BZAN 535 - Homework 2

Kevin Gardner

September 7, 2016

Problem 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.

part a.

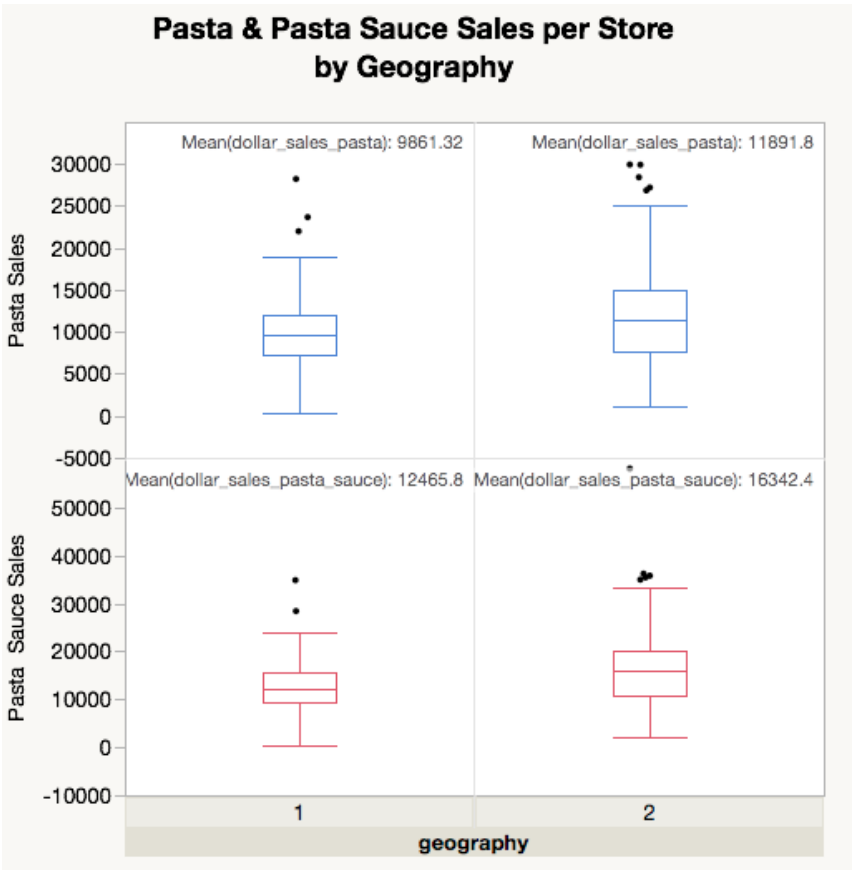
```
select t1.dollar_sales_pasta_sauce, t2.dollar_sales_pasta, t1.store, t1.geography
from
    # dollar sales per store for pasta sauce
    (select sum(dollar_sales*units) as dollar_sales_pasta_sauce, store, geography
    from carbo_transactions
    where upc IN (select distinct upc from carbo_product_lookup where commodity = 'pasta sauce')
    group by store) as t1

    # dollar sales per store for pasta
    , (select sum(dollar_sales*units) as dollar_sales_pasta, store, geography
    from carbo_transactions
    where upc IN (select distinct upc from carbo_product_lookup where commodity = 'pasta')
    group by store) as t2

where t1.store = t2.store and t1.geography = t2.geography;
```

part b.

