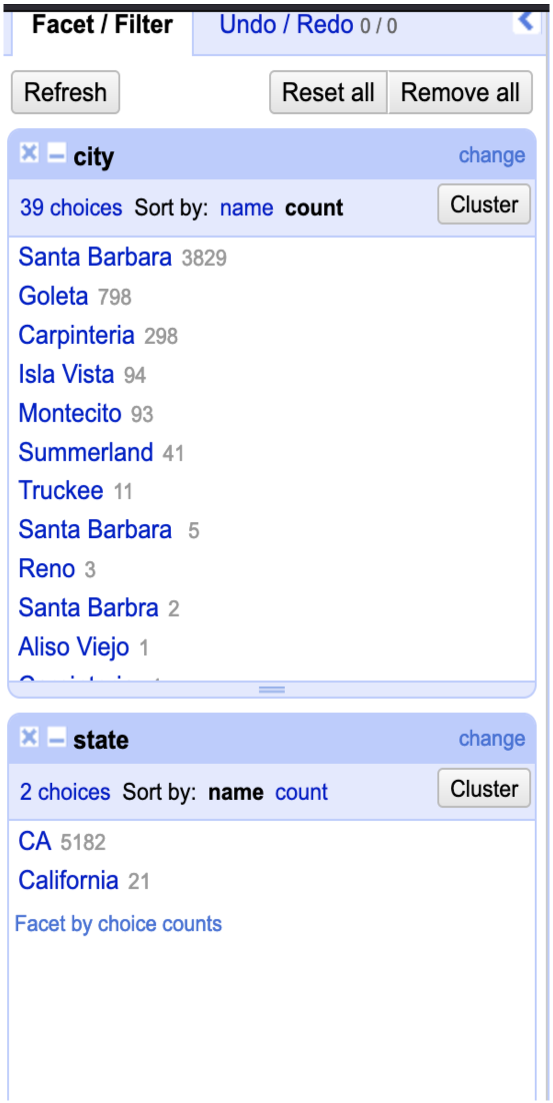
HW1 –

Part 1: Data cleaning and EDA

Q1 : 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 1.1): Post opening the csv file in openrefine. Applied the text facet on the "city" column to get a cluster of cities and based on that find out all the different groups formed in the data. Repeated the same for "state" data.



city

39 choices Sort by: name count

Santa Barbara	3829
Goleta	798
Carpinteria	298
Isla Vista	94
Montecito	93
Summerland	41
Truckee	11
Santa Barbara	5
Reno	3
Santa Barbra	2
Aliso Viejo	1

state

2 choices Sort by: name count

CA	5182
California	21

Ans 1.2): There are about 39 entries in the "city" category apart from the expected "Santa Barbara". As for "states" we have CA and California.

Text facet screenshot on city and state with dirty data.

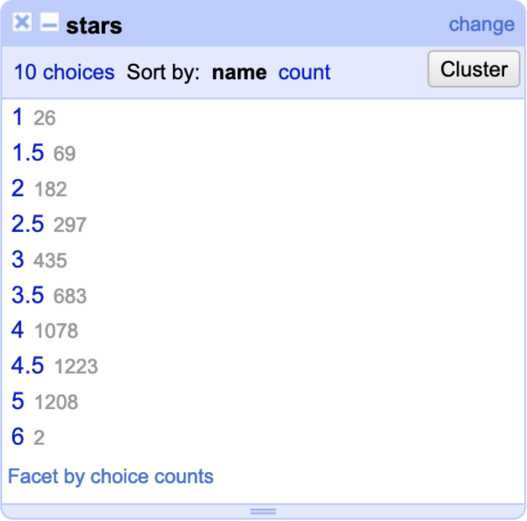
Ans 1.3): a) Dirty data in cities come under three categories apart from the primary **Santa Barbara**:

- Typos/different formats of Santa Barbara.
- Counties/cities part of Santa Barbara county, these are the places that will be categorised and merged under Santa Barbara.
- Places not in Santa Barbara officially will be discarded from the dataset. Listed below:
- Truckee, Reno, Cerritos, Costa mesa, Eagle, kings beach, LA, oxnard, port hueneme, salinas, tampa, valencia, ventura, west hill, meridian, santa clara.

