

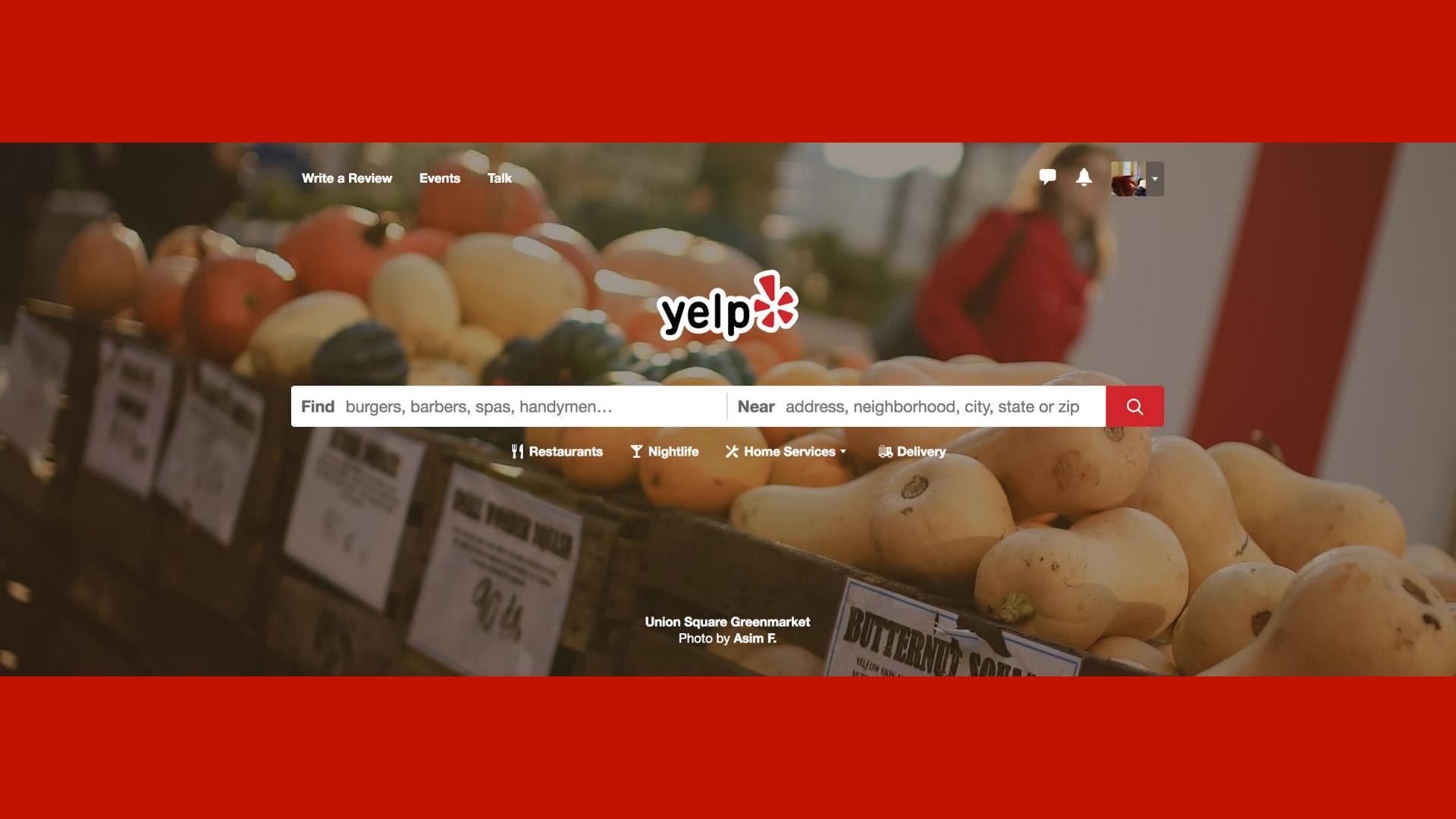
Yelp Dataset Investigation

Caelan O'Sullivan

University of Texas
Data Analytics and Visualization

April 2018



The background of the page is a blurred photograph of a vibrant outdoor market. In the foreground, there are several boxes filled with various fruits and vegetables, including apples, oranges, and what look like butternut squash or pumpkins. A sign in one of the boxes clearly reads "BUTTERSCOTCH SQUASH". In the background, a person wearing a red jacket is visible, and the overall atmosphere is bright and colorful.

Write a Review

Events

Talk



Find burgers, barbers, spas, handymen...

Near address, neighborhood, city, state or zip



🍴 Restaurants

▀ Nightlife

✖ Home Services

🚚 Delivery

Union Square Greenmarket
Photo by Asim F.

Ramen Tatsu-Ya

Claimed



1121 reviews

[Details](#)[★ Write a Review](#)[Add Photo](#)[Share](#)[Bookmark](#)

\$\$ · Ramen

[Edit](#)

📍 1234 S Lamar Blvd
Austin, TX 78704
Bouldin Creek, South Lamar District,
78704 (South Austin)

[Edit](#)[Get Directions](#)[\(512\) 893-5561](#)[ramen-tatsuya.com](#)[Message the business](#)[Send to your Phone](#)[See all 1359](#)

"I get the og (veggie when it's offered on Sunday's) with a **spicy bomb** and it is amazing!" in 73 reviews



"In total, We ordered the **Tonkotsu Original**, the Tsukemen w/ extra chashu, Gyoza and the Gotcha Matcha dessert." in 38 reviews



Today 11:00 am - 10:00 pm
Closed now

[Full menu](#)

\$\$\$\$ Price range \$11-30

Yelp Dataset Challenge

Discover what insights lie hidden in our data.



What is the dataset challenge?

The challenge is a chance for students to conduct research or analysis on our data and share their discoveries with us. Whether you're trying to figure out how food trends start or identify the impact of different connections from the local graph, you'll have a chance to win cash prizes for your work! See some of the [past winners](#) and [hundreds of academic papers written](#) using the dataset.

The Dataset



5,200,000 reviews



174,000 businesses



11 metropolitan areas

1,100,000 tips by 1,300,000 users

Over 1.2 million business attributes like hours, parking, availability, and ambience

Aggregated check-ins over time for each of the 174,000 businesses

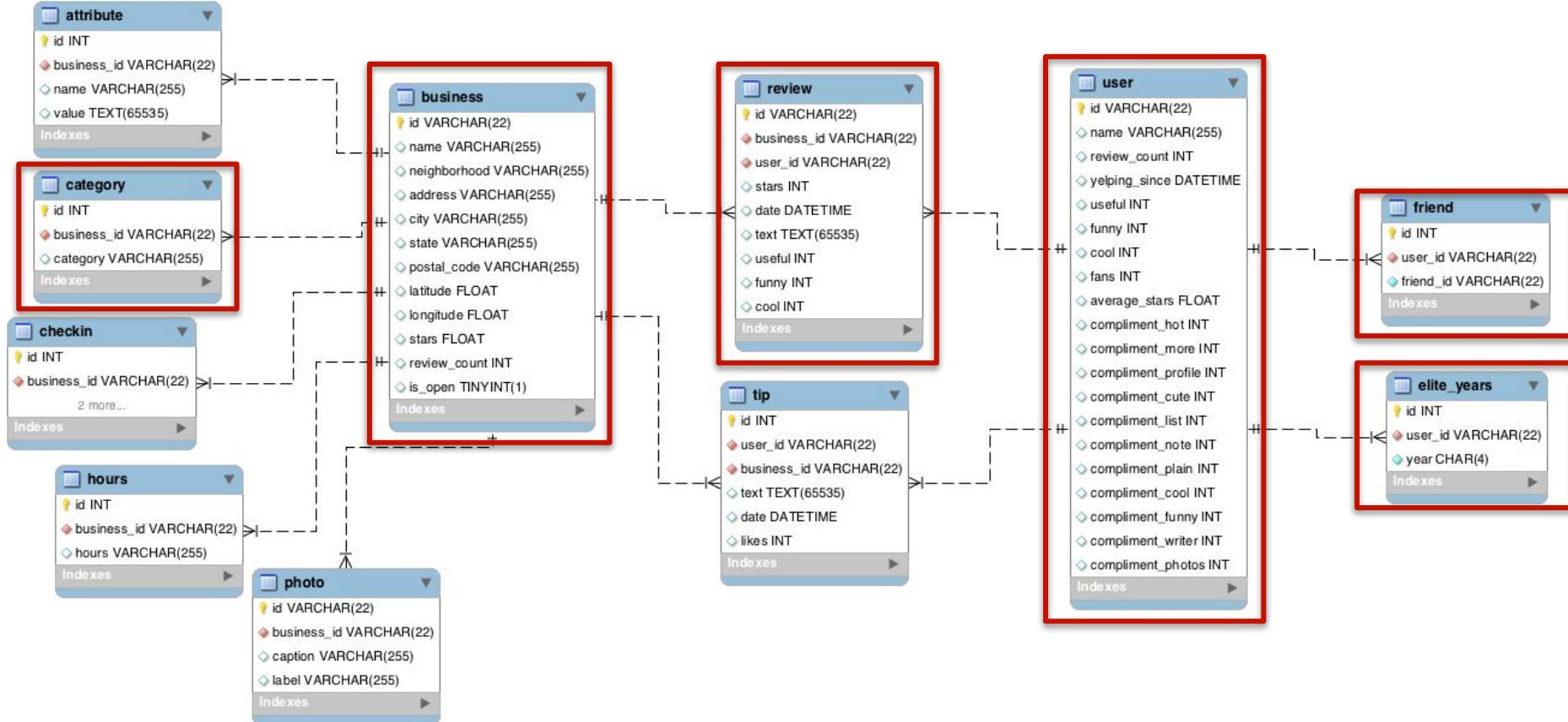


Project Goals

- Drive toward Yelp's KPIs
- Sharpen SQL DBA skills
- Explore Tableau
- Create business-driven, dynamic dashboards (all-time and periodic roll-ups)



Dataset





Dataset

Big data, big problems

- 7.5 GB
- 1.3 million user rows
- 5.2 million review rows
- Tableau is slow, so:
 - Forget joins
 - Create custom pre-aggregated tables in MySQL and visualize
 - Utilize extracts



Dataset

```
USE yelp_db;

-- Aggregated table with friend counts
CREATE TABLE user_copy AS SELECT
    user.*,
    COALESCE(friend.friend_count,
    0) friend_count,
    IF(elite.user_id IS NULL,
    0,
    1) elite_status
FROM
    user AS user
LEFT JOIN
(
    SELECT
        user_id,
        COUNT(*) friend_count
    FROM
        friend
    GROUP BY
        1
) friend
    ON friend.user_id = user.id
LEFT JOIN
(
    SELECT
        DISTINCT user_id
    FROM
        elite_years
) elite
    ON friend.user_id = elite.user_id;

-- Per-day review roll-up table
CREATE TABLE review_daily_rollup SELECT
    user_id,
    date,
    AVG(stars) AS avg_stars,
    COUNT(*) AS review_count
FROM
    review
GROUP BY
    date,
    user_id
ORDER BY
    date DESC;
```

- Created custom pre-agged tables
- Connected MySQL database to Tableau
- Plus:
 - Flag columns (0/1)
 - Self-joins (Categories)
 - Daily/weekly rollup tables based on aggregations



Dataset

```
280 WHERE date >= MAX(date)
281 GROUP BY date
282 -- INTERVAL 7 DAY AND date < MAX(date)
283 -- ORDER BY Running query...
284
285
286 -- MySQL roll-up table
287 CREATE TABLE review_daily_rollup
288 SELECT user_id, date, AVG(stars) AS avg_stars, COUNT(*) AS review_count
289 FROM review
```

