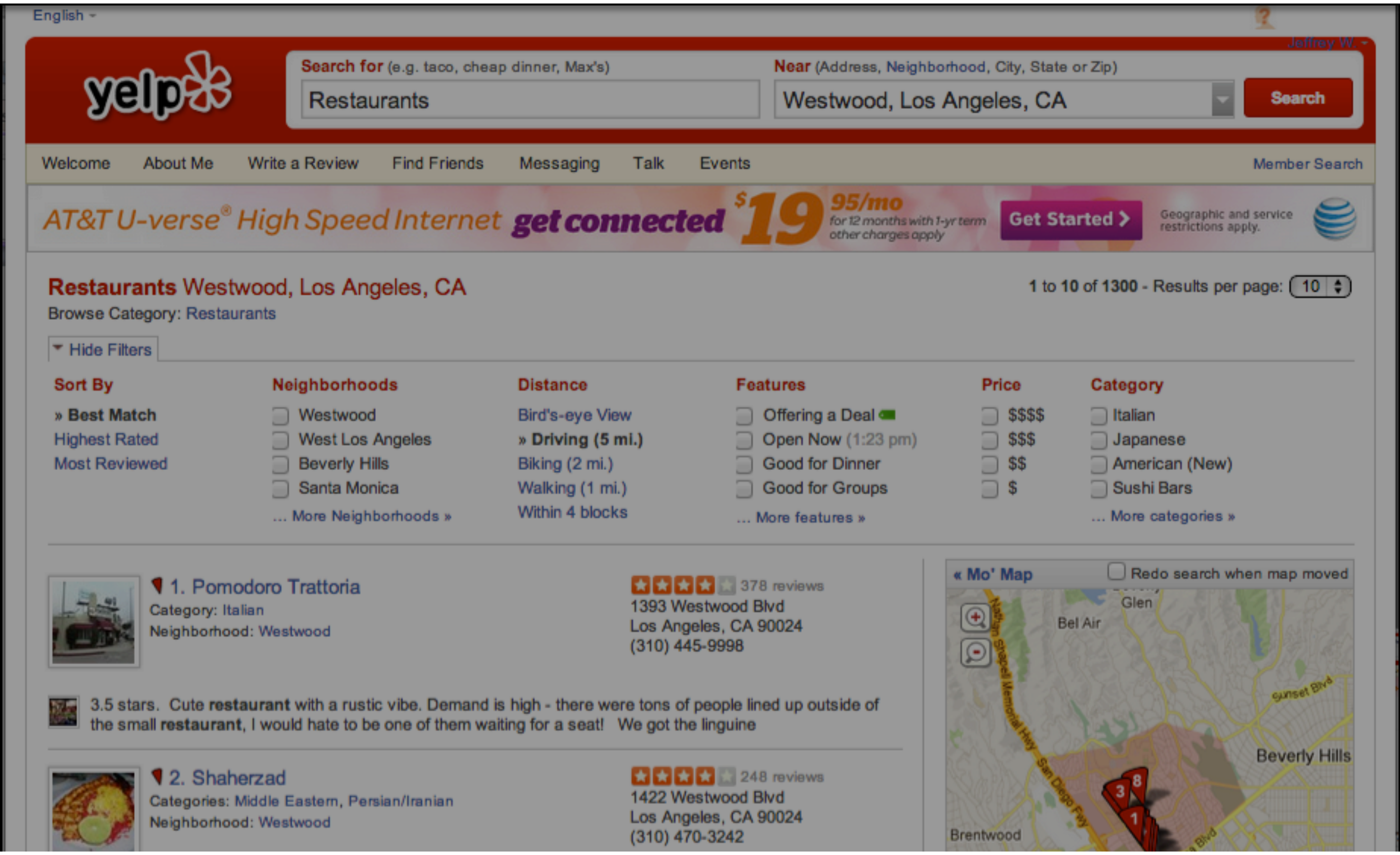
Predicting Restaurant Review Counts near Universities using the yelp Academic Dataset

by Jeffrey Wolf, Jonathan Arfa, Yang Lu



English ▾

Jeffrey W. ▾

Search for (e.g. taco, cheap dinner, Max's): Restaurants

Near (Address, Neighborhood, City, State or Zip): Westwood, Los Angeles, CA

Restaurants Westwood, Los Angeles, CA

Browse Category: Restaurants

1 to 10 of 1300 - Results per page: 10

▾ Hide Filters

Sort By
- » **Best Match**
- Highest Rated
- Most Reviewed

Neighborhoods
- □ Westwood
- □ West Los Angeles
- □ Beverly Hills
- □ Santa Monica
- ... More Neighborhoods »

Distance
- Bird's-eye View
- » **Driving (5 mi.)**
- Biking (2 mi.)
- Walking (1 mi.)
- Within 4 blocks

