YELP DOCUMENTATION

Firstly I created the default schema and relationship by the documentation provided by the YELP. Here, I have not used any kind of constraints.

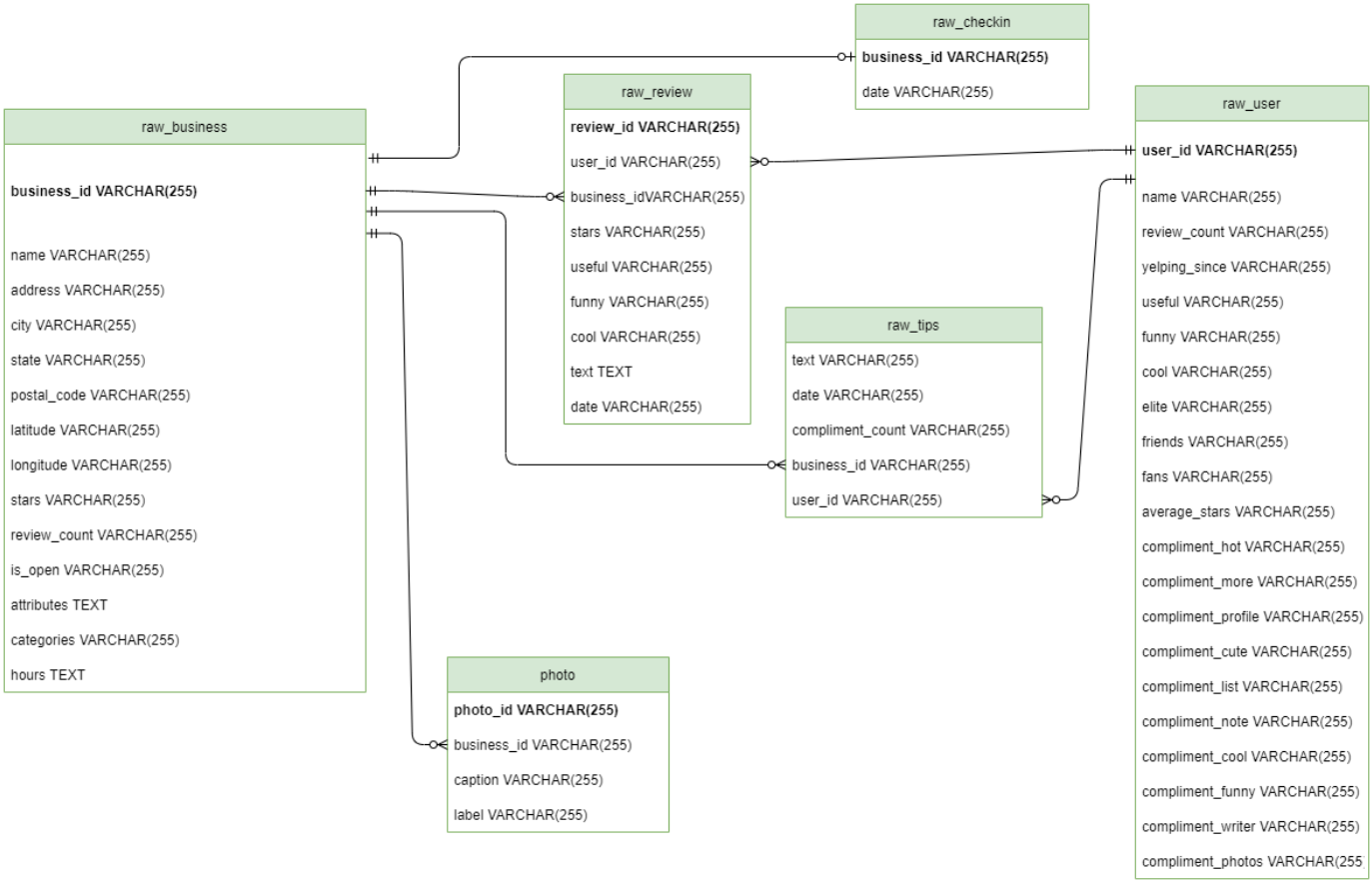


Fig:ER diagram of Raw tables

Logical Modelling:

After the data was pushed into the raw table, I had a lot of insight about the data and I started further exploring its entities and attributes and the data types for it.