Stop query

Business Tables

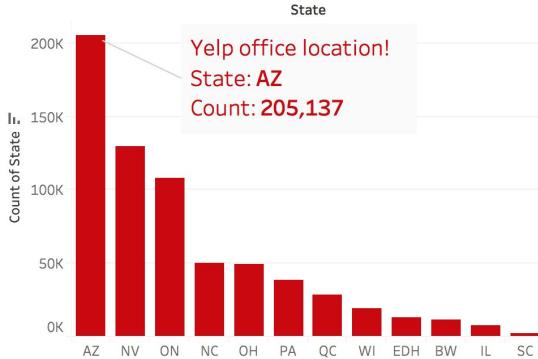


Dataset

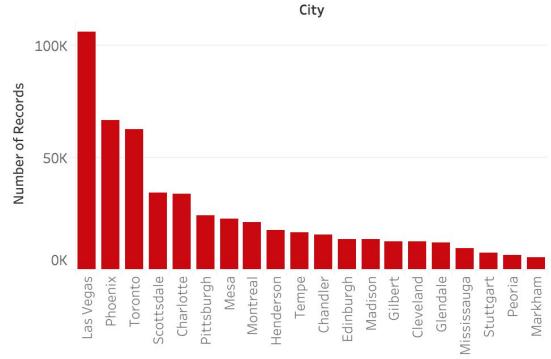
Business Dashboard

Drilling by location

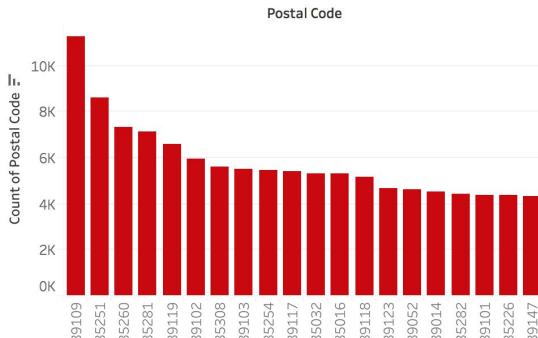
Most Businesses by State



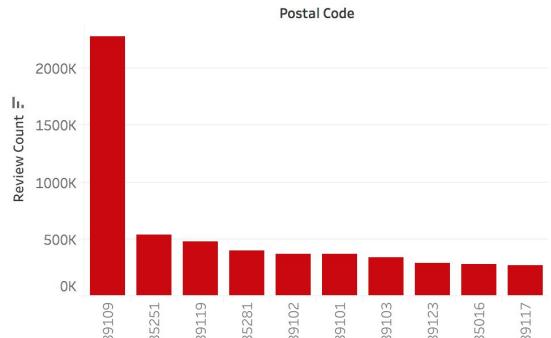
Most Businesses by City



Most Businesses by Postal Code



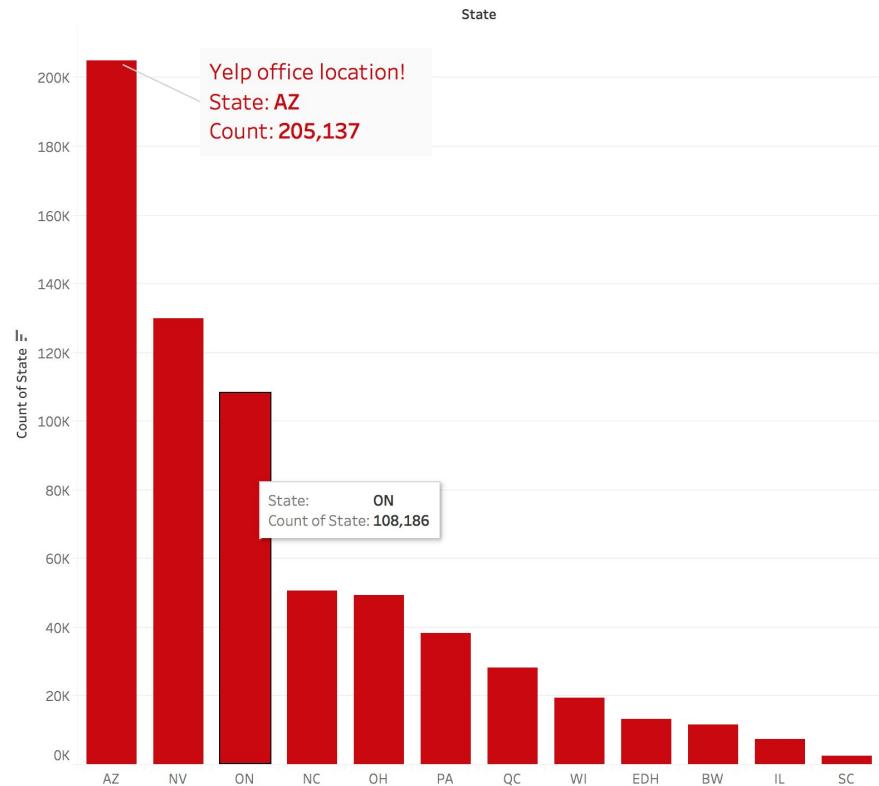
Most Reviews by Postal Code



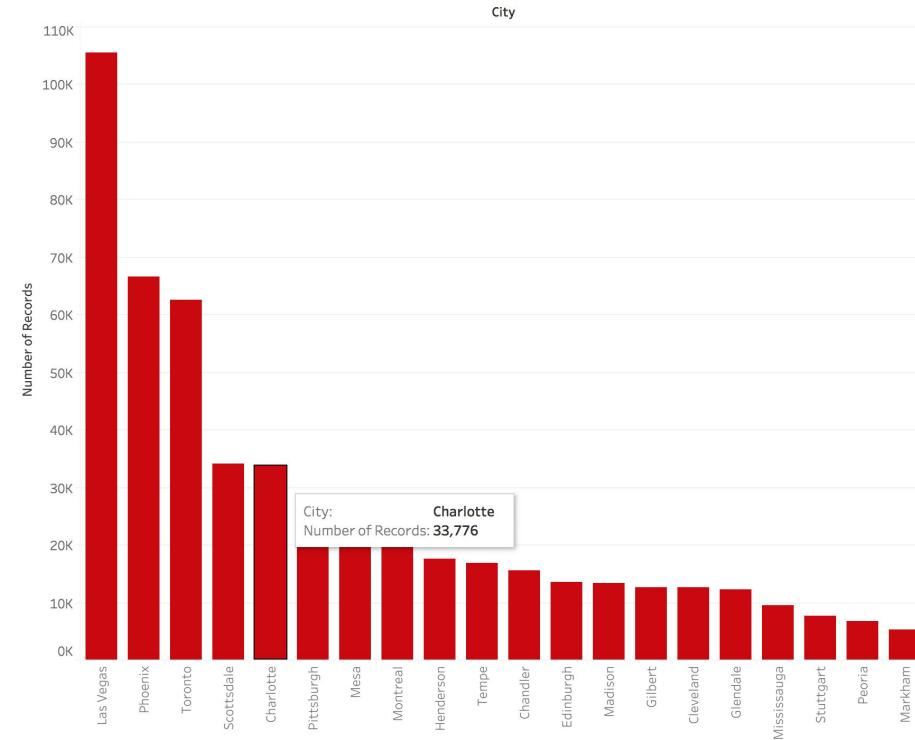


Dataset

Most Businesses by State



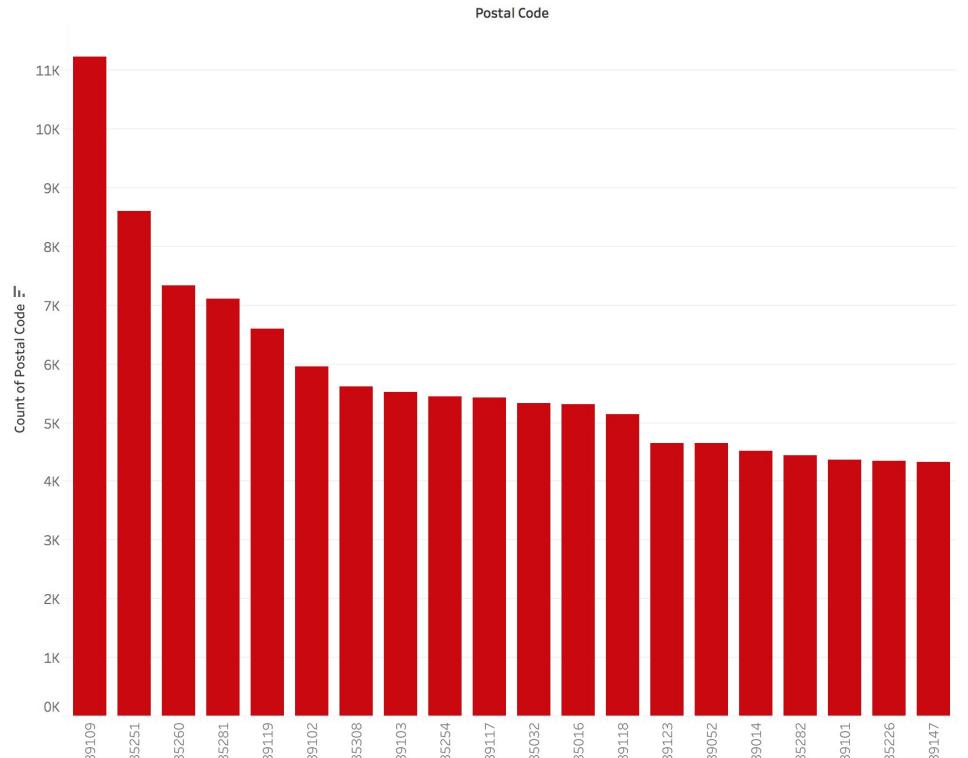
Most Businesses by City



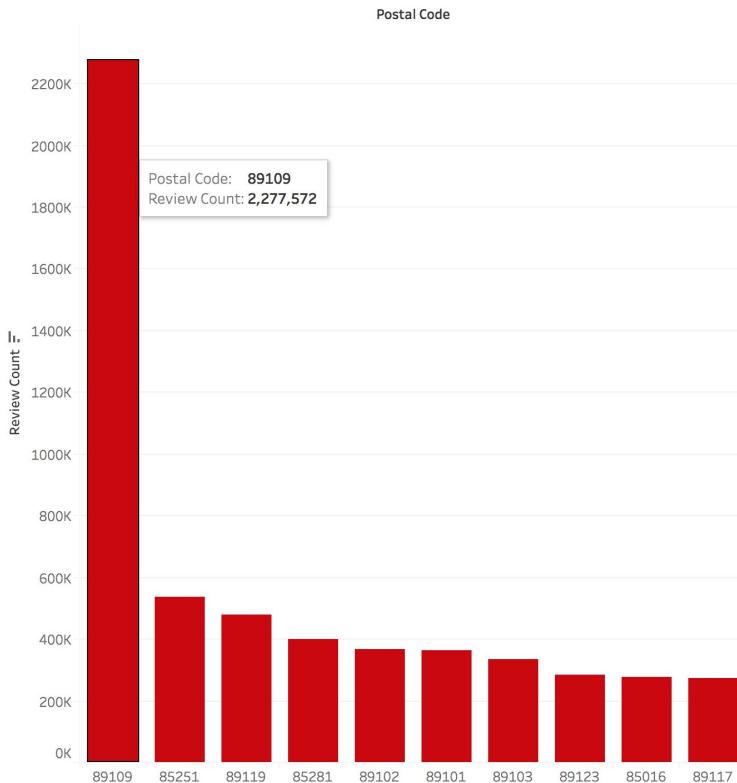


Dataset

Most Businesses by Postal Code



Most Reviews by Postal Code

