Springboard Capstone Project - CEI

James Hamilton

March 24, 2016

# The Question

What useful insights can be drawn from the clients' customer and invoice data? - Are some clients "over purchasing" because they don't understand the licensing scheme? - What market segments are the largest / smallest and which market segments are growing / shrinking?

To answer the first question the strategy will be to calculate a median purchase quantity using the quantity attribute in the order data. The median can then be used to compare with specific orders and determine outliers. The sum of the orders of the outliers minus the "expected" buy represents the amount of over purchased dollars.

The second question can be answered by the sum of the orders across an industry segment displayed graphically.

# The Data

The data was provided by the client in the form of a text CSV export from their Salesforce database.

I focused primarily on the Account, Order, and Order Line tables. Invoice data was provided but I found that it essentially duplicated the order data so I ignored it. Some cleanup of the data was needed - a detailing of that activity is included at the end of this report.

### Account

The account data included the customer name and demographic data: location, employee counts, annual revenue and various industry classifiers. After data cleaning there were 2134 rows / observations in the "accounts" table.

*Account.Number* - simply a unique identifier was used to join the account data to the order data  
*Annual.Revenue* - a dollar figure estimate of the customers annual revenue  
*Employees* - estimated number of employees of the customer  
*Industry* - broad categorization of the industry that the customer operates in

A correlation test indicates that the number of employees and Annual revenue are highly correlated (.93 at 99% confidence) across all customers indicating that using employee count alone as an indicator of the "size" of the customer would be accurate.

### Order data

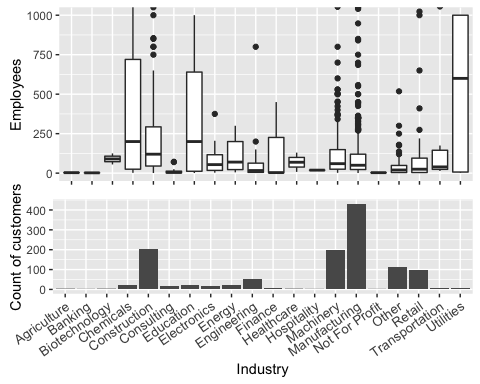
These tables included details about individual orders including unique identifiers, order date, amounts and quantities, discounts applied, terms, status, product identifiers, list price, etc.

*Net.Amount* - Total dollar amount for the order  
*Order.Date* - Date the order was placed  
*Payment.Method* - Cash, check, credit, etc  
*ProductAmount* - Dollar amount charged for each product in the order  
*Quantity* - Number of licenses ordered for the product  
*SalesOrder* - Unique identifier

The account and order data was merged to produce a working data set that maps each line of the order to an account and industry.

Therefore, the following visuals are based on that conclusion and use the employee count as the primary indicator of client size.

Finally, a working data set is generated containing only customers that have an Employee count > 0 - this yields a final account table with 1225 rows.



Primary conclusions here are that the Machinery, Manufacturing and Construction industries represent the majority of the client base and those companies are predominately < 250 employees in size. However, several other industries are represented and more significantly what is missing here are the companies with no Industry designation at all.

Therefore, the remaining plots will focus on the following Industry categories:  
Chemicals, Construction, Consulting, Education, Electronics, Energy, Engineering, Finance, Machinery, Manufacturing, Other, and Retail.

## Industry Employees   
## Manufacturing:432 Min. : 1   
## Construction :203 1st Qu.: 18   
## Machinery :199 Median : 50   
## Other :114 Mean : 414   
## Retail : 96 3rd Qu.: 145   
## Engineering : 52 Max. :168000   
## (Other) :109

Here we see that these industries are represented by companies that are generally in the range of 0 - 145 employees. Specifically, across all industries 75% of all clients have <= 145 employees and 95% of all clients have <= 1000 employees. Essentially, the largest customers make up about 5% of the customer base.

### Orders

This data spanned two tables and included primarily dollar amounts, dates, quantities, payment methods and keys into other tables including the invoices and accounts tables.

A working data frame containing the fields necessary joined from the Accounts, Order and Order Product Line tables was created. Specifically, I needed these elements:  
- SalesOrder (Order, Order Product Line)  
- ProductAmount (Order Product Line)  
- Quantity (Order Product Line)  
- Order.Date (Order)  
- Order.Type (Order)  
- Account.Number (Order, Account)  
- Employees (Account)  
- Industry