Below is the list of entities , description and domain of the model.

Entity	Description	Domain
dim_category	The categories of the business	
Attributes: id	Identifier for the entity,SK,FK	Auto generated,Serial

name	The name of the category	Text
link_fact_business_dim_c ategoty	The table which link entity fact_business and dim_cateogty which has the many to many relationship with each other	
Attributes: business_id category_id	 Identifier of the entity fact_business, FK Identifier of the entity category	 Id which references the table fact_business Valid Id which references the table dim_cateogty
dim_photo	The photos id of the business	
Attributes: photo_id business_id caption label	 The identifier which represents the photo FK which references the fact_business The caption of the photo The label of the photo	 Valid business_id from table fact_business Text Text
dim_photo_count	The counts of photos associated with a business.	
Attributes: business_id photo_count	 PK and FK The count of photos of a specific business	 Valid Id from table fact_business Natural Number
fact_business	The details about the business.	
Attributes: business_id name address city state postal_code location stars review_count is_open hours_monday hours_tuesday hours_wednesday hours_thursday hours_friday hours_saturday hours_sunday wheelchairaccessible parking_garage parking_street parking_validation parking_lot parking_valet businessacceptscreditcar ds outdoorseating noiselevel restaurantsdelivery wifi	 Identifier of the entity , PK The name of the business The address of the business The city in which the business is located The state in which the business is located The postal code of that area Latitude,Longitude of the business,GPS The star rating,rounded to half-stars The total count of the reviews Boolean, True or False for closed or open Start_time and end_time separated by dash Start_time and end_time separated by dash Start_time and end_time separated by dash Start_time and end_time separated by dash Start_time and end_time separated by dash Start_time and end_time separated by dash Start_time and end_time separated by dash Boolean value showing true or false Boolean value showing true or false Boolean value showing true or false Boolean value showing true or false Boolean value showing true or false Boolean value showing true or false Boolean value showing true or false Boolean value showing true or false Text Value telling the noise Boolean value showing true or false Boolean value showing true or false	 22 Character Text Text Text State Code Text GPS,POINT Numeric Int Boolean Text Text Text Text Text Text Text Boolean Boolean Boolean Boolean Boolean Boolean Boolean Boolean Text Boolean Boolean