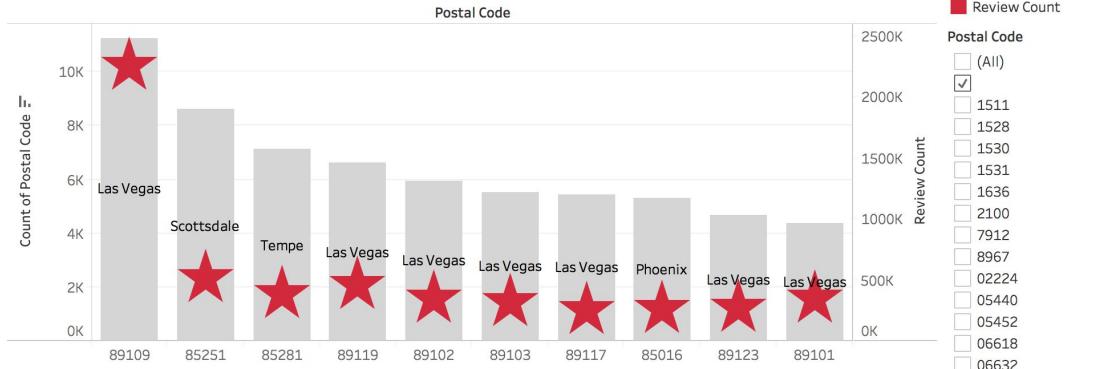




Dataset

Business dashboard

Top 10 Most Businesses by Postal Code - Review Count/City



Top 20 Most Businesses by Postal Code - Open/Closed

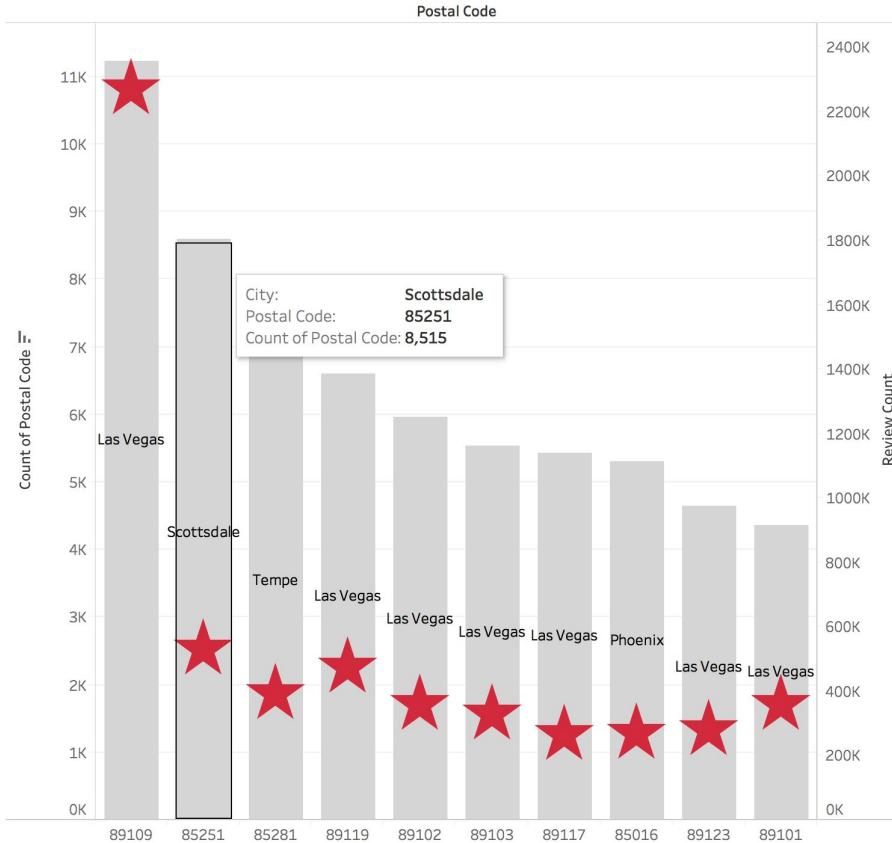




Dataset

Review count doesn't
exactly pace with
business presence

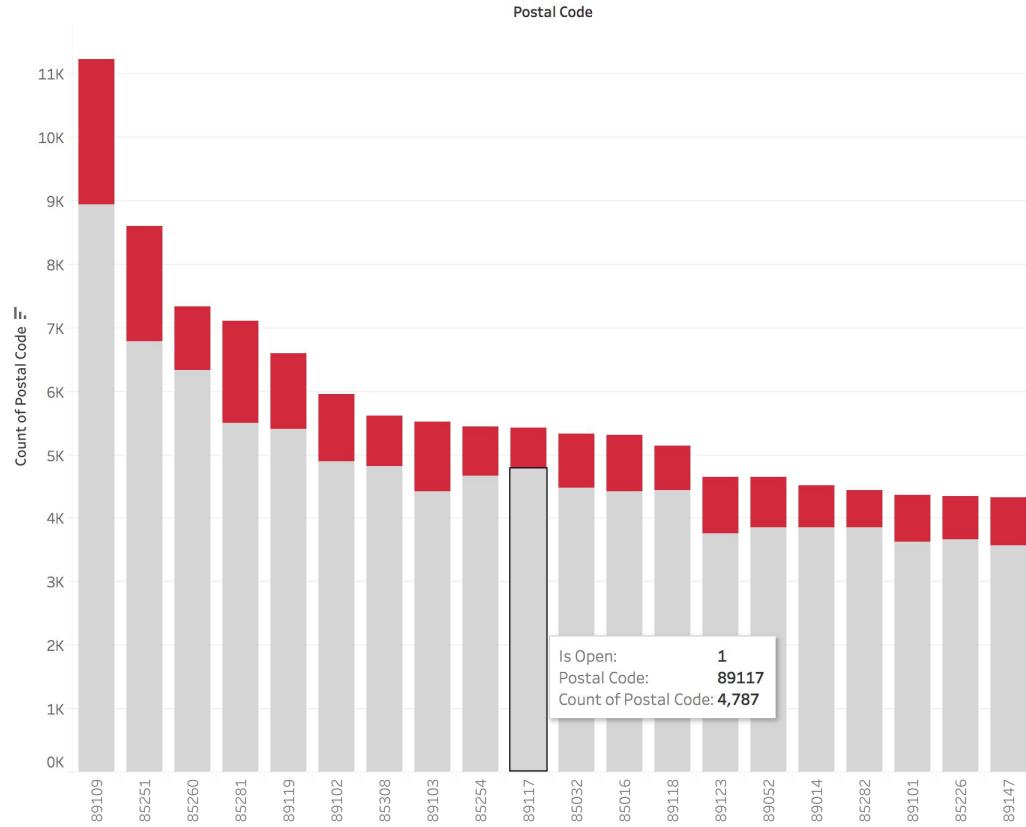
Top 10 Most Businesses by Postal Code - Review Count/City





Dataset

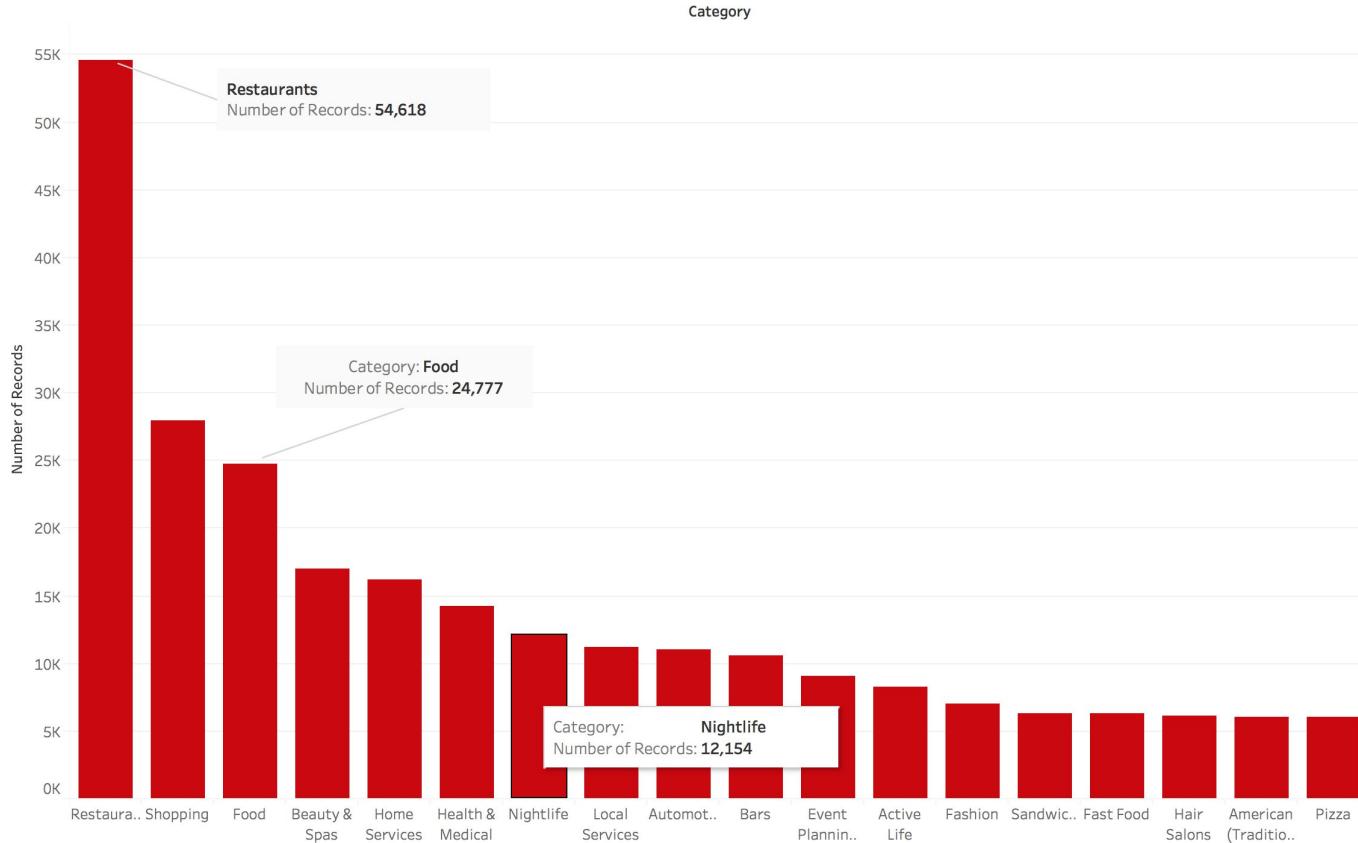
Top 20 Most Businesses by Postal Code - Open/Closed





Dataset

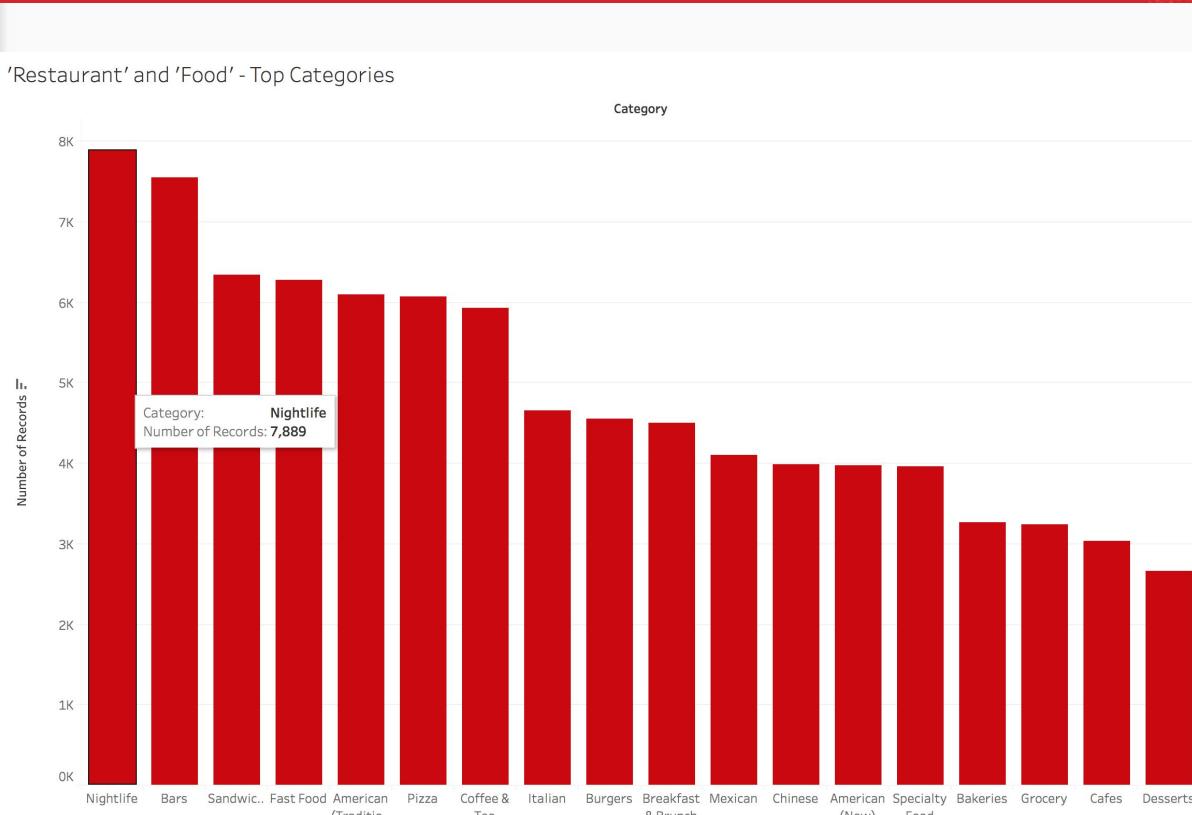
Category Frequency





Within Restaurants and Food, top categories are:

- Nightlife
- Bars
- Sandwiches
- Fast Food
- American (Traditional)