Features
- □ Offering a Deal
- □ Open Now (1:23 pm)
- □ Good for Dinner
- □ Good for Groups
- ... More features »

Price
- □ $$$$
- □ $$$
- □ $$
- □ $

Category
- □ Italian
- □ Japanese
- □ American (New)
- □ Sushi Bars
- ... More categories »

1. Pomodoro Trattoria
Category: Italian
Neighborhood: Westwood
378 reviews
1393 Westwood Blvd
Los Angeles, CA 90024
(310) 445-9998

3.5 stars. Cute **restaurant** with a rustic vibe. Demand is high - there were tons of people lined up outside of the small **restaurant**, I would hate to be one of them waiting for a seat! We got the linguine

2. Shaherzad
Categories: Middle Eastern, Persian/Iranian
Neighborhood: Westwood
248 reviews
1422 Westwood Blvd
Los Angeles, CA 90024
(310) 470-3242

« Mo' Map

□ Redo search when map moved

Data - Yelp's Academic Dataset

•Data for the businesses closest to each of 31 universities in the US & Canada, including UCLA. We only studied those classified as open restaurants (n=4597.)

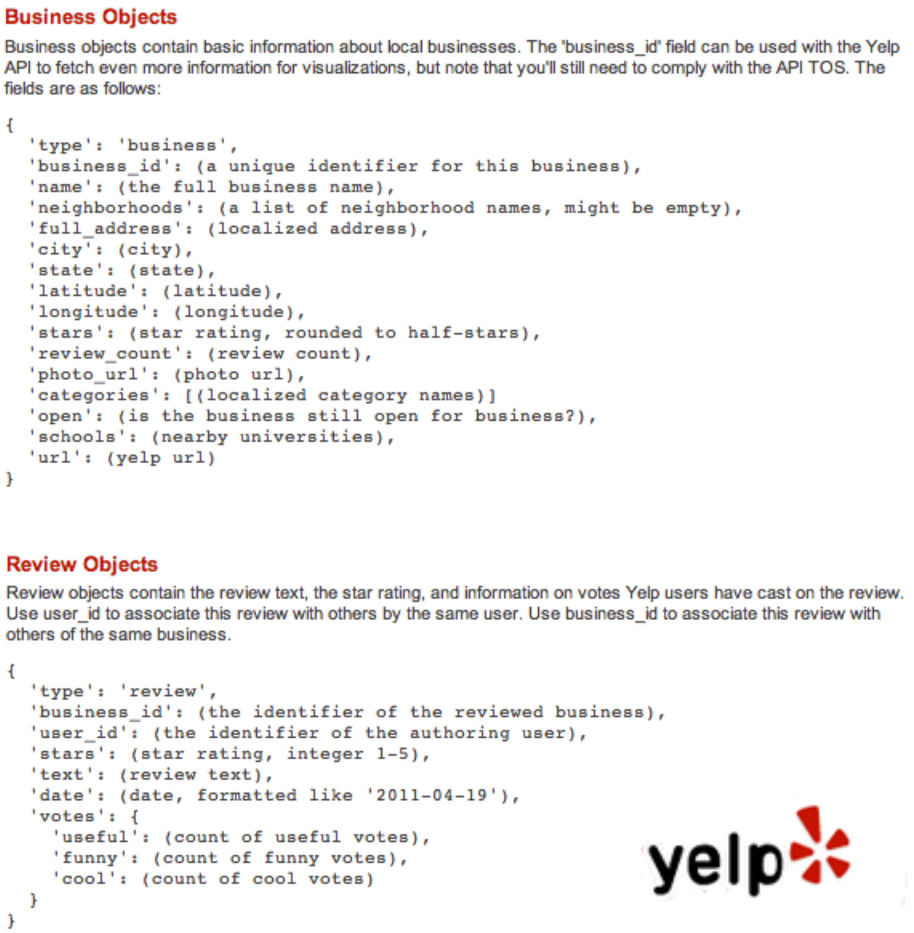
•Separate data for businesses, yelp users, and individual reviews.

•Stored as a collection of JSON objects, which we parsed with a Python program.

•We extracted review count, latitude, longitude, stars, business id, nearest university, and date of first review from the JSON objects.

•Our Python program also computed distance to campus and days since first review for each restaurant.

• In R, we calculated restaurant density using 2D kernel density estimation.



