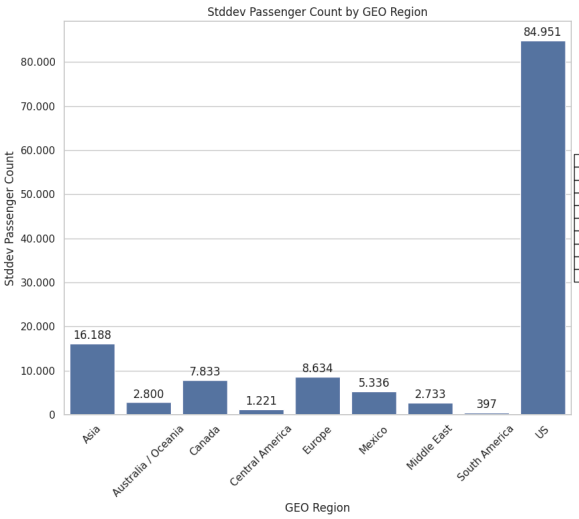
Passenger Count per GEO Regions



GEO Region	mean_passenger_count	stddev_passenger_count	count_passenger_count
Asia	13435	16188	3273
Australia / Oceania	6417	2800	737
Canada	9778	7833	1418
Central America	4947	1221	274
Europe	12756	8634	2089
Mexico	7174	5336	1115
Middle East	8659	2733	214
South America	2786	397	90
US	58330	84951	5797

En términos de media de pasajeros, Estados Unidos (US) destaca con la media más alta, alcanzando 58,330 pasajeros, lo que indica una demanda considerablemente mayor en comparación con otras regiones. Asia y Europa también muestran medias elevadas, con 13,435 y 12,756 pasajeros respectivamente, reflejando una fuerte demanda en estas áreas.

Canadá se sitúa en una posición intermedia con una media de 9,778 pasajeros. Medio Oriente y México presentan medias de 8,659 y 7,174 pasajeros, respectivamente, indicando una demanda moderada. Australia/Oceanía tiene una media de 6,417 pasajeros, mientras que Centroamérica y Sudamérica presentan las medias más bajas, con 4,947 y 2,786 pasajeros, respectivamente, lo que refleja una menor demanda en estas regiones.



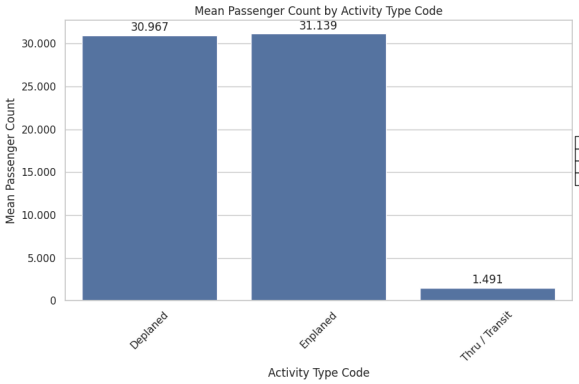
GEO Region	mean_passenger_count	stddev_passenger_count	count_passenger_count
Asia	13435	16188	3273
Australia / Oceania	6417	2800	737
Canada	9778	7833	1418
Central America	4947	1221	274
Europe	12756	8634	2089
Mexico	7174	5336	1115
Middle East	8659	2733	214
South America	2786	397	90
US	58330	84951	5797

En cuanto a la variabilidad del número de pasajeros, Estados Unidos nuevamente destaca con la mayor desviación estándar, 84,951, lo que sugiere una alta fluctuación en el número de pasajeros transportados. Asia presenta una desviación estándar considerable de 16,188, seguida por Europa con 8,634, indicando variabilidades significativas pero menores en comparación con Estados Unidos.

Canadá y México muestran desviaciones estándar de 7,833 y 5,336 respectivamente, reflejando una variabilidad moderada. Medio Oriente y Australia/Oceanía tienen desviaciones estándar relativamente bajas de 2,733 y 2,800, respectivamente, sugiriendo una consistencia mayor en el número de pasajeros. Finalmente, Centroamérica y Sudamérica presentan las desviaciones estándar más bajas, de 1,221 y 397 respectivamente, indicando una variabilidad mínima en el conteo de pasajeros.

En resumen, Estados Unidos lidera tanto en la media como en la variabilidad del número de pasajeros, mientras que Centroamérica y Sudamérica tienen las cifras más bajas y consistentes. Las demás regiones se sitúan en posiciones intermedias, con variabilidades y demandas diversas.