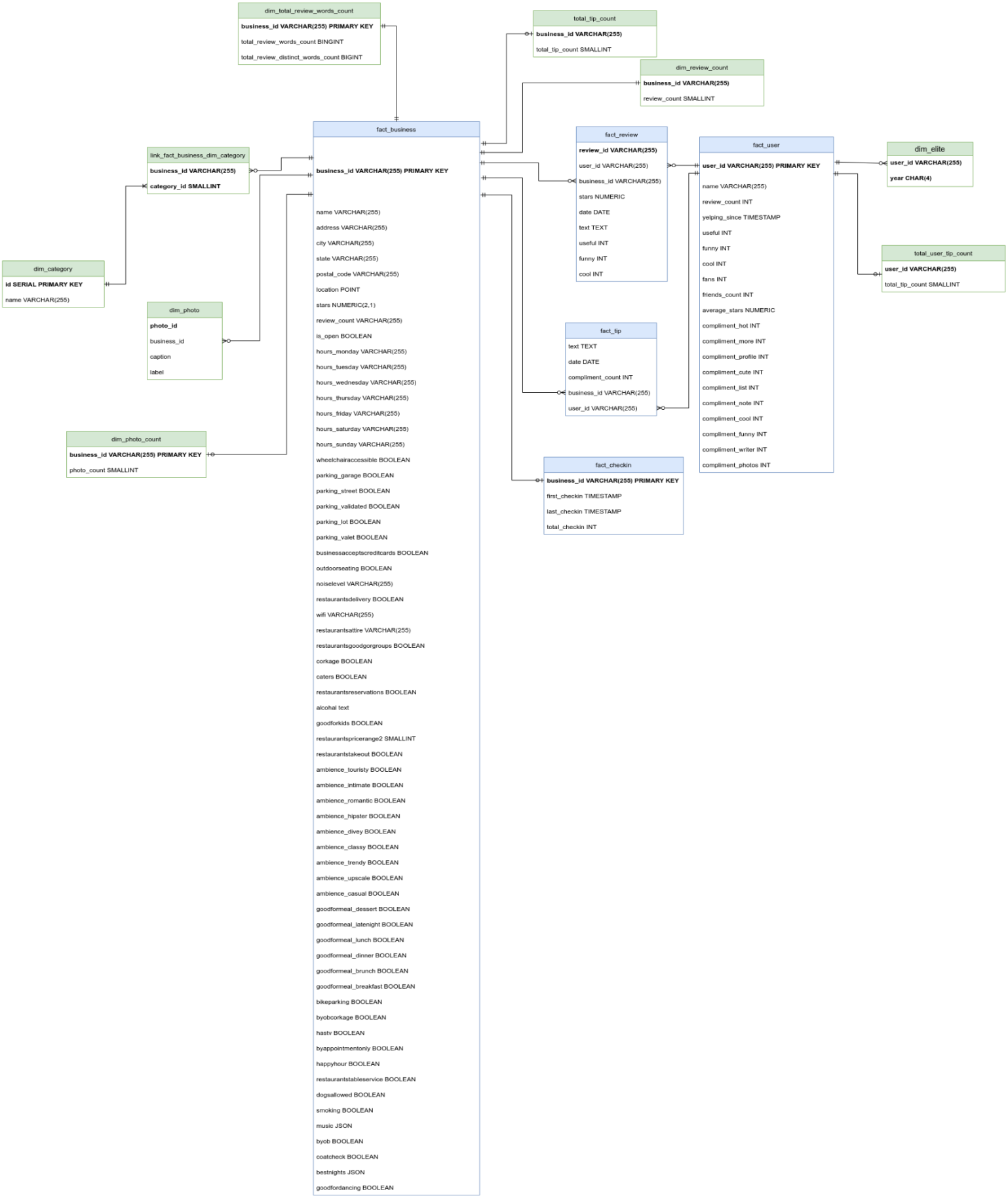
restaurantattire	The text value showing the attire	Text
restaurantsgoodforgroups	Boolean value showing true or false	Boolean
corkage	Boolean value showing true or false	Boolean
caters	Boolean value showing true or false	Boolean
restaurantsresercations		
alcohal	Text value telling the info about alcohol	Text
goodforkids	Boolean value showing true or false	Boolean
restauranrspricerange2	Boolean value showing true or false	Boolean
restaurantstakeout	Boolean value showing true or false	Boolean
ambience_touristy	Boolean value showing true or false	Boolean
ambience_intimate	Boolean value showing true or false	Boolean
ambience_romantic	Boolean value showing true or false	Boolean
ambience_hipster	Boolean value showing true or false	Boolean
ambience_divey	Boolean value showing true or false	Boolean
ambience_classy	Boolean value showing true or false	Boolean
ambience_trendy	Boolean value showing true or false	Boolean
ambience_upscale	Boolean value showing true or false	Boolean
ambience_casual	Boolean value showing true or false	Boolean
goodformeal_dessert	Boolean value showing true or false	Boolean
goodformeal_latenight	Boolean value showing true or false	Boolean
goodformeal_lunch	Boolean value showing true or false	Boolean
goodformeal_dinner	Boolean value showing true or false	Boolean
goodformeal_bunch	Boolean value showing true or false	Boolean
goodformeal_breakfast	Boolean value showing true or false	Boolean
bikeaparking	Boolean value showing true or false	Boolean
byobcorkage	Boolean value showing true or false	Boolean
hastv	Boolean value showing true or false	Boolean
byappointmentonly	Boolean value showing true or false	Boolean
happyhour	Boolean value showing true or false	Boolean
restaurantstableservice	Boolean value showing true or false	Boolean
dogsallowed	Boolean value showing true or false	Boolean
smoking	Boolean value showing true or false	Boolean
music	JSON showing the music keys and values,Taking this as JSON as only 7,198 values are not null	JSON
byob	Boolean value showing true or false	Boolean
coatcheck	Boolean value showing true or false	BooleanBoolean
bestnights	Boolean value showing true or false	Boolean
goodfordancing	JSON showing the music keys and values,Taking this as JSON as only 5,526 values are not null	JSON
dim_total_review_words_count	The table showing the total review words count and the total review distinct words count	
Attributes:		
business_id	Identifier which represents the business, PK,FK	Valid ID from table fact_business.
total_review_word_count	The total word counts of all reviews	INT
total_review_distinct_words_count	The total distinct words counts of all reviews of a business	INT
total_tip_count	The total tip count of the business	
Attributes:		
business_id	Identifier which represents the business,PK,FK	Valid ID from the table fact_business
total_tip_count	The total tip count of a business	INT
fact_review	The details about the reviews made by the user to the business	

