Yelp Data Wrangling Project

Rich Dean

22 October 2015

This document covers my initial exploration of the Yelp Dataset Challenge data. This dataset from the online review site Yelp, and is part of a data anlysis competition, now in its fifth run. The data contains (from the Yelp website):

- 1.6M reviews and 500K tips by 366K users for 61K businesses
- 481K business attributes, e.g., hours, parking availability, ambience.
- Social network of 366K users for a total of 2.9M social edges.
- Aggregated check-ins over time for each of the 61K businesses

Initialisation

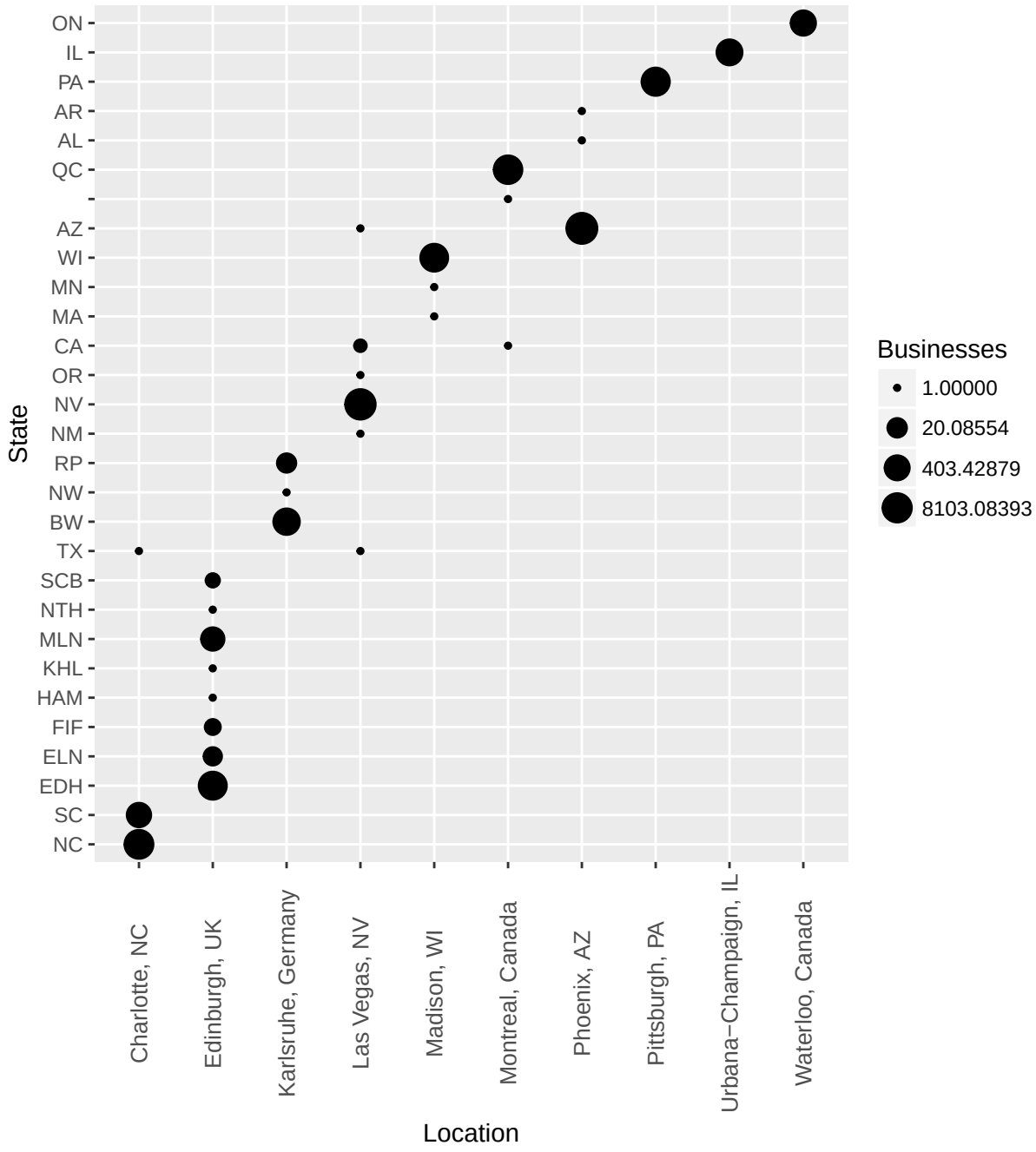
To start, I'll import my libraries, include files, and set globals variables. After that, I'll load in the data (from cache if available).

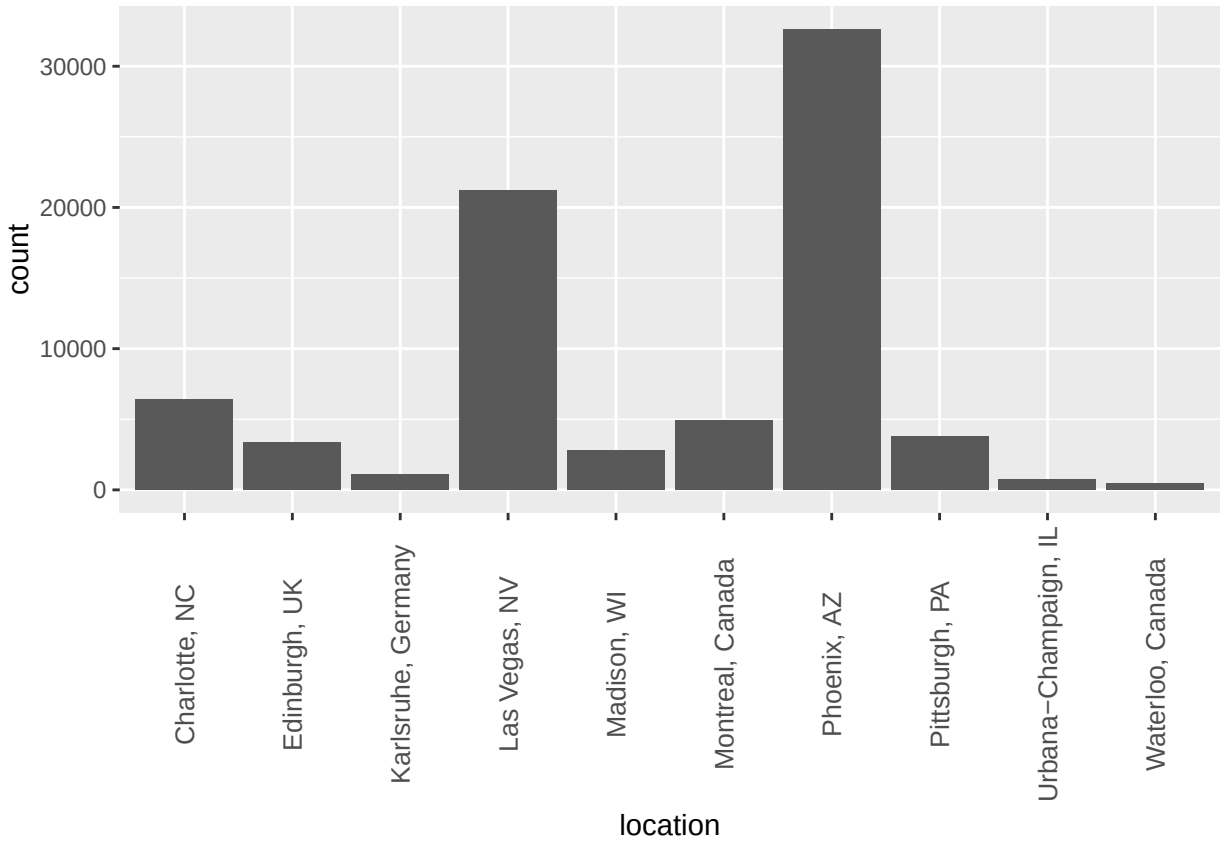
Now that the data is loaded, I'll quickly verify the Yelp statements about the dataset size, and take a quick look at the structure of the **y.business** data frame.

```
##         Stat    Rows
## 1    Reviews 2225213
## 2       Tips  591864
## 3      Users  552339
## 4 Businesses   77445
## 5   Checkins   55569
```

Geography

I know that the businesses are located in 10 distinct areas. I have their names, and the **latitude**/**longitude** are in the **y.business** object. I'll use k-means clustering to label the groups, which will allow for some comparative geo-analysis. I have the list of actual locations which the data should be from, so i will seed the kmeans with the lat/long of those ten locations.

