BZAN 535 – HW#2 – Due September 7 2016

The purpose of this assignment is to help you become acquainted the dunnhumby data, to begin practice using JMP, and to continue practice with mysql.

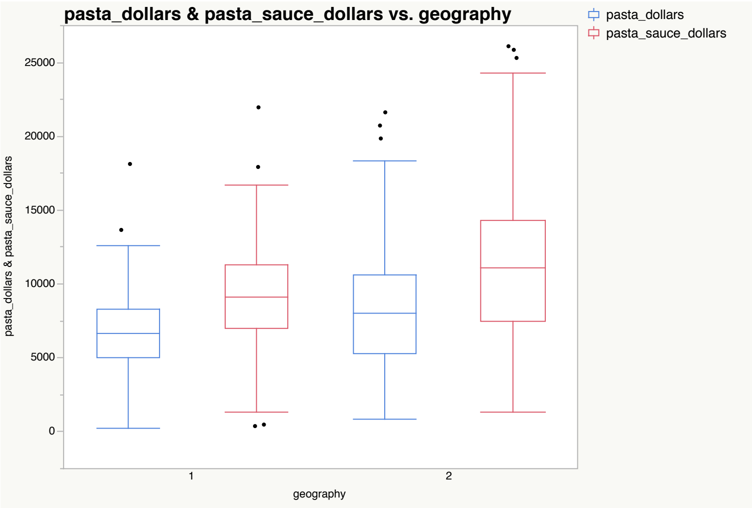
1. CARBO: Using the dh carbo data, compute dollar sales per store for pasta sauce and separately the dollar sales for pasta.
   a) Present the SQL query you used for this calculation.
   b) Using JMP's Graph Builder, create side-by-side box plots of these totals per store, separated by geography. (You will have 4 box plots.) Copy the graph and comment on what insights this graph provides.

1a

```
SELECT geography, store,
sum(case when commodity = 'pasta' then dollars else 0 end) pasta_dollars,
sum(case when commodity = 'pasta sauce' then dollars else 0 end)
pasta_sauce_dollars from
(Select geography, store, upc, sum(dollar_sales) as dollars from carbo_transactions
GROUP BY geography, store, upc) T
JOIN
carbo_product_lookup P
ON T.upc=P.upc
GROUP BY geography, store;
```

Your query can be different and still be correct. It can produce twice as many lines, with one column for Sales $ and another column identifying commodity.