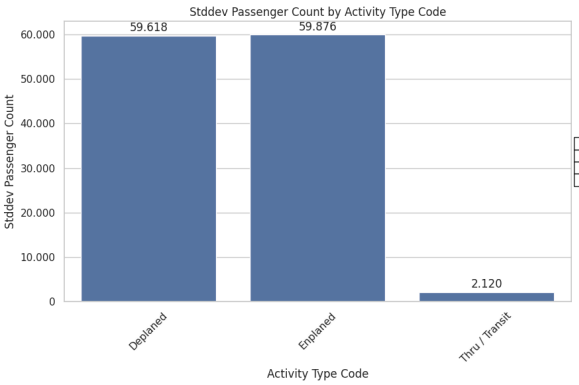
Passenger Count per Activity Type Code



Activity Type Code	mean_passenger_count	stddev_passenger_count	count_passenger_count
Deplaned	30967	59618	7071
Enplaned	31139	59876	7016
Thru / Transit	1491	2120	920

En términos de la media de pasajeros, la actividad Enplaned (embarque) presenta la media más alta con 31,139 pasajeros, lo que indica que un mayor número de pasajeros están abordando vuelos. La actividad Deplaned (desembarque) sigue de cerca con una media de 30,967 pasajeros, mostrando un volumen casi igual de pasajeros que desembarcan.

Por otro lado, la actividad Thru / Transit (pasajeros en tránsito) tiene una media considerablemente menor, con solo 1,491 pasajeros, lo que refleja una menor cantidad de pasajeros en tránsito en comparación con los que embarcan o desembarcan.



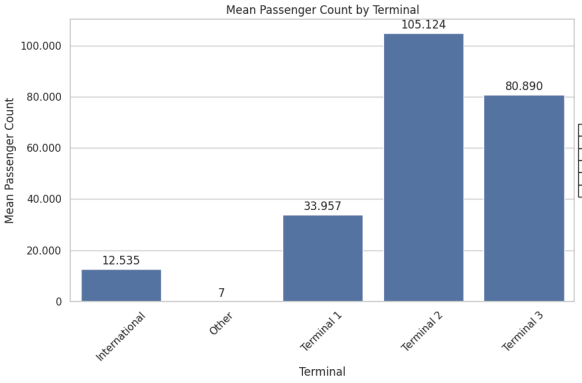
Activity Type Code	mean_passenger_count	stddev_passenger_count	count_passenger_count
Deplaned	30967	59618	7071
Enplaned	31139	59876	7016
Thru / Transit	1491	2120	920

La variabilidad en el número de pasajeros también muestra diferencias significativas. La actividad Enplaned presenta la desviación estándar más alta con 59,876, indicando una gran fluctuación en el número de pasajeros que embarcan. De manera similar, la actividad Deplaned tiene una desviación estándar de 59,618, reflejando también una alta variabilidad en el número de pasajeros que desembarcan.

En contraste, la actividad Thru / Transit muestra una desviación estándar mucho menor, de 2,120, lo que sugiere una mayor consistencia en el número de pasajeros en tránsito en comparación con las actividades de embarque y desembarque.

En resumen, las actividades de embarque y desembarque muestran medias altas y alta variabilidad en el número de pasajeros, mientras que la actividad de tránsito tiene una media y una variabilidad significativamente menores.

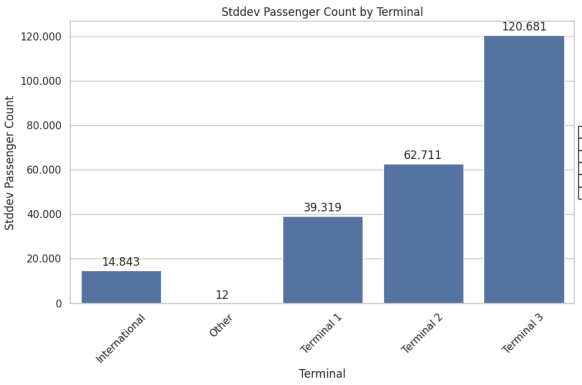
Passenger Count per Terminal



Terminal	mean_passenger_count	stddev_passenger_count	count_passenger_count
International	12535	14843	9197
Other	7	12	27
Terminal 1	33957	39319	3241
Terminal 2	105124	62711	324
Terminal 3	80890	120681	2218

En términos de la media de pasajeros, la Terminal 2 se destaca con la media más alta, registrando 105,124 pasajeros, lo que indica que es la terminal con el mayor volumen de tráfico. Le sigue la Terminal 3 con una media de 80,890 pasajeros, también mostrando una alta actividad.