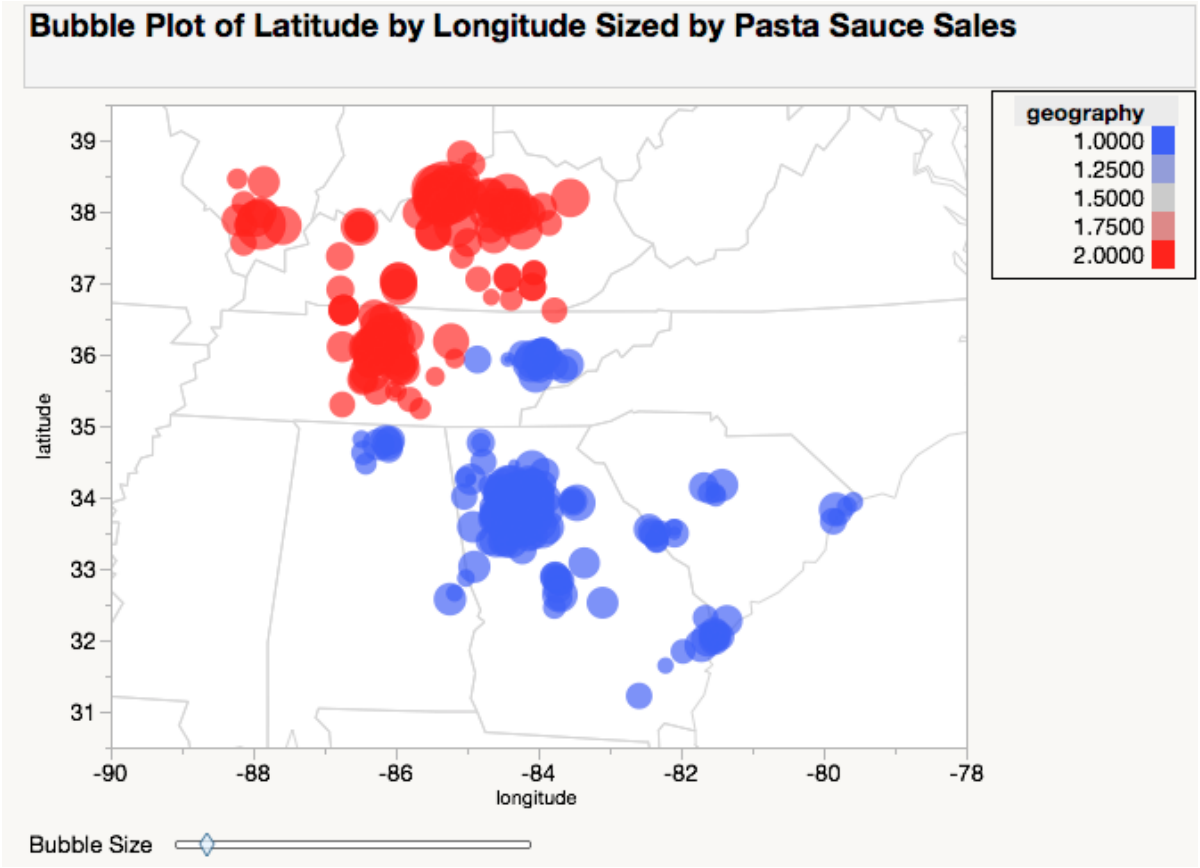
From the graph we see there are a handful of stores which sell a lot more pasta and pasta sauce than the rest of stores. These observations are likely having a strong influence on the mean. In regards to geography, we observe that sales are higher for both products in geography `2` than in geography `1`.



Problem 2

> Join the data from question 1 to the store_lookup table so that you have the zipcode for each store and 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. Are the differences in sales due to regional (or urban vs. rural) differences?

From the graph we see there are clusters of high selling stores near urban areas, and that the geographic difference noted in part 1 can be attributed to stores being further North, perhaps because they are closer to Cincinati where Kroger's is headquartered.



Problem 3

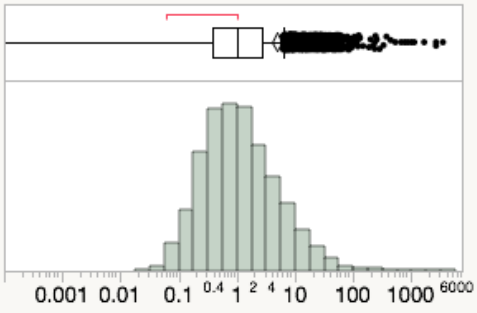
Slide 20 from Lecture 5 contains a query that computes the number of orders per household by zipcode. Multiply this number by 1000 to get the number of orders per 1000 households. a) Examine the distribution of penetration. Describe the distribution in terms of its histogram and box plot. What are the values for the 25th, 50th, and 75th percentiles? b) The analysis in 4a used zipcode as the geographic unit. Repeat the computation, using scf instead (that is, group by the first 3 digits of zipcode). 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?

part a.

From the figure we see the values for the `25th`, `50th` and `75th` percentile are 0.38975, 0.9846 and 2.78775, respectively.