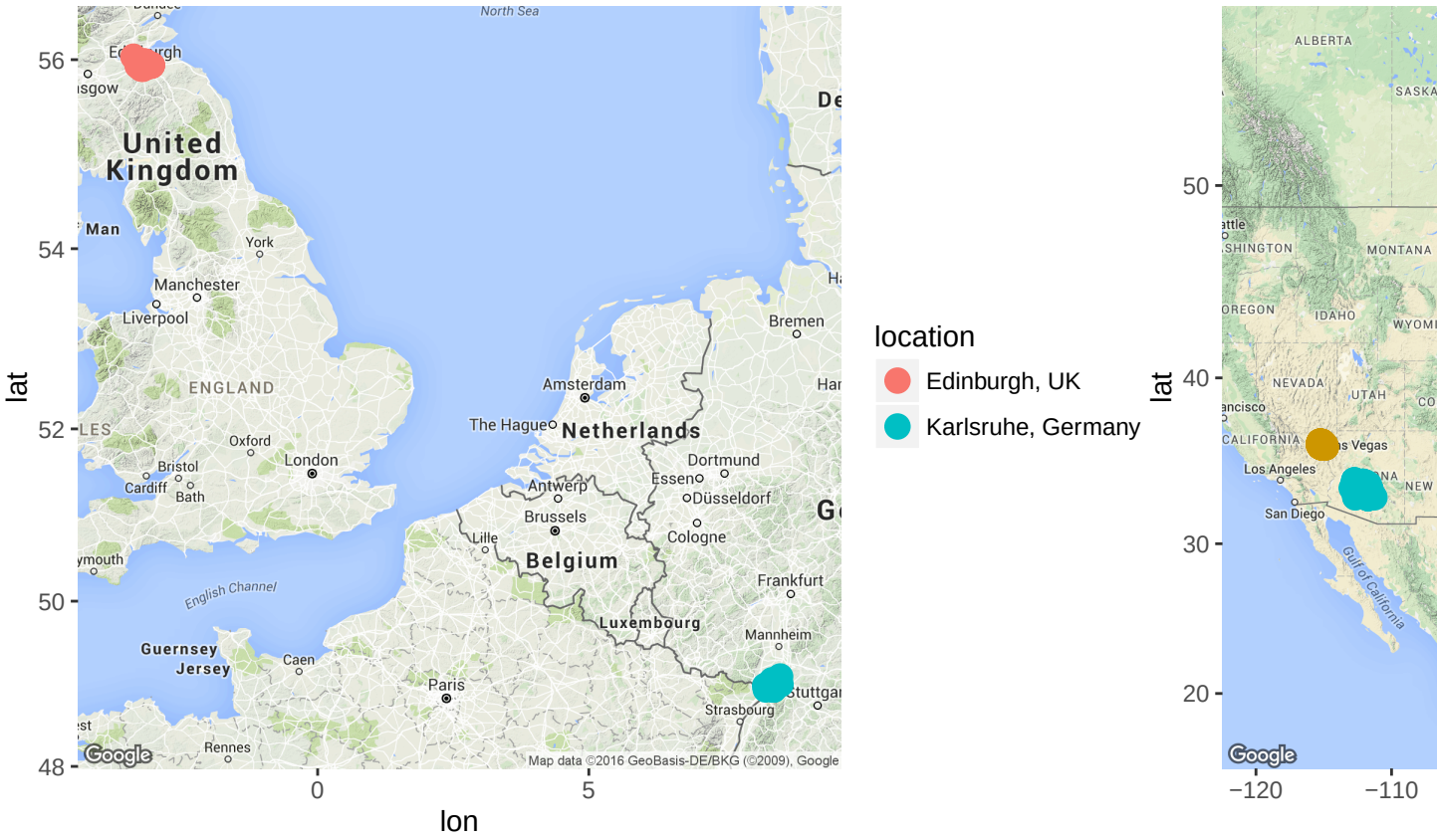

ON
IL
PA
AR
AL
QC
AZ
WI
MN
MA
CA
OR
NV
NM
RP
NW
BW
TX
SCB
NTH
MLN
KHL
HAM
FIF
ELN
EDH
SC
NC
State
Charlotte, NC
Edinburgh, UK
Karlsruhe, Germany
Las Vegas, NV
Madison, WI
Montreal, Canada
Phoenix, AZ
Pittsburgh, PA
Urbana−Champaign, IL
Waterloo, Canada
Location
Businesses
1.00000
20.08554
403.42879
8103.08393



These results show a good grouping, where each state is entirely represented by one cluster, with the exception of a few erronously identified businesses. For instance, 'Cavalia' in Montreal (based on it's latitude/longitude), has been given the address details of Cavalia in Burbank California. Note that Edinburgh covers many Scottish counties - these are much smaller than US states.

```
## [1] "Outremont\nBurbank, CA 91502"
```

I'll do a pair of map plots - one for the US, and one for Europe - to show the distribution of the business, coloured by the cluster.

