Yelp's academic data set:

- 400 MB
- Businesses with their average ratings ("stars"), latitude, longitude
- Reviews (text, business ID,...)
- Many reviews in the Las Vegas metropolitan area

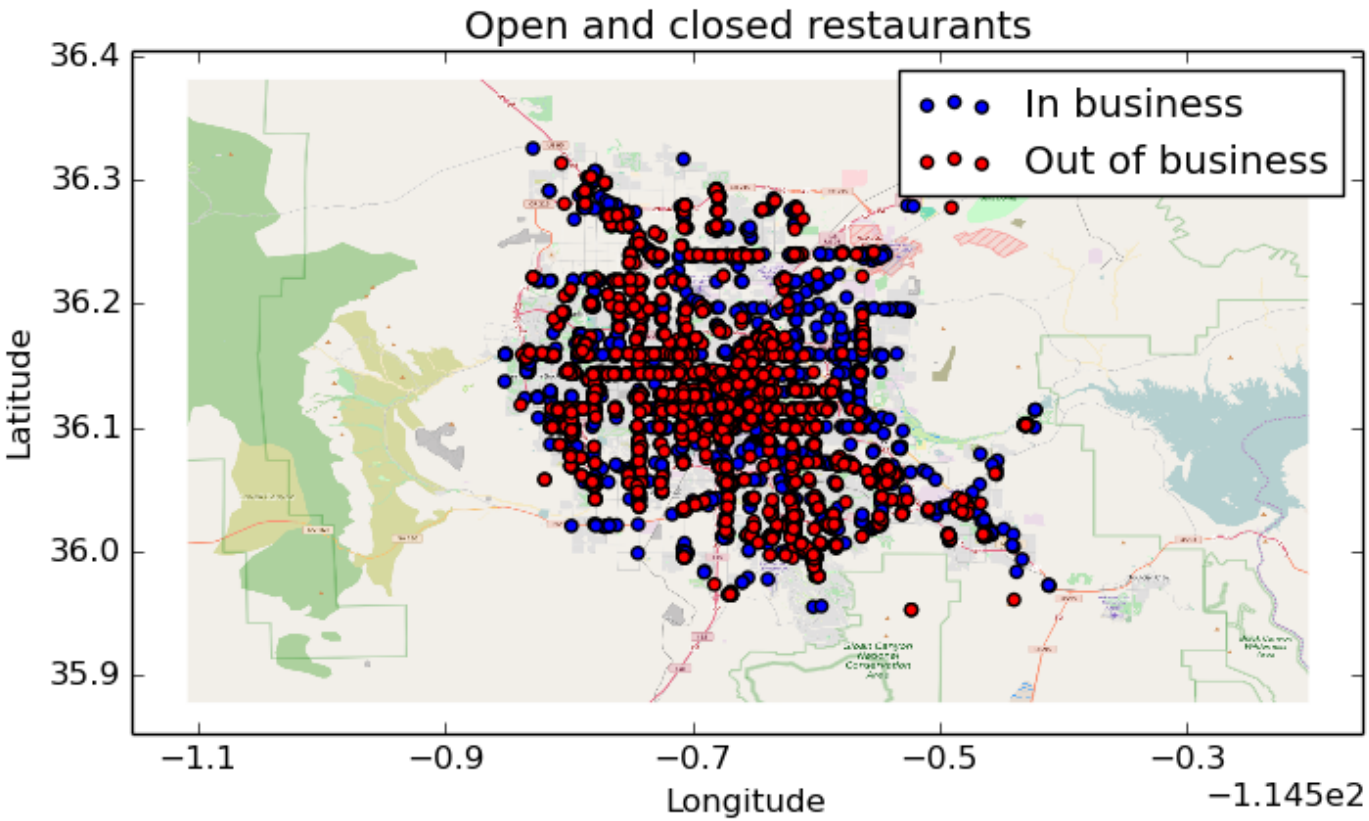
Goal: to design an algorithm to "predict" whether or not a business will go out of business based on location.