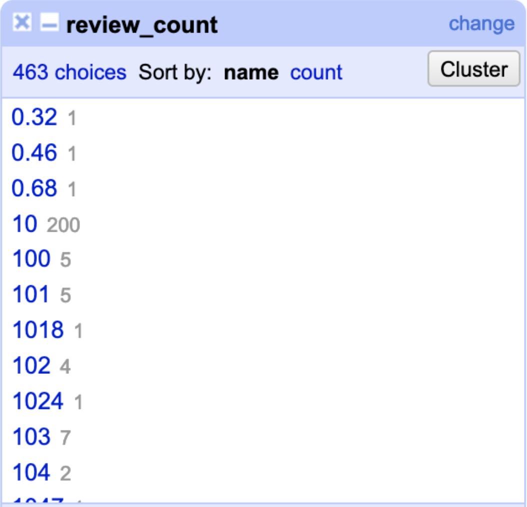
b) For states we merge "**California**" into **CA**, and our job is done here.

Q2 : 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 2.1) Add the text facet to “**stars**” column and sort by names filter. Similarly repeat for “**review_count**” column as well.



1	26
1.5	69
2	182
2.5	297
3	435
3.5	683
4	1078
4.5	1223
5	1208
6	2

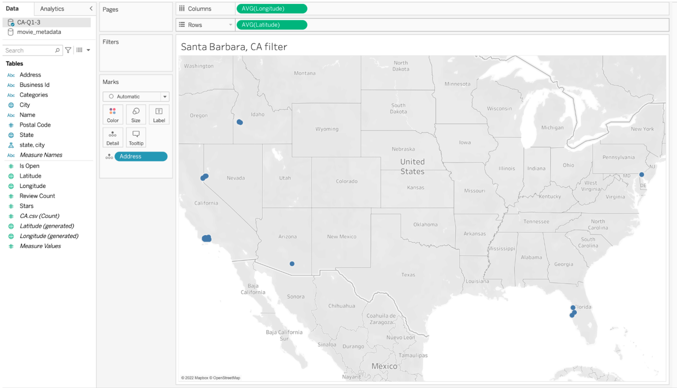


0.32	1
0.46	1
0.68	1
10	200
100	5
101	5
1018	1
102	4
1024	1
103	7
104	2

Ans 2.2) a) Stars column has 2 entries with 6 as a rating. Although rating systems can have arbitrary upper limit on the number of stars but for the sake of assumptions, let’s set a cap to the ratings at 5.0 and drop the rows with 6 rating. Although, we could have rounded them down to 5 but 2 rows no harm.

b) Review_count column has 3 entries with fractional numbers, since its impossible for fractional number of people to review a place, we drop the rows with fractional numbers in review_count. **To drop just select the three outlier columns and then goto “all” column dropdown -> edit rows -> remove matching rows.**

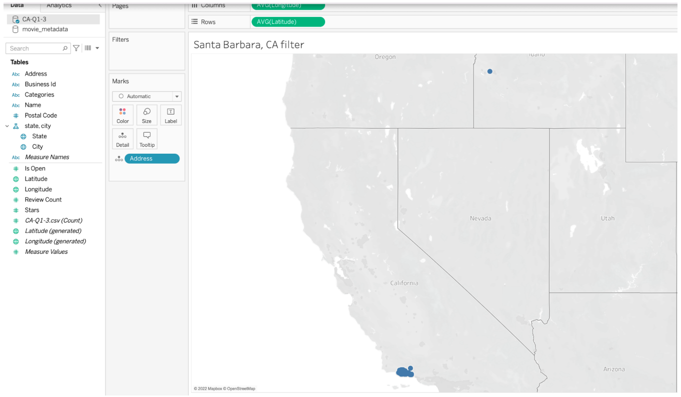
Q3 : With the help of Tableau, identify if there are any dirty data in their geographic locations of all businesses. 1) where are the dirty data located? Show the screenshot from Tableau and also list all state that contains dirty data; 2) Now that you know where are the dirty data, continue using OpenRefine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


Columns AVG(Longitude)
Rows AVG(Latitude)
Santa Barbara, CA filter

Ans 3.1) Dirty data seems to be seeping into the following states:
Idaho, Nevada, Arizona, Florida, NJ.

Screenshot 1: Tableau filter on geographic data of the original data.

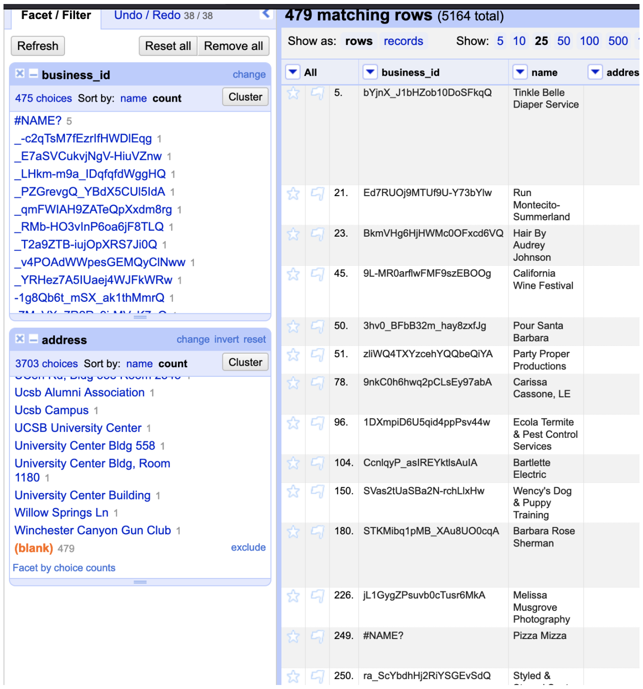
Screenshot 2: Tableau filter on geographic data of the cleaned data from previous questions.