Review Tables

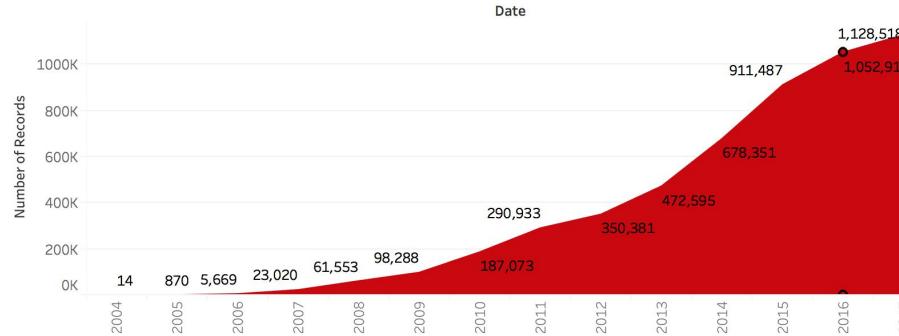


Dataset

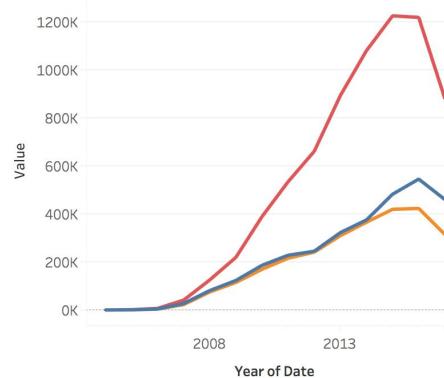
Overall Reviews Dashboard

All-Time Reviews and Compliments

Review Count over Time



Review Compliments



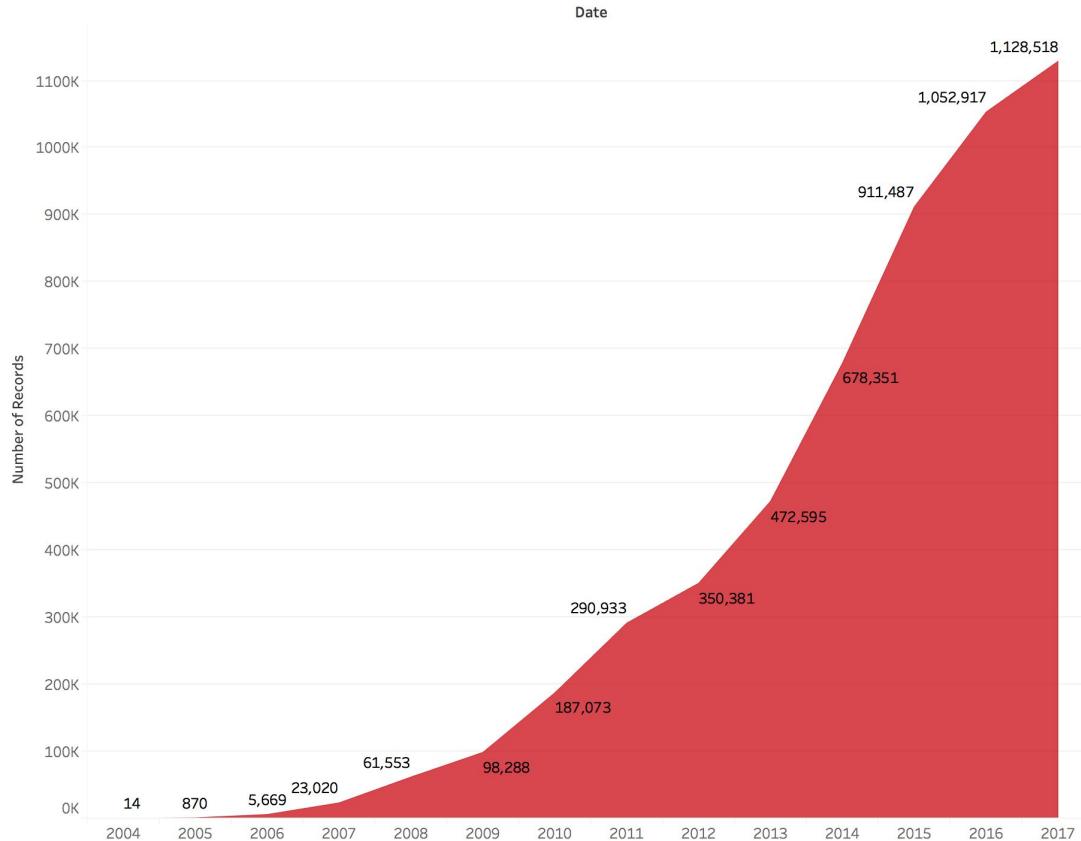
Review Compliments - Quarterly





Dataset

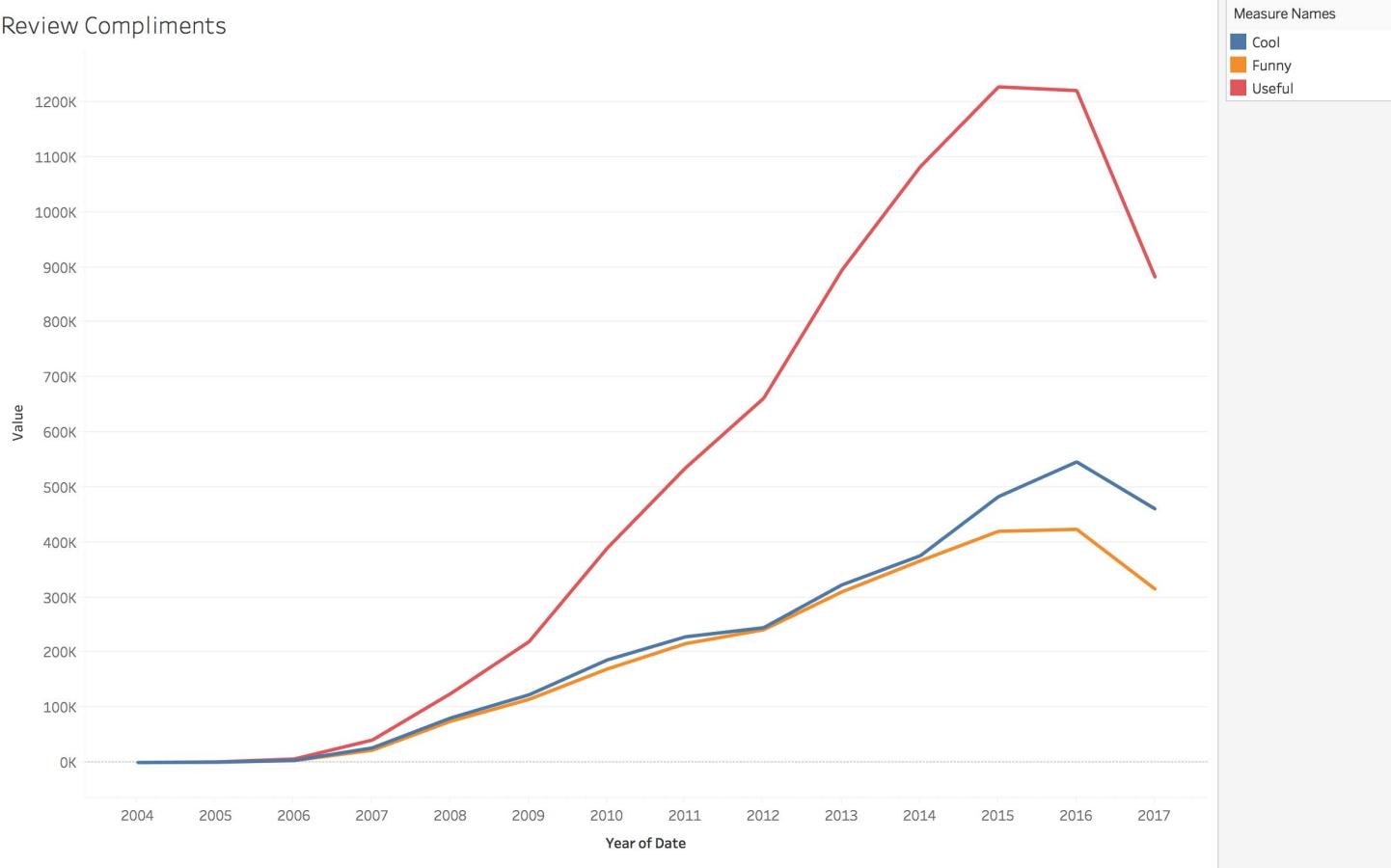
Review Count over Time





Dataset

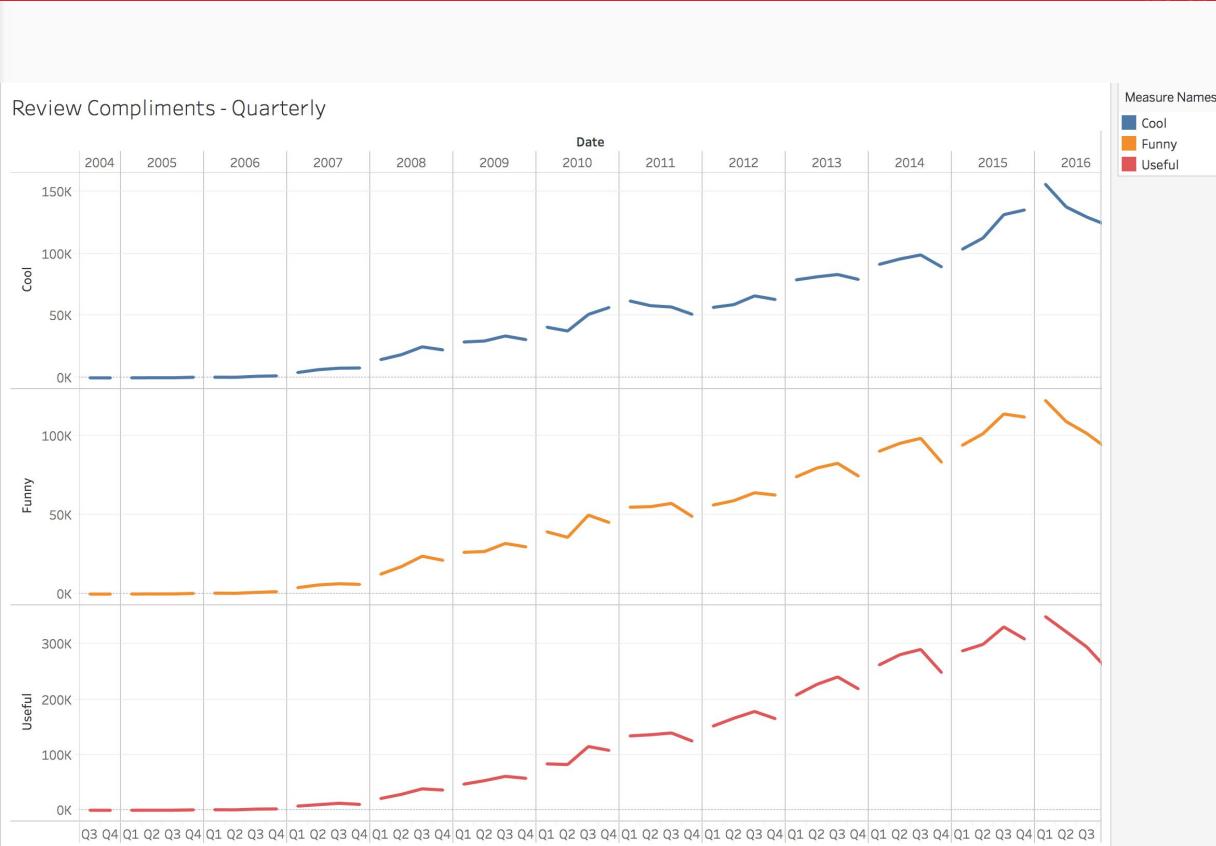
Review Compliments





Dataset

Mysterious dip in
compliments in 2016.
Is review count
affected?