La Terminal 1 tiene una media de 33,957 pasajeros, lo que representa un volumen moderado de tráfico en comparación con las terminales 2 y 3. La Terminal Internacional presenta una media de 12,535 pasajeros, reflejando un volumen significativamente menor en comparación con las terminales nacionales. Finalmente, Otras terminales tienen una media extremadamente baja de solo 7 pasajeros, lo que sugiere una actividad mínima.



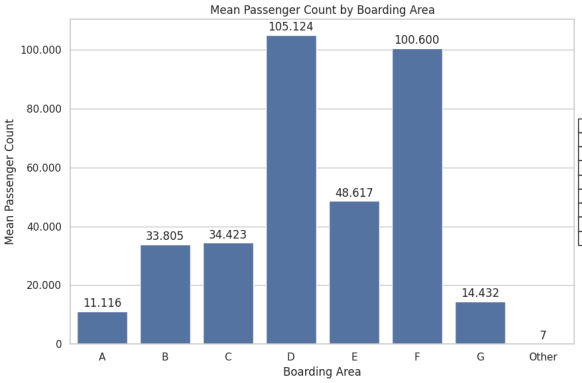
Terminal	mean_passenger_count	stddev_passenger_count	count_passenger_count
International	12535	14843	9197
Other	7	12	27
Terminal 1	33957	39319	3241
Terminal 2	105124	62711	324
Terminal 3	80890	120681	2218

En cuanto a la variabilidad en el número de pasajeros, la Terminal 3 muestra la mayor desviación estándar con 120,681, indicando una gran fluctuación en el número de pasajeros que la utilizan. La Terminal 2 tiene una desviación estándar de 62,711, reflejando una variabilidad significativa, aunque menor en comparación con la Terminal 3.

La Terminal 1 presenta una desviación estándar de 39,319, lo que sugiere una variabilidad moderada en el número de pasajeros. La Terminal Internacional tiene una desviación estándar de 14,843, indicando una menor variabilidad en comparación con las terminales nacionales. Otras terminales presentan una desviación estándar muy baja de 12, lo que sugiere una alta consistencia en el escaso número de pasajeros que las utilizan.

En resumen, las terminales 2 y 3 son las más concurridas y presentan una alta variabilidad en el número de pasajeros. La Terminal 1 muestra una actividad moderada, mientras que la Terminal Internacional y otras terminales tienen volúmenes y variabilidades mucho menores.

Passenger Count per Boarding Area



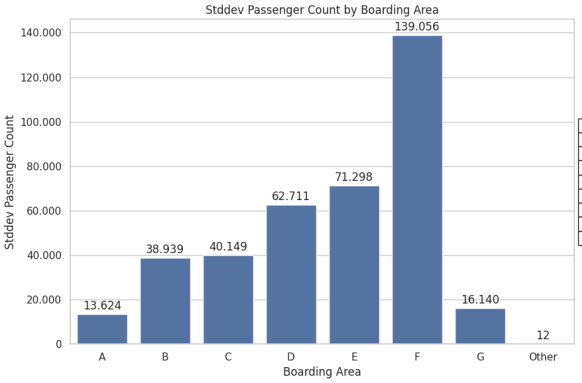
Boarding Area	mean_passenger_count	stddev_passenger_count	count_passenger_count
A	11116	13624	5225
B	33805	38939	1993
C	34423	40149	1228
D	105124	62711	324
E	48617	71298	841
F	100600	139056	1377
G	14432	16140	3992
Other	7	12	27

El Área F tiene la media más alta con 100,600 pasajeros, lo que indica que es la zona con el mayor volumen de tráfico. Le sigue el Área E con una media de 48,617 pasajeros, mostrando también una actividad considerable.

El Área C presenta una media de 34,423 pasajeros, mientras que el Área B tiene una media de 33,805 pasajeros, reflejando una actividad moderada pero significativa en comparación con las áreas más concurridas.

El Área D tiene una media de 105,124 pasajeros, destacándose como el área con el mayor tráfico. El Área A muestra una media de 11,116 pasajeros, y el Área G tiene una media de 14,432 pasajeros, indicando volúmenes más bajos en comparación con las áreas principales.

Finalmente, el Área Other tiene una media extremadamente baja de solo 7 pasajeros, lo que sugiere una actividad mínima en esa categoría.



