La Quinta is Spanish for next to Denny's

Colin Rundel and Mine Çetinkaya-Rundel

Introduction

This observation is a joke made famous by the late comedian Mitch Hedberg. Last year, John Reiser on his *new jersey geographer* blog wrote up an analysis of this joke using data scraped from the respective websites of Denny's and La Quinta Inns and Suites. We have used John's work as inspiration for developing a team based homework assignment in both our undergraduate and master's level statistical computing courses at Duke University. As part of this assignment students replicate John's analysis by scraping location information from both Denny's and La Quinta's websites using R (instead of Python) and then analyzing and visualizing the validity of Mitch Hedberg's joke.

There are a few reasons why we like this application:

- It is a good way to introduce modern techniques like data scraping.
- It is an opportunity to expose students to geospatial data. Our experience from a few previous DataFests is that students find working with data that varies spatially and temporally very challenging and that they need further training in these applications.
- The joke brings a fun element to it. This is an involved and challenging assignment both statistically and computationally, so the light-heartedness of the application helps keep the students engaged.
- The application, in addition to being fun, is easily accessible. It doesn't require additional technical expertise in another discipline, and hence, students can easily check their own work as they go. They can even confirm their findings using Google Street View.

Depending on the statistical level and the computational focus of the course one could use this assignment in many different ways. It could be just an application in scraping data off the web or one could provide the data to the students and ask them to work on the analysis tasks instead. As such, we hope that this activity is a good fit for many different courses from introductory data science to a higher level (undergraduate or graduate) statistical computing course.

In this article we will focus on the statistical analysis and visualization of the data as these tasks can be adapted to a widest range of statistical courses. At the end, we will briefly discuss the web scraping aspect of the assignment, which is excellent add-on for statistical computing and data science courses.

The data

We have provided two datasets (available on the Chance website) for Denny's and La Quinta that contain available location data for both chains, as of early December 2015. Both of these datasets consist of 6 columns (`Address`, `City`, `State`, `Zip`, `Longitude`, `Latitude`) with each row reflecting a distinct location of each chain within the United States. When these data were collected there were 1643 Denny's and 851 La Quinta locations available.

Spatial analysis

Below we outline details of the spatial analysis of these two datasets. The printed article contains limited R code, however all source code for reproducing the figures and the tables included in the article can be found in the GitHub repository for the article.

Identifying nearest neighbors

In order to perform an analysis on the relationship between Denny's and La Quinta location, it is necessary to first identify all possible Denny's / La Quinta pairs and their distances from one another. However the analysis will focus only on the pairs that are the nearest neighbors. It is important to note that this subset is made up of two sets of pairs that are similar but not identical:

1. The set of pairs created by finding the La Quinta that is closest to any given Denny's location.
2. The set of pairs created by finding the Denny's closes to any given La Quinta location.

For the provided data, the first set contains 1643 pairs and the second set contains 851 pairs, which are the number of locations in each dataset, respectively. It may not be initially obvious to the students that these two sets are different and why. One can show simple graphical examples with two locations of each chain to illustrate this point, such as in Figure 1.