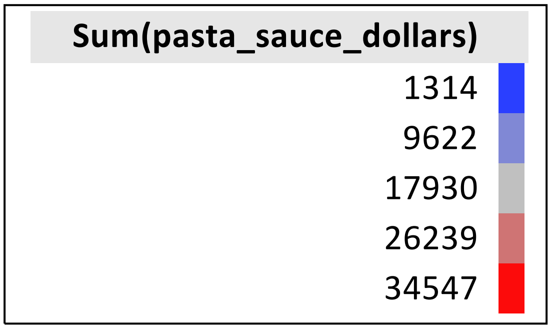
1b

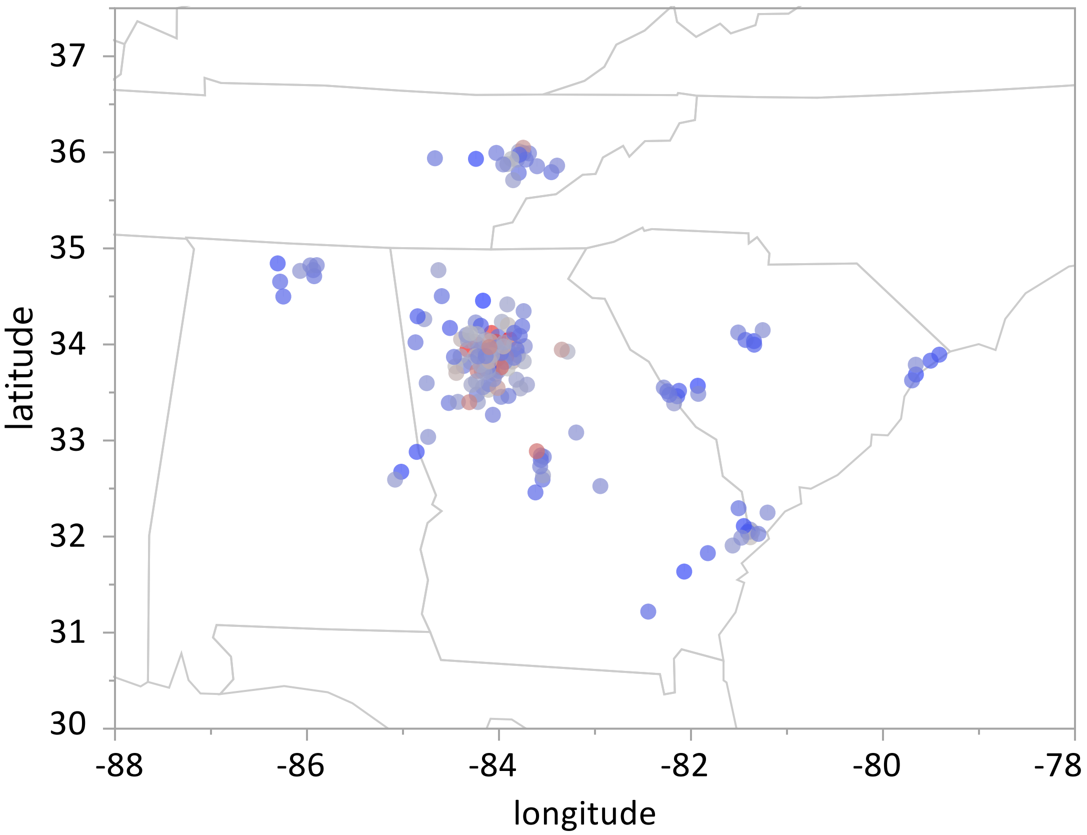


The distributions are nearly symmetric. Pasta_sauce dollars are greater than pasta_dollars. And geography 2 stores sell more than geography 1 stores on a per-store basis.

2. CARBO: The stores differ in size and location. In JMP, join the data from question 1 to the store_lookup table so that you have the zipcode for each store. Using JMP, create a map to display the total pasta sauce sales for each zipcode. Distinguish between Geography 1 and Geography 2, either with separate maps or by some other means of your choice. Are the differences in sales due to regional (or urban vs. rural) differences?