Why?

- Average number of stars for open restaurants in the Las Vegas area: 3.4
- Average number of stars of closed businesses: 3.4(!)

Ratings seem to have little to do with whether or not a restaurant in Las Vegas stays in business. My next guess for a predictor: location.

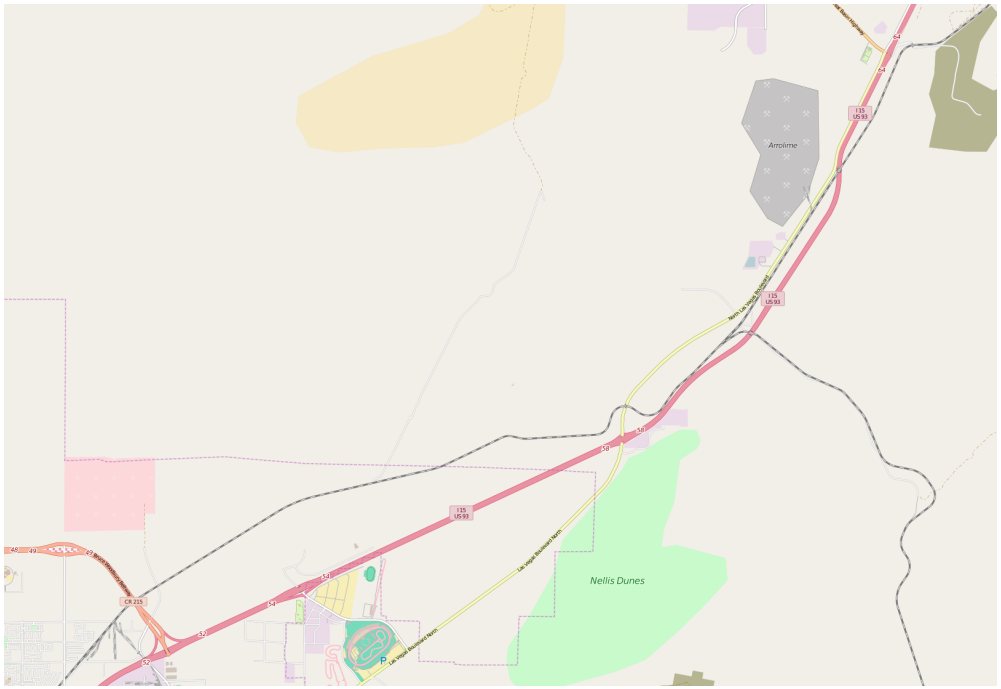
The restaurants of Las Vegas



N=5583 restaurants in the "training set."

Preliminary analysis: can one find bad locations?

The Las Vegas area was broken up into a 80×80 grid of longitudes and latitudes. The fraction of closed businesses was determined across this grid. The locations with the highest fraction of closed businesses were sorted out. Here is the worst location, with 83% of restaurants out of business:



One feature: number of nearby reviews

For each restaurant i, I determine $x_i =$ the number of reviews of restaurants a distance $d^2 < 10\ \mathrm{km}^2$ away from restaurant i:

- For restaurants still in business, the average number of nearby reviews $= 225$.
- For restaurants out of business, it is 5332(!)