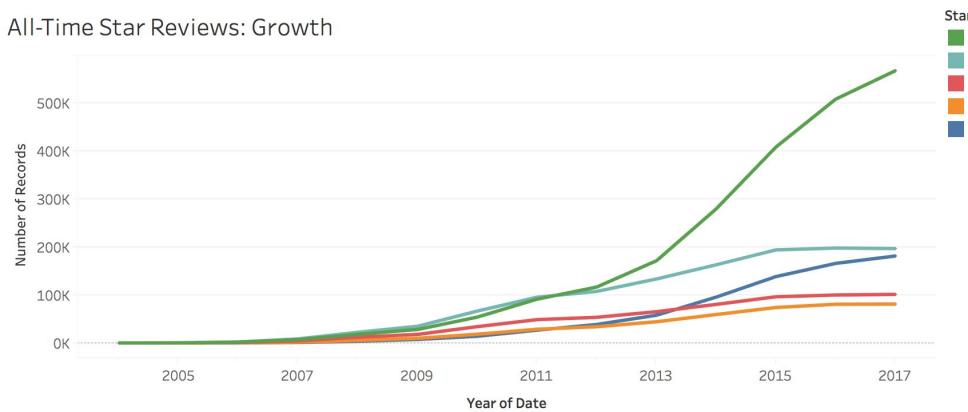


Dataset

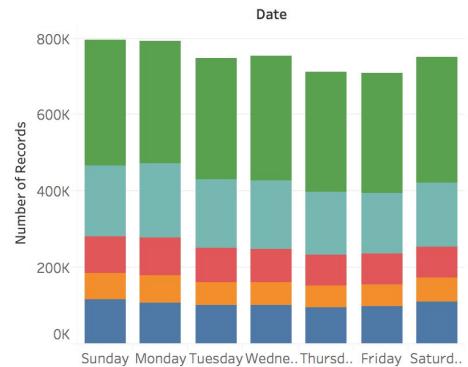
Stars Dashboard

Star Reviews: Weekly Drill-Down

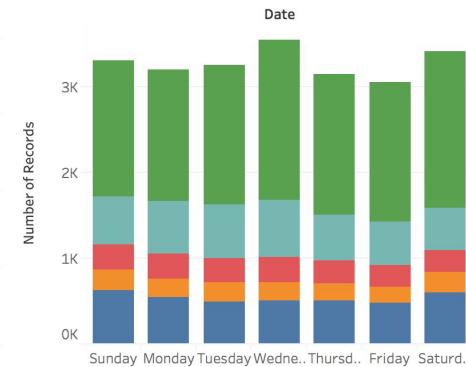
All-Time Star Reviews: Growth



All-Time Star Reviews: Day of Week



Weekly Star Reviews: Day of Week



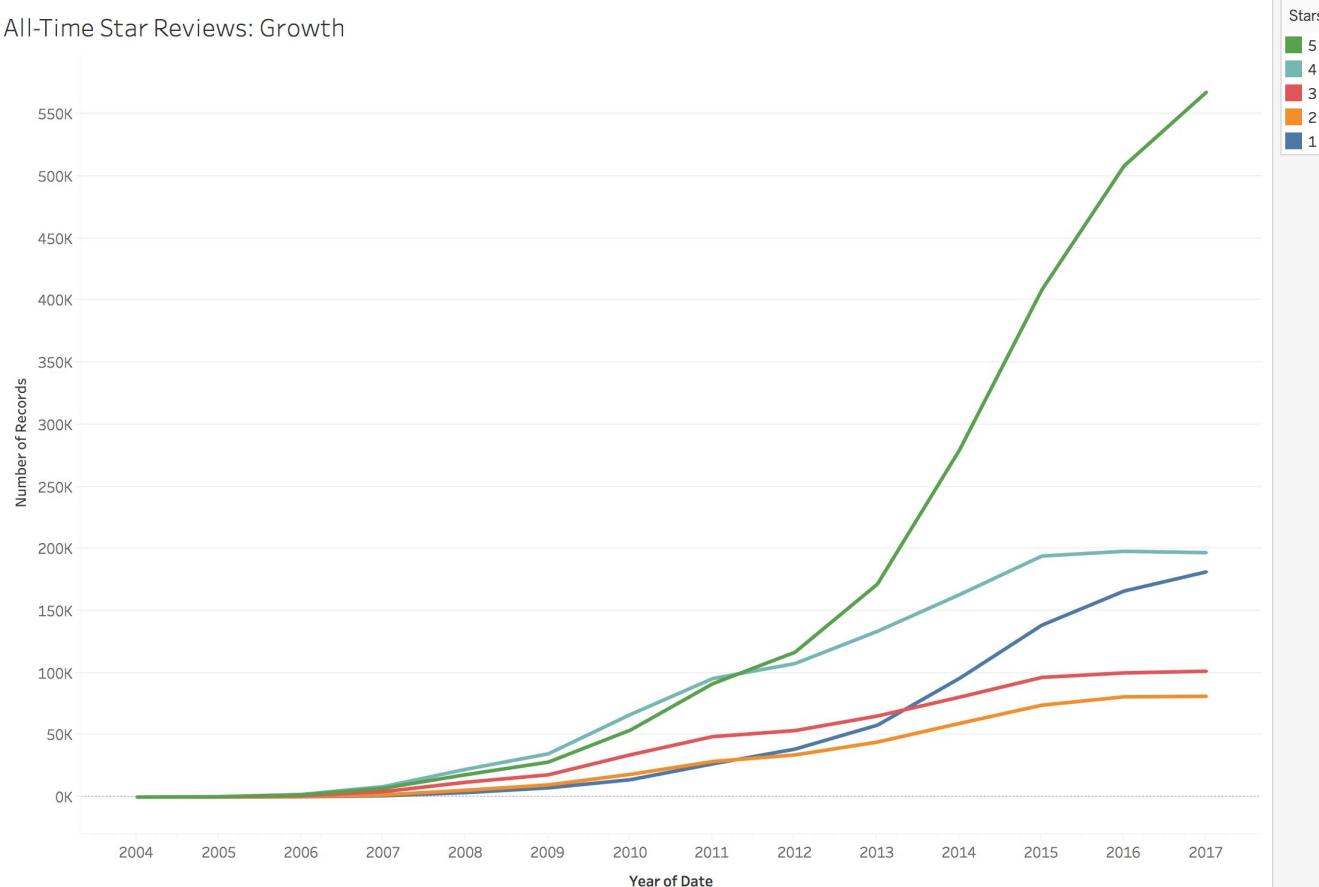


Dataset

Most reviews?

5 STARS!

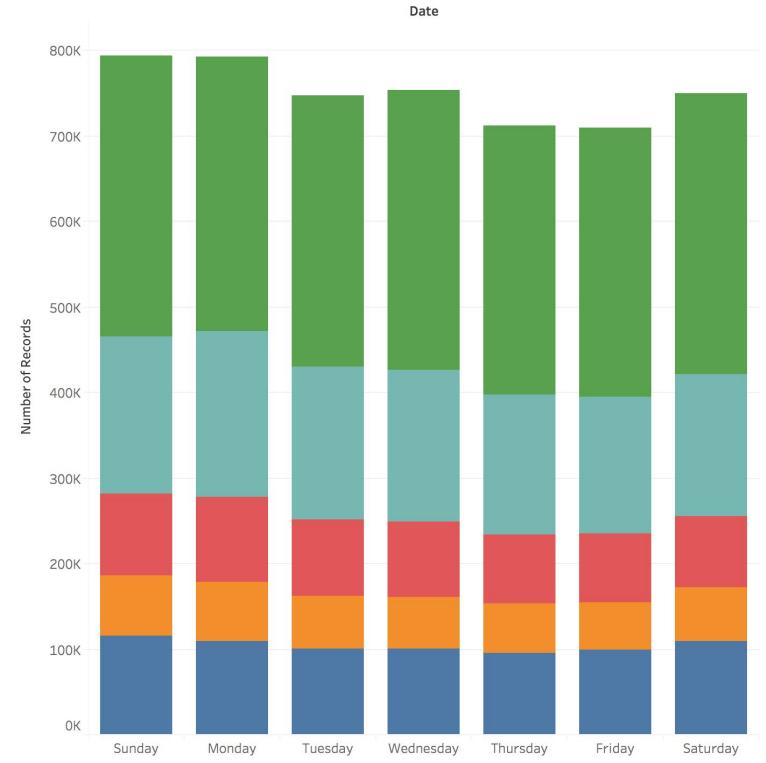
All-Time Star Reviews: Growth



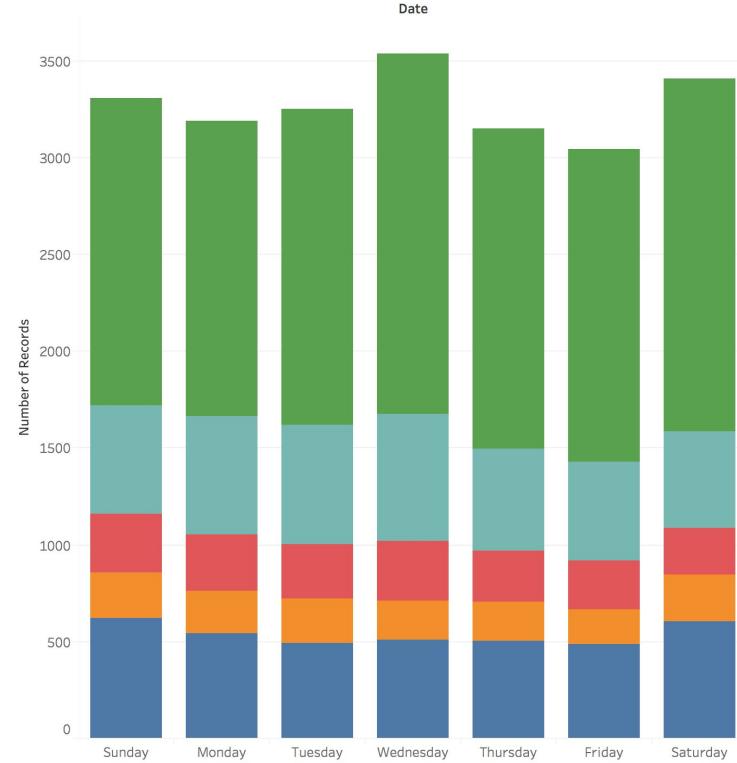


Dataset

All-Time Star Reviews: Day of Week



Weekly Star Reviews: Day of Week



Stars

- 5
- 4
- 3
- 2
- 1

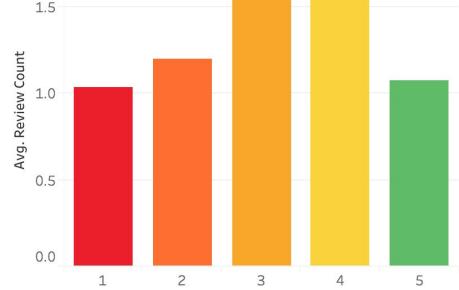


Dataset

Weekly Stars-Reviews Dashboard

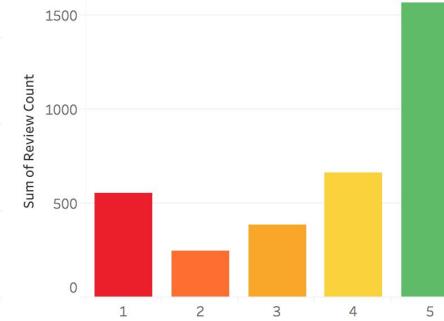
Daily Avg Review Count per Avg Stars

Avg Stars (bin)



Daily Review Count per Avg Stars

Avg Stars (bin)



Avg Stars (bin)

1

2

3

4

5

Daily Average Star Ratings by Review Count (1-6 Reviews)

Review Cou..

1

2

3

4

5

6

Daily Average Star Ratings by Review Count (2-4 Reviews)

Review Cou..

2

3

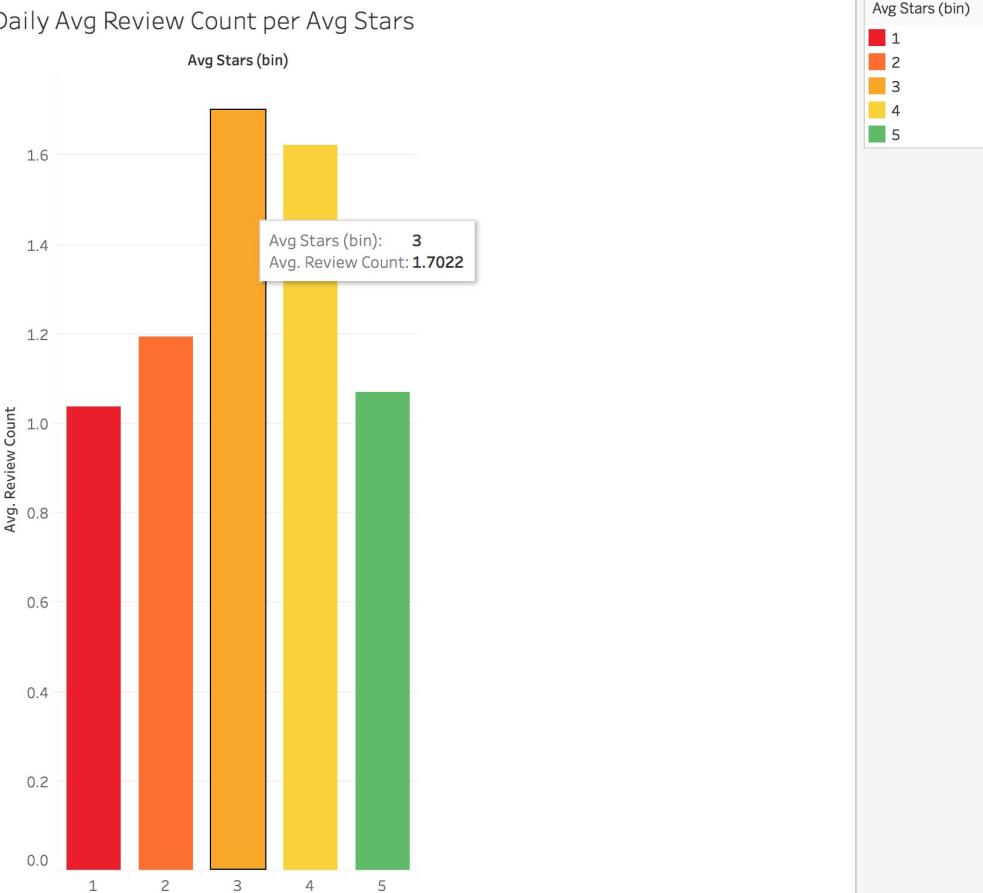
4



Dataset

Users who averaged 3 stars left the most reviews

Daily Avg Review Count per Avg Stars





Dataset

Most reviews?