Business Objects

Business objects contain basic information about local businesses. The 'business_id' field can be used with the Yelp API to fetch even more information for visualizations, but note that you'll still need to comply with the API TOS. The fields are as follows:

```
{
  'type': 'business',
  'business_id': (a unique identifier for this business),
  'name': (the full business name),
  'neighborhoods': (a list of neighborhood names, might be empty),
  'full_address': (localized address),
  'city': (city),
  'state': (state),
  'latitude': (latitude),
  'longitude': (longitude),
  'stars': (star rating, rounded to half-stars),
  'review_count': (review count),
  'photo_url': (photo url),
  'categories': [(localized category names)]
  'open': (is the business still open for business?),
  'schools': (nearby universities),
  'url': (yelp url)
}
```

Review Objects

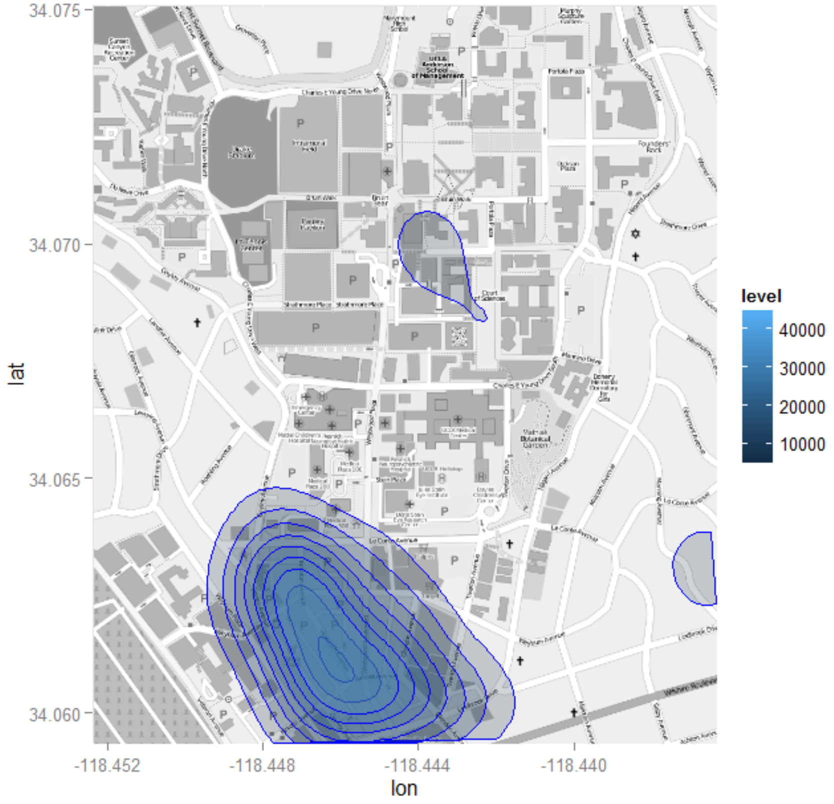
Review objects contain the review text, the star rating, and information on votes Yelp users have cast on the review. Use user_id to associate this review with others by the same user. Use business_id to associate this review with others of the same business.

```
{
  'type': 'review',
  'business_id': (the identifier of the reviewed business),
  'user_id': (the identifier of the authoring user),
  'stars': (star rating, integer 1-5),
  'text': (review text),
  'date': (date, formatted like '2011-04-19'),
  'votes': {
    'useful': (count of useful votes),
    'funny': (count of funny votes),
    'cool': (count of cool votes)
  }
}
```

