

Name: Shalini Roy

Q1 (10 pt): Clean “city” and “state”. Describe 1) what steps did you do to find the dirty data; 2) what are the dirty data entries; 3) how did you clean them?

Ans.

- Perform Text Facet filtering to look at the distribution of the data
- Correcting spelling errors, lowercase/uppercase inconsistencies, extra punctuations, etc, by clustering the same city names together and renaming to the correct name.

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
6	3838	<ul style="list-style-type: none">• Santa Barbara (3829 rows)• Santa Barbara (5 rows)• SANTA BARBARA• Santa Barbara• Santa Barbara,• santa Barbara	<input type="checkbox"/>	<input type="text" value="Santa Barbara"/>
2	299	<ul style="list-style-type: none">• Carpinteria (298 rows)• Carpinteria	<input type="checkbox"/>	<input type="text" value="Carpinteria"/>

- Correcting remaining few spelling errors or a different representation of the same city/state name manually

[Santa Barbra](#) 2

[SANTA BARBARAAP](#) 1

[CA](#) 5182


[California](#) 21

- Some of the mapping of city to state were incorrect. For example, if we click on the city Reno from the Facet/Filter section, Reno's state is listed as California, however Reno is a city in Nevada. Since an overwhelming majority of the cities in data are California based, I decided to remove the cities which are from other US states.

▼ city	▼ state
Reno	CA
Reno	CA
Reno	CA

Reno, Nevada

city



Src: Wikipedia

Similar logic for other cities with incorrect city/ state mapping.

city	state
Tampa	CA

- I have a feeling the “Eagle” slipped in there due to the hit song Hotel California by The Eagles (the band). Highly unsure if anyone’s calling up home services in a ghost town. I removed these rows too. Applied the same logic for other incorrect entries.



All	business_id	name	address	city	state	postal_code	latitude	longitude	stars	review_count	is_open	categories
	1949	pOUpRcmhDWrewp_IVsCQ	Gilliland Masonry	Eagle	CA	93616	43.6954424	-118.3540138	4	27	1	Home Services, Masonry/Concrete, Contractors

Real Goleta 1

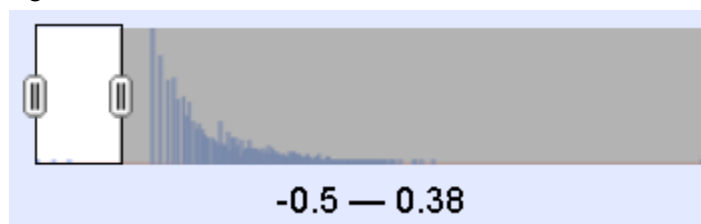
Q2 (10 pt): Clean “stars” and “review_count”. Describe 1) what steps did you do to find the dirty data; 2) list the dirty data entries you found; Please drop all the dirty data for this question.




Ans.

- After looking at majority of the rating, I think the ratings are on 1-5 stars scale and “6” is an incorrect record.

All	business_id	name	address	city	state	postal_code	latitude	longitude	stars	review_count	is_open	categories	
	48.	GCU4EGXALp7A-dut7yTMQ	Upholstery Decor	5788 Hollister Ave	Goleta	CA	93117	34.4359941	-119.826305	6	24	1	Local Services, Furniture Reupholstery, Antiques, Home & Garden, Shopping, Furniture Stores, Mattresses
	53.	PaqDZUu78IZIMvcYqCEBtg	Rincon Brewery	205 Santa Barbara St	Santa Barbara	CA	93101	34.4161821	-119.6897762	6	32	1	Breweries, Brewpubs, Food

- After converting the review_counts to their numeric counterpart and looking at its logarithmic scale it’s clear that there are some faulty records due to the negative logarithms. Review counts cannot be a fraction.



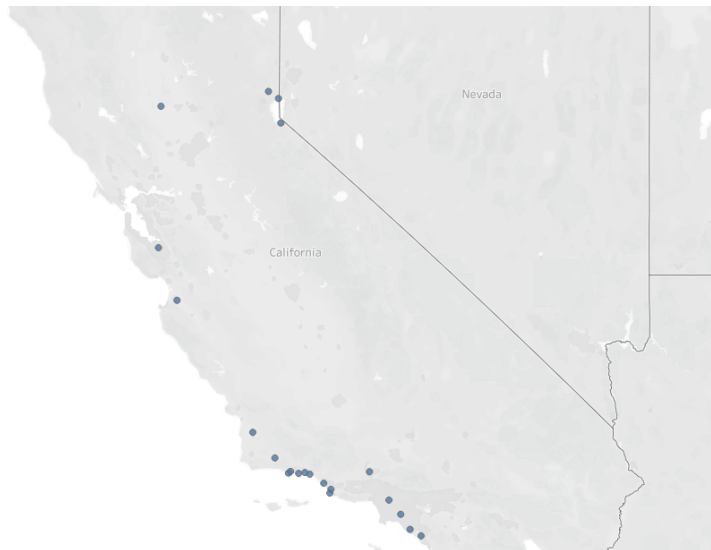
	All	business_id	name	address	city	state	postal_code	latitude	longitude	stars	review_count	is_open	categories
	7.	QZU7TcrztBb3bPaPbVCKXg	805 Ink	1228 State St	Santa Barbara	CA	93101	34.4241297	-119.7053211	4.5	0.68	1	Beauty & Spas, Tattoo
	21.	#NAME?	Dune Coffee Roasters - Anacapa	528 Anacapa St	Santa Barbara	CA	93101	34.4189939	-119.6950694	4	0.32	1	Coffee & Tea, Coffee Roasteries, Food
	22.	18eWJFJbXyR8j_5xfCRLYA	Siam Elephant	508 Linden Ave	Carpinteria	CA	93013	34.3965102	-119.5216815	4.5	0.46	1	Restaurants, Thai