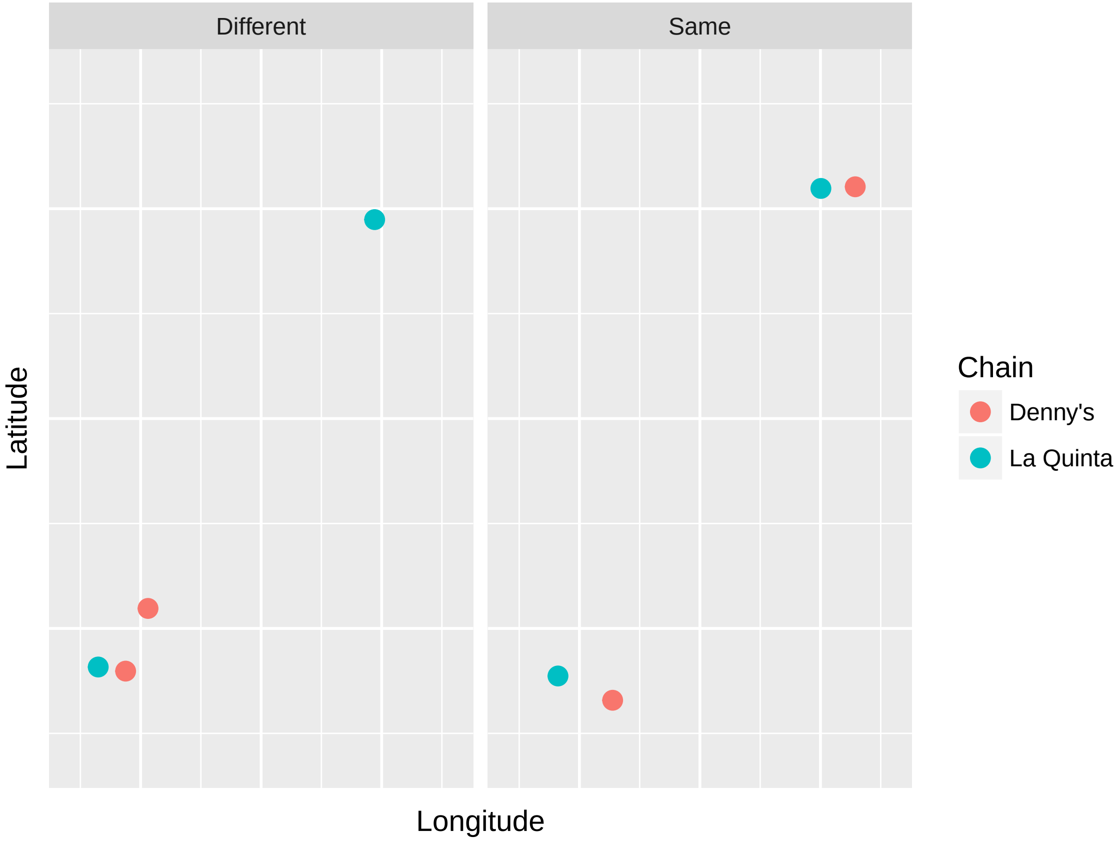

Figure 1: Figure 1: Illustration of pairs of points with different and same differences between the two chains.

In Figure 1 the left-hand panel shows a setting where the Denny's-La Quinta pairs will have short distances (both Denny's are nearest to the same La Quinta) while the La Quinta-Denny's pairs will have one short and one long distance. Conversely, in the case of the right-hand panel, both the Denny's-La Quinta pairs and the La Quinta-Denny's pairs will have the identical distances.

The most straight forward way for identifying these pairs involves first constructing a pairwise distance matrix between the Denny's and La Quinta locations. This involves nested iterations through each Denny's location and La Quinta location producing 1643 x 851 unique location pairs. For each of these pairs we calculate a distance for and store the results in a matrix, with 1643 rows and 851 columns. There are a couple of important considerations for this particular step, perhaps the most important being the choice of the distance function to use. Most students are likely to naively use a Euclidean distance for these calculations - this is an incorrect choice since ignores the fact that latitude/longitude are coordinates for points on the earth's surface, which is a sphere and not a plane. As such, the length of one unit of longitude varies depending on the latitude at which you are making the measurement.

This particular aspect of the analysis provides an excellent opportunity to discuss the idea of coordinate systems and the dependence of fundamental measures like distance and area on this choice. Possible solutions include adopting a distance function that is appropriate for locations of a sphere or reprojecting our locations into a new coordinate system where Euclidean distance is appropriate. For the sake of simplicity, we suggest the former. The haversine distance formula, shown below, which is a reliable way of calculating the great circle distance is simple enough for the students to implement themselves.

$$d = R \times 2 \arcsin \sqrt{\sin^2\left(\frac{t_2 - t_1}{2}\right) + \cos(g_1)\cos(g_2)\sin^2\left(\frac{g_2 - g_1}{2}\right)}$$

This calculation also offers an opportunity to discuss optional topics like optimization and profiling. For example, in R, students will often opt to create the distance matrix using nested for-loops, where each element is calculated by separate calls to the haversine function. This particular approach, and most approaches using for loops in R, tend to be quite slow. Instead, the haversine formula can be vectorized such that the distance from each Denny's location is calculated for all of the La Quinta locations at the same time, thereby calculating a whole row.