Attributes: review_id user_id business_id stars date text useful funny cool	 Identifier of the entity,PK Identifier of the entity user,FK Identifier of the entity business,FK The stars given by the user to the business The date at which the review was given The review text The number of useful reaction by users The number of funny reaction by users The number of cool reaction by users	 Text Text Text Numeric Date Text INT INT INT
fact_tip	The tip given by the user to the business.	
Attributes: text date compliment_count business_id user_id	 The text about the tip given to the business by the user The date at which the tip was given The count of the compliment The business to which the tip is given The user who gives the tips .	 Text Date INT Valid ID from table fact_business Valid ID from table fact_user
fact_checkin	The checkin done on the business.	
Attributes: business_id first_checkin last_checkin total_checkin	 The identifier which references the business,FK The TIMESTAMP at which the first checkin was done The TIMESTAMP at which the last checkin was done The total number of checkin done	 Valid ID from the table fact_business TIMESTAMP TIMESTAMP INT
fact_user	The info about the user	
Attributes: user_id name review_count useful funny cool fans friends_count average_stars compliment_hot compliment_more compliment_cute compliment_list compliment_note compliment_cool compliment_funny Compliment_writer compliment_photos	 Identifier of the entity fact_user,PK The name of the user The total review count of the user The total number of useful reaction received to their reviews The total number of funny reaction received to their reviews The total number of cool reaction received to their reviews The total number of fans of the user The total number of friends of the user The number of compliments_hot The number of compliments_hot The number of compliment_more The number of compliment_cute The number of compliment_list The number of compliment_note The number of compliment_cool The number of compliment_funny The number of Compliment_writer The number of compliment_photos	 PK Text INT INT INT INT INT NUMERIC INT INT INT INT INT INT INT INT INT INT
dim_elite	The year in which the user was elite.	

Attributes: user_id year	 Identifier which references the user.PK The valid year at which the user was elite,PK	 Valid ID referencing the fact_user table CHAR
total_user_tip_count		
Attributes: user_id total_tip_count	 Identifier which references the user.PK The total number of tips count by user	 Valid ID from fact_user table. INT

Proposed ER model

Below is the proposed ER diagram of the warehouse.