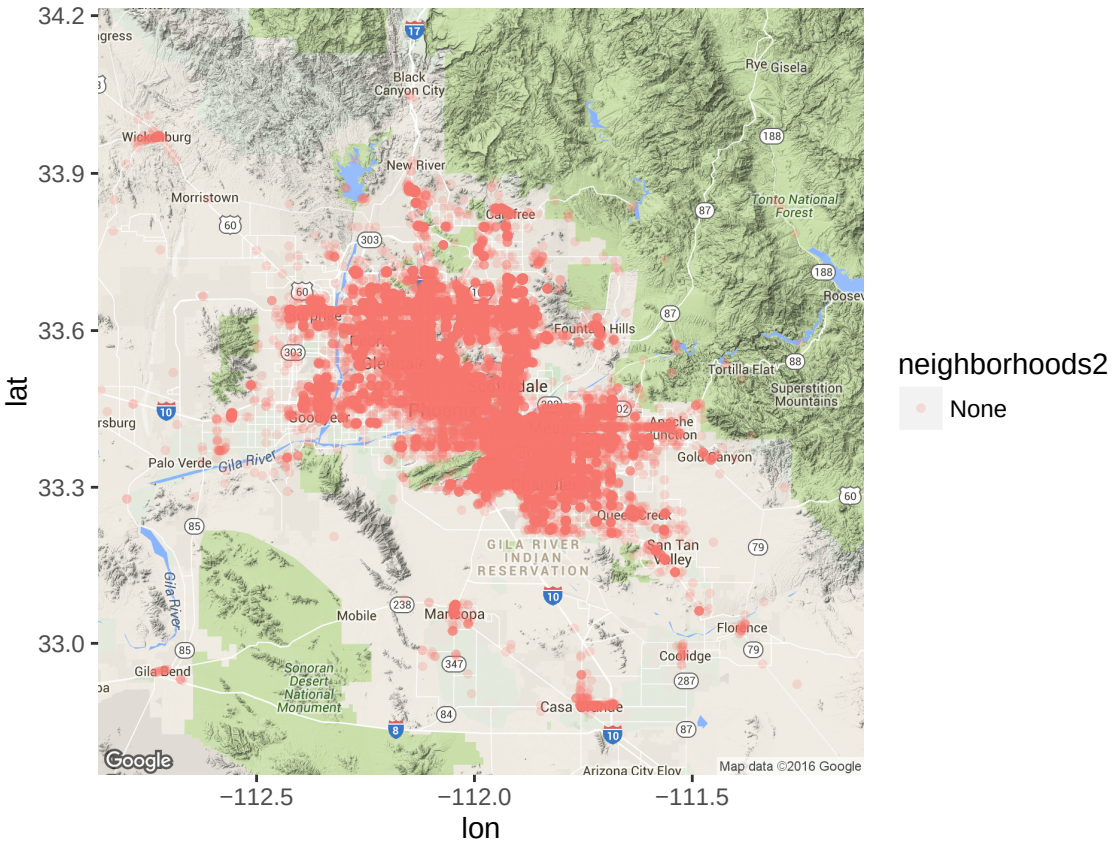


The maps show that the clustering worked nicely, and also shows the geographic spread of the businesses in the dataset. Despite the scale of the USA map, you can see something of the dispersal of businesses within each cluster - the Phoenix AZ cluster is the largest of the groups on the map, while the Urbana-Champaign IL group is smallest.

I'll add a distance to the cluster center too - I can then do some analysis of the area covered by each cluster This process uses the Haversine formula for distance on a sphere, given two pairs of coordinates in radians. The resulting figures are in kilometres.

```
## Source: local data frame [10 x 3]
##
##                location        mn        sd
##                   (chr)     (dbl)     (dbl)
## 1         Charlotte, NC 10.192273  6.152191
## 2         Edinburgh, UK  1.727740  2.023256
## 3    Karlsruhe, Germany  3.165571  3.220315
## 4         Las Vegas, NV  9.040471  5.505476
## 5           Madison, WI  6.605847  4.517044
## 6      Montreal, Canada  4.859244  4.337261
## 7           Phoenix, AZ 19.680806 12.066621
## 8        Pittsburgh, PA  4.628722  2.861922
## 9  Urbana-Champaign, IL  1.743124  1.250232
## 10     Waterloo, Canada  3.740925  2.766424
```

This shows that Urbana-Champaign has the tightest group of businesses (low mean and standard deviation), while Phoenix has the most dispersed, as seen on the map plots. I'll take a closer look at Phoenix - this will be a good location to investigate, as it is both the most dispersed, and has the most business points.



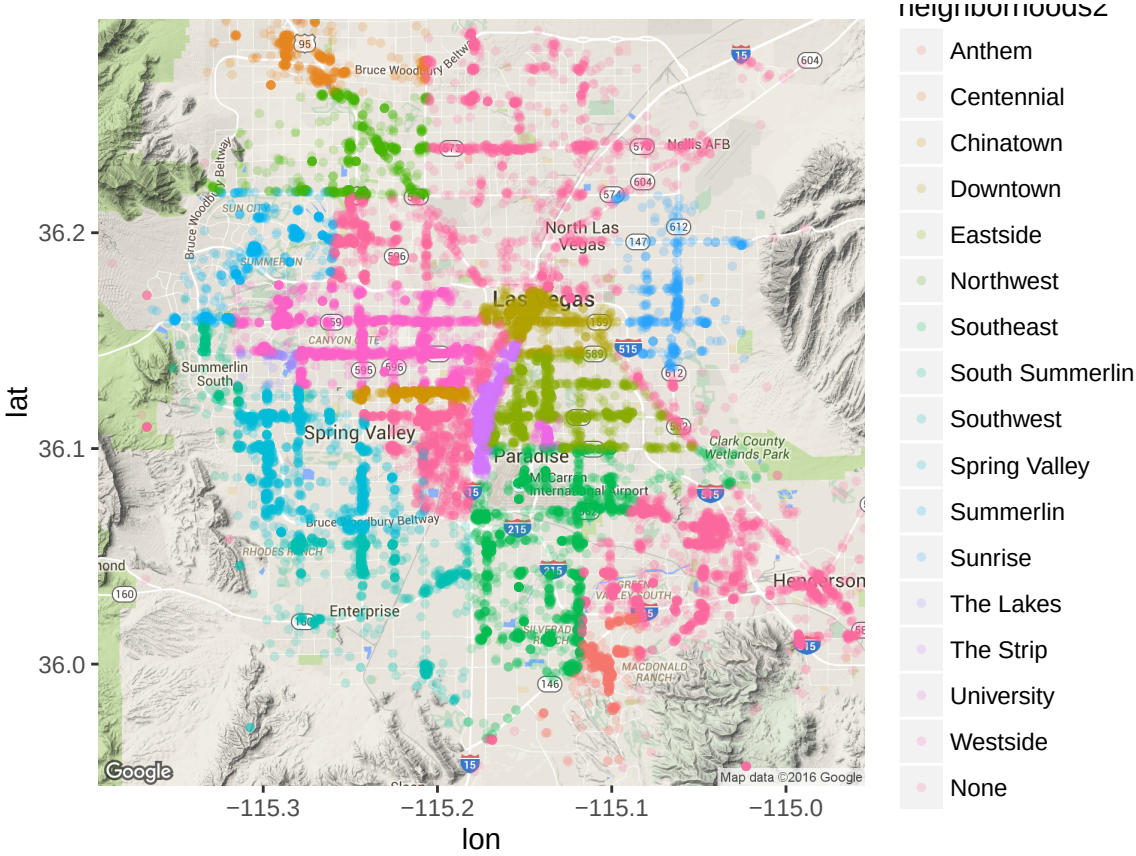
From the above plot, you can clearly see the distribution of the businesses along the grid pattern of the roads with clustering around Phoenix itself, and the satellite towns including Scottsdale, Glendale and Tempe. I tried to use the *neighborhoods* column to highlight each area, however in Phoenix, no actual neighbourhoods were indicated. Checking the businesses that have neighbourhood data:

```
##
##                       Has Neighborhood No Neighbourhood
##   Charlotte, NC                   3699             2735
##   Edinburgh, UK                   2807              569
##   Karlsruhe, Germany                 0             1067
##   Las Vegas, NV                  15411             5828
##   Madison, WI                     1377             1427
##   Montreal, Canada                4552              392
##   Phoenix, AZ                        0            32616
##   Pittsburgh, PA                  3104              650
##   Urbana-Champaign, IL               0              737
##   Waterloo, Canada                   0              474
```

The neighbourhood data is very spotty, with four locations having no neighbourhood data at all. This can't really be used for analysis - though viewing a specific location could be interesting.

```
## Warning: Removed 262 rows containing missing values (geom_point).
```

```
## Warning: Removed 4 rows containing missing values (geom_point).
```



The neighbourhoods are clearly grouped together on this plot. Where we do have neighhood data, clustering could be used to label the unidentified neighbourhoods - although this could fail - as seen by the area east of Spring Valley, where there are a large group of unlabeled businesses. This looks like it should be a neighbourhood, but is inexplicably blank.