Once the distance matrix is calculated, the next task is to identify the nearest La Quinta for each Denny's and the nearest Denny's for each La Quinta. This can be accomplished by finding either the row-wise minimum distances or column-wise minimum distances, respectively. In R this is easily achieved using the `apply` function, for example `apply(dist, 1, min)` will return a vector of length 1643 with the distance from each Denny's to the nearest La Quinta. By changing `1` to `2` in the second argument we instead get a vector of length 851 containing the distance from each La Quinta to the nearest Denny's. Note that the reported distances will have units that match the units used for the radius of the earth in the haversine formula, which is kilometers for our implementation. If instead of the minimum distance between pairs we wanted to know which particular locations of Denny's and La Quinta are closest to each other, we can use `which.min` instead of `min` in the `apply` function. This will return the row or column index of the minimum distance.

We usually recommend that our students work with data frames in R whenever possible. This is a good place to combine information from the two original location datasets and the distance matrix into two data frames summarizing the Denny's-La Quinta and La Quinta-Denny's nearest neighbor pairs. The resulting data frames should look something like the following for the Denny's-La Quinta pairs.

dist	dn.Address	dn.City	dn.State	lq.Address	lq.City	lq.State
0.011	607 Avenue Q	Lubbock	TX	601 Ave Q	Lubbock	TX
0.014	160 N 10th Street	Fowler	CA	190 N 10th Street	Fowler	CA
0.014	3321 Milan Road	Sandusky	OH	3304 Milan Rd	Sandusky	OH
0.019	5910 Veterans	Metairie	LA	5900 Veterans Memorial Blvd	Metairie	LA
0.022	2801 No Black Canyon	Phoenix	AZ	2725 North Black Canyon Hwy	Phoenix	AZ
0.022	3026 Washington Rd	Augusta	GA	3020 Washington Rd	Augusta	GA
0.023	1110 So 10th St	Mcallen	TX	1100 South 10th St	McAllen	TX

Table 1: Denny's-La Quinta nearest neighbor pairs. Note that zip code, longitude, and latitude have been omitted for the sake of space and the results are ordered by distance.

The first 7 rows are identical in the La Quinta-Denny's table (not shown), and it is not until the 70th row (after sorting by distance) that the two tables diverge. We can spot check either data frame using Google Street View. For example, the Lubbox, TX pair can be seen in Figure 2 below and at https://goo.gl/maps/yqHSnApXFdz.

Figure 2: Lubbock, TX - Denny's and La Quinta pair street view.

Exploring the distance distributions

Our next goal is to understand the distribution of these distances. A quick look at numerical summaries can provide some preliminary insight into the data:

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
DN-LQ	0.011	3.962	10.435	42.80868	29.237	3867.907
LQ-DN	0.011	1.307	5.019	18.29447	17.092	281.586

Table 2: Summary of Denny's-La Quinta and La Quinta-Denny's distances.

What stands out is that the maximum of the Denny's-La Quinta distances is huge, 3868 km, compared to the more reasonable 281.6 km for the La Quinta-Denny's distances. Why is this? Plotting the locations of both chains on a map reveals that while Denny's has locations in both Alaska and Hawaii La Quinta is only within the continental United States. It is important to note that if we have only looked at the La Quinta-Denny's distance we might have missed this issue completely. Thankfully, there are only a handful of locations in these two states outside of the continental United States, therefore excluding them only reduces number of observations by 9 for the Denny's-La Quinta pairs, and 0 for the La Quinta-Denny's pairs

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
DN-LQ (CONUS)	0.011	3.93825	10.3285	24.86966	28.31075	235.918
LQ-DN (CONUS)	0.011	1.30700	5.0190	18.29447	17.09200	281.586

Table 3: Summary of Denny's-La Quinta and La Quinta-Denny's distances for the continental US.