On the other hand we have a giant number here, 2380000 (exceeding the city’s population of 91k).

▼ All	▼ business_id	▼ name	▼ address	▼ city	▼ state	▼ postal_code	▼ latitude	▼ longitude	▼ stars	▼ review_count	▼ is_open	▼ categories	
★	30.	#NAME?	Lama Dog Tap Room	116 Santa Barbara St	Santa Barbara	CA	93101	34.4157471	-119.8884801	4	2380000	1	Beer, Wine & Spirits, Pubs, Arts & Entertainment, Kombucha, Nightlife, Wineries, Food, Bars, Beer Bar

Q3 (10 pt): With the help of Tableau, identify if there are any dirty data in their geographic locations of all businesses. 1) where are the dirty data located? Show the screenshot from Tableau and also list all state that contains dirty data; 2) Now that you know where are the dirty data, continue using OpenRefine to drop them; 3) Looking into the dirty data, how would you clean them if we didn't want to drop them?

Ans.



		Mission Canyon	Unrecognized
		Montecito	Unrecognized
Spring Hill	Unrecognized	Santa Barbara & Ventura ...	Unrecognized
Tampa	Unrecognized	Sparks	Unrecognized
West Hill	Unrecognized		

Via Tableau

Dirty datas: Mission Canyon, Montecito, Santa Barbara & Ventura counties, Sparks, Spring Hill, Tampa, West Hill. These are all located in California.

- I think some of these were unrecognized because we are technically looking for cities, but these particularly are towns, residential groups, counties, etc.

Q4 (10 pt): Clean other fields in the file. Describe 1) what steps did you do to find them; 2) what are the businesses that need cleaning and how did you clean them?

Ans.

- Some of the same Business names had variations in casing, so I performed clustering. Ex:

3	7	<ul style="list-style-type: none"> Blenders In the Grass (5 rows) Blenders In The Grass Blenders in the Grass 	<input type="checkbox"/>	Blenders In the Grass
2	2	<ul style="list-style-type: none"> lululemon Athletica lululemon athletica 	<input type="checkbox"/>	lululemon Athletica
2	2	<ul style="list-style-type: none"> ANGL Angl 	<input type="checkbox"/>	ANGL
2	5	<ul style="list-style-type: none"> Carl's Jr (3 rows) Carl's Jr. (2 rows) 	<input type="checkbox"/>	Carl's Jr
2	2	<ul style="list-style-type: none"> il Fustino il fustino 	<input type="checkbox"/>	il Fustino

- For the purpose of maintaining a better database, I think records without a unique business_id should be removed (or if possible generate a random id if the situation requires)