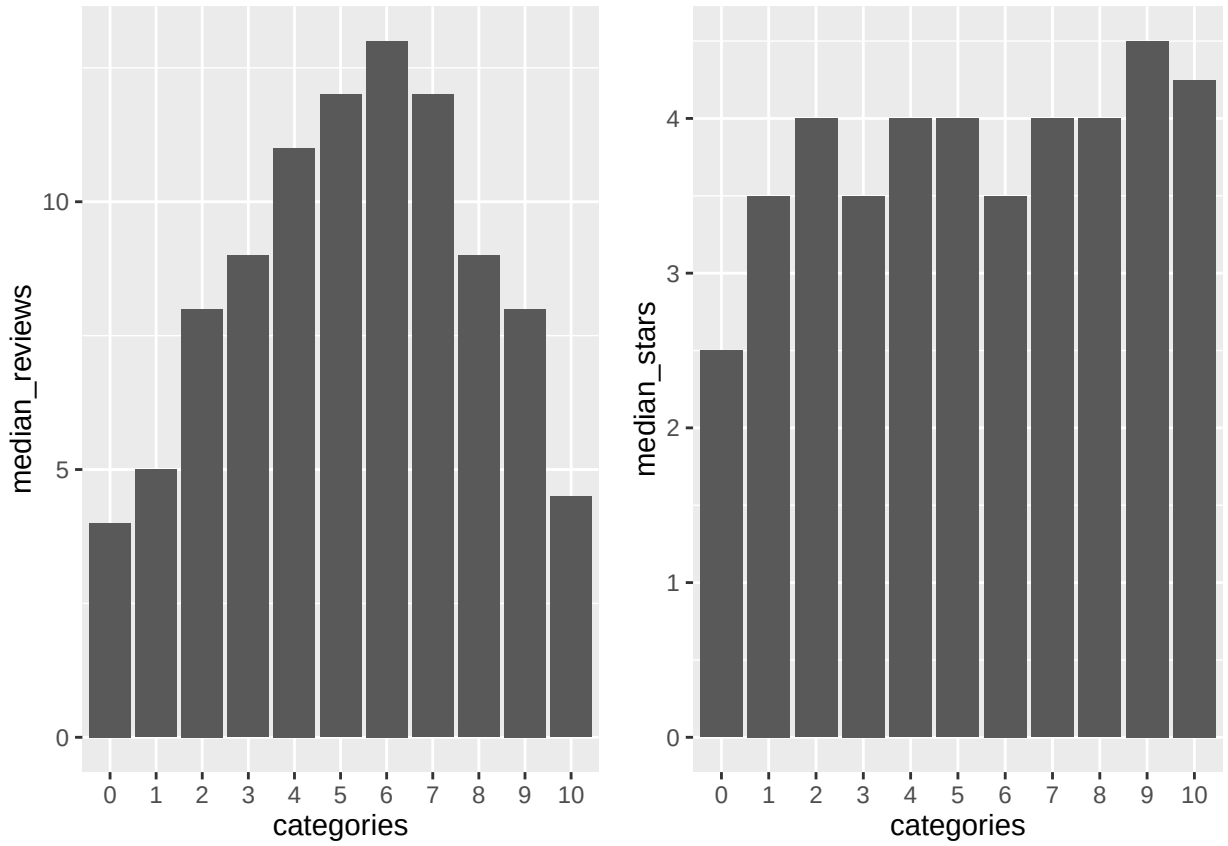
This kind of study shows the unreliability of the details for businesses in this dataset. While there is a vast amount of data available, the high level of inconsistency means that it would be difficult to rely on any of it. With this in mind, I'll move on to the data generated internally by Yelp.

Business Categories

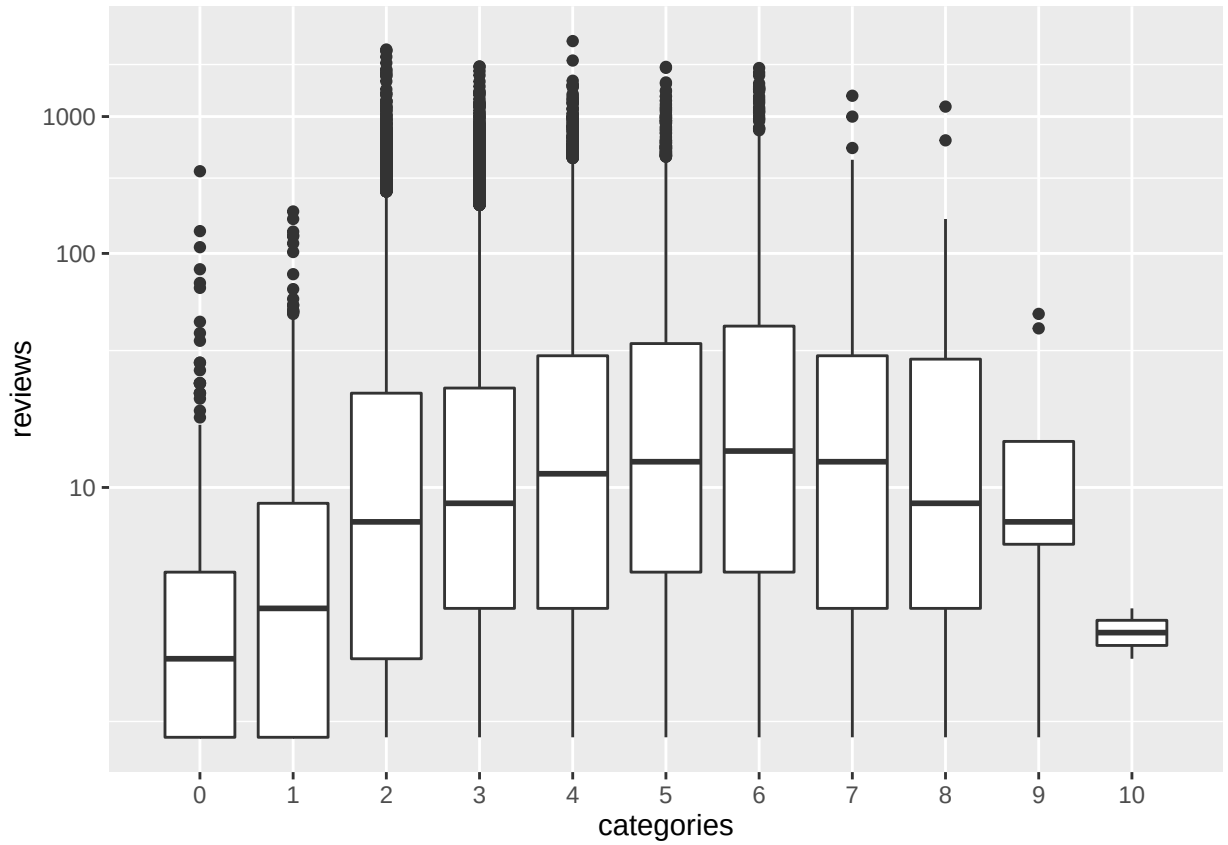
I'd like to get a primary category for each business - this will also simplify analysis. The primary category list is taken from https://www.yelp.com/developers/documentation/v2/all_category_list.

```
## [1] 346
```

There are 346 combinations of primary categories! This makes any specific comparison of primary categories tricky, as businesses could belong to several primary categories. Taking an slightly different tack led me to think more deeply about the categories and review counts. Do businesses with more categories have more reviews and/or higher ratings?