5 STARS!

Daily Review Count per Avg Stars



User Tables



Dataset

<73 Friends Dashboard

<73 Friends: Outlier Detection and Reviews

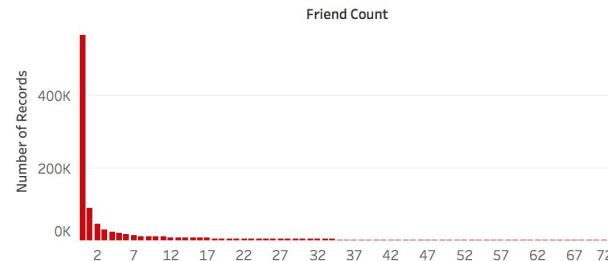
Friend Count Quartiles



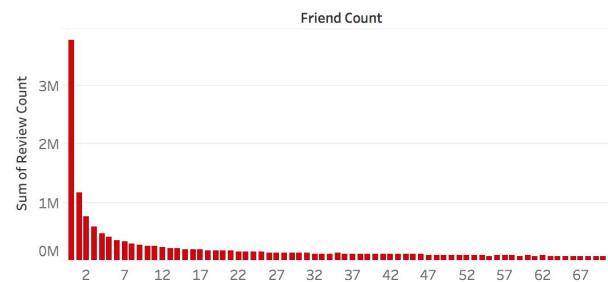
FC Quartiles Table

Number of Records	1,326,101
PERCENTILE([Friend Count],0.25)	0
PERCENTILE([Friend Count],0.50)	2
PERCENTILE([Friend Count],0.75)	29

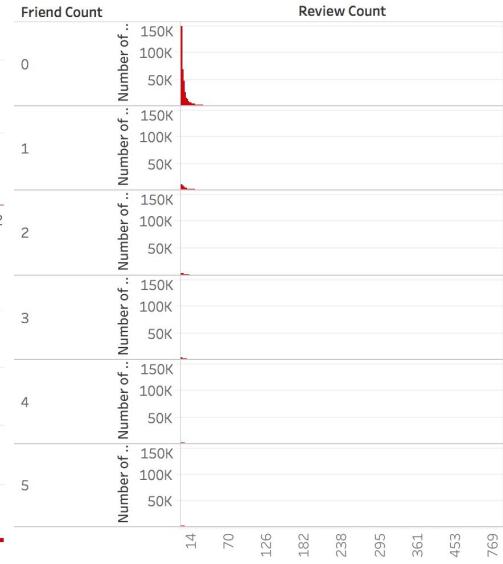
Friend Count Rows (<73 Friends)



Friend Count vs. Review Share (<73 Friends)

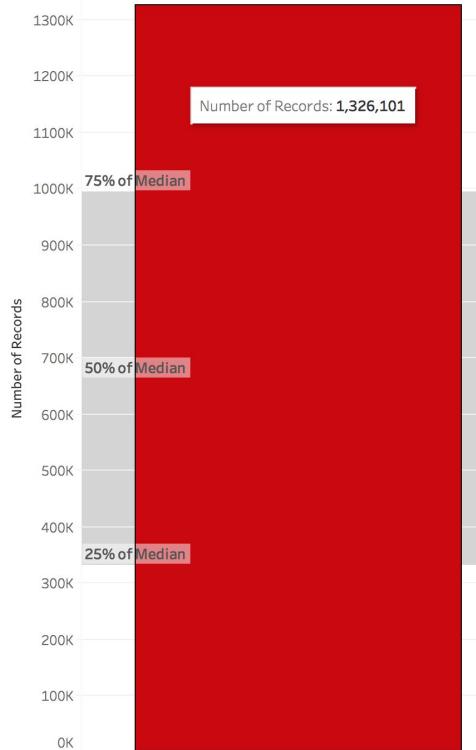


Review Counts per Friend Count (0-5 Friends)





Friend Count Quartiles



Friend Count Quartiles Table

Number of Records	1,326,101
PERCENTILE([Friend Count],0.25)	0
PERCENTILE([Friend Count],0.50)	2
PERCENTILE([Friend Count],0.75)	29

Interquartile Range = $29 - 0 = 29$

Outlier fences = $1.5 \times IQR$

→ Greater than 72 friends



Dataset

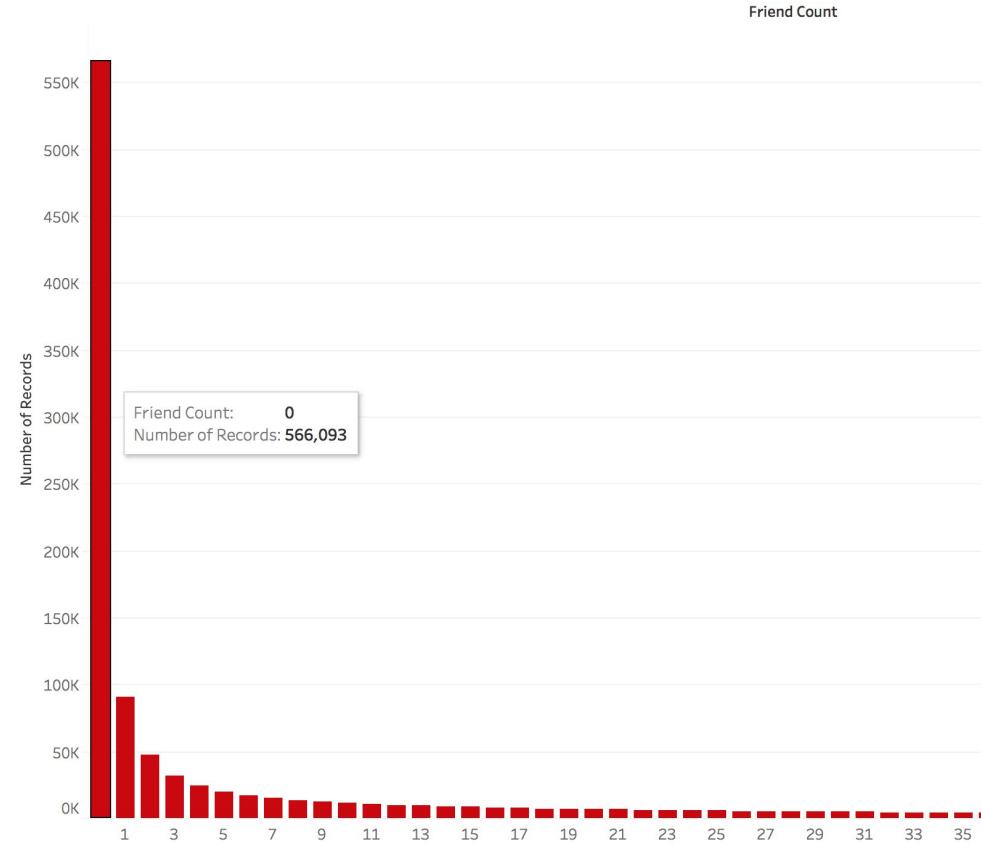
Understanding user behavior

- Most users have zero friends
- Median friend count is 2
- Most reviews come from zero-friend users
- Outliers with a shocking number of friends (>72) are responsible for some 15-20% of all reviews
 - Yelp as a social media platform?
 - Spammers?
 - Bots?

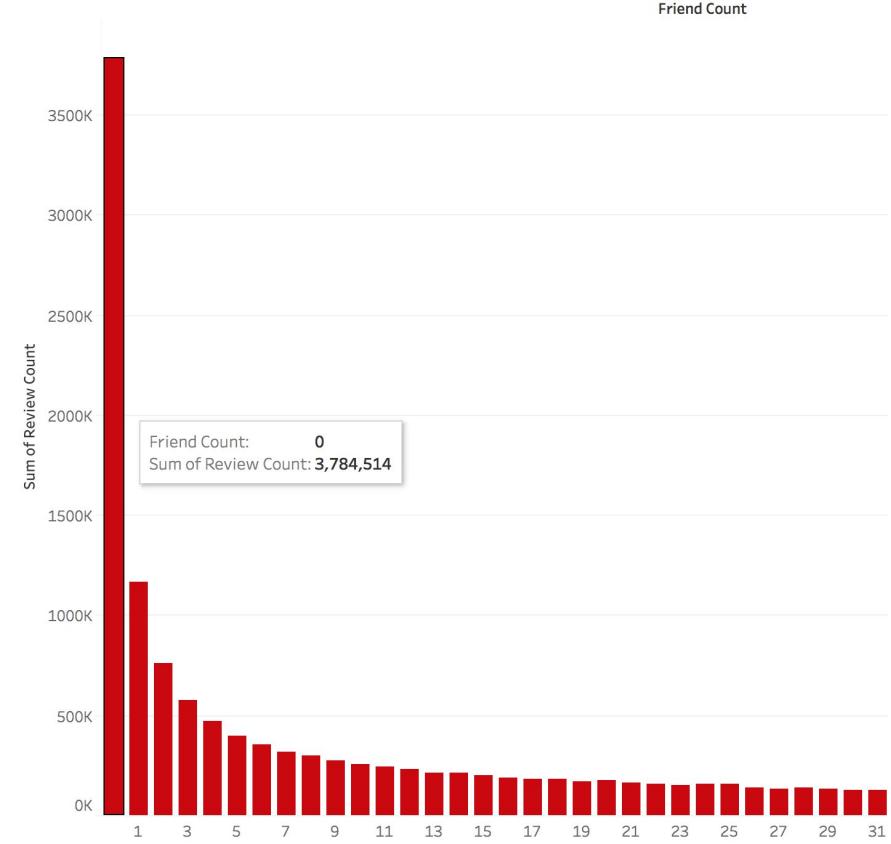


Dataset

Friend Count Rows (<73 Friends)



Friend Count vs. Review Share (<73 Friends)



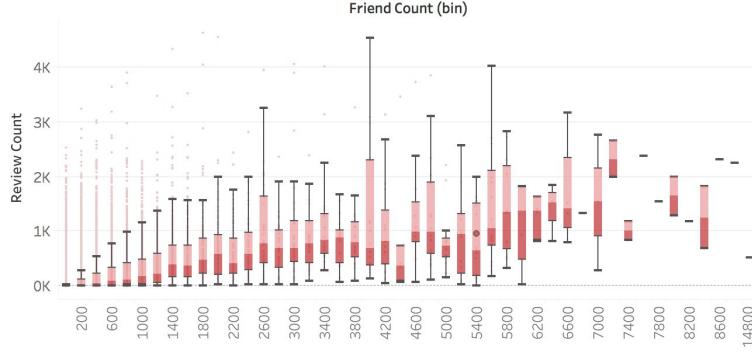


Dataset

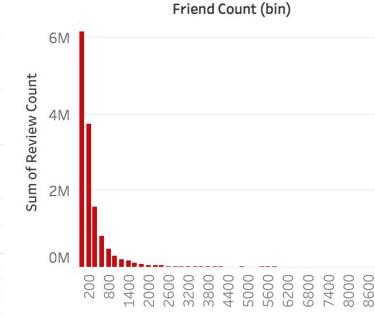
>72 Friends (Outliers) Dashboard

> 72 Friends: Networkers, Bad Actors, Bots?

