Data Dictionary

US Visitors

visit_fact table

Field Name	Type	Key	Description	Lineage	Examples	Missing Data
arrivaldate_id	date	FK	Foreign key linking to date dimension. The arrival date in the US.	Extracted from i94 SAS data and transformed to YYYY-MM-dd format.	2016-02-01	
port_id	text	FK	Foreign key linking to port dimension. The arrival port in the US or pre-clearance in another country	The i94port code extracted from i94 SAS data combined with the i94mode which is an integer -1,2,3,9.	1_MIA - the port in Miami for air arrivals 2_FTL - the port in Fort Lauderdale for sea arrivals	If the port is not identified then the port_id will end with_XXX. If the mode of arrival is not recorded then the port_id will start with 9_. If the mode is not recognised the port_id will start with -1_
arrivalstate	text		The full name of the state in which the port of arrival is located.	Extracted from the enriched port data set against which the i94 data is joined by i94port during processing.	Florida	Unknown arrival states are marked as 'unknown'
arrivalstate_abbr	text		2 character abbreviation for the state in which the port of arrival is located	Extracted from the enriched port data set against which the i94 data is joined by i94port during processing.	FL	Unknown arrival states are marked as 'unknown'
mode	text		One of: • Air • Sea • Land • Not reported • unknown	Extracted from a join against the mode data set defined manually as: 1 = 'Air' 2 = 'Sea' 3 = 'Land' 9 = 'Not reported' -1 = 'unknown'	Air	Any data for which the mode was not reported is marked 'Not reported'. Any data for which the mode was missing or invalid is marked 'unknown'

visitpurpose	text		One of: • Business • Pleasure • Student	Extracted from a join against the mode data set defined manually as: 1 = 'Business' 2 = 'Pleasure' 3 = 'Student' -1 = ‘unknown’	Business	Any data for which the mode is missing or invalid is marked ‘unknown’
firstdeststate	text		The full name of the state specified as the first destination by the visitor.	Extracted from the data set provided as a definition of the states by Udacity and joined against the i94addr value.	Pennsylvania	Any data for which i94addr equals ’99’ are marked as ‘All Other Codes’. For invalid i94addr values the field is marked ‘unknown’
firstdeststate_abbr	text		2 character abbreviation for the state specified as the first destination state.	The i94addr value from the original SAS data set.	PA	Any invalid or unrecognised i94 addr values are marked as ‘unknown’
visitduration_id	integer	FK	Foreign key linking to the duration dimension	Identifies the range of days, if known, spent by the visitor in the US.	2	Any invalid durations (i.e. negative values) have the duration_id 0. Any unknown durations (i.e. departure date is null) have the duration_id 999.
age_id	integer	FK	Foreign key linking to the age dimension	Identifies the age bracket of the visitor	2	Any invalid ages (i.e. negative values) have the age_id 0. Any unknown values (i.e. i94bir is null) have the age_id 999.
countryofresidence	text		Name of visitor country of residence	Identified by joining the i94res field against the list of codes and countries provided by Udacity.	UNITED KINGDOM	Any invalid or unrecognised i94res codes result in a value of ‘unknown’ in this field.
gender	text		The gender of the visitor as represented by ‘M’, ‘F’, ‘U’, or ‘unknown’	Extracted from the gender field in the original SAS data set.	F	Any null or unrecognised gender in the original data set results in a value of ‘unknown’
year	integer		The year of arrival. Included as a potential partition field but not used as such.	i94yr value	2016	
month	integer		The month of arrival. Included as a potential partition field but not used as such.	i94mon value	2	