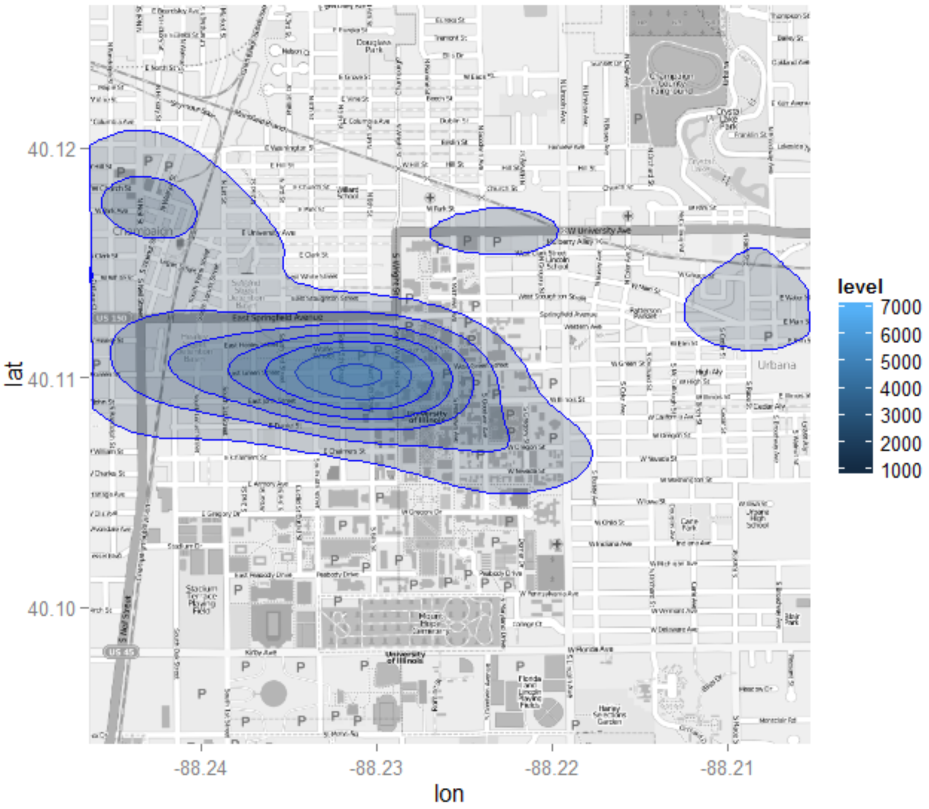
yelp

Restaurant Density

UCLA



U. Illinois –Urbana-Champaign



Can we predict variation in review count?

We expected the review count to be related to the following variables:

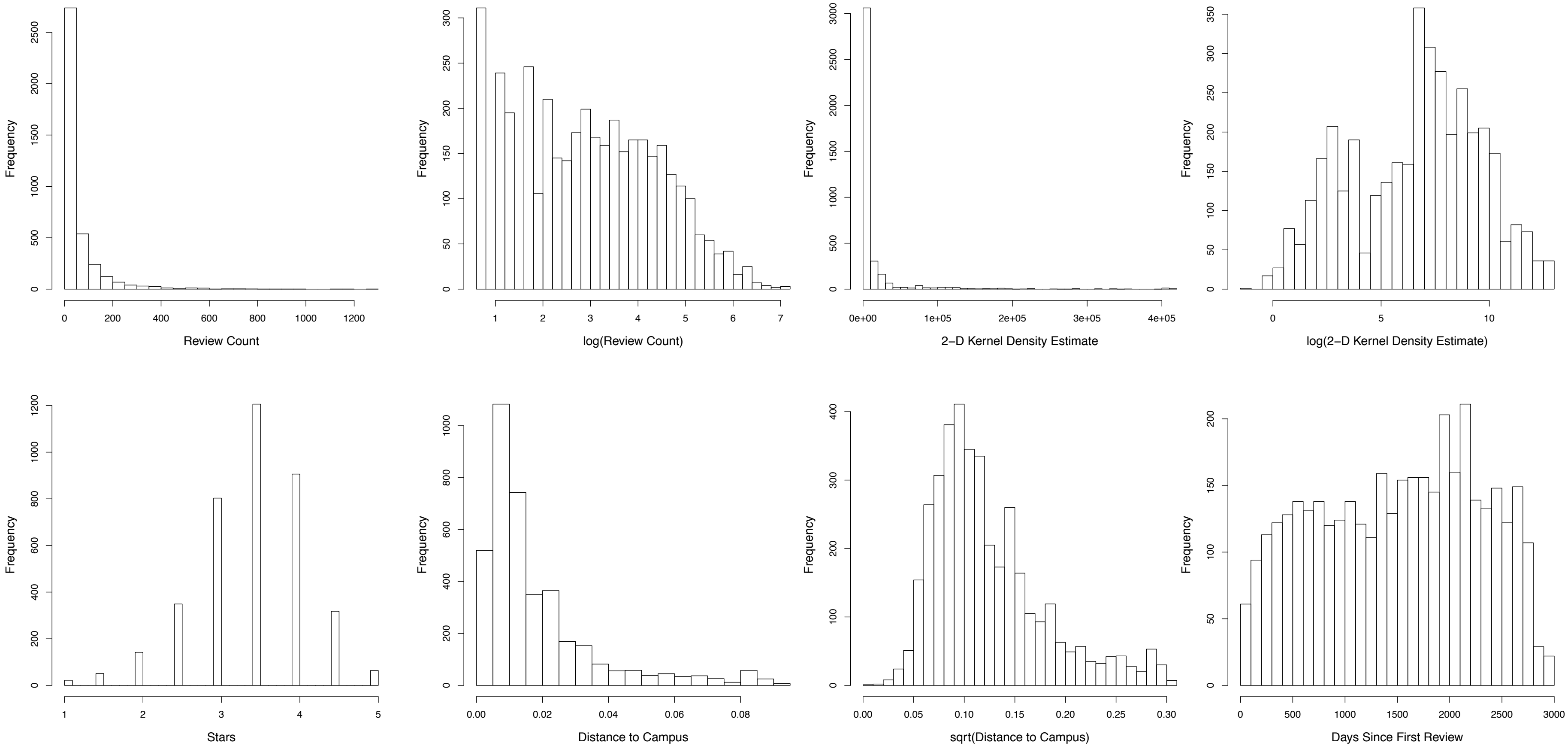
1) **Positive** relationship with restaurant **density.**