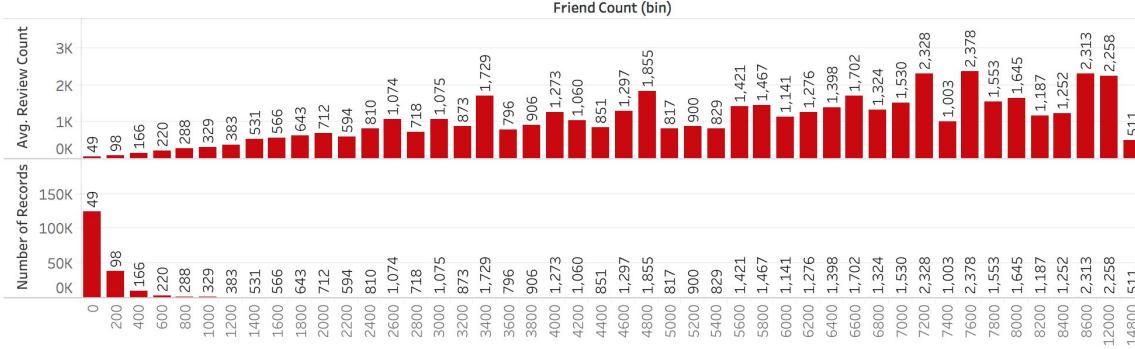
Binned Boxplots: <5k Reviews



Power Users: Friend Count vs.
Review Share (> 72)



Power Users: Friend Count vs. Reviews (> 72)

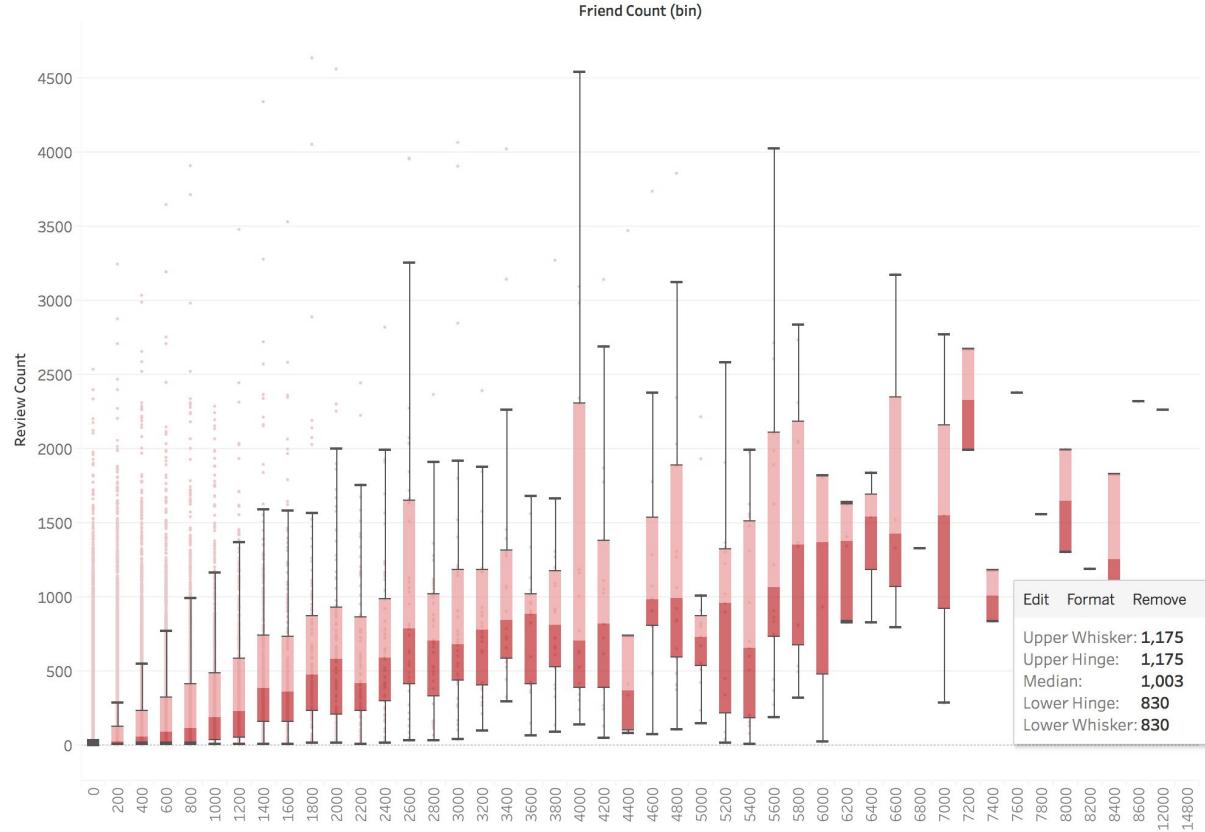




Dataset

Big buckets: 200 users

Binned Boxplots: <5k Reviews

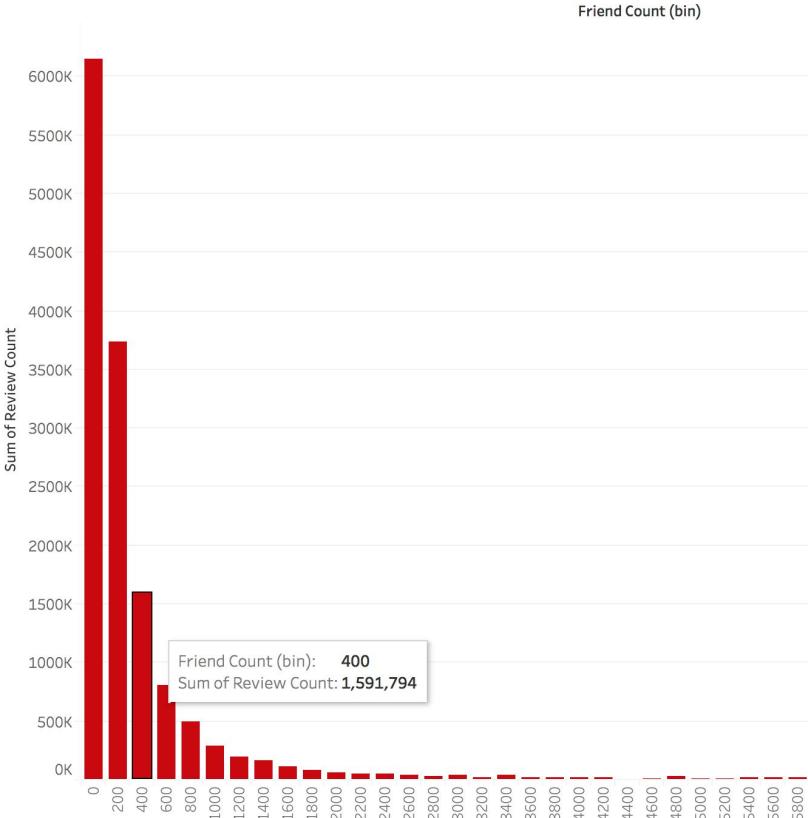




Dataset

Low-friend buckets produce the majority of reviews, but outlier high-friend users account for a significant portion of reviews

Power Users: Friend Count vs. Review Share (> 72)

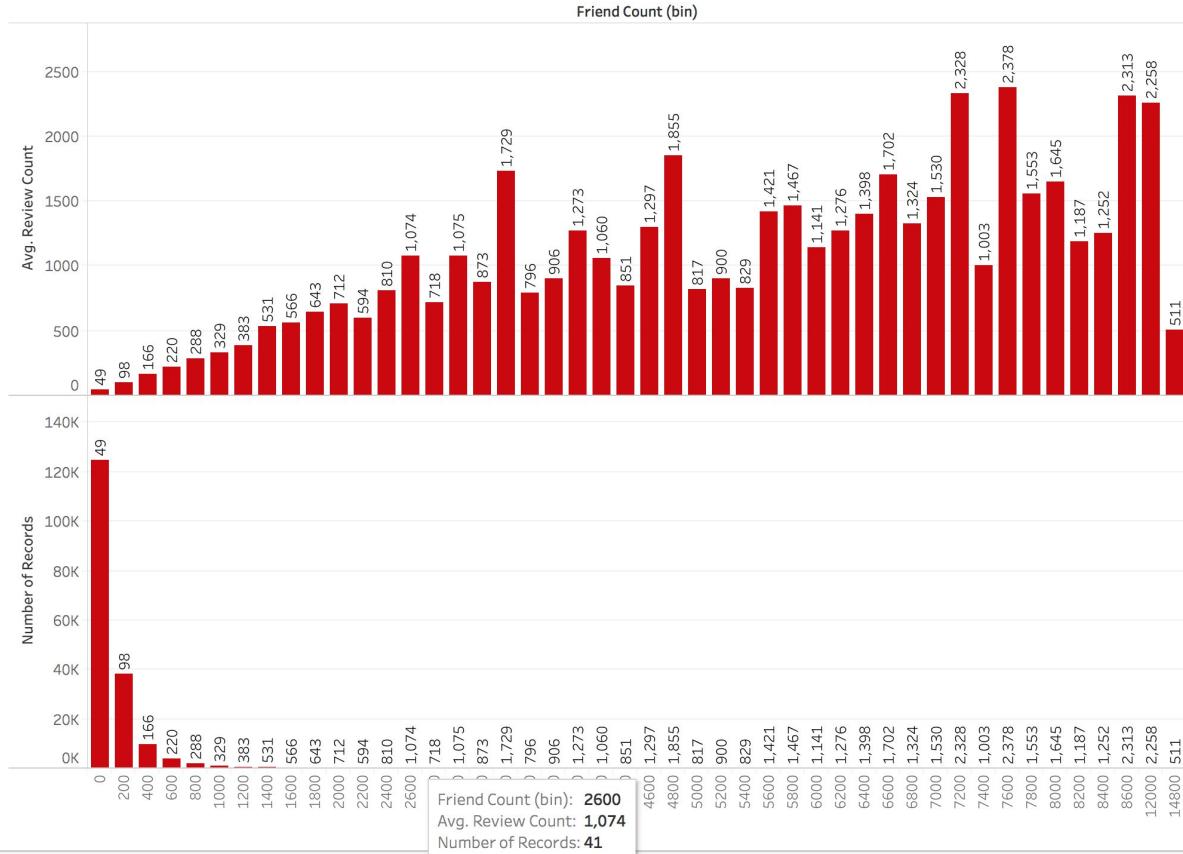




Dataset

A prolific few

Power Users: Friend Count vs. Reviews (> 72 Friends)





Extra time, extra Yelp

- Even better pre-agged tables; more interesting pivots
- Move to Postgres
- Chi-Square test of independence on average stars by weekday (TabPy)
- Automate top city dashboards (users, review activity, businesses)
- Business category vs. review count
- Descriptive → predictive analytics

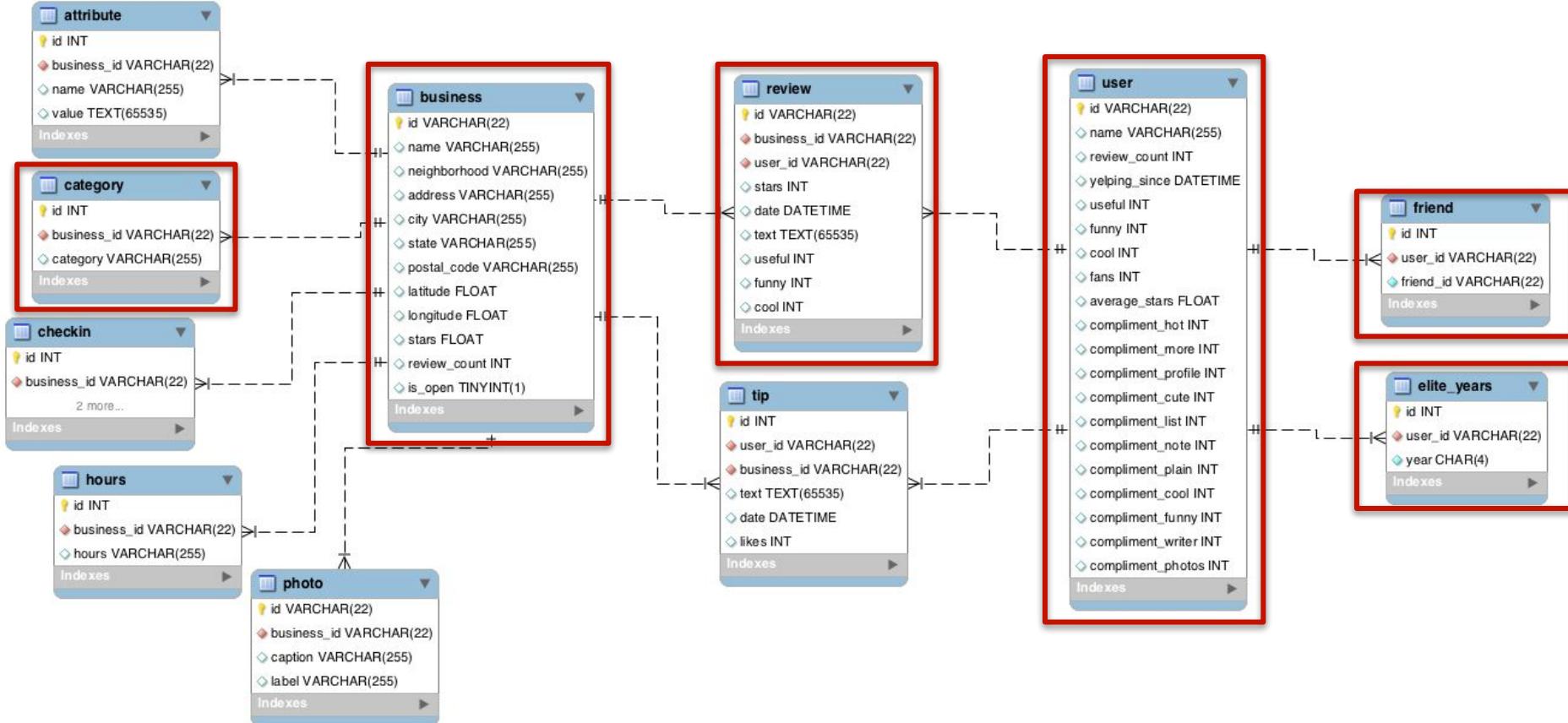


Extra time, extra Yelp

- Map and analyze lat/long of users' reviewed businesses
- Where are trendsetters going before it's popular?
 - What is a 'trendsetter'? What is 'popular'?
 - Review/check-in throughput
- Reviews: NLP, text analysis, prediction modeling
 - Review stars
 - Compliments
 - Dubious accounts
- Image analysis