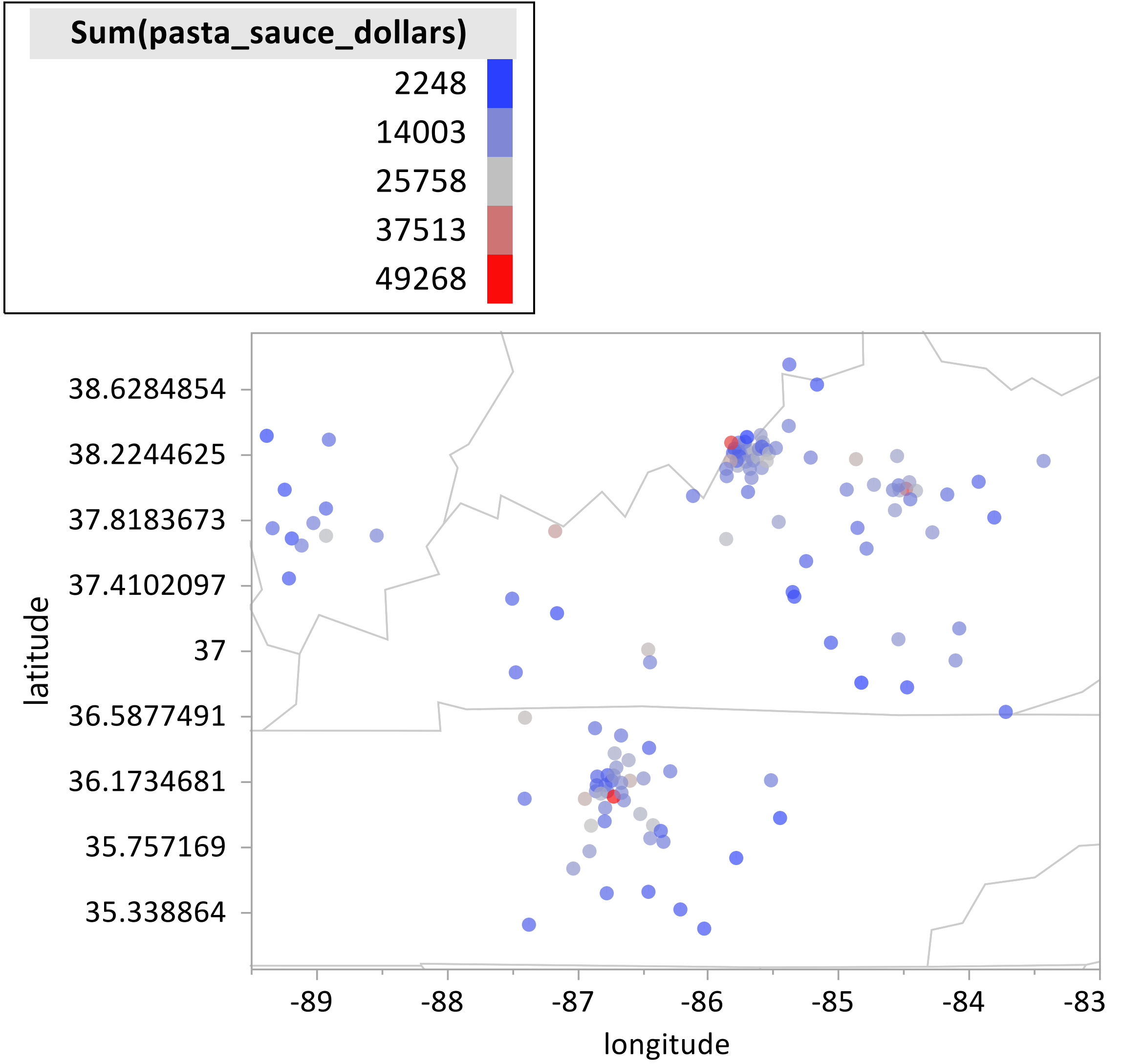
Geography 1:





For the most part, the highest sales zipcodes are for suburban areas, not smaller towns. Note the red around Atlanta, Knoxville, and in the map below around Louisville and Nashville.

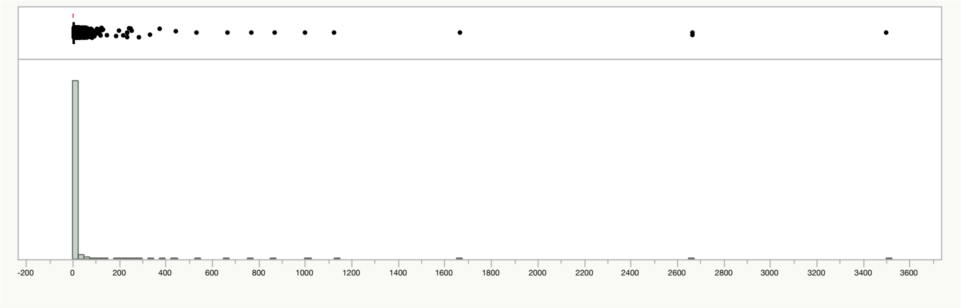
Geography 2:



3. Slide 20 from Lecture 5 contains a query that computes the number of orders per household by zipcode. Multiply this number by 1000 to get the number of orders per 1000 households.
a) Examine the distribution of penetration. Describe the distribution in terms of its histogram and box plot. What are the values for the 25th, 50th, and 75th percentiles?

Using the query from Lecture 5, we obtain a total of 185418 orders from 11537 zipcodes. The orders table contains 192983 orders; we hope that these 7565 orders that were dropped were for international and overseas military addresses.