Columns AVG(Longitude)
Rows AVG(Latitude)
Santa Barbara, CA filter

Ans 3.2) Dropped caroline rife photography to remove outlier from previously cleaned data.

Ans 3.3) Ideally would check if the address added and the coordinates match, in case there is a mistake, update the value but if there is no match no choice but to drop them.

Q4 : 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?



business_id — change
475 choices Sort by: name count — Cluster

business_id	count
#NAME?	5
_-c2qTsM7fEzrIfHWDlEqg	1
_E7aSVCukvjNgV-HiuVZnw	1
_LHkm-m9a_IDqfqfdWggHQ	1
_PZGrevgQ_YBdX5CUI5ldA	1
_qmFWIAH9ZATeQpXxdm8rg	1
_RMb-HO3vlnP6oa6jF8TLQ	1
_T2a9ZTB-iujOpXRS7Ji0Q	1
_v4POAdWWpesGEMQyClNww	1
_YRHez7A5lUaej4WJFkWRw	1
-1g8Qb6t_mSX_ak1thMmrQ	1

address — change invert reset
3703 choices Sort by: name count — Cluster

address	count
Ucsb Alumni Association	1
Ucsb Campus	1
UCSB University Center	1
University Center Bldg 558	1
University Center Bldg, Room 1180	1
University Center Building	1
Willow Springs Ln	1
Winchester Canyon Gun Club	1
(blank)	479

exclude
Facet by choice counts

479 matching rows (5164 total)

Show as: rows records Show: 5 10 25 50 100 500

All	business_id	name	addres
5.	bYjnX_J1bHZob10DoSFkqQ	Tinkle Belle Diaper Service	
21.	Ed7RUOj9MTUf9U-Y73bYlw	Run Montecito-Summerland	
23.	BkmVHg6HjHWMc0OFxcd6VQ	Hair By Audrey Johnson	
45.	9L-MR0arflwFMF9szEBOOg	California Wine Festival	
50.	3hv0_BFbB32m_hay8zxfJg	Pour Santa Barbara	
51.	zliWQ4TXYzcehYQQbeQiYA	Party Proper Productions	
78.	9nkC0h6hwq2pCLsEy97abA	Carissa Cassone, LE	
96.	1DXmpiD6U5qid4ppPsv44w	Ecola Termite & Pest Control Services	
104.	CcnlqyP_asIREYktlsAuIA	Bartlette Electric	
150.	SVas2tUaSBa2N-rchLlxHw	Wency's Dog & Puppy Training	
180.	STKMibq1pMB_XAu8UO0cqA	Barbara Rose Sherman	
226.	jL1GygZPsuvb0cTusr6MkA	Melissa Musgrove Photography	
249.	#NAME?	Pizza Mizza	
250.	ra_ScYbdhHj2RiYSGEvSdQ	Styled &	

Ans 4.1) Applied text facet on the three remaining columns which would be considered necessary for a serious analysis.

- **Business-id :** About 68 rows have missing business-ids, since we are short on data we can either look them ids up or just set a random different ids for them and carry on. Easiest way is dropping them as this is a unique identifier for each business and not easy to source.
- **Address**: About 479 rows have missing ids but since we have coordinates we can narrow their addresses near perfectly but this seems overkill as they arent needed immediately. I'd keep them blank unless needed.
- **Categories**: One blank field here as well, but the name of the business "**Kennedy Accounting Systems**" helps us categorise it with good approximation. Although, since its a singular row and the name is relevant/simple enough, it's easier to do but otherwise I would have dropped it.