In the blog post John Reiser was able to find several pairs of Denny's and La Quintas that are next to each other, and hence partially "confirm" the joke. We can use our data in the same way, but as good statisticians, we like to look at the entire distribution. We like to leave this as an open ended task for the students, however two good examples of this type of analysis are density estimation of the distribution and empirical cumulative distributions. From these we can answer questions like how many Denny's are within 50 meters of a La Quinta. which happens to be 13. We can just as easily ask how many are within 1 kilometer, which happens to be 189 for Denny's to La Quinta, and 188 for La Quinta to Denny's. We can also easily address quantiles of the distance distribution, which tell us that 50% of Denny's have a La Quinta within 5 kilometers, which is also true of La Quintas.

Data collection

So far in this article we have focused on the data analysis aspects of this assignment. Typically this has been a smaller secondary aspect of this exercise, and the usual focus is instead on scraping and collecting the data directly from Denny's and La Quinta's websites. The choice of these two specific chains is driven by the Mitch Hedberg joke, but it turns out that they are particularly well suited for introducing web based data collection/scraping since the location data for each chain is available through representative and distinct interfaces.

Denny's location information is provided by a third party service, Where2GetIt, that uses an undocumented XML API. In order to obtain data from this service it is necessary that students understand the query format - including obtaining an individual API key, finding valid parameter ranges, and handling errors and other unexpected behavior. Once they master the API they then need to be able to parse and organize the XML data into a workable format for their later analyses.

For La Quinta John Reiser was able to locate a javascript file that contained latitude and longitudes for all La Quinta locations. We differ slightly in that we ask students to instead use La Quinta's hotel listing page

