DATA DICTIONARY

TABLE: reviews

Column	Datatype	Data Source	Description
review_id	bigint identity(0, 1)	Generated in Redshift	Primary Key: unique identifier for a review
hotel_id	smallint	generated in hotels table in Spark, fetched as a FK.	Unique identifier for a hotel
reviewer_nationality	varchar(50)	515k reviews data set	Reviewer's country -- it is a noun (Germany) rather than a verb (German).
review_date	date	515k reviews data set	Date when the review was written
num_reviews_by_reviewer	smallint	515k reviews data set	Total number of reviews the reviewer has written.
reviewer_score	real	515k reviews data set	The score (1-10, decimal) that the reviewer has given to the hotel
negative_review	varchar(3000)	515k reviews data set	The negative portion of the review. This text can be quite large.
review_negative_words_count	smallint	515k reviews data set	The negative words extracted out of the negative review.
positive_review	varchar(3000)	515k reviews data set	The positive portion of the review. This text can be quite large.
review_positive_words_count	smallint	515k reviews data set	The positive words extracted out of the positive review.
tags	varchar(256)	515k reviews data set	string tags used by the reviwers while writing their review.

TABLE: hotelreviewsmetadata

Column	Datatype	Data Source	Description
hotel_id	smallint	Generated in Spark	Primary key
total_hotel_reviews	smallint	515k reviews	Total number of reviews that the hotel has from customers
total_hotel_ratings	smallint	515k reviews	Total number of ratings by customers (ratings without reviews)
average_score	real	515k reviews	Average rating for the hotel across the whole data set

TABLE: hoteladdresses

Column	Datatype	Data Source	Description
hoteladdress_id	smallint	515k reviews	Primary key
hotel_address	varchar(256)	515k reviews	The complete hotel address
google_address	varchar(256)	Google Local data set	The address from the Google data set, can be None in this table
latitude	numeric(18,8)	515k reviews	Latitude from reviews data set
longitude	numeric(18,8)	515k reviews	Longitude from reviews data set
Country	varchar(50)	Extracted from hotel_address	Country is in title case
gplus_place_id	varchar(25)	Google Local	Unique xGoogle ID for the business

TABLE: date

Column	Datatype	Data Source	Description
review_date	date	515k reviews	Primary key, date of review
day	Int4	Extracted from review_date	Day (1-31)
week	Int4	Extracted from review_date	Week (1-52)
Month	Int4	Extracted from review_date	Month (1 to 12)
year	Int4	Extracted from review_date	Year
weekday	Int4	Extracted from review_date	Day of the week (1-7)

TABLE: Hotels

Column	Datatype	Data Source	Description
hotel_id	smallint	Generated in Spark	Primary Key: unique identifier for a hotel
hotel_name	varchar(60)	Google Local if available, otherwise 515k reviews	The name of the hotel
nearest_airport_id	varchar(10)	Airport codes	The code of the nearest airport, obtained by fuzzy match
country	varchar(50)	515k reviews (extracted from address)	Country where hotel is located
phone	varchar(20)	Google Local	Phone number if available
price	varchar(10)	Google Local	Indicator of how expensive the hotel is: $, $$, $$$.
original_hotel_name	varchar(60)	515k reviews	The hotel name from the 515k reviews. It sometimes has missing Unicode characters
monday_hours	varchar(30)	Google Local (extracted from nested array)	Working hours for Monday
tuesday_hours	varchar(30)	Google Local (extracted from nested array)	Working hours for Tuesday
wednesday_hours	varchar(30)	Google Local (extracted from nested array)	Working hours for Wednsday
thursday_hours	varchar(30)	Google Local (extracted from nested array)	Working hours for Thursday
friday_hours	varchar(30)	Google Local (extracted from nested array)	Working hours for Friday
saturday_hours	varchar(30)	Google Local (extracted from nested array)	Working hours for Saturday
sunday_hours	varchar(30)	Google Local (extracted from nested array)	Working hours for Sunday

TABLE: airports

Column	Datatype	Data Source	Description
airport_id	varchar(10)	Airport codes	Primary Key: unique string identifier for an airport
type	varchar(20)	Airport codes	Restricted to 3 types: small, medium, large
airport_name	varchar(100)	Airport codes	The full name of the airport
municipality	varchar(60)	Airport codes	Municipality where the airport is located
country	varchar(50)	Country List ISO	ISO Mapped to country code
iso_country	char(2)	Airport codes	ISO country code (2-digit)
continent	char(2)	Airport codes	Always 'EU' for this project
iso_region	varchar(10)	Airport codes	ISO code of the region
latitude	numeric(18,8)	Airport codes	Airport's latitude
longitude	numeric(18,8)	Airport codes	Airport's longitude
elevation_in_feet	integer	Airport codes	Elevation of the airport in feet
gps_code	varchar(10)	Airport codes	Global positioning code for the airport
Iata_code	varchar(10)	Airport codes	International Air Transport Association airport code
local_code	varchar(10)	Airport codes	Local code of neigbhourhood

TABLE: countryindicators

Column	Datatype	Data Source	Description
Country	varchar(50)	Airport codes	Primary Key: all the indicators are based on the country field
Iso_code	char(2)	Country List ISO	2-letter ISO code for the country
tourism_expenditure_millions	float	Tourist-Visitors Arrival and Expenditure (UNWTO/UN)	The amount the country spends on tourism (in million $)
tourist_arrivals_thousands	float	Tourist-Visitors Arrival and Expenditure (UNWTO/UN)	Number of inbound tourists (in thousands)
currency	varchar(30)	Exchange rates (IMF/UN)	Country's official currency
exchange_rate_end_of_period	float	Exchange rates (IMF/UN)	Exchange rate at the end of the period (for e.g., end of year)
gni_per_capita	float	UNDP	Measure of country's income
gdp_per_capita	float	UNDP	Measure of country's produced goods and services
mobile_phone_subscriptions	real	UNDP	Number of cell phone subscriptions, an indicator of well-being.
net_migration_rate	real	UNDP	Difference between no. of immigrants and emigrants
population	numeric(15,5)	UNDP	Total population in millions
urban_population_percent	float	UNDP	Percentage of population which lives in urban areas, a sign of development
hdi_rank	integer	UNDP	Rank according to the human development index calculations
hdi	float	UNDP	Human development indicator
internet_users_percent	Real	UNDP	Percentage of population which has access to internet. Correlation with no. of reviews
political_rights_freedom_score	Real	Freedom House	1-7 scale score for political freedom

civil_liberties_freedom_score	Real	Freedom House	1-7 scale score for civil liberies
freedom_status	varchar(15)	Freedom House	Free/Not Free/Partly free
democracy_or_not	Boolean	Freedom House	Simple True of False it indicate if the country is a democracy.
political_regime_type_score	float	Our World In Data	-10 (autocracy) to 10 (totally free democracy)
human_rights_score	Float	Our World In Data	Higher the better