Distributions

penetration



Quantiles		
100.0%	maximum	3500
99.5%		64.429892
97.5%		24.929035
90.0%		7.82602
75.0%	quartile	2.78775
50.0%	median	0.9845
25.0%	quartile	0.38975
10.0%		0.19428
2.5%		0.0915
0.5%		0.040969
0.0%	minimum	0

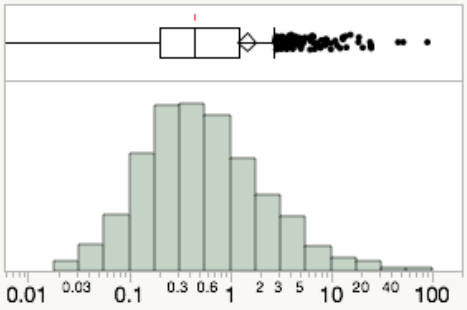
Summary Statistics	
Mean	4.9978108
Std Dev	55.231007
Std Err Mean	0.5142053
Upper 95% Mean	6.0057403
Lower 95% Mean	3.9898812
N	11537

part b.

From the figure we see the values for the `25th`, `50th` and `75th` percentile are 0.1974, 0.443 and 1.2324, respectively.

Distributions

penetration

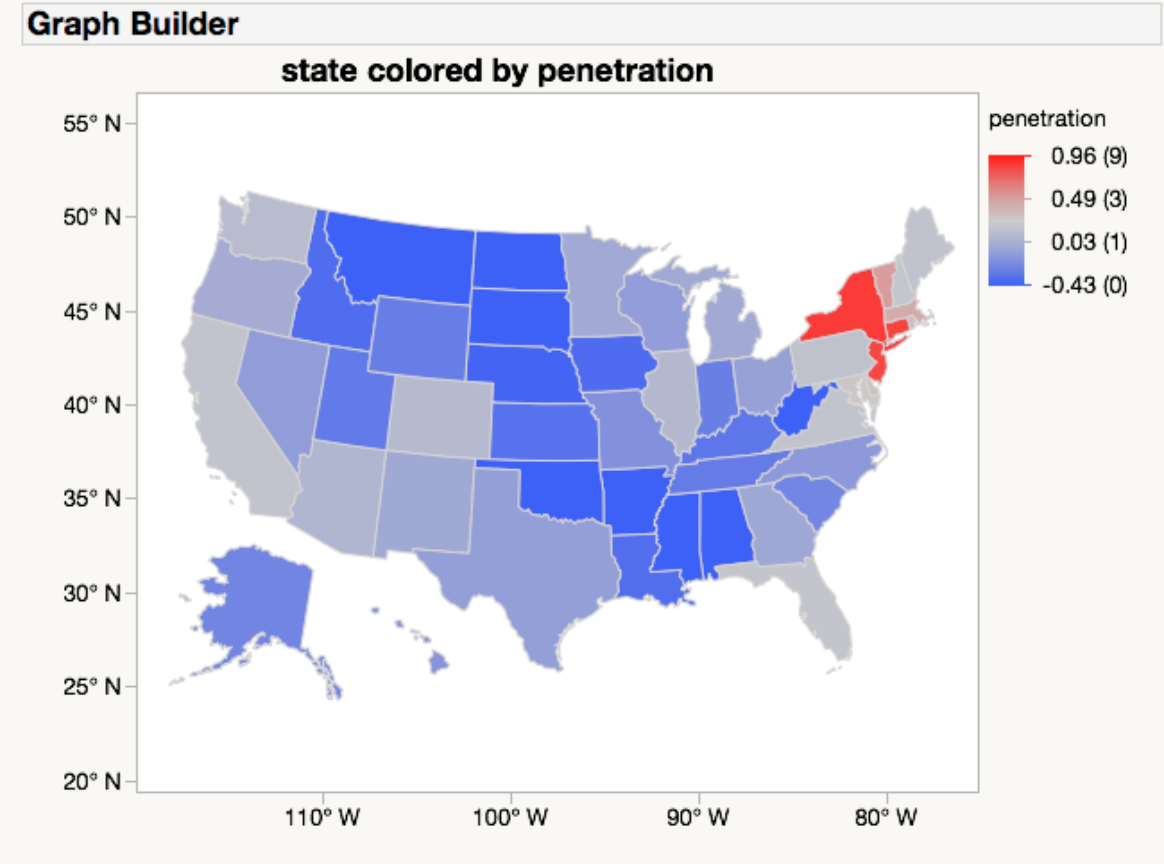


Quantiles		
100.0%	maximum	90.9091
99.5%		25.188236
97.5%		9.11144
90.0%		3.17742
75.0%	quartile	1.2324