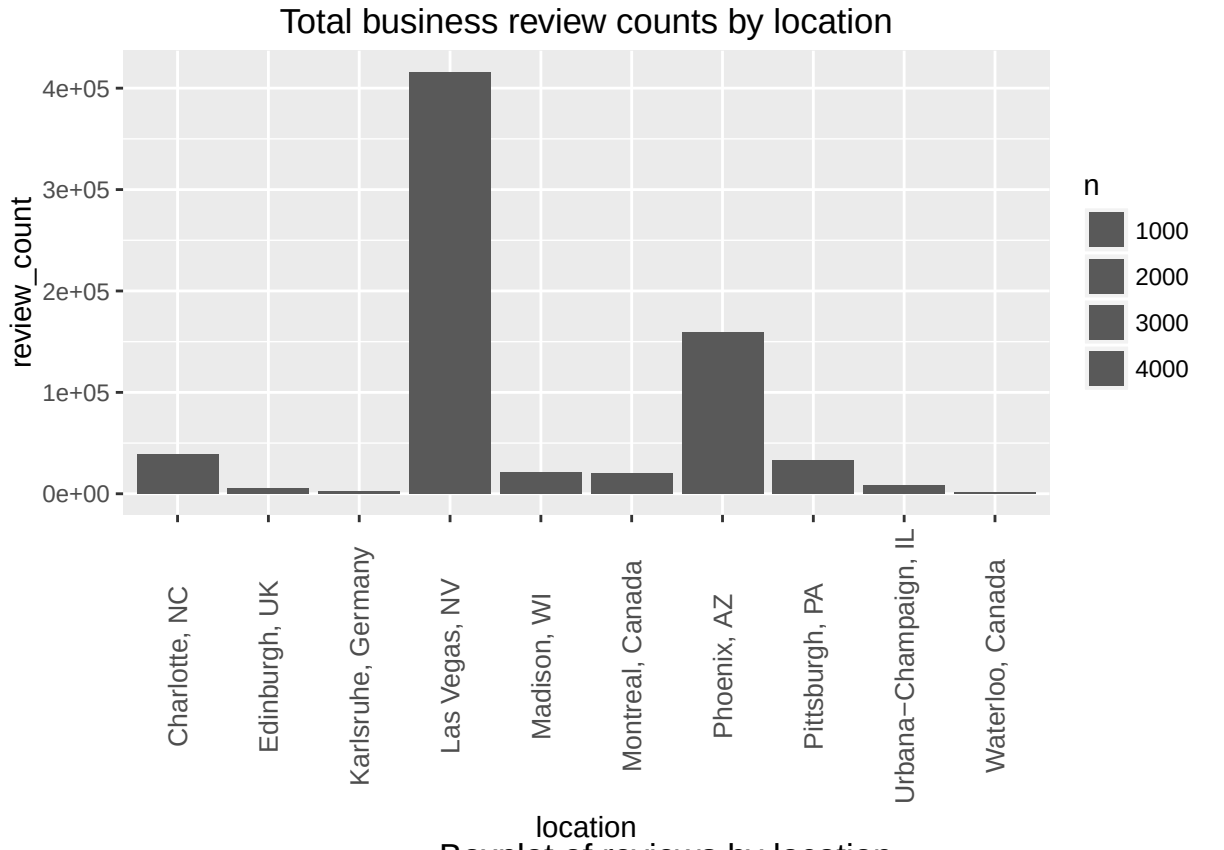
There isn't a clear link between the ratings and number of categories, but there appears to be a clear link between categories and number of reviews

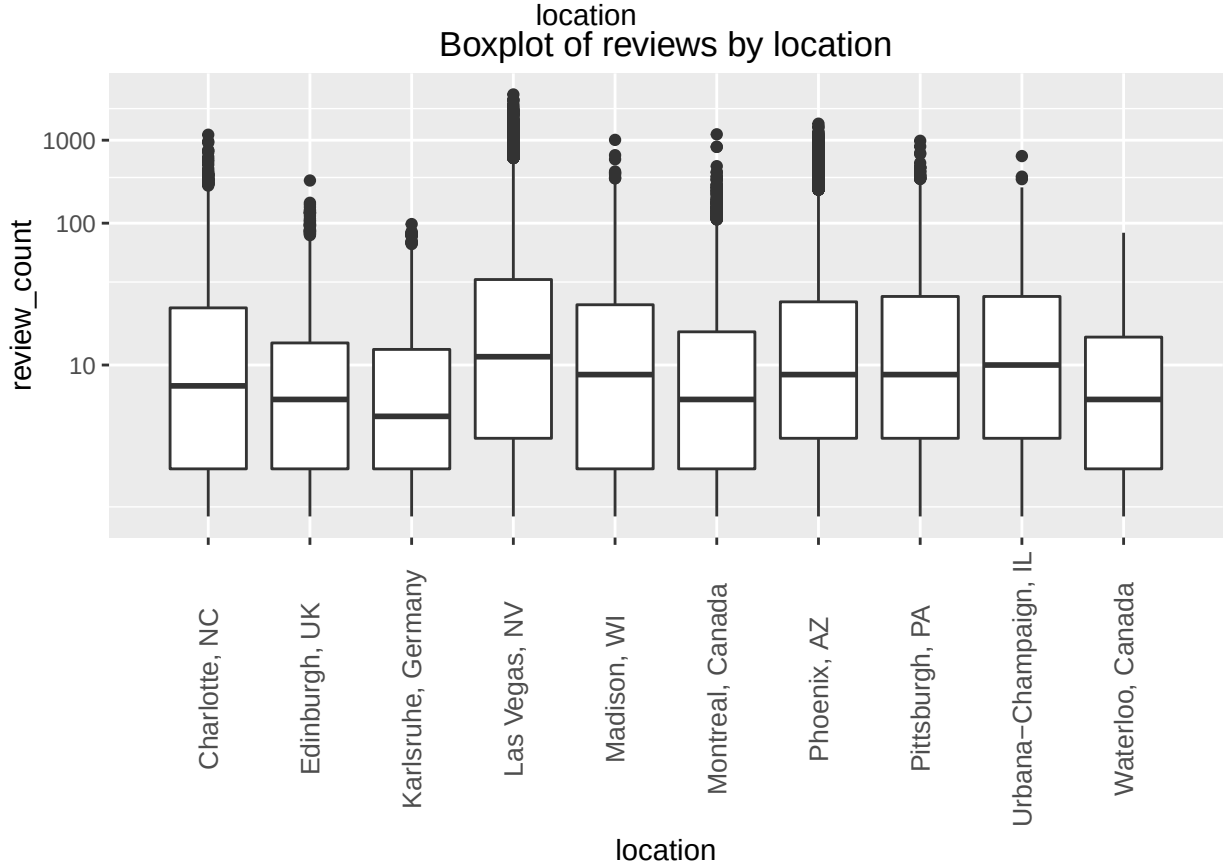


Using a log plot of reviews shows there are actually few businesses with very high numbers of reviews. The point density shows the bulk of businesses have between 2 and 5 categories, and less than 100 reviews. Let's have a look at some of the high review businesses that form those outliers.

```
##                             name      location review_count stars
## 1                   Mon Ami Gabi Las Vegas, NV         5642   4.0
## 2                   Wicked Spoon Las Vegas, NV         4558   3.5
## 3               Earl of Sandwich Las Vegas, NV         4452   4.5
## 4               Bacchanal Buffet Las Vegas, NV         4390   4.0
## 5            Gordon Ramsay BurGR Las Vegas, NV         3811   4.0
## 6                 Serendipity 3 Las Vegas, NV         3478   3.0
## 7                     The Buffet Las Vegas, NV         3281   3.5
## 8         The Buffet at Bellagio Las Vegas, NV         3014   3.5
## 9             Hash House A Go Go Las Vegas, NV         3014   4.0
## 10 The Cosmopolitan of Las Vegas Las Vegas, NV         2988   4.0
```

Every one of the top ten business by *review_count* is in Las Vegas! (In fact, 89 of the top 100 reviewed businesses are in Las Vegas). I'll examine the total review counts and the means, to see if this is a trend.