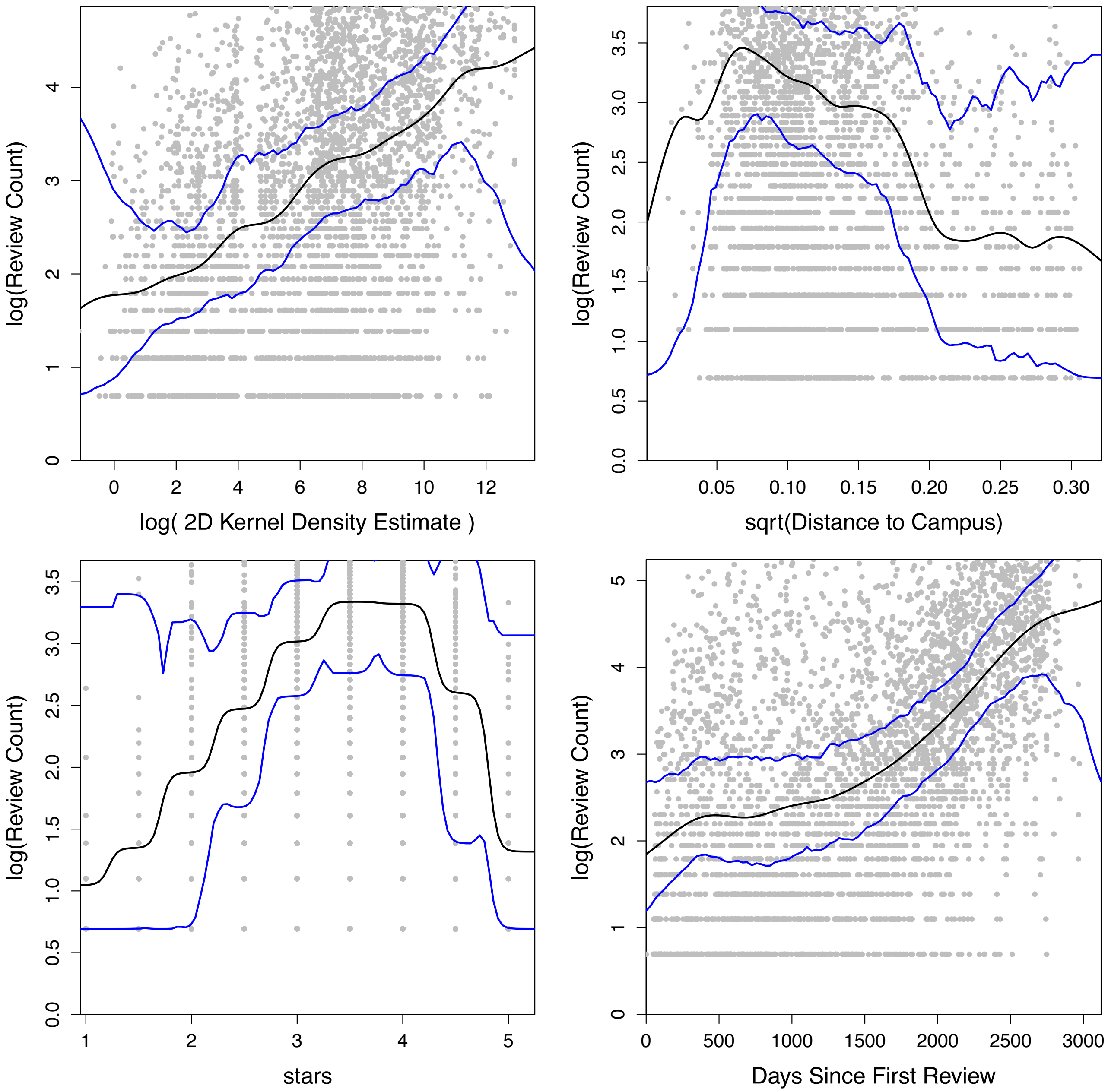
2) **Negative** relationship with **distance** from university campus.

3) **Positive** relationship with **average star rating** of the restaurant.

4) **Positive** relationship with time since first review (proxy for the **restaurant's age**.)

5) Additional **regional variation** in review counts.