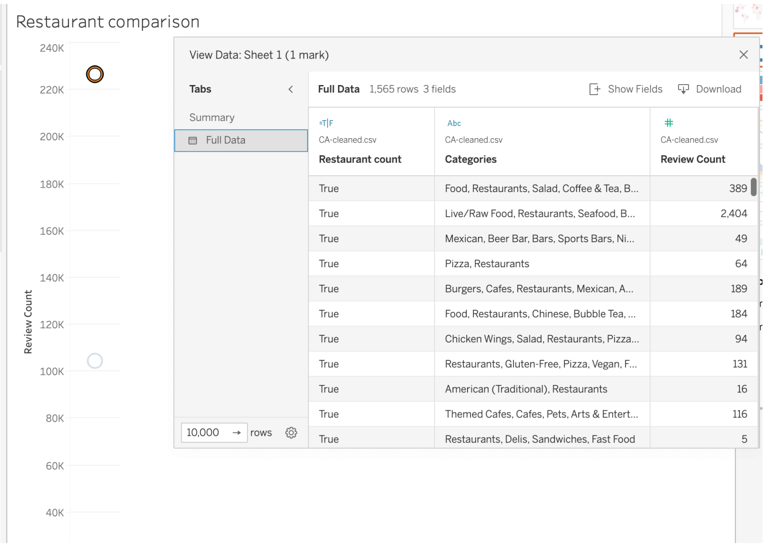
Ans 4.2) For address I believe keeping them as is should be good for our analysis since we also have coordinates for further location, but for business-id we have no source of finding out the correct ids unless we dig deep. So for the 68 rows of absent business-ids, I'll drop them from the data, since 68 rows dont diminish our overall quality and quantity of data by much..

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

- Pandas would be my first choice to clean data. Playing around with data in python and performing complex filtering based on multiple columns with correlated invalid data. Suppose a business might having a missing address, not a big deal, but if they were missing coordinates and other fields then they are essentially dirty data. Pandas also has beautiful visualisation libraries that would help with a more clear cut visualisation of how outliers are skewing the data helping us further. Ultimately, the customisation around lambda functions and conversion of CSV to JSON definitely gives the user a more powerful cleaning/EDA tool.
- Another underrated aspect of data cleaning would be to read the source of truth, if there is any. Knowing how the data was sourced and its references would help us gather more insight on how the data came to be. Knowing downstream sources of truth would help us cross verify in a sense the data we believe can be skewed and if they are different prevent us from working further on the “misleading” data.
- Not sure if there is a tool, but something similar to what geographic visualisation tableau has but with features that would help me auto remove locations post a certain radius of miles/kms.

Part II. EDA

Q6) a) Compare the review count between restaurants and non-restaurants qualitatively. Report your findings.



View Data: Sheet 1 (1 mark)

Tabs

Summary

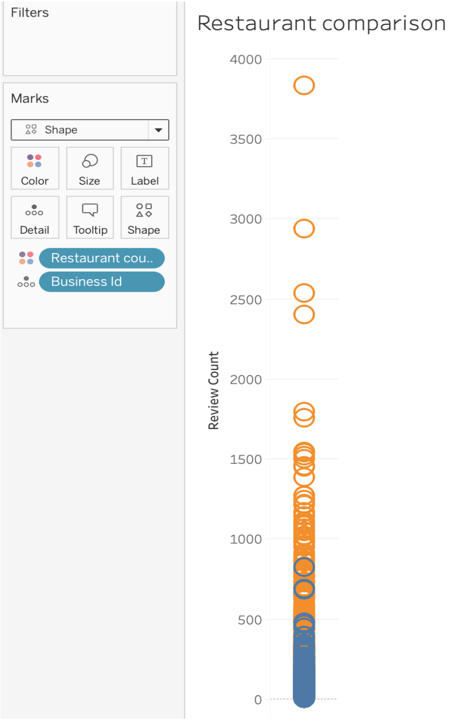
Full Data

Full Data 1,565 rows 3 fields

Show Fields Download

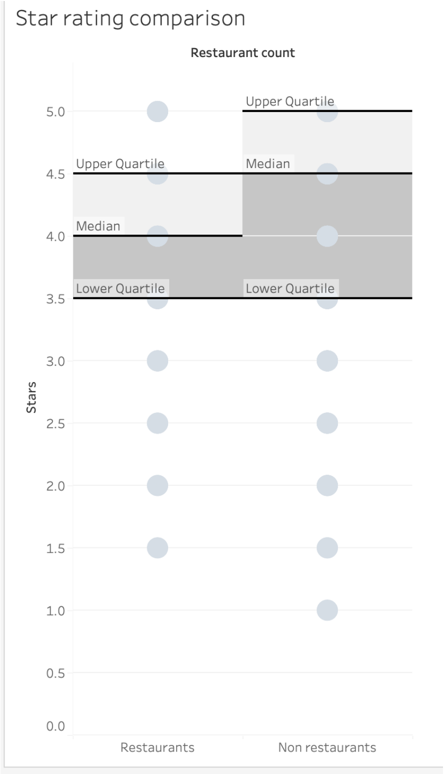
CA-cleaned.csv Restaurant count	CA-cleaned.csv Categories	CA-cleaned.csv Review Count
True	Food, Restaurants, Salad, Coffee & Tea, B...	389
True	Live/Raw Food, Restaurants, Seafood, B...	2,404
True	Mexican, Beer Bar, Bars, Sports Bars, Ni...	49
True	Pizza, Restaurants	64
True	Burgers, Cafes, Restaurants, Mexican, A...	189
True	Food, Restaurants, Chinese, Bubble Tea, ...	184
True	Chicken Wings, Salad, Restaurants, Pizza...	94
True	Restaurants, Gluten-Free, Pizza, Vegan, F...	131
True	American (Traditional), Restaurants	16
True	Themed Cafes, Cafes, Pets, Arts & Entert...	116
True	Restaurants, Delis, Sandwiches, Fast Food	5