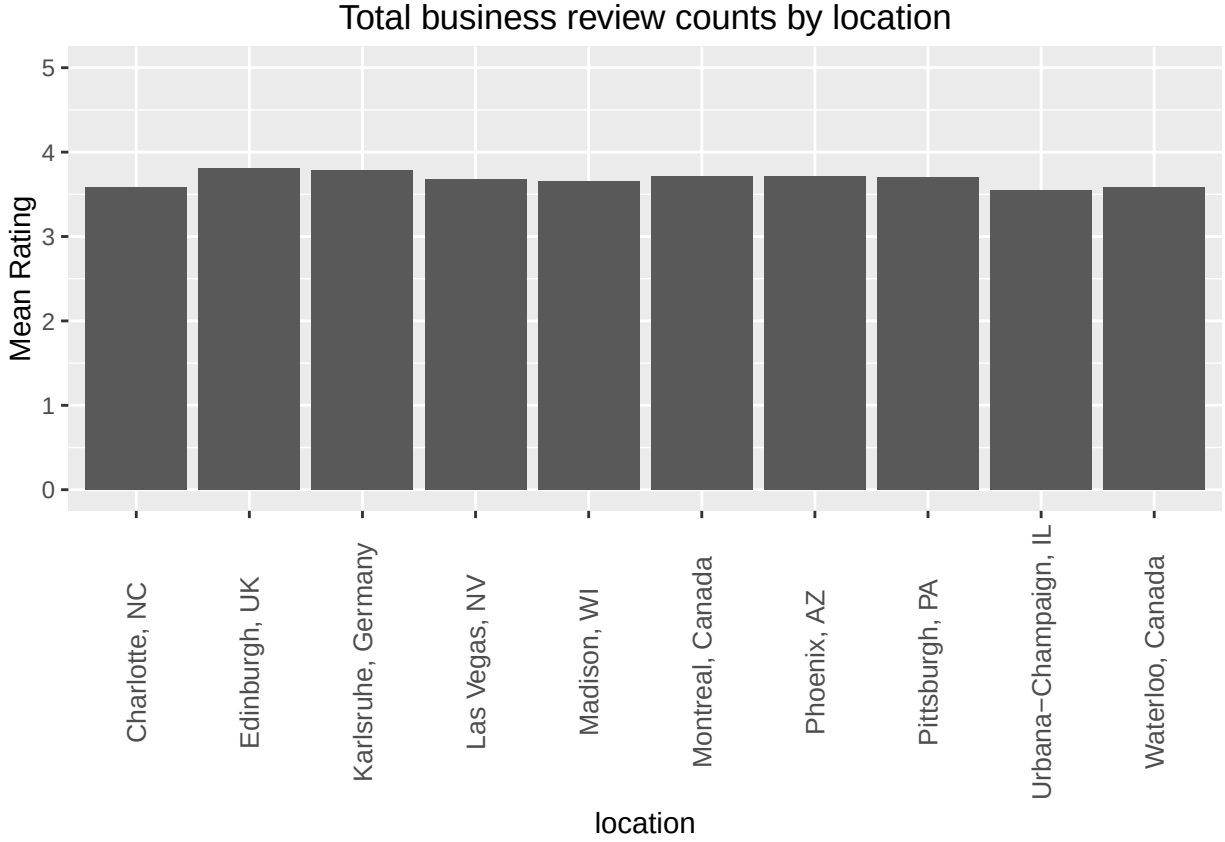




While the total reviews in Las Vegas is significantly greater than the other locations, the boxplot shows

only a small increase in the median *review_count*. This implies that a small group of very highly reviewed businesses are present in Las Vegas, skewing the figures.

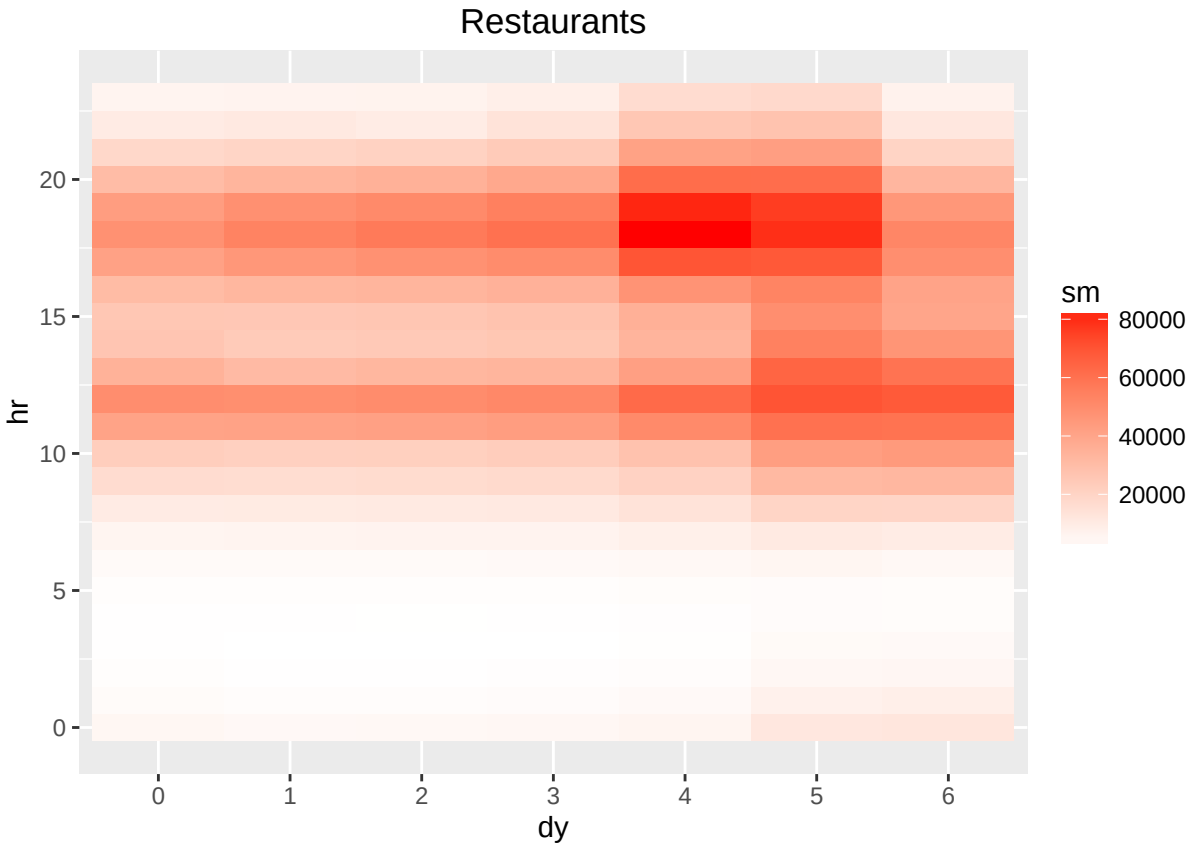
I would theorise that this skewing is due to Las Vegas being the most touristic of the ten locations. Many more people will visit Las Vegas than the other locations, thus increasing the number of reviews.