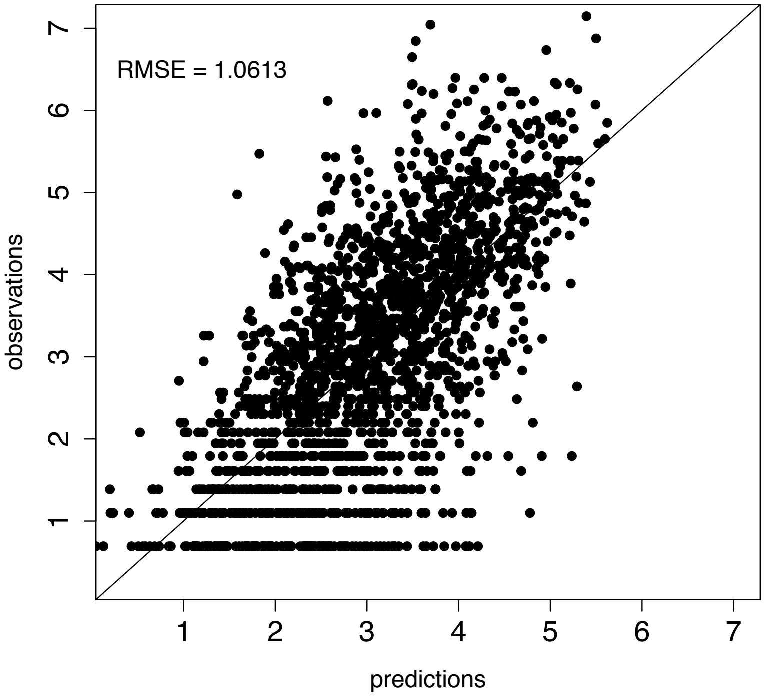

Frequency
0
500
1000
1500
2000
2500
0
200
400
600
800
1000
1200
Review Count
Frequency
0
50
100
150
200
250
300
1
2
3
4
5
6
7
log(Review Count)
Frequency
0
500
1000
1500
2000
2500
3000
0e+00
1e+05
2e+05
3e+05
4e+05
2−D Kernel Density Estimate
Frequency
0
50
100
150
200
250
300
350
0
5
10
log(2−D Kernel Density Estimate)
Frequency
0
200
400
600
800
1000
1200
1
2
3
4
5
Stars
Frequency
0
200
400
600
800
1000
0.00
0.02
0.04
0.06
0.08
Distance to Campus
Frequency
0
100
200
300
400
0.00
0.05
0.10
0.15
0.20
0.25
0.30
sqrt(Distance to Campus)
Frequency
0
50
100
150
200
0
500
1000
1500
2000
2500
3000
Days Since First Review

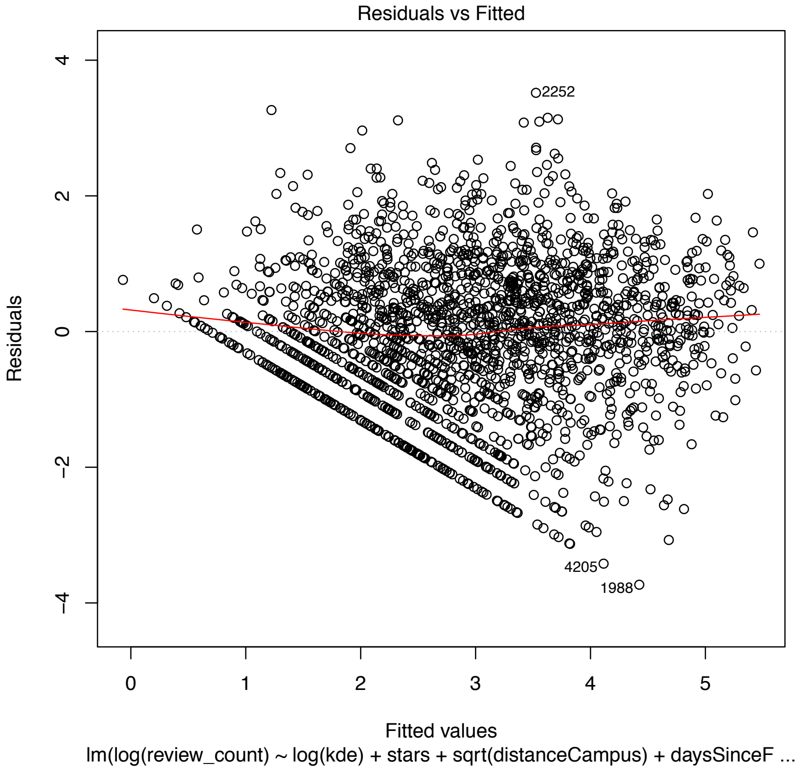

log(Review Count)
0
1
2
3
4
0
2
4
6
8
10
12
log(2D Kernel Density Estimate)
log(Review Count)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
0.05
0.10
0.15
0.20
0.25
0.30
sqrt(Distance to Campus)
log(Review Count)
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
1
2
3
4
5
stars
log(Review Count)
0
1
2
3
4
5
0
500
1000
1500
2000
2500
3000
Days Since First Review