10,000 → rows



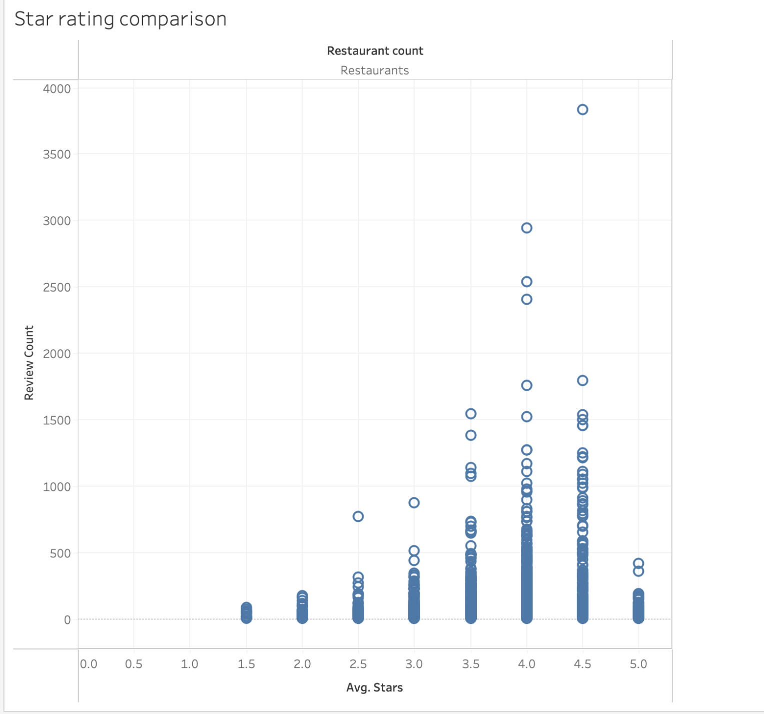
Ans 6) 1) Restaurants count for about 1673 rows of reviews, this is with the filter 'restaurants', 'food' and 'bars' for added assurance. Despite, non restaurants outnumbering restaurants by almost 2x (3423 vs 1673), the number of reviews for restaurants are far more in number (225,573 vs 104,512)). Not only are the total number of reviews dominated by restaurants, the number of reviews left on average for a business tend to skew higher for restaurants as compared to non-restaurants. Second graph depicts the same. It seems the restaurants tend to get more feedback and reviews compared to other business which shows that reviewers are more vocal about leaving feedback on food related businesses compared to other companies in the area.

Q6 b) Compare the star rating between restaurants and non-restaurants qualitatively. Report your findings.



Ans 6) 2) Here it seems the restaurants take a major share in terms of relatively bad reviews. The median of restaurants lie at 4.0 compared to non restaurants with a median of 4.5, although it is surprising that restaurants and non restaurants share the same lower quartile, this give some insights on customers majorly reviewing businesses within the 3.5-5 range.

Q7): 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.