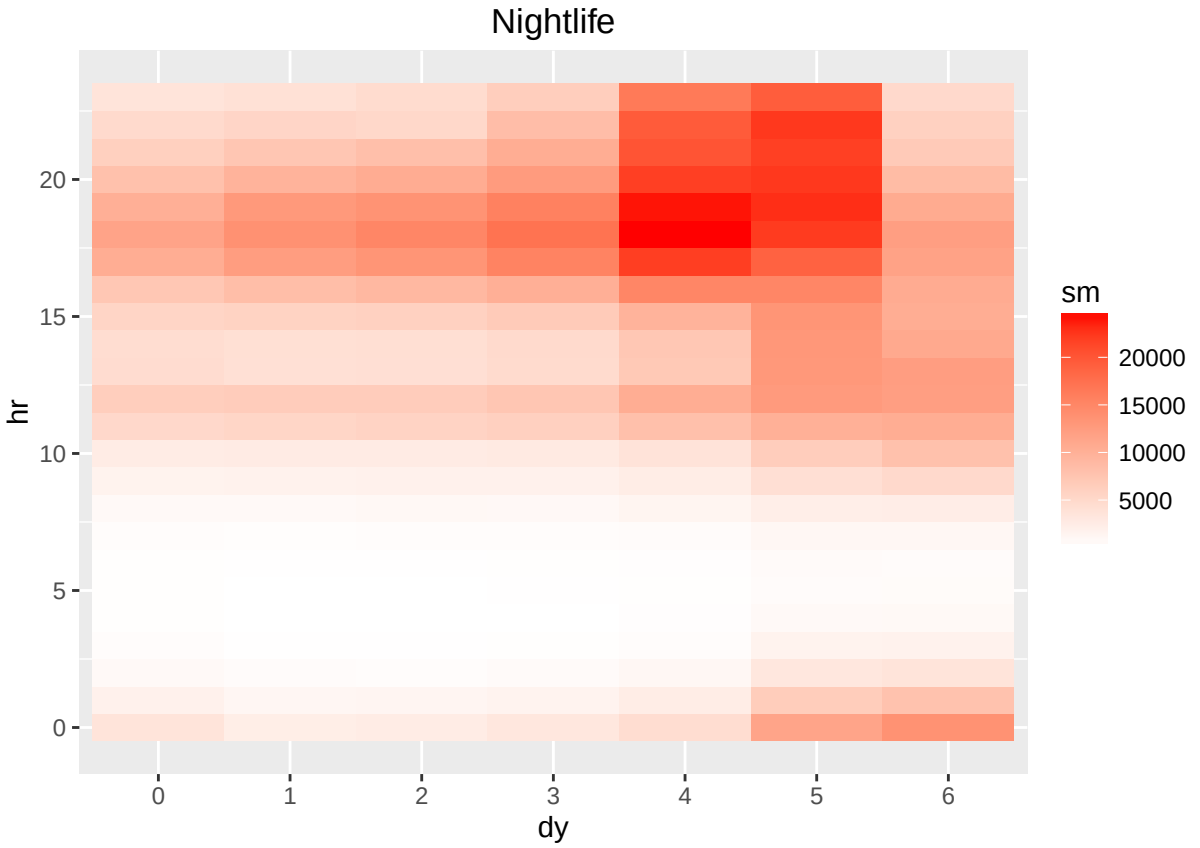
Tourism doesn't seem to affect the mean rating, however - every location has roughly the same mean.

Checkins

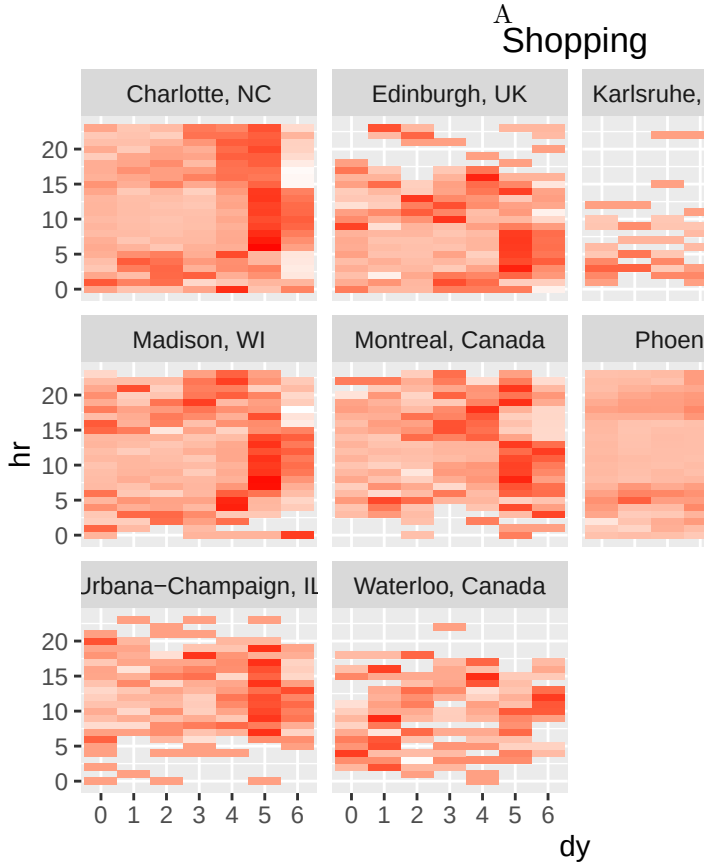
I had hoped to perform more category-based analysis, but with multiple primary categories, I would have to sift the combinations and derive so kind of hierarchy to determine a single primary for each business. Instead, I chose to look into a different view of the data - checkins. This shows when businesses are visited, by week-hour. The **y.checkin** data frame contains my core data - I'll add some of the columns from **y.business** to provide a more useful single frame. I will use heatmaps to visualise this data.



Here we can see a clear pattern in restaurant checkins - more frequent at lunchtimes and evenings, and also on Friday and Saturdays. The *nightlife* category should clearly show a similar pattern...



A

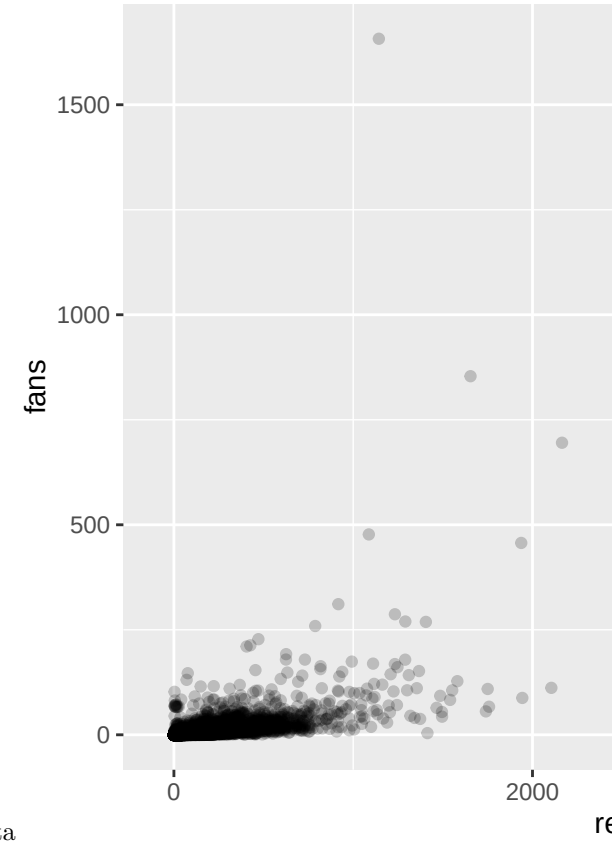


similar pattern, but with more emphasis on the weekend evenings.