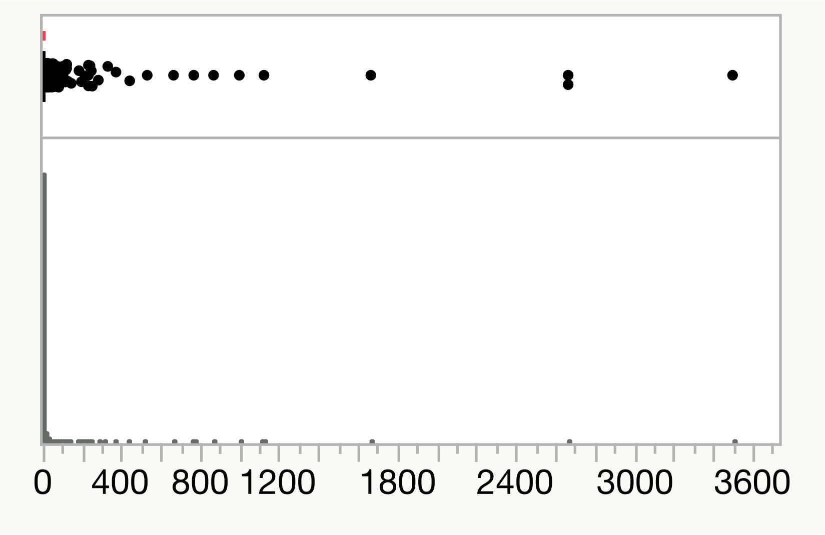
The quartiles are 0.39, 0.98, and 2.79, respectively. Given that the maximum is 3500, the original histogram and box plot are of little use.



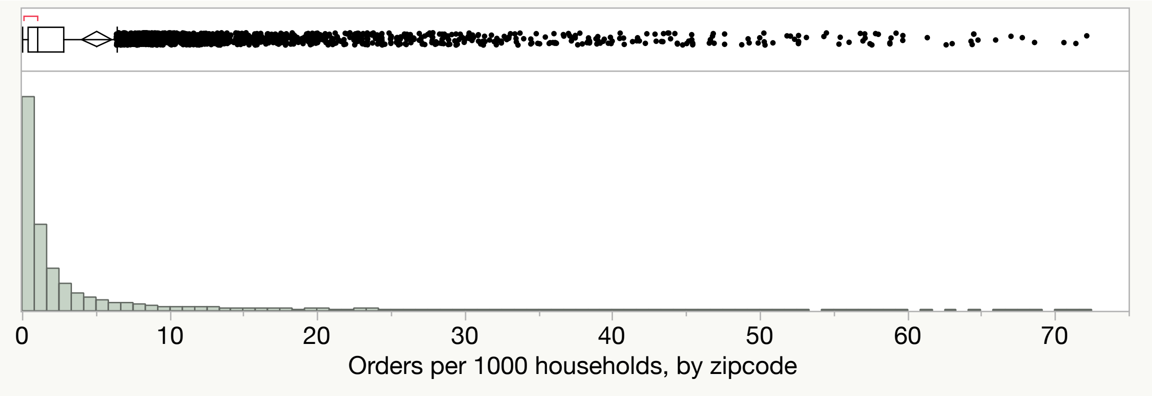
There is actually a problem with this answer in that it includes some orders that are international, which have five digit zipcodes but no state. An improved answer is obtained using the following query:

```
use rmee_sqlbook;
SELECT zc.zipcode, longitude, latitude, numords,
(CASE WHEN hh = 0 THEN 0.0 ELSE numords*1000/hh END) as penetration
FROM zipcensus zc  JOIN
(SELECT zipcode, COUNT(*) as numords      FROM orders  WHERE state NOT IN (' ')
GROUP BY zipcode) o
ON zc.zipcode = o.zipcode ;
```

With this inner join, excluding orders with no state listed, we get 11,499 zipcodes. The quartiles are 0.3888, 0.9823, and 2.7733. Note that the 75th percentile is slightly smaller.



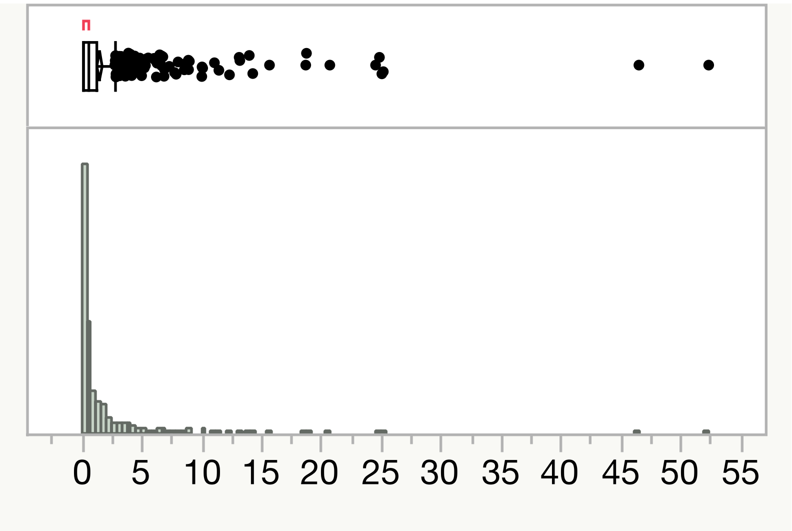
If we exclude outliers above 75, we can see more detail (see next plot).