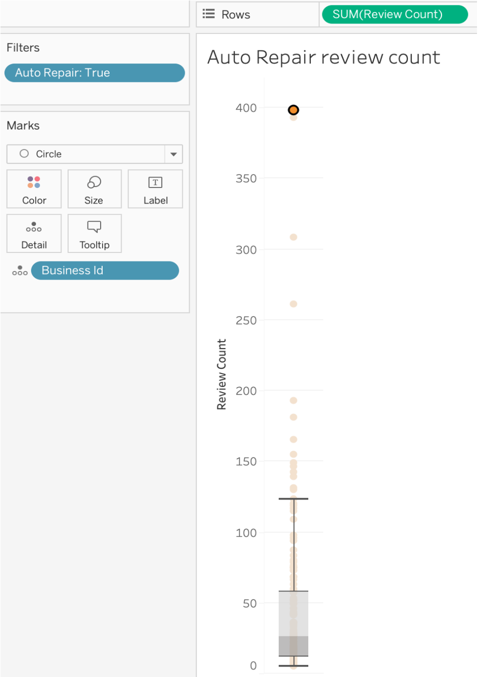


Ans 7) This conjecture works at a certain level of review count and what we would consider high and above high. At 1000 reviews and above we see the restaurants getting rated 3.5 and above, certainly high enough compared to lower reviewed restaurants proving the conjecture. If we consider restaurants with 500 reviews and more we see the distribution stretching across 2.5 to 4.5, except a few outliers we still see quite a few restaurants at a higher rating in line with the conjecture. Finally, based on observing the graph, I believe the conjecture works for the most part but is subject to our considered threshold. When we consider a lower limit like 200-300 as high, we can counter argue that we find a lot of restaurants with a 5 rating in this range of reviews, and an almost equal number above and below 3.5. So conjecture works but with caveats.

Q8: 1) Using boxplots to explore the distribution of review count of different types of business.

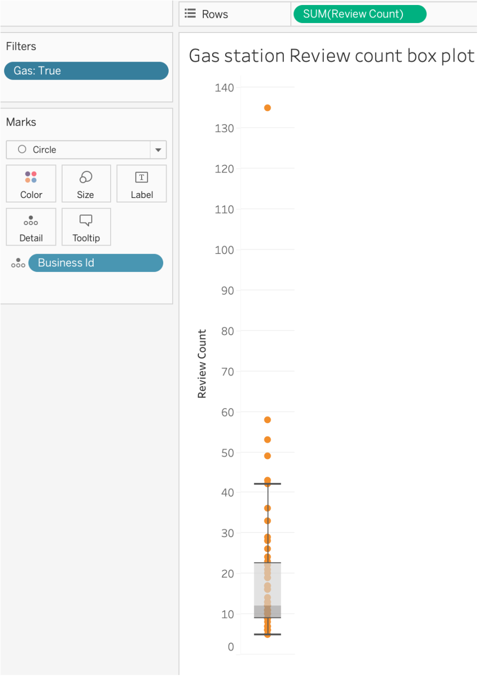
2) Using boxplots to explore the distribution of star rating of different types of business.

1) **"Review Count" box plots for different businesses:**



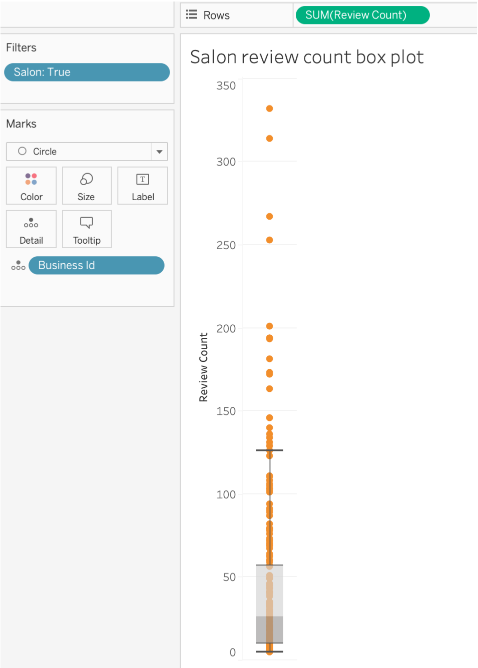
Auto Repair repair review count box plot

- Median: 26
- Upper Whisker: 123
- Lower Whisker: 5
- Minimum count: 5
- Maximum count: 398



Gas station review count box plot

- Median: 12
- Upper Whisker: 42
- Lower Whisker: 5
- Minimum count: 5
- Maximum count: 135



Rows: SUM(Review Count)

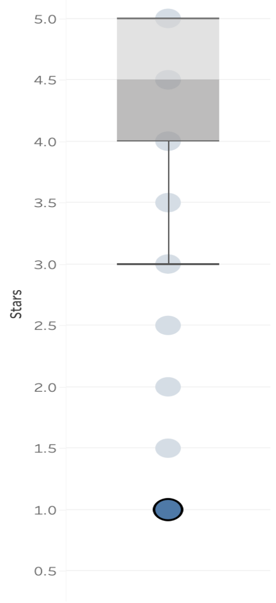
Filters: Salon: True

Salon review count box plot

- Median: 26
- Upper Whisker: 126
- Lower Whisker: 5
- Minimum count: 5
- Maximum count: 332

