Sourcing Open Data

Chinook Music Store

1.DATA

1a. Data Source:

This data is publicly available open-source data. It was downloaded from Kaggle.com (Dataset)

1b. Data Collection:

This database includes tables: invoice, invoice_line, playlist, playlist_track, track, album, artist, genre, employee and customer information related to the store's sales. We will use this database and the sqlite3 module in order to explore and answer some questions.

1c. Data Contents:

This dataset can be useful for a variety of purposes, such as analyzing sales trends, identifying successful products or marketing campaigns, and developing strategies for future sales.

1d. Data Profile:

- Original data set:
  employee 8 rows, 15 cols
  customer 59 rows, 13 cols
  invoice 614 rows, 9 cols
  invoice_line 4757 rows, 5 cols
  track 3503 rows, 9 cols
  media_type 5 rows, 2 cols
  genre 25 rows, 2 cols
  album 347 rows, 3 cols
  artist 275 rows, 2 cols
  playlist_track 347 rows, 3 cols
  playlist 18 rows, 2 cols
- Found:
  employee (1 missing)
  customer (49 missing data at company, 29 missing data at state, 47 missing data at fax))
  invoice (0 missing)
  invoice_line (0 missing)
  track (978 missing data at composer)
  media_type (0 missing)
  genre (0 missing)
  album (0 missing)
  artist (0 missing)
  playlist_track (0 missing)
  playlist (0 missing)

- Fixed data mixed types in these columns: customer, track
- No duplicates Values

Table	Column	Data Types	Description	Data Type	Time (yes=variant, no = invariant)
invoice	invoice_id	integer (32)	a unique identifier for invoice	Quantitative, Discrete	no
	customer_id	integer (32)	a unique identifier for customer	Quantitative, Discrete	no
	Invoice_date	timestamp without time zone (6)	the date and time when the invoice is made	Qualitative, Discrete	Yes
	billing_address	character varying(50)	a unique identifier bill for address	Qualitative, Nominal	no
	billing_city	character varying(50)	a unique identifier bill for city	Qualitative, Nominal	no
	billing_state	character varying(50)	a unique identifier bill for state	Qualitative, Nominal	no
	billing_counry	character varying(50)	a unique identifier bill for country	Qualitative, Nominal	no
	billing_postal_code	character varying(20)	a unique identifier bill for address		
	total	float (32)	amount paid by the customer	Quantitative, Discrete	no
Invoice_line	Invoice_line _id	integer (32)	a unique identifier for invoice line	Quantitative, Discrete	no
	invoice_id	integer (32)	a unique identifier for invoice	Quantitative, Discrete	no
	track_id	integer (32)	a unique identifier for track	Quantitative, Discrete	no
	unit_price	float(32)	a unique identifier for price	Qualitative, Discrete	Yes

	quantity	integer (32)	a unique amount of product	Quantitative, Discrete	no
track	track_id	integer (32)	a unique identifier for track	Quantitative, Discrete	no
	name	character varying(50)	first name of album	Qualitative, Nominal	no
	album_id	integer (32)	a unique identifier for album	Quantitative, Discrete	no
	media_type_id	integer (32)	a unique identifier type for media	Quantitative, Discrete	no
	genre_id	integer (32)	a unique identifier for genre	Quantitative, Discrete	no
	composer	character varying(50)	a name of composer	Qualitative, Nominal	no
	milliseconds	integer (32)	A time in milliseconds for track	Quantitative, Discrete	no
	bytes	integer (32)	A unit for each track	Quantitative, Discrete	no
	unit_price	float (32)	A unit price for each track	Quantitative, Discrete	no

Dimension Table

Table	Column	Data Types	Description	Data Type	Time (yes=variant, no = invariant)
customer	customer_id	integer(32)	a unique identifier for actor	Quantitative, Discrete	no
	first_name	character varying(45)	first name of the actor	Qualitative, Nominal	no
	last_name	character varying(45)	last name of the actor	Qualitative, Nominal	no

	company	character varying(45)	Name for the company	Qualitative, Nominal	no
	address	character varying(45)	Address of customer	Qualitative, Nominal	no
	city	character varying(45)	Name of the city	Qualitative, Nominal	no
	state	character varying(45)	the postal code of an address	Qualitative, Nominal	no
	country	character varying(45)	Name of the country	Qualitative, Nominal	no
	postal_code	smallint(16)	the postal code of an address	Qualitative, Nominal	no
	phone	smallint(16)	the phone number for the address	Qualitative, Nominal	no
	fax	smallint(16)	the fax for the address	Quantitative, Discrete	no
	email	character varying(45)	the email for the address	Qualitative, Nominal	no
	support_rep_id	integer(32)	a unique identifier for support customer	Quantitative, Discrete	no
employee	employee_id	integer(32)	the date and time when the last update is made	Quantitative, Discrete	no
	last_name	character varying(45)	a unique identifier for address	Qualitative, Nominal	no
	first_name	character varying(50)	the first line of an address	Qualitative, Nominal	no

	title	character varying(50)	an optional second line of an address	Qualitative, Nominal	no
	report_to	character varying(20)	n/a	Qualitative, Nominal	no
	birthdate	smallint(16)	The birthdate of employee	Qualitative, Nominal	no
	address	character varying(45)	Address of employee	Qualitative, Nominal	no
	city	character varying(45)	Name of the city	Qualitative, Nominal	no
	state	character varying(10)	the postal code of an address	Qualitative, Nominal	no
	country	character varying(20)	Name of the country	Qualitative, Nominal	no
	postal_code	smallint(16)	the postal code of an address	Qualitative, Nominal	no
	phone	timestamp without time zone(6)	the phone number for the address	Qualitative, Nominal	no
	fax	smallint(16)	the fax for the address	Qualitative, Nominal	no
	email	character varying(20)	the email for the address	Qualitative, Nominal	no
media_type	media_type_id	integer(32)	a unique identifier type for media	Quantitative, Discrete	no
	name	character varying(25)	name of media type	Qualitative, Nominal	no

genre	genre_id	integer(32)	a unique identifier for genre	Quantitative, Discrete	no
	name	character varying(50)	name of genre	Qualitative, Nominal	no
playlist_track	playlist_track_id	integer(32)	a unique identifier for playlist_track	Quantitative, Discrete	no
	track_id	character varying(50)	a unique identifier for playlist_track	Quantitative, Discrete	no
album	album_id	integer(32)	a unique identifier for album	Quantitative, Discrete	no
	title	character varying(20)	Name of the album	Qualitative, Nominal	no
	artist_id	integer(32)	a unique identifier for artist	Quantitative, Discrete	no
playlist	playlist_id	integer(32)	a unique identifier for playlist	Quantitative, Discrete	no
	name	character varying(255)	the name of the playlist	Qualitative, Nominal	no
artist	artist _id	integer(32)	a unique identifier for artist	Quantitative, Discrete	no
	name	character varying(255)	The name of the artist	Qualitative, Nominal	no

2. Limitations and Ethics:

- Limitation: The data contains incorrect data types for analysis

- Ethical issue: The data contains personal Information. PLA Security is required.

3. Questions to explore:

- Which region has the most total sales?
- What are the top 10 countries for the most sales?
- Which sale agent made the most sales over all?
- What is the top 5 most selling for track?
- What is the top 5 most selling for artist?
- What is the top 5 most selling for album?
- What is the top 5 most selling for genre?