Boarding Area	mean_passenger_count	stddev_passenger_count	count_passenger_count
A	11116	13624	5225
B	33805	38939	1993
C	34423	40149	1228
D	105124	62711	324
E	48617	71298	841
F	100600	139056	1377
G	14432	16140	3992
Other	7	12	27

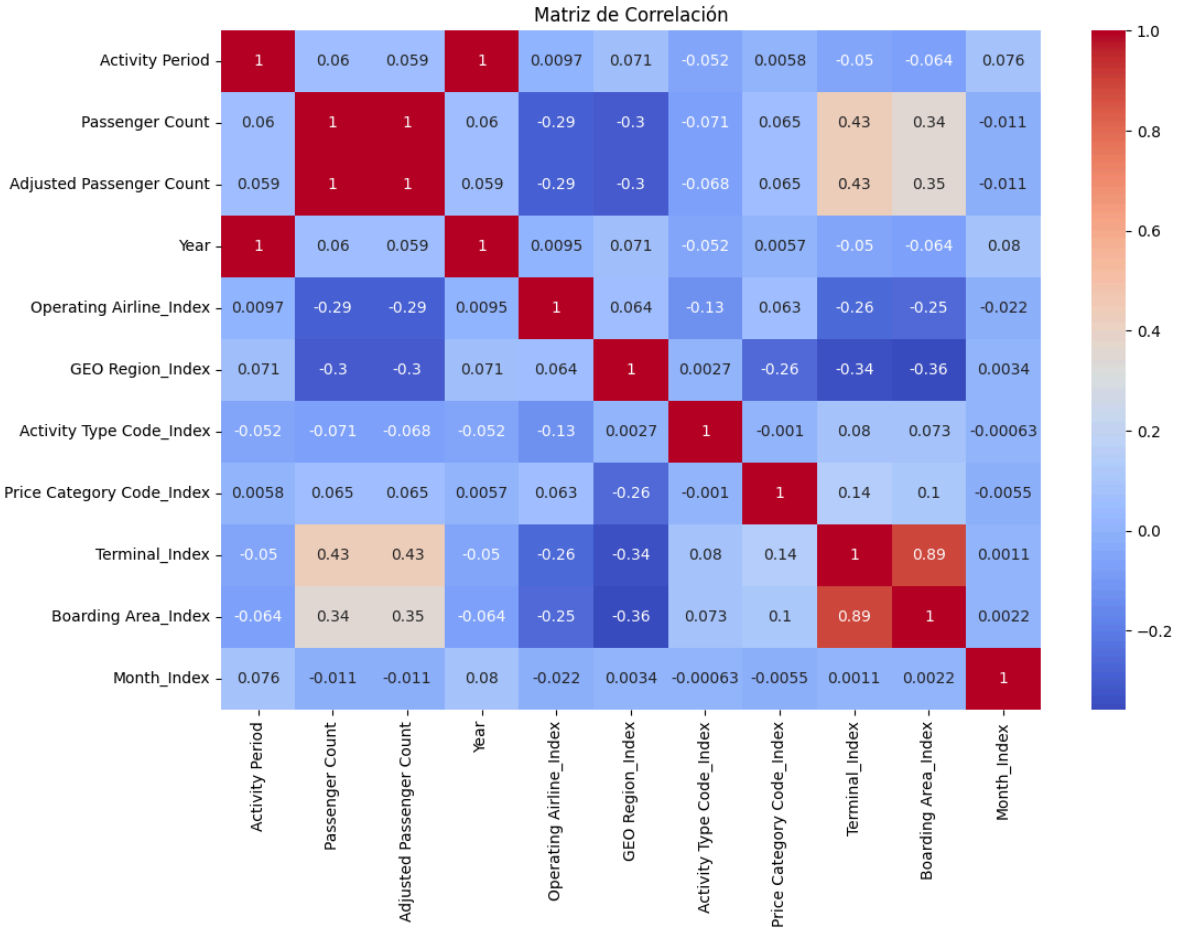
En términos de variabilidad, el Área F presenta la desviación estándar más alta con 139,056, indicando una gran fluctuación en el número de pasajeros. El Área E sigue con una desviación estándar de 71,298, mostrando también una alta variabilidad en el tráfico.

El Área D tiene una desviación estándar de 62,711, mientras que el Área C presenta una desviación estándar de 40,149. El Área B muestra una desviación estándar de 38,939, reflejando una variabilidad considerable en el número de pasajeros.