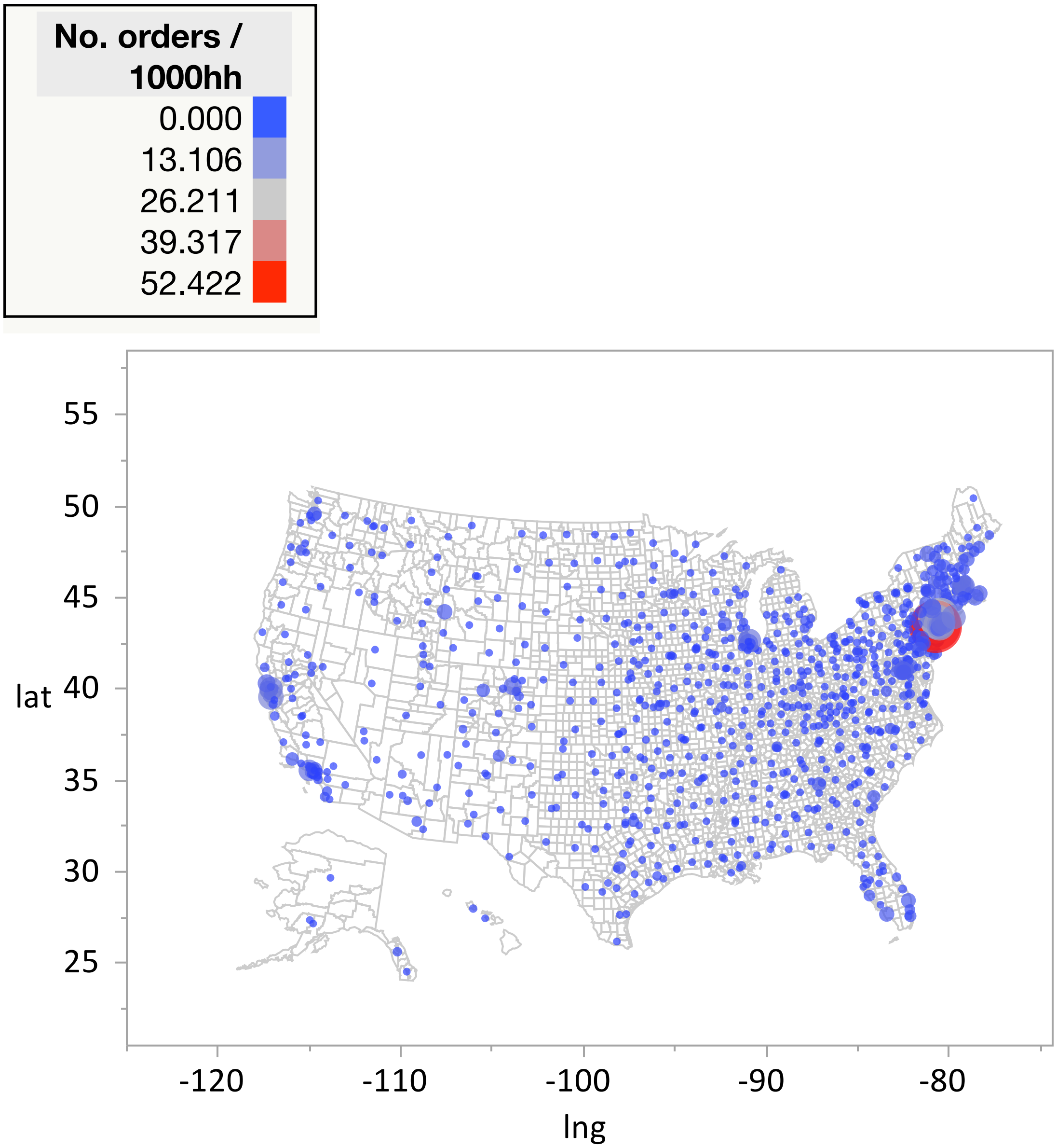


b) The analysis in 3a used zipcode as the geographic unit. Repeat the computation, using scf instead (that is, group by the first 3 digits of zipcode).

```
use rmee_sqlbook;
SELECT z.scf, z.lat, z.lng, z.nhh, o.norders from
(SELECT substr(zipcode,1,3) as scf, avg(latitude) as lat, avg(longitude) as lng,
sum(hh) as nhh from zipcensus