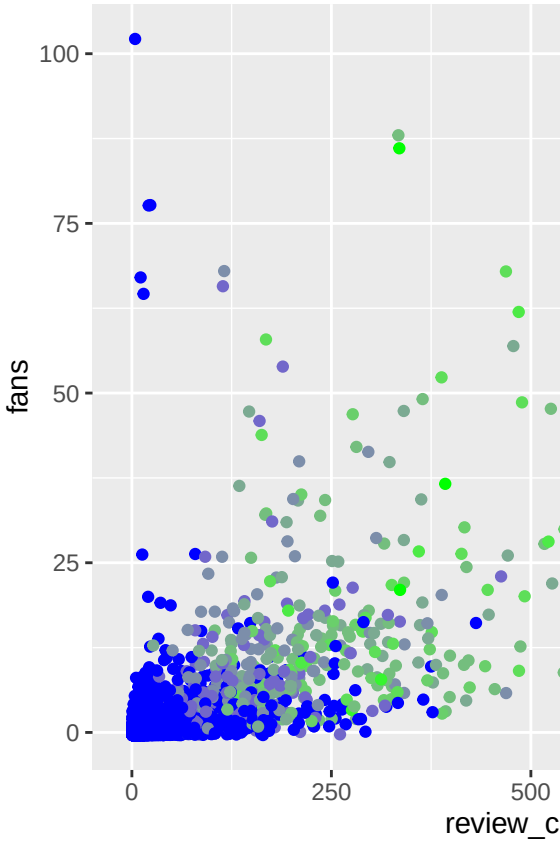
Shopping patterns show that Saturday is the most popular day across the ten locations. Sunday trading

shows an interesting pattern, particularly in Karlsruhe where there is almost no activity - Germany has strict Sunday trading laws, and most shops will be closed.

Users

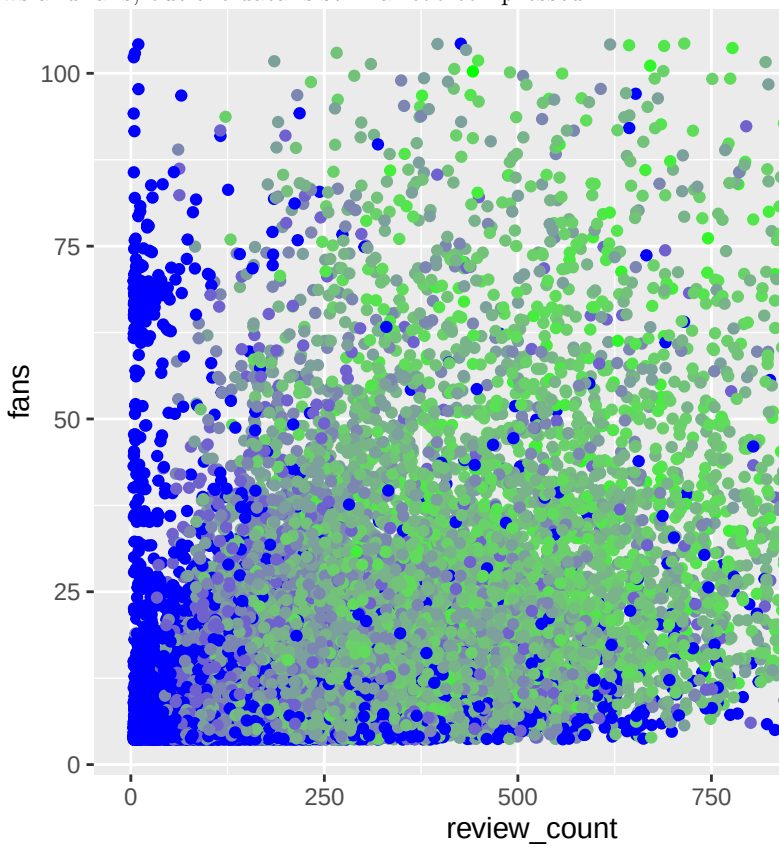


********** TODO Expand on the text Initial look at the user data

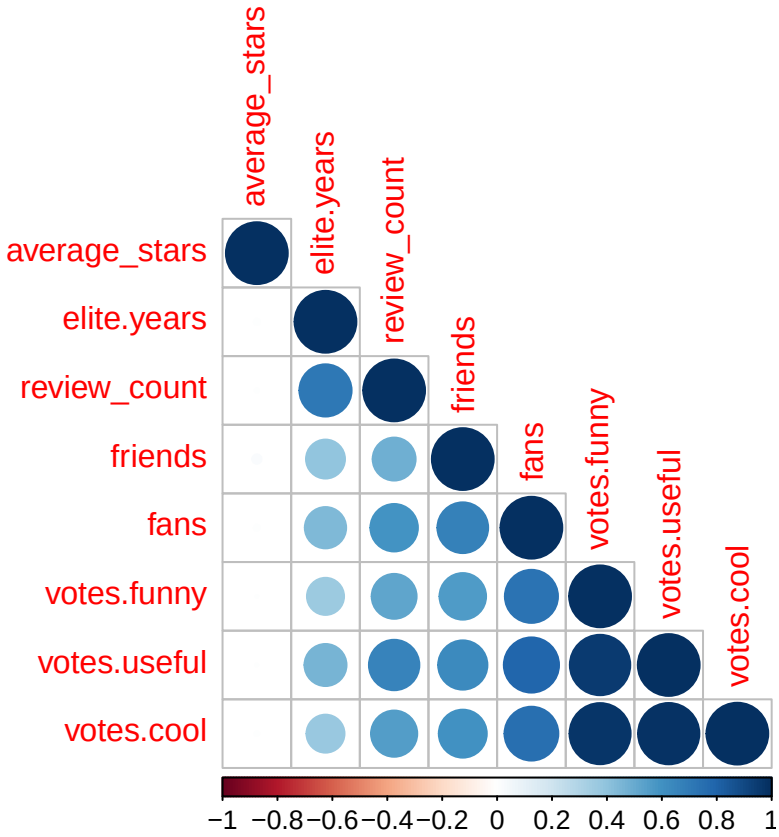


Improve - remove extreme outliers, log the scales, add the elite.years as a colour

This plot appears to show a mild correlation between reviews and fans, but the data is still far too compressed.



I'll remove the low-fan users, and see what impact that has.
Now we can see the correlation between *review_count* and *fans* more clearly - and additionally we can see a peculiar grouping of high *fan*, low *review_count* users. This is a strange grouping of outliers, particularly as they seem to be clustered around 70 fans. I could speculate many reasons for this, but without sufficient supporting data I can't make any confident statements. I do, however, consider them to be an anomaly, and will exclude them from further analysis (*fans*>60, *review_count*<30).



Look at the correlations in the user stats:

We can see that there is no correlation between *average_stars* and any of the other values, but there is a strong correlation between the *votes* columns, particularly *votes.funny* and *votes.useful*.

****** TODO Further user analysis - 2-3 more vis

Final Plots and Summary

****** TODO Select my 3 plots, make them perfect. Suggest: ****** TODO USA map, heat map of religion in Phoenix, one other TBD

Reflections

******* TODO 200 words Large multi-layered dataset. Primarily categorical/ text based. Well suited for a NLP/prediction project.