2) Star rating box plots for different businesses

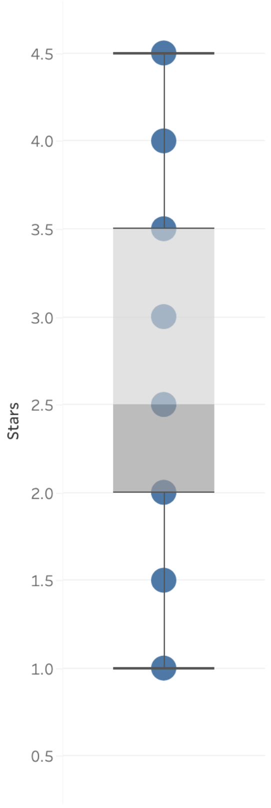
Auto repair star count box plot



Auto Repair star rating box plot

- Median: 4.5
- Upper Whisker: 5
- Lower Whisker: 3
- Minimum count: 1
- Maximum count: 5

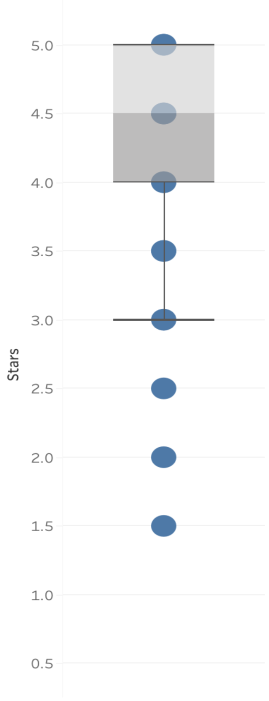
Gas star count box plot



Gas station star rating box plot

- Median: 2.5
- Upper Whisker: 4.5
- Lower Whisker: 1
- Minimum count: 1
- Maximum count: 4.5