group by 1) z
left join
(SELECT substr(zipcode,1,3) as scf, count(*) as norders from orders
where state not in (' ')
group by 1) o
on z.scf=o.scf;
```

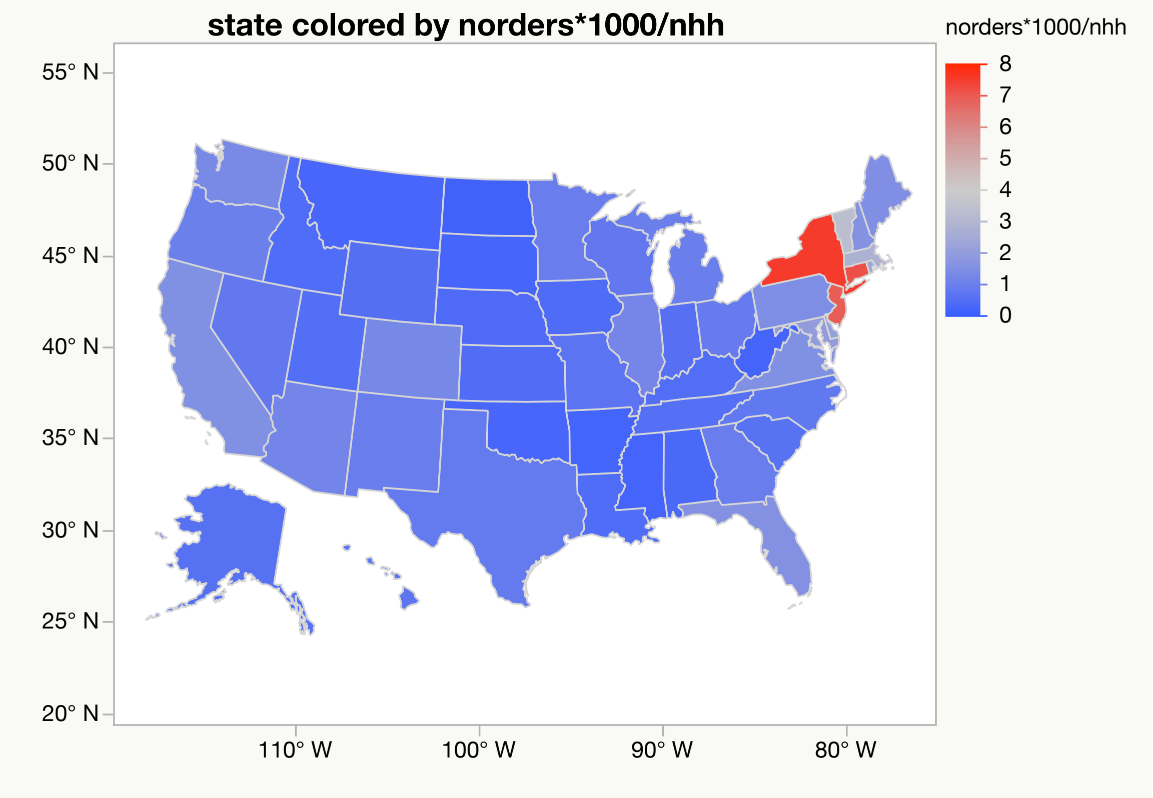
The correct result has 34 zeros for scf and the following legend for color. There are only two values above 30. The quartiles are 0.18, 0.44, and 1.22.





