Trifacta Cleaning

1. cleaned column with all NA or empty or just several useless values: thumbnail_url, medium_url, xl_picture_url, neighbourhood_group_cleansed, experiences_offered, host_acceptance_rate, license, jurisdiction_names
2. cleaned column with meaningless same value throughout the sheet: scrape_id, neighbourhood, city, market, smart_location, country_code, street, host_location, host_neighbourhood, host_listings_count, calendar_last_scraped, has_availability, requires_license, is_business_travel_ready
3. delete all text description and pictures(we might add them back in next stages analysis if we decide to go fancy and applying some NLP, for no we only look at those numeric variables: name, summary, space, description, neighborhood_overview, notes, transit, access, interaction, house_rules, host_about
4. Delete all urls which we might also add back later but for now just useless: picture_url, host_thumbnail_url, host_picture_url
5. Delete listing_url(= https://www.airbnb.com/rooms/listing_id), host_url(=https://www.airbnb.com/users/show/host_id)totally meaningless.
6. Although we only leave numbers and deleted both, just a reminder, summary and description are very similar column.
7. Change the data type of “reply rate” to numbers and replace N/A to average value.
8. Make all empty or 0 value in host_total_listings_count / square_feet /amenities/square_feet as NULL
9. Delete “” and space in the amenities text and host_verifications
10. Clean the empty and “” value in zip code to NULL, for some zip code longer than 5 digit, only leave the first five digit, all Boston area zip code should start with 0, make those wired zip code NULL as well.
11. I don think we care about the maximum_night we should only care about the minimum night at this case(if I am wrong we can add it back anyway, but for now let’s just delete other column.
12. Calculate the average price by host and add as a new column by left join.
13. Still looking for a way to turn amenities and host_verifications into factors but I think that’s pretty much good for now.