Salon star count box plot

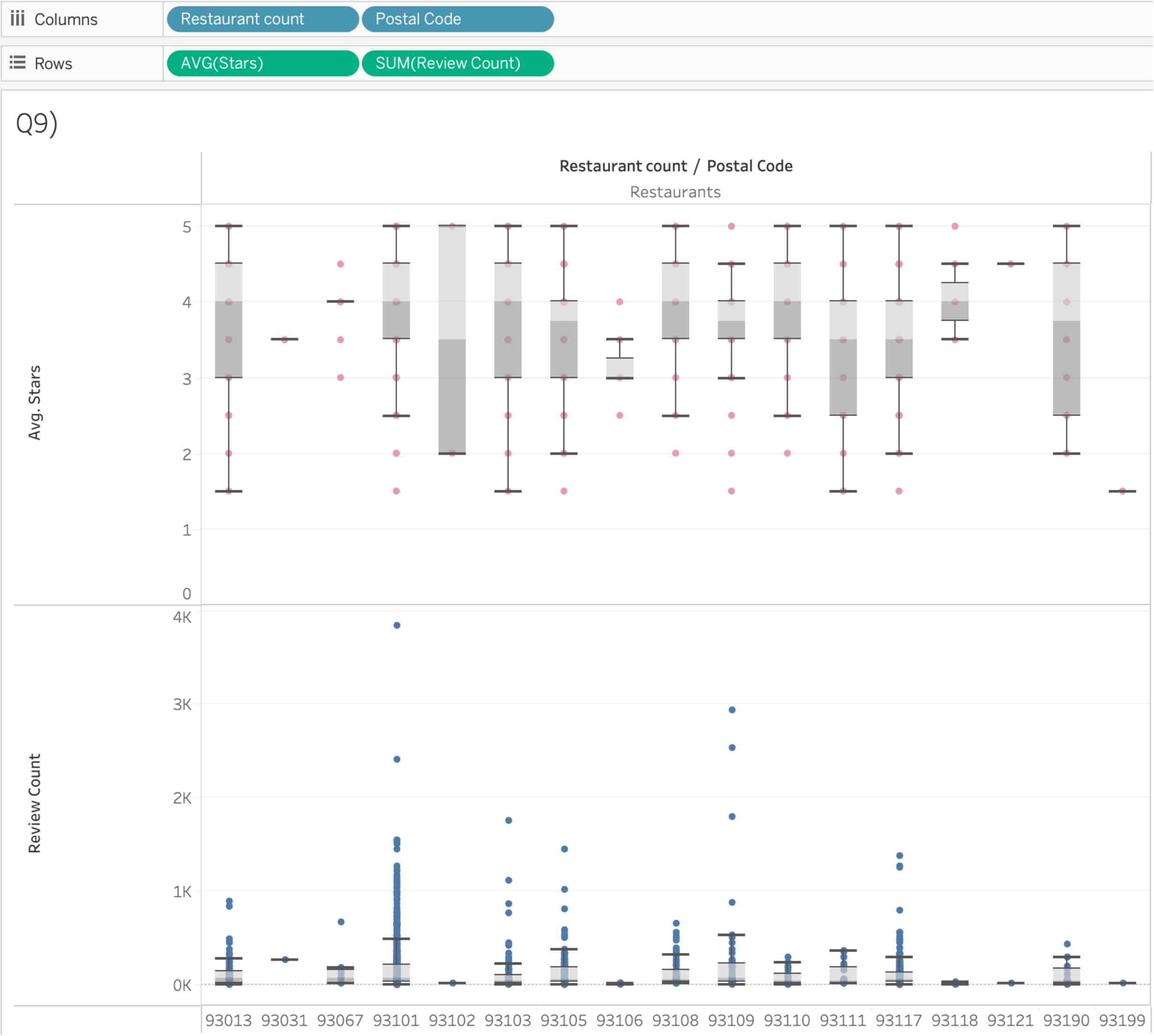


Salon star rating box plot

- Median: 4.5
- Upper Whisker: 5
- Lower Whisker: 3
- Minimum count: 1.5
- Maximum count: 5

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

Q) 9.1) a) Which areas have the highest number of restaurants on average and among those areas which areas have the highest rated restaurants on median?



Columns Restaurant count Postal Code

Rows AVG(Stars) SUM(Review Count)

Ans 9.1 a) Areas with following pincodes have the most restaurants (Noting top 5):

- 93101: 611
- 93117: 338

- 93105: 162
- 93013: 108
- 93103: 106

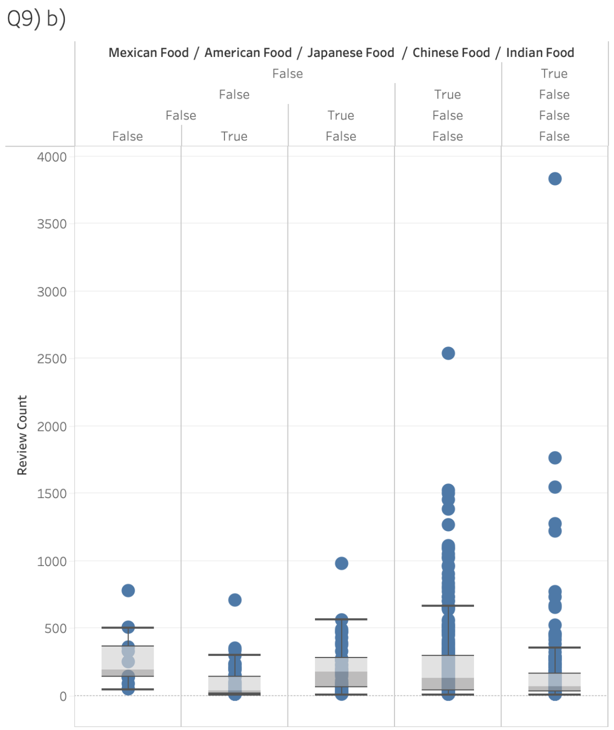
These pincodes would be the preferred places if you wanted to go to places with a good amount of diverse restaurants.

b) Areas with following pincodes have on median the highest rated restaurants (Noting top 5): **(Note: Considered the top 5 from above for comparison, since we want high density of good restaurants in an area code).**

- 93101: Median - 4.0
- 93103: Median - 4.0
- 93013: Median - 4.0
- 93105: Median - 3.75
- 93117: Median - 3.5

Based on the above graph, if we were to decide which places to explore in SB in terms of restaurants, we'd on average find higher quality in 93101 with the most restaurants and among the best median star rating, especially "Los Agaves" in this area code is the most popular restaurant in SB as a whole by far, with 4.5 rating and over 3500 reviews, this place is a must try for people visiting SB. We should also avoid area code 93199, which has singular 1.5 rated starbucks, yikes!.

Q) 9.2) Which cuisine is the most popular in SB?



Ans) Cuisines considered here Mexican, Chinese, Japanese, Indian and American. Cuisines most common here

- American - 272
- Mexican - 154
- Chinese - 69
- Japanese - 39
- Indian - 10

American cuisine is by and far the most commonly found food in SB with Mexican as a second. We can also find the most reviewed restaurants for each cusine, this can be correlated to their popularity. Boathouse at Hendry's is the most popular eatery serving american food. Los Agaves is the most reviewed for mexican food, Arigato Sushi for Japanese, Empty Bowl Gourmet Noodle Bar and Tamira for Chinese and Indian respectively. Also to note here is all the above 5 restaurants have above 4.0 rating, so these are definitely highly recommended restaurants for your cuisines of choice.