This is still too skewed to be useful, except one can see the outliers in NY and that the Northeast is denser. It is useful to zoom in on the NE.

c) Now compute orders per 1000 households by state and produce a map.

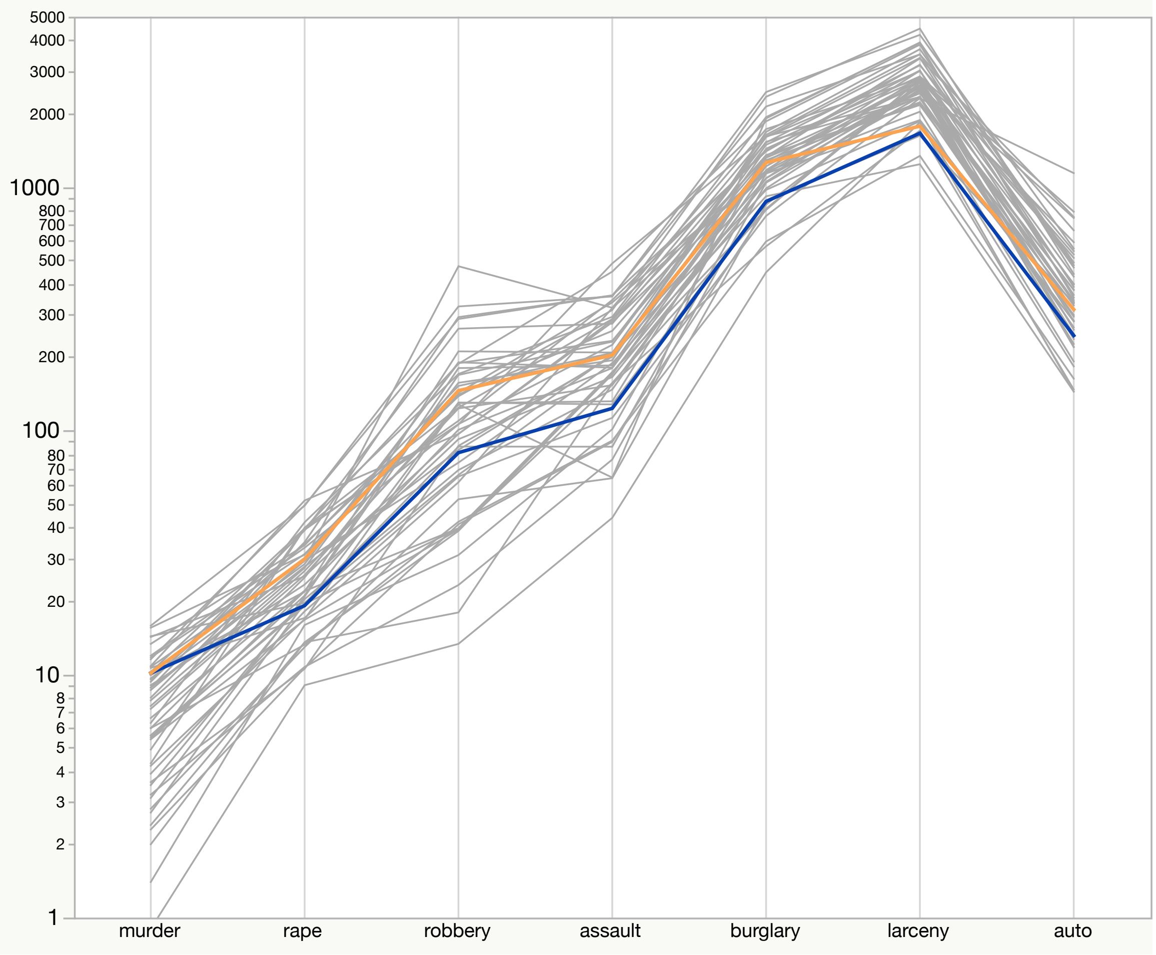


d) Discuss how you would present insights from these analyses in a meeting of this company. What analysis would you use (a, b, and/or c) and in what order?

Show the US map by states first. Then SCFs, zooming in on the northeast, to see which areas are high. The individual zipcode map seems too granular, although the Tableau version with filled map could be used to study a small geographic area.