50.0%	median	0.443
25.0%	quartile	0.1974
10.0%		0.08918
2.5%		0.0041
0.5%		0
0.0%	minimum	0

Summary Statistics	
Mean	1.4995032
Std Dev	4.6493769
Std Err Mean	0.1561108
Upper 95% Mean	1.8058932
Lower 95% Mean	1.1931131
N	887

part c.

Below is a map of U.S. states colored by market penetration.



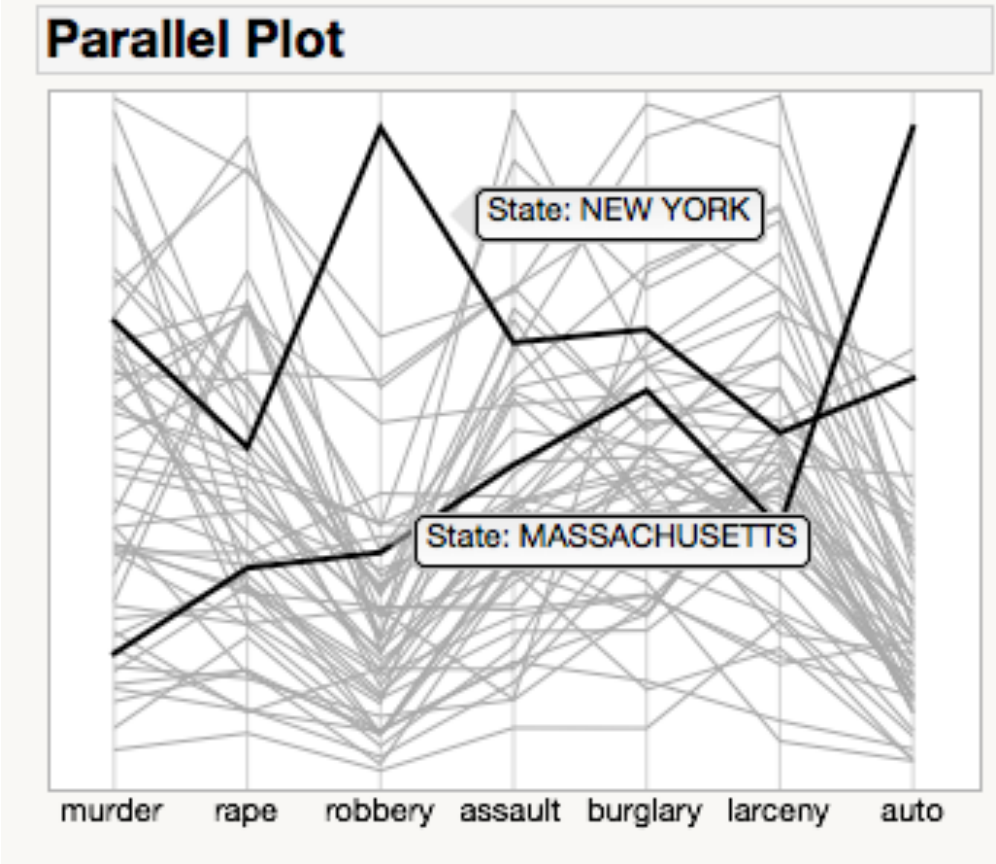
part d.

I would start with analysis c as an overview of market penetration, then use analysis b and perhaps analysis a to provide a more granular view of where efforts should be made to grow market penetration.

Problem 4

> In Lecture 5, you were introduced to a parallel plot using the file crime.jmp. a) 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.)