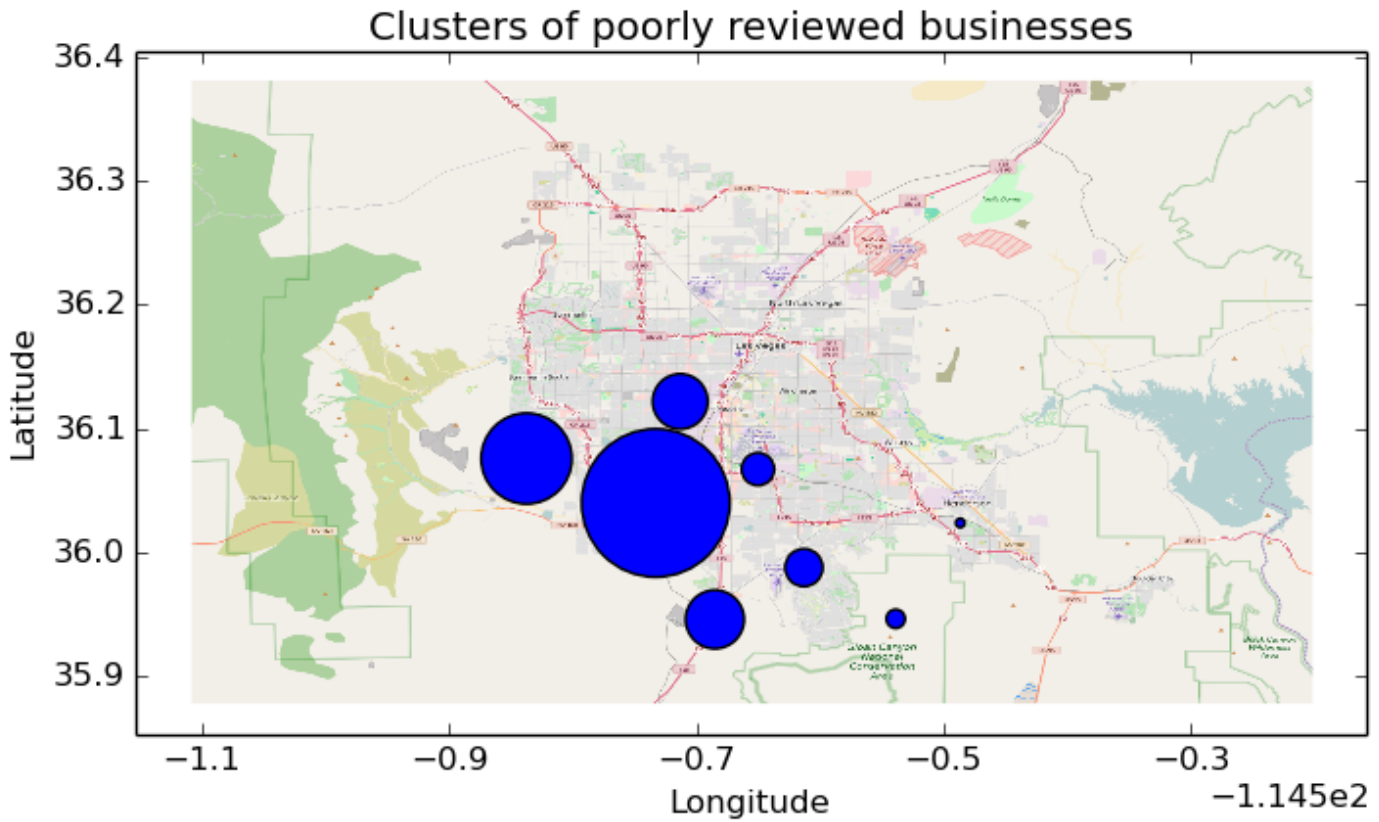
While the worst locations are remote, businesses seem to come and go frequently from highly competitive locations. The ideal location is somewhere in the middle of the range of x.

My project

- I identified a feature x which is on average large for closed businesses.
- The relationship between x and the probability of remaining in business is non-linear.
- I will perform supervised learning (logistic regression) to "predict" whether or not a business will stay open. My feature space is x, x^2, x^3...

Possible additional feature: where are there unsatisfied customers looking for a better option?

K-means clustering of bad reviews:



The largest cluster has 54806 unsatisfied customers.

Side project: unreliable reviews

"The pasta was cooked perfectly, al dente. The espresso was too strong. All in all, a decent meal. 3.2 stars"

VS.

"WORST EXPERIENCE OF MY LIFE."

What happens to the variance of reviews, for one given restaurant, when the reviews with superlatives are filtered out?