☆	3161	#NAME?	Finch & Fork	31 W Camillo St	Santa Barbara	CA	93101	34.4203608	-119.7024752	4	1405	1	Breakfast & Brunch, American (New), Restaurants, American (Traditional), Nightlife, Bars
☆	3167	#NAME?	Beachside Bar-Cafe	6905 Sandpitt Rd	Goleta	CA	93117	34.4170513	-119.8292189	3.5	840	0	American (New), Restaurants, Nightlife, Seafood, Bars
☆	2161	#NAME?	Benchmark Eatery	1201 State St	Santa Barbara	CA	93101	34.4233679	-119.7047765	4	544	1	American (Traditional), American (New), Breakfast & Brunch, Restaurants, Seafood, Vegetarian, Nightlife, Event Planning & Services, Bars, Venues & Event Spaces
☆	3652	#NAME?	Edomasa	2710 De La Vina St	Santa Barbara	CA	93105	34.4359425	-119.7253428	3.5	406	1	Restaurants, Sushi Bars, Japanese
☆	1076	#NAME?	Renaud's Patisserie & Bistro	1324 State St Ste N, Arlington Plaza	Santa Barbara	CA	93101	34.42536822	-119.7059072	4	379	1	Coffee & Tea, Breakfast & Brunch, Restaurants, French, Salad, Sandwiches, Desserts, Bakeries, Food
☆	1913	#NAME?	Apna Indian Kitchen	719 State St	Santa Barbara	CA	93101	34.4192398	-119.6993869	4.5	246	1	Gluten-Free, Gastropubs, Vegan, Restaurants, Indian
☆	1301	#NAME?	Tinker's Burgers	2275 Ortega Hill Rd, Ste C	Summerland	CA	93067	34.4215736	-119.6012172	4	201	1	Burgers, Restaurants
☆	2013	#NAME?	Taffy's Pizza	2026 De La Vina St	Santa Barbara	CA	93105	34.4295503	-119.7178051	4	192	1	Restaurants, Pizza
☆	255	#NAME?	Pizza Mizza		Santa Barbara	CA	93105	34.4208305	-119.6991901	3.5	189	0	Food Delivery Services, Food, Restaurants, Pizza, Italian
☆	3904	#NAME?	Segway of Santa Barbara	122 Gray Ave	Santa Barbara	CA	93101	34.4149987	-119.6889359	5	165	1	Arts & Entertainment, Scooter Rentals, Bike Rentals, Hotels & Travel, Active Life, Local Services, Scooter Tours, Tours

- Also removed records with empty addresses (too many to currently look for without resources)
- I filled in an obvious category

All	business_id	name	address	city	state	postal_code	latitude	longitude	stars	review_count	is_open	categories	
	1721	FhhbGFzG3w7qZcUJFA0g	Kennedy Accounting Systems	1332 De La Vina St	Santa Barbara	CA	93101	34.4149987	-119.6889359	5	165	1	Accounting

Date type: text

Accounting

Apply

Apply to All Identical Cells

Cancel

Enter

Ctrl-Enter

Exit

- Also came across a few incorrect information such as NLC production is_open = 1, however, according to google and their social media they are permanently closed.

Q5 (10 pt): If we want to perform deeper cleaning based on this result, what external resource can you think of that can be useful?

Ans. I think web crawling through an external API would help in keeping the data updated.

Q6 (10 pt): a) Compare the review count between restaurants and non-restaurants qualitatively. Report your findings. b) Compare the star rating between restaurants and non-restaurants qualitatively. Report your findings.

Ans.

a. As evident from the longer tail and smaller concentration at 0 compared to the non-restaurant businesses, the Restaurants have higher review counts. Despite there existing a greater number of non-restaurant businesses (around 3.4+), the most review counts is around 400+. However, there are only 1k+ Restaurants, some of the businesses must be really big as they report as high as 2000-3800 review counts while others showing a downward curve from 5 review counts to 2000 review counts.

b. On average, Restaurants get more 4 stars while Non-Restaurants get more 5 stars. Restaurants show more of a bell curve while Non-restaurants do not.

Q7 (10 pt): We conjecture those popular restaurants (restaurants with more reviews) tend to be rated higher. Make a plot to check if this is true. Report your findings and why.

Ans. It is true that restaurants with more review counts are rated high. Possibly because these restaurants consistently are so good, that there are more customers and therefore more reviews. Perhaps bigger restaurants are also capable of implementing effective ways to get customer feedback, such as through online apps and so on.

Q8 (10 pt): a) Using boxplots to explore the distribution of review count of different types of business. b) Using boxplots to explore the distribution of star rating of different types of business. For both questions, you should compare at least 'Auto Repair', 'Gas' and 'Salon'. Report your findings.

Ans.

- a. We see that nightlife, restaurants and salons have the highest average and median review count than others. There are quite a lot of outliers. The restaurant shows the biggest leap in terms of review counts. There are also some big nightlife businesses. Auto repairs and Doctors have the lowest review counts.
- b. Businesses like Auto Repair, doctors, salon get between 50-100 4 or 5 stars. Low rating is quite low in comparison. Home services one of the few to get more 5 stars than 4, besides professional services. However, professional services has a

lower median than home services. Gas business shows not much difference among the stars they receive, they mostly get 3-stars but that is also very very little (about 11) which is closely followed by other rating. Nightlife and Restaurants are both very well received, as evident from the high number of 4 stars.

Q9 (10 pt): Formulate a meaningful question in this dataset and answer it yourself using one or more plots.

Question - How does the city, number of reviews, and ratings affect each other?

Ans -

- a) Here City color shows details about star ratings. Size shows maximum of Review Count. The marks are labeled by city.

Here we see that the cities in the big bubbles Santa Barbara, Santa Vista, Goleta, Carpinteria have very high number of review counts but none of them reach 5 stars. Perhaps Increased number of reviews allows for more diversity in ratings. We see several small bubbles, indicating few reviews have 5 star ratings.