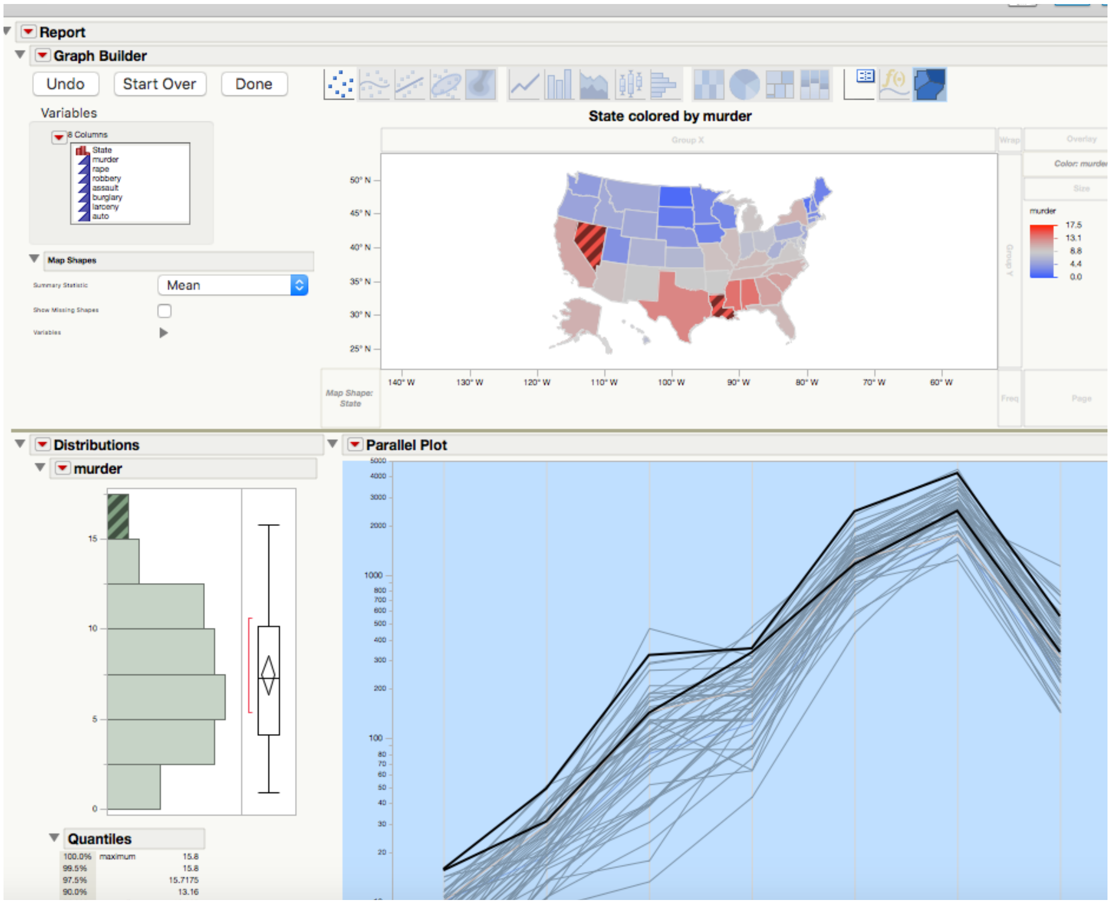
4. In Lecture 5, you were introduced to a parallel plot using the file crime.jmp. Choose and highlight two states, display a parallel plot and discuss what you learn from this comparison. (Also, explain why you thought a comparison of those two states would be of interest.)

I chose TN and KY, since they are neighbors. Kentucky has lower crime rates for all crimes except for murder. TN is much higher for robbery and assault.



5. In Lecture 5, we joined windows analyzing burglary rates to create an application. Create an application with a state map and a histogram using a crime other than burglary. Include a parallel plot as well. Highlight either the low end of the high end for that rate and take a screen shot showing your application. Provide a brief discussion of insights based on the graph.

   Louisiana and Nevada have the highest murder rates. Nevada is close to highest of several other crime rates as well, whereas Louisiana is generally not near the highest.



Summary Statistic: Mean

Quantiles

100.0%	maximum	15.8
99.5%		15.8
97.5%		15.7175
90.0%		13.16