El Área G tiene una desviación estándar de 16,140, indicando una variabilidad menor en comparación con las áreas principales. Finalmente, el Área A presenta una desviación estándar de 13,624, y el Área Other tiene la desviación estándar más baja de solo 12, reflejando una alta consistencia en el número de pasajeros.

En resumen, las áreas F y E son las más concurridas y muestran una alta variabilidad en el número de pasajeros, mientras que las áreas A, G y Other tienen volúmenes y variabilidades menores.

Matriz de correlación



Variable 1	Variable 2	Correlation
Passenger Count	Adjusted Passenger Count	0.9999408877427202
Adjusted Passenger Count	Passenger Count	0.9999408877427202
Activity Period	Year	0.9999398696979982
Year	Activity Period	0.9999398696979982
Terminal_Index	Boarding Area_Index	0.8922236543226197
Boarding Area_Index	Terminal_Index	0.8922236543226197
Terminal_Index	Adjusted Passenger Count	0.434731425969817
Adjusted Passenger Count	Terminal_Index	0.434731425969817
Passenger Count	Terminal_Index	0.43338800952860984
Terminal_Index	Passenger Count	0.43338800952860984

Contador de Pasajeros y Contador de Pasajeros Ajustado: La correlación entre el Contador de Pasajeros y el Contador de Pasajeros Ajustado es extremadamente alta, con un valor de 0.999941. Esto indica una relación casi perfecta, lo que sugiere que el ajuste realizado no altera significativamente el conteo original de pasajeros.

Periodo de Actividad y Año: Existe una correlación igualmente alta entre el Periodo de Actividad y el Año, con un valor de 0.999940. Este resultado muestra que ambos factores están muy alineados, indicando una fuerte relación entre ellos en los datos.

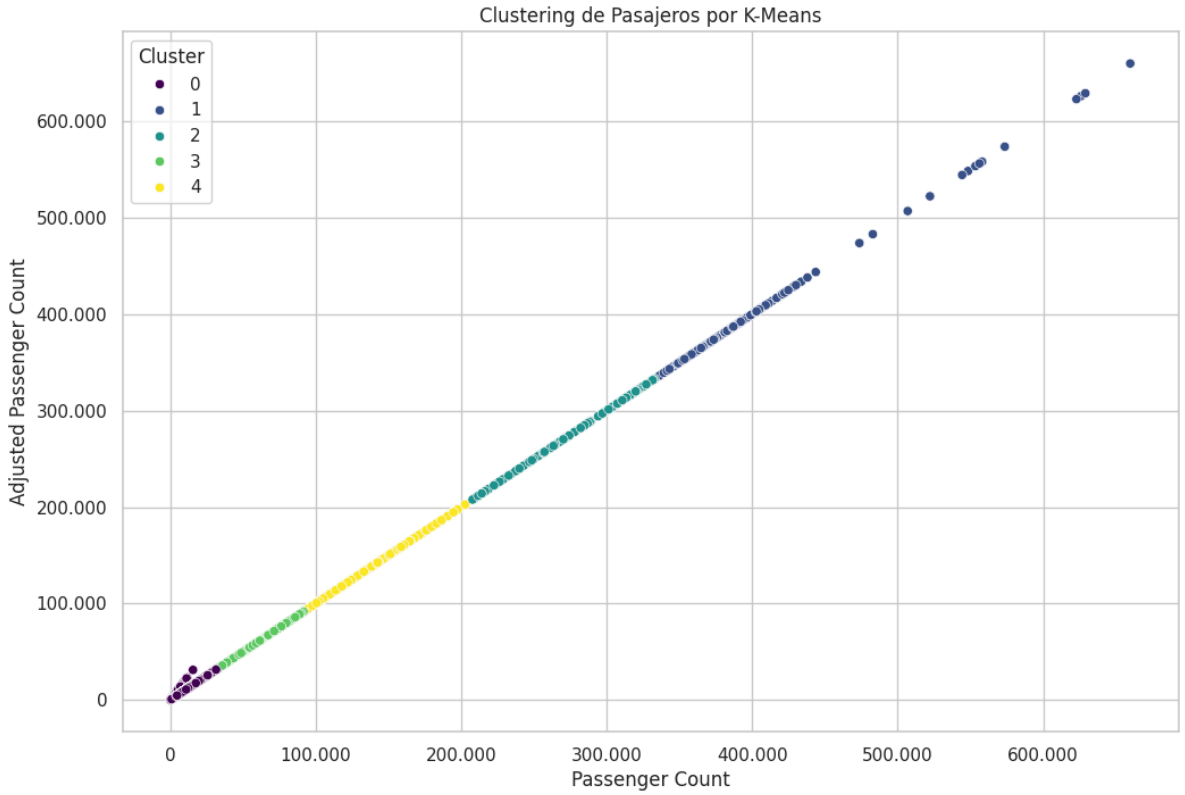
Índice de Terminal e Índice de Área de Embarque: La correlación entre el Índice de Terminal y el Índice de Área de Embarque es de 0.892224, lo que sugiere una relación fuerte. Esto indica que las áreas de embarque están estrechamente relacionadas con las terminales correspondientes.