Dataset



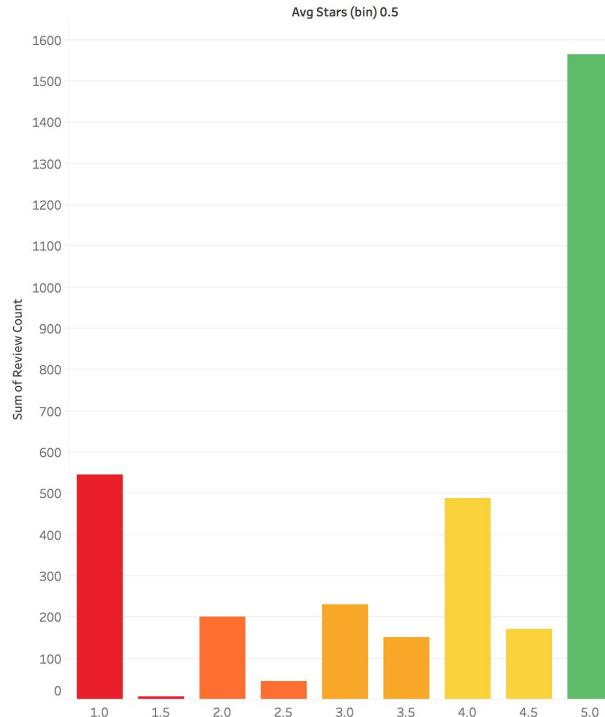
Some very bad graphs



Dataset

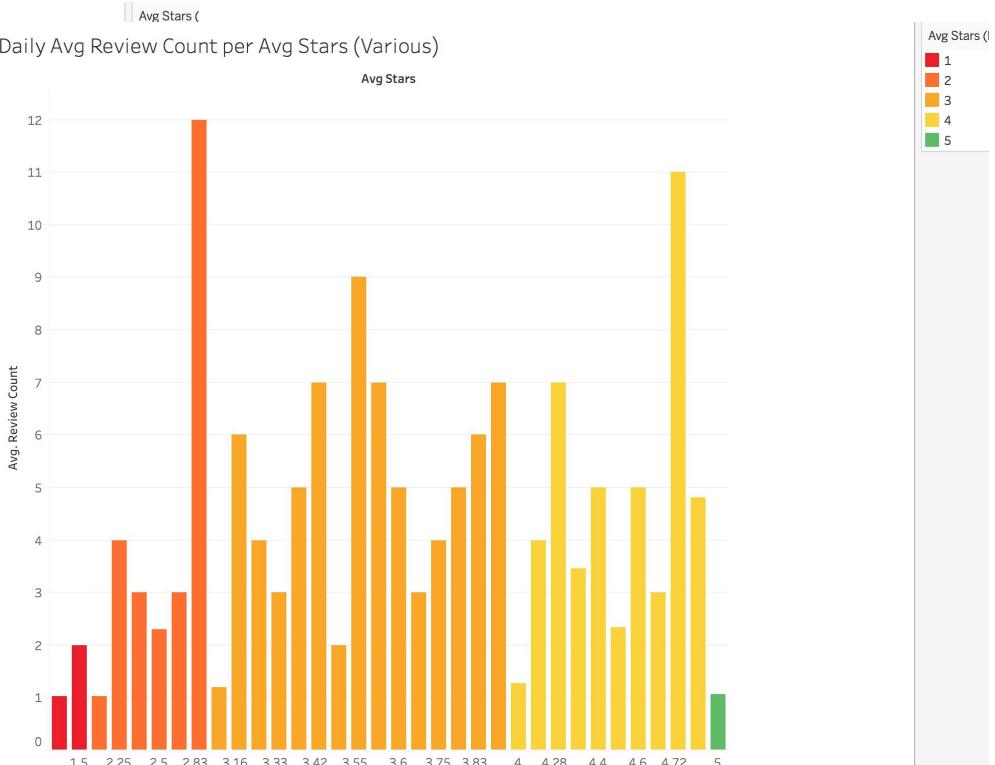
Misleading bins

Daily Reviews per Avg Stars (0.5 Bins)



Too many bins

Daily Avg Review Count per Avg Stars (Various)

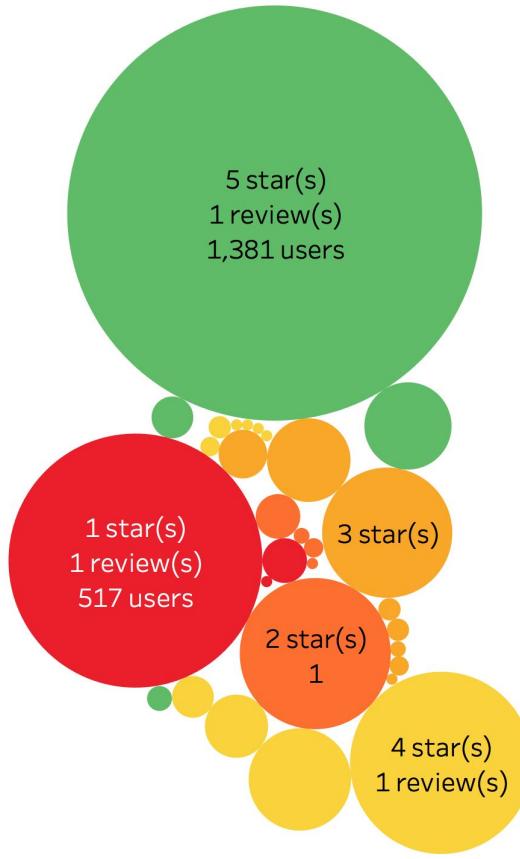




Dataset

Not helpful

Daily Review Counts Per Avg Star Bin



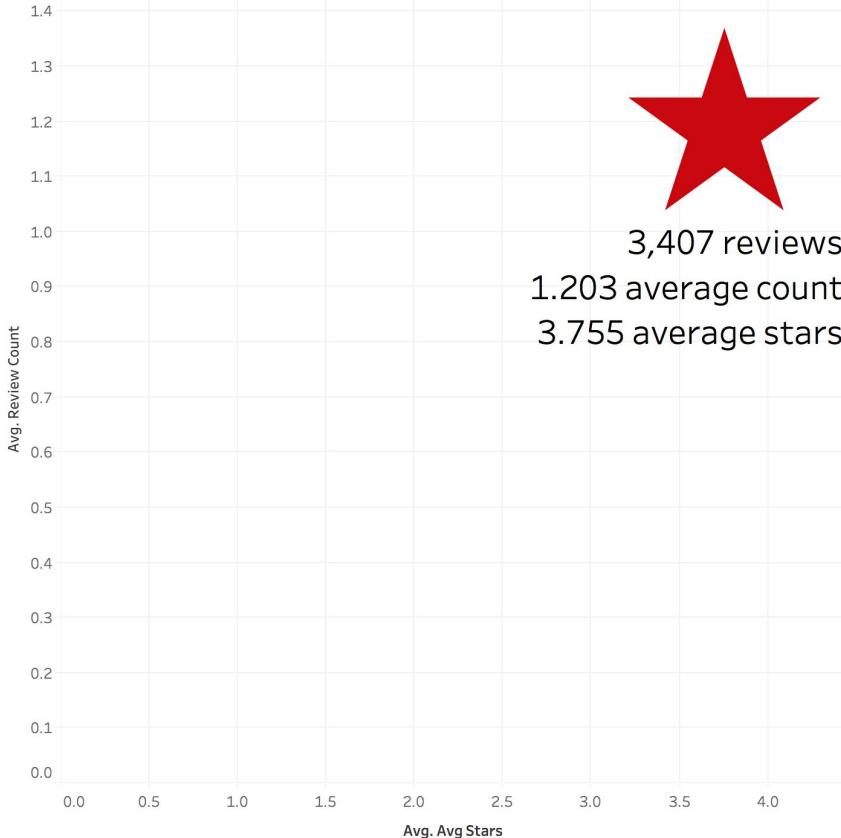
Avg Stars
1
2
3
4
5



Dataset

The self-congratulatory one-dot scatterplot

Weekly Reviews - Top-Level Metrics

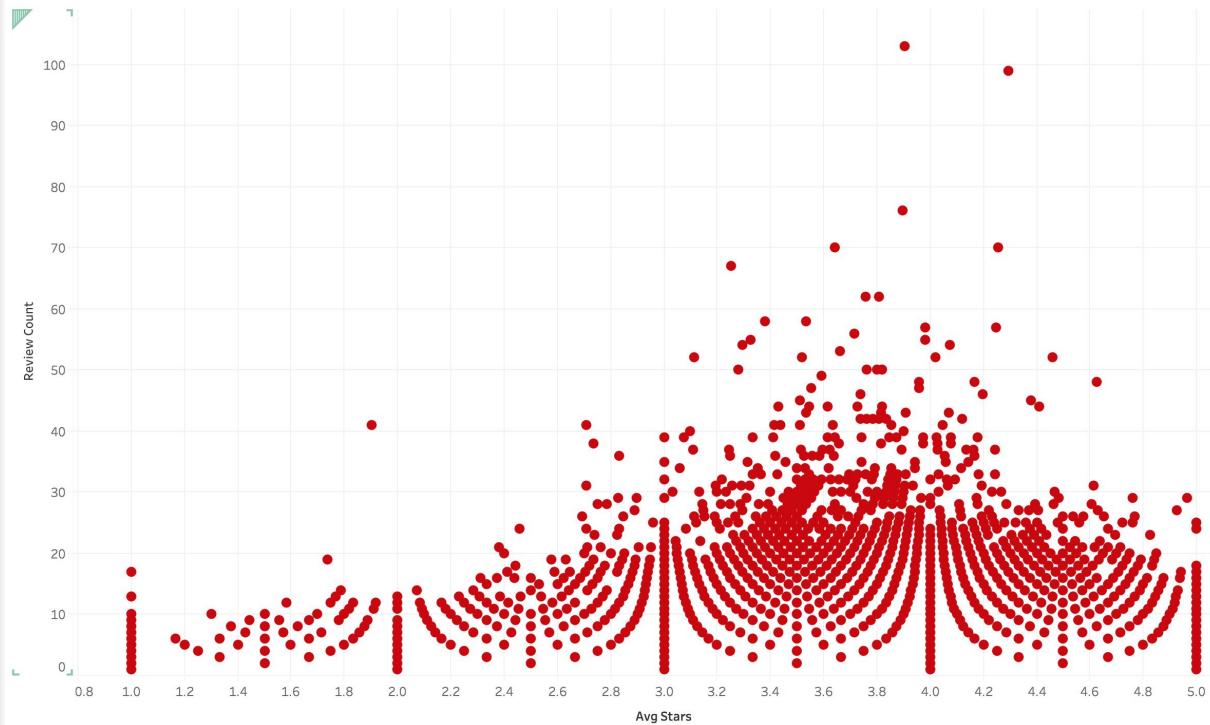




Dataset

Rorschach?

Sheet 2





Questions?



Woohoo! As good as it gets!

Share on



Post Review

Want to work with me?



[linkedin.com/in/caelanosullivan](https://www.linkedin.com/in/caelanosullivan)



github.com/caelanosullivan