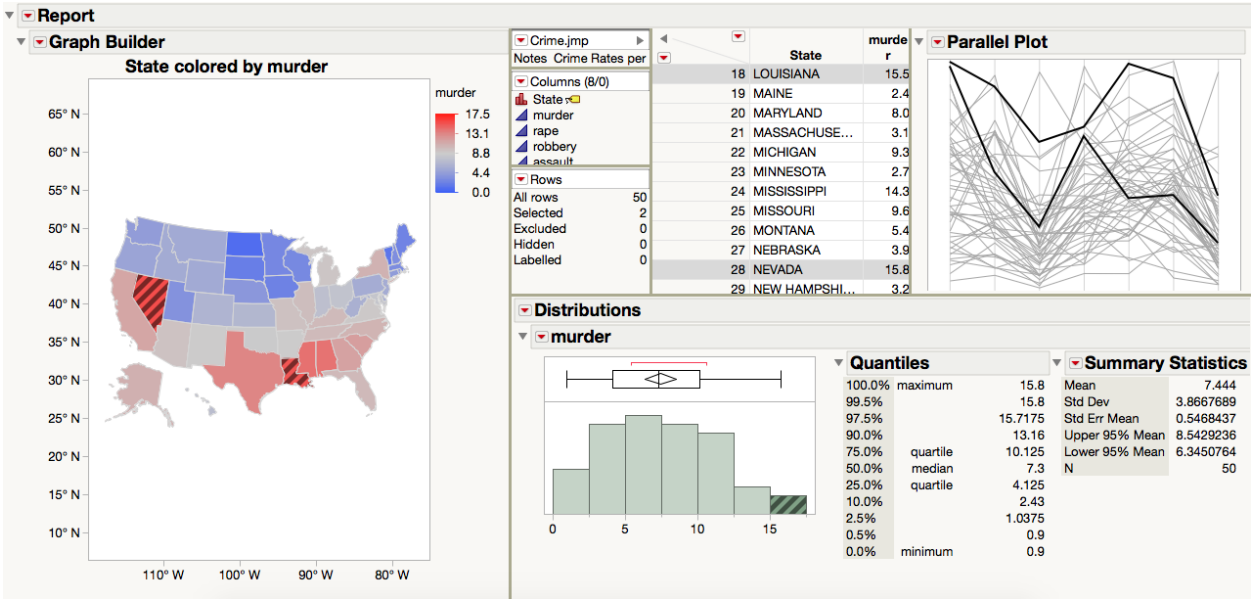
The plot below compares New York and Massachusetts with respect to crime rates. I thought a comparison of the two states would be of interest because both have high population density and are located in the northeast.



Problem 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.

From the graph we see that Louisiana and Nevada have murder rates within the upper tail of the distribution. In the parallell plot we see these two states are nearly off the chart in murder rate, which is represented by the leftmost segment of the line.


Report
Graph Builder
State colored by murder
murder
17.5
13.1
8.8
4.4
0.0
65° N
60° N
55° N
50° N
45° N
40° N
35° N
30° N
25° N
20° N
15° N
10° N
110° W
100° W
90° W
80° W
Crime.jmp
Notes Crime Rates per
Columns (8/0)
State
murder
rape
robbery
assault
Rows
All rows 50
Selected 2
Excluded 0
Hidden 0
Labelled 0
State
murder
18 LOUISIANA 15.5
19 MAINE 2.4
20 MARYLAND 8.0
21 MASSACHUSE... 3.1
22 MICHIGAN 9.3
23 MINNESOTA 2.7
24 MISSISSIPPI 14.3
25 MISSOURI 9.6
26 MONTANA 5.4
27 NEBRASKA 3.9
28 NEVADA 15.8
29 NEW HAMPSHI... 3.2
Parallel Plot
Distributions
murder
0
5
10
15
Quantiles
100.0% maximum 15.8
99.5% 15.8
97.5% 15.7175
90.0% 13.16
75.0% quartile 10.125
50.0% median 7.3
25.0% quartile 4.125
10.0% 2.43
2.5% 1.0375
0.5% 0.9
0.0% minimum 0.9
Summary Statistics
Mean 7.444
Std Dev 3.8667689
Std Err Mean 0.5468437
Upper 95% Mean 8.5429236
Lower 95% Mean 6.3450764
N 50