Índice de Terminal y Contador de Pasajeros Ajustado: La correlación entre el Índice de Terminal y el Contador de Pasajeros Ajustado es de 0.434731. Esto muestra una relación moderada, indicando que hay una cierta conexión entre el índice de terminal y el número de pasajeros ajustado, aunque no tan fuerte como otras correlaciones.

Índice de Terminal y Contador de Pasajeros: La correlación entre el Índice de Terminal y el Contador de Pasajeros es de 0.433388, lo que también indica una relación moderada. Esto sugiere que el índice de terminal está algo relacionado con el número de pasajeros, pero con una fuerza similar a la observada con el contador de pasajeros ajustado.

En resumen, las correlaciones más altas se encuentran entre los contadores de pasajeros y sus versiones ajustadas, así como entre el periodo de actividad y el año. Las relaciones entre los índices de terminal y las áreas de embarque son fuertes, mientras que las correlaciones entre el índice de terminal y los contadores de pasajeros muestran una relación más moderada.

Modelo K-Means



Para los conteos de pasajeros de 27,271 y 29,131, los valores ajustados coinciden exactamente con los conteos originales, es decir, 27,271 y 29,131 respectivamente. En ambos casos, la predicción es 0, lo que indica que no se espera ninguna anomalía o cambio significativo en estos conteos.

Para el conteo de pasajeros de 5,415 con un valor ajustado de 10,830, hay una discrepancia notable entre el conteo original y el ajustado. A pesar de esta diferencia, la predicción es 0, sugiriendo que el modelo no anticipa cambios en el conteo de pasajeros para este caso específico.

Para los conteos de pasajeros de 35,156 y 34,090, cuyos valores ajustados coinciden con los originales, las predicciones son 3 en ambos casos. Esto podría indicar una previsión de anomalías o eventos adicionales que afectan a los conteos ajustados.

En resumen, la mayoría de los datos muestran coincidencias entre los conteos de pasajeros y sus valores ajustados, con algunas predicciones que sugieren la posibilidad de cambios o anomalías, especialmente en los casos donde el conteo ajustado difiere del original.

CONCLUSIONES

- Los datos muestran que los conteos de pasajeros son significativamente más altos en las actividades de embarque y desembarque en comparación con los pasajeros en tránsito. Esto sugiere que la mayor parte del tráfico de pasajeros se concentra en estos procesos, con una notable variabilidad en los conteos.
- Las Terminales 2 y 3 presentan los mayores conteos medios de pasajeros y una alta variabilidad en los datos, lo que indica que estas terminales manejan el mayor volumen de tráfico. Por otro lado, la Terminal Internacional y otras terminales muestran volúmenes mucho menores y menos variabilidad.
- El Área F destaca por tener el conteo medio más alto de pasajeros y la mayor desviación estándar, lo que indica un alto tráfico y variabilidad. En comparación, las áreas A y G, así como otras, presentan volúmenes y variabilidades menores.
- Se observa una alta correlación entre el conteo de pasajeros y su versión ajustada, así como entre el periodo de actividad y el año. Las relaciones entre el índice de terminal y el índice de área de embarque son fuertes, mientras que la conexión entre el índice de terminal y los contadores de pasajeros es moderada.
- Las predicciones para los conteos de pasajeros ajustados en su mayoría coinciden con los valores originales, sugiriendo estabilidad en los datos. Sin embargo, algunas predicciones sugieren posibles cambios o anomalías en los conteos ajustados.

En resumen, el análisis revela patrones claros en el tráfico de pasajeros, destacando una alta correlación entre conteos y ajustes, así como una variabilidad significativa en terminales y áreas de embarque. Las predicciones indican en su mayoría estabilidad, aunque algunas áreas pueden experimentar cambios.