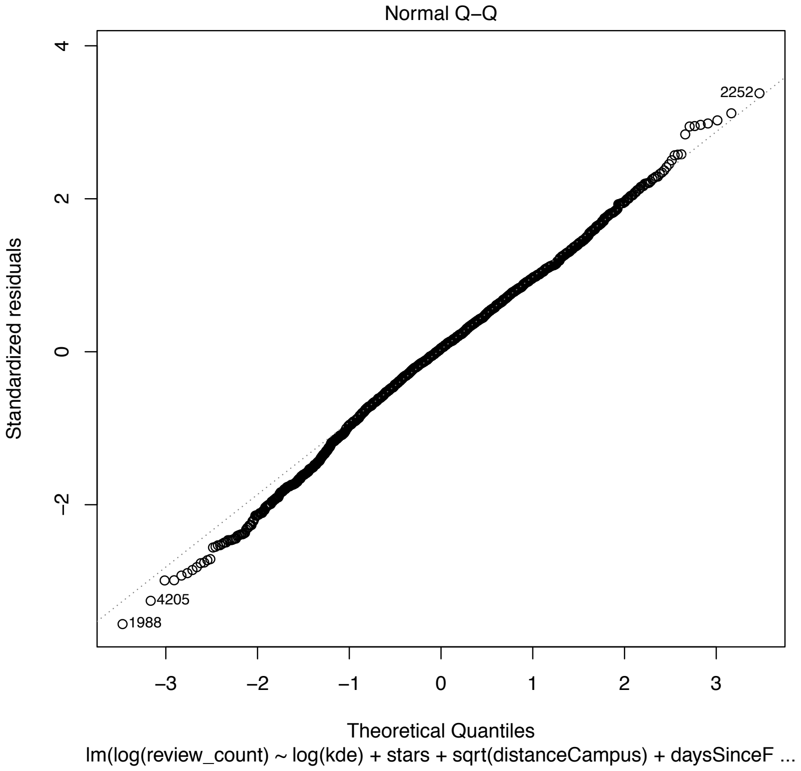
			95 % Confidence Interval	
Predictor	Estimate	Standard Error	Upper	Lower
2-D Kernel Density Estimate (log tranformed)	0.325	0.0289	0.381	0.268
Star-Rating of Restaurant	0.382	0.0334	0.447	0.316
Distance to Campus (square root transformed)	1.86	0.636	3.1	0.608
Days Since the First Review	0.000837	0.0000322	0.0009	0.000774
27 Schools… binary variables that were significant at *p* < 0.001				
…UCLA	-1.29	0.258	-0.733	-1.85
Intercept	-0.945	0.224	-0.505	-1.39