dim_total_review_words_count
business_id VARCHAR(255) PRIMARY KEY
total_review_words_count BINGINT
total_review_distinct_words_count BIGINT
total_tip_count
business_id VARCHAR(255)
total_tip_count SMALLINT
dim_review_count
business_id VARCHAR(255)
review_count SMALLINT
fact_business
business_id VARCHAR(255) PRIMARY KEY
name VARCHAR(255)
address VARCHAR(255)
city VARCHAR(255)
state VARCHAR(255)
postal_code VARCHAR(255)
location POINT
stars NUMERIC(2,1)
review_count VARCHAR(255)
is_open BOOLEAN
hours_monday VARCHAR(255)
hours_tuesday VARCHAR(255)
hours_wednesday VARCHAR(255)
hours_thursday VARCHAR(255)
hours_friday VARCHAR(255)
hours_saturday VARCHAR(255)
hours_sunday VARCHAR(255)
wheelchairaccessible BOOLEAN
parking_garage BOOLEAN
parking_street BOOLEAN
parking_validated BOOLEAN
parking_lot BOOLEAN
parking_valet BOOLEAN
businessacceptscreditcards BOOLEAN
outdoorseating BOOLEAN
noiselevel VARCHAR(255)
restaurantsdelivery BOOLEAN
wifi VARCHAR(255)
restaurantsattire VARCHAR(255)
restaurantsgoodgorgroups BOOLEAN
corkage BOOLEAN
caters BOOLEAN
restaurantsreservations BOOLEAN
alcohal text
goodforkids BOOLEAN
restaurantspricerange2 SMALLINT
restaurantstakeout BOOLEAN
ambience_touristy BOOLEAN
ambience_intimate BOOLEAN
ambience_romantic BOOLEAN
ambience_hipster BOOLEAN
ambience_divey BOOLEAN
ambience_classy BOOLEAN
ambience_trendy BOOLEAN
ambience_upscale BOOLEAN
ambience_casual BOOLEAN
goodformeal_dessert BOOLEAN
goodformeal_latenight BOOLEAN
goodformeal_lunch BOOLEAN
goodformeal_dinner BOOLEAN
goodformeal_brunch BOOLEAN
goodformeal_breakfast BOOLEAN
bikeparking BOOLEAN
byobcorkage BOOLEAN
hastv BOOLEAN
byappointmentonly BOOLEAN
happyhour BOOLEAN
restaurantstableservice BOOLEAN
dogsallowed BOOLEAN
smoking BOOLEAN
music JSON
byob BOOLEAN
coatcheck BOOLEAN
bestnights JSON
goodfordancing BOOLEAN
fact_review
review_id VARCHAR(255)
user_id VARCHAR(255)
business_id VARCHAR(255)
stars NUMERIC
date DATE
text TEXT
useful INT
funny INT
cool INT
fact_user
user_id VARCHAR(255) PRIMARY KEY
name VARCHAR(255)
review_count INT
yelping_since TIMESTAMP
useful INT
funny INT
cool INT
fans INT
friends_count INT
average_stars NUMERIC
compliment_hot INT
compliment_more INT
compliment_profile INT
compliment_cute INT
compliment_list INT
compliment_note INT
compliment_cool INT
compliment_funny INT
compliment_writer INT
compliment_photos INT
dim_elite
user_id VARCHAR(255)
year CHAR(4)
total_user_tip_count
user_id VARCHAR(255)
total_tip_count SMALLINT
link_fact_business_dim_category
business_id VARCHAR(255)
category_id SMALLINT
dim_category
id SERIAL PRIMARY KEY
name VARCHAR(255)
dim_photo
photo_id
business_id
caption
label
dim_photo_count
business_id VARCHAR(255) PRIMARY KEY
photo_count SMALLINT
fact_tip
text TEXT
date DATE
compliment_count INT
business_id VARCHAR(255)
user_id VARCHAR(255)
fact_checkin
business_id VARCHAR(255) PRIMARY KEY
first_checkin TIMESTAMP
last_checkin TIMESTAMP
total_checkin INT

Validation:

Fan-Trap

It looks like there is a fan trap everywhere throughout the ER model, but the above fan trap is not going to affect our model as per my design.

Chasm-Trap

Since, the pathway exists between all of the entities, so there's no occurrence of the chasm-trap.

Physical Implementation:

1. Making raw tables
2. Pushing the data from the pipeline into the raw tables.
3. Data cleaning of the table.
    - Making the 'None' value as NULL.
    - Changing the data types of the attributes as per the model proposed.
    - Changing the values like 'no' , 'uno', uyes_corkage','uyes_free','yes_free' to 0 and 1 and casting it to Boolean.
    - While splitting the words by comma on the category field of the business entity, spaced at the beginning was trimmed off to make consistency among the same category name.
    - Making the elite years '20,20' as the 2020.

4. Creating the schemas as proposed above in the ER model.
5. Pushing the data into the fact and dimension table by further cleaning the data if necessary.
6. Validation of different aspects of the data, such as
    - Checking the total photo_id and the total distinct photo_id is 0.
    - Checking all the unique photos are associated with the unique business_id
    - Checking if all the friends are not the user
    - Checking if the yelping_since is not in the future
    - Checking if the average_stars is not in between 0 and 5
    - The review count of a business is not equal to the provided reviews_count at fact_businesss.

7. Exporting the data into flat files for the visualization of data on power BI as I have DBMS on the linux system on my computer, and PowerBI is only supported on the windows.
8. Finally uploading the data on PowerBI for visualization.

CODE DESCRIPTION LINK:
https://github.com/callingsandesh/yelp/blob/final_project/docs/code_explanation.md

Visualization:
LINK:
https://github.com/callingsandesh/yelp/blob/final_project/docs/Dashboard%20and%20Report.pdf