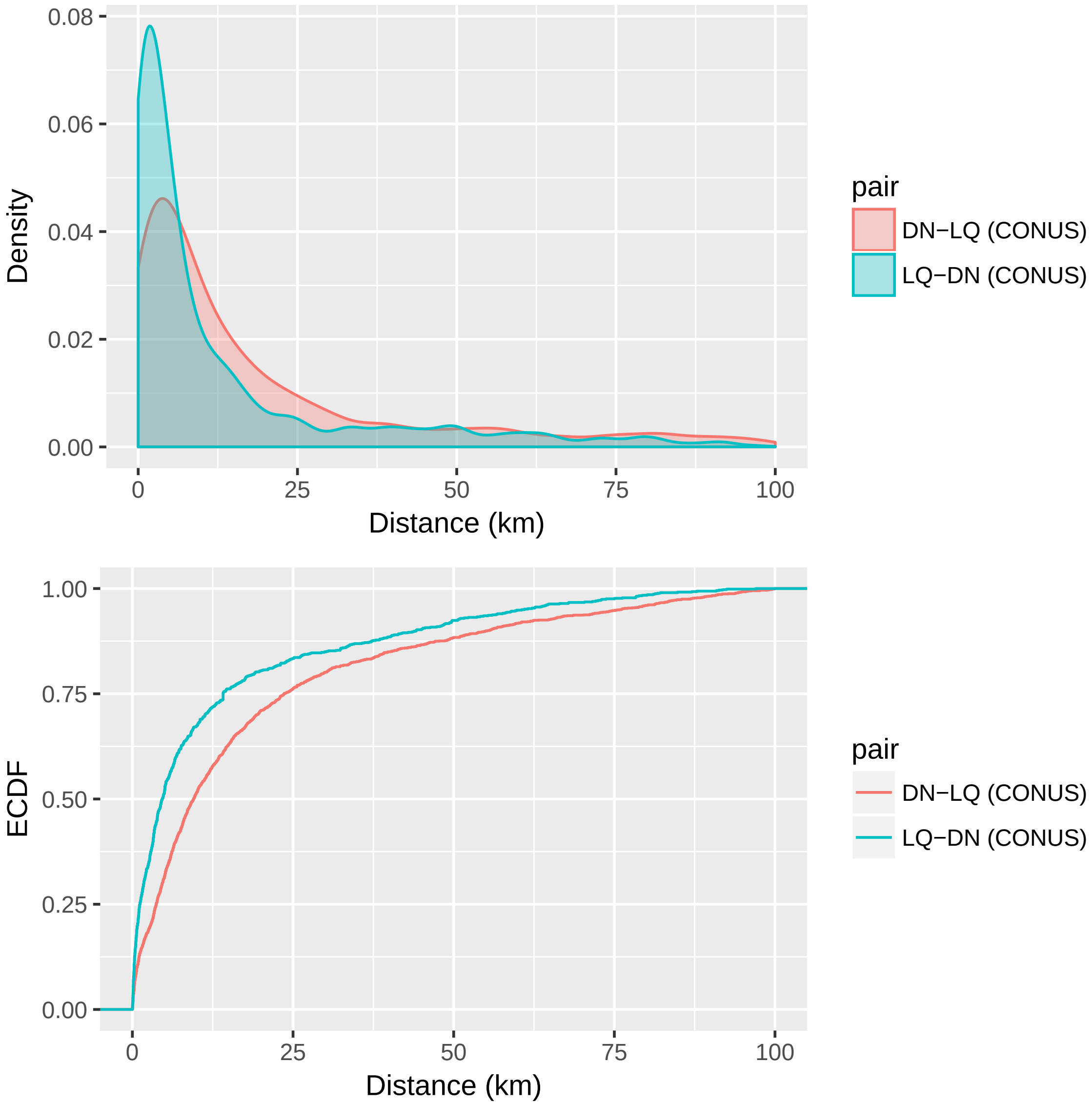

Figure 2: Figure 3: Distribution of Denny's-La Quinta and La Quinta-Denny's distances for the continental US.

(http://www.lq.com/en/findandbook/hotel-listings.html). Students scrape each location's link from the main hotel listings page, and then visit each sub-page individually where they then scrape location specific details such as name, address, amenities. In looking at the sub-pages it may not be obvious where latitude and longitude data can be obtained, but a closer examination of the page's HTML source code reveals that the embedded Google Map image contains this critical detail, and hence the latitude and longitude can also be extracted as part of the scraping. Since the pages are consistently structured, students can create a generic function for scraping this information, and apply it over all sub-pages.

In this way, the assignment exposes students to both web APIs as well as well structured HTML to obtain the necessary data for the sunsequent analysis task.

Conclusion

In the introduction we listed a few reasons why we like this application. Another, very important, reason is that it is a great way of bridging modern data analysis skills with fundamentals of statistical science. The gist of this data story is that thinking about distributions is a lot more informative than just focusing on point estimates, and this is a good lesson for students to encounter in a statistics course at any level. The assignment can be given as a whole (from data scraping to data analysis) in a higher level computational course or can be broken into smaller parts for a lower level course where the students would benefit from completing the tasks with some check points along the way.

There is a lack of emphasis on spatial data in traditional statistics curricula, and this application helps address this issue as well. Lots of interesting problems rely on spatial data, and additionally, people generally like maps. However working with type of data is not trivial. While the tools are getting better, students still need to have a good grounding to be proficient in saying something interesting about this type of data.

The focus of this particular application may not be of interest to many (distances between Denny's and La Quinta locations), however the general skills of (1) data scraping, (2) mapping, (3) calculating spatial distances, (4) finding minima/maxima, (5) describing distributions, and (6) drawing meaningful conclusions about real data based on results of statistical analysis are widely applicable.

Note

You can the R Markdown file for this article, including all source code for the figures and tables at https://github.com/rundel/Dennys_LaQuinta_Chance.

Further reading

- J Reiser. (2014, Jan 30). new jersey geographer: Mitch Hedberg and GIS. Retrieved from http://njgeo.org/2014/01/30/mitch-hedberg-and-gis/.
- What Is ASA DataFest? http://www.amstat.org/education/datafest/.