$R^2 = 0.5123$, $F_{31,1899} = 64.34$, $p < 0.001$

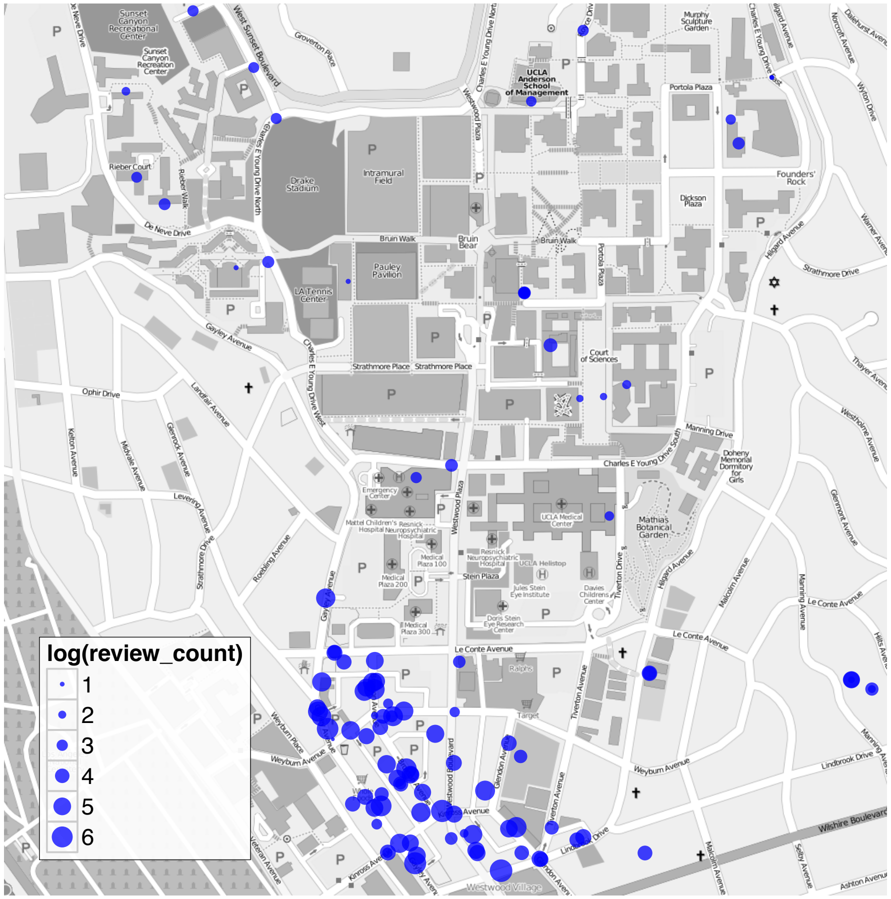
Cross-validated RMSE = 1.06



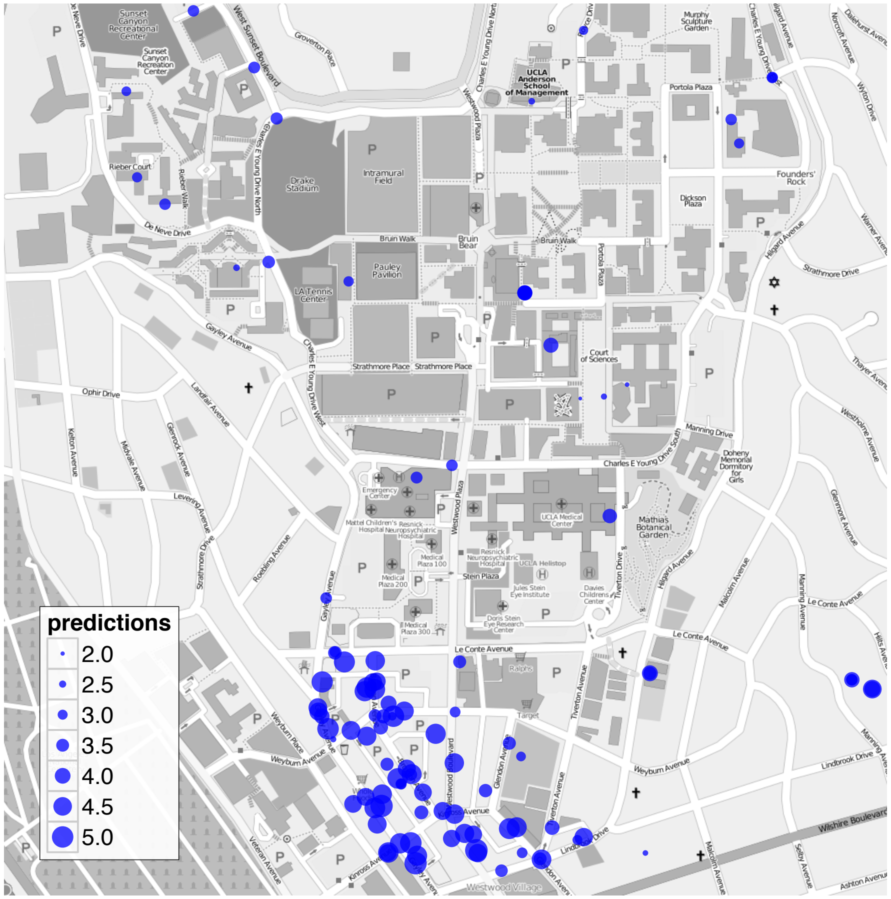

Residuals vs Fitted
Residuals
4
2
0
−2
−4
2252
4205
1988
0
1
2
3
4
5
Fitted values
lm(log(review_count) ~ log(kde) + stars + sqrt(distanceCampus) + daysSinceF ...


Normal Q−Q
Standardized residuals
4
2
0
−2
2252
4205
1988
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(log(review_count) ~ log(kde) + stars + sqrt(distanceCampus) + daysSinceF ...

Observations



Predictions



Summary

Can we predict variation in review count? **YES,** accurate to within about 1 unit on log scale based on split-sample cross-validation. R^2 was better than we expected, statistically explaining 51 % of the variance in review counts.

Revisiting Our Expectations

1) **Positive** relationship with restaurant **density.** ✓

2) ~~**Negative**~~ **Positive** relationship with **distance** from campus. ✗

Three explanations

1. Limitations in our dataset
2. Students are not as active on Yelp as non-students.
3. Limitations of the model.

3) **Positive** relationship with **average star rating** of the restaurant. ✓

4) **Positive** relationship with time since first review (proxy for the **restaurant's age**.) ✓

5) Additional **regional variation** in review counts. ✓

Potential Future Directions

- Spatial Regression
- Machine Learning
- Study additional problems using the Yelp Academic Dataset.