port table

Field Name	Type	Key	Description	Lineage	Examples	Missing Data
port_id	text	PK	Primary key for this arrival port entity.	The i94port code extracted from i94 SAS data combined with the i94mode which is an integer -1,2,3,9.	1_MIA - the port in Miami for air arrivals 2_FTL - the port in Fort Lauderdale for sea arrivals	If the port is not identified then the port_id will end with_XXX. If the mode of arrival is not recorded then the port_id will start with 9_. If the mode is not recognised the port_id will start with -1_
i94port_code	text		The original 3 character port code from the SAS i94 data set	The original port code from the SAS i94 data set	SRQ	
port_mode	text		One of: • Air • Sea • Land • Not reported • unknown	Extracted from a join against the mode data set defined manually as: 1 = 'Air' 2 = 'Sea' 3 = 'Land' 9 = 'Not reported' -1 = 'unknown'	Air	Any data for which the mode was not reported is marked 'Not reported'. Any data for which the mode was missing or invalid is marked 'unknown'
port_latitude	float		The latitude of the most accurate geographical location for the port. In some cases this is accurate to the place where a road crosses the border, in others it is is the middle point of a large city.	Mixed: In some cases the location was extracted as a city from the port definition supplied by Udacity and joined against data sets from simplemaps. In other cases, the port was identified via online data sources and the geographical data completed manually.	24.5636	port_latitude will be null for any visitor arriving at any of the 5 ports for which geographical information could not be identified: TST, LIN, FOP, NWN and FER.
port_longitude	float		As above but for longitude	As above but for longitude	-80.2102	As above but for longitude.
port_place	text		Certain land crossings, particularly on the border with Canada, have no nearby city. These crossings are named in the port_place field.	These values are manually entered following online research into the border crossing.	Daltons Cache/ Pleasant Camp Border Crossing	The majority of ports do not have a 'port_place' value and are marked as 'unknown'

port_city	text		The closest city to the port of arrival.	Mixed: In some cases the city was extracted from the port definition supplied by Udacity and joined against data sets from simplemaps. In other cases, the port was identified via online data sources and the geographical data completed manually.	Miami	Any unknown city is marked ‘unknown’
port_county	text		The county in which the port is situated.	As above for county.	Duval	Unknown counties are marked as ‘unknown’
port_state	text		The full name of the state in which the port of arrival is located.	Extracted from the enriched port data set against which the i94 data is joined by i94port during processing.	Florida	Unknown arrival states are marked as ‘unknown’
port_state_abbr	text		2 character abbreviation for the state in which the port of arrival is located	Extracted from the enriched port data set against which the i94 data is joined by i94port during processing.	FL	Unknown arrival states are marked as ‘unknown'
port_country	text		The full name of the country in which the port of arrival is located. Note that due to pre-clearance and other reasons the port country is not always the United States.	Extracted from the enriched port data set against which the i94 data is joined by i94port during processing.	United States	Unknown arrival countries are marked as ‘unknown’
port_country_abbr	text		The 2 character abbreviation of the country in which the port of arrival is located.	Extracted from the enriched port data set against which the i94 data is joined by i94port during processing.	MX	Unknown arrival countries are marked as ‘unknown’

date table

Field Name	Type	Key	Description	Lineage	Examples	Missing Data
date_id	date	PK	Primary Key for the date entity.	Manually generated	2016-03-24	
day	integer		The day part of the date		24	
week	integer		The week of the year		36	
month	integer		The month part of the date		3	

year	integer		The year part of the date		2016	
weekday	text		The day of the week		Saturday	

duration table

Field Name	Type	Key	Description	Lineage	Examples	Missing Data
duration_id	integer	PK	Primary Key for the duration entity.	Manually generated	1	Any invalid durations (i.e. negative values) have the duration_id 0. Any unknown durations (i.e. departure date is null) have the duration_id 999.
duration_days	text		The range of ages represented by this age entity. Specified as • invalid • 0-3 days • 4-7 days • 8-10 days • 11-14 days • 15-21 days • 22-28 days • age 29+ • unknown	Manually generated. Assumption is that these age ranges are specified by the Data Analysts.	8-10	Any invalid durations (i.e. negative values) have the duration_days value of ‘invalid’. Any unknown durations (i.e. departure date is null) have the duration_days value of ‘unknown’.

age table

Field Name	Type	Key	Description	Lineage	Examples	Missing Data
age_id	integer	PK	Primary Key for the age entity.	Manually generated	1	Any invalid ages (i.e. negative values) have the age_id 0. Any unknown values (i.e. i94bir is null) have the duration_id 999.

age_range	text		The range of ages represented by this age entity. Specified as • invalid • age 0-1 • age 2-10 • age 11-15 • age 16-20 • age 21-25 • age 26-35 • age 36-45 • age 46-55 • age 56-65 • age 66+ • unknown	Manually generated. Assumption is that these age ranges are specified by the Data Analysts.	11-15	Any invalid ages (i.e. negative values) have the age_range ‘invalid’. Any unknown values (i.e. i94bir is null) have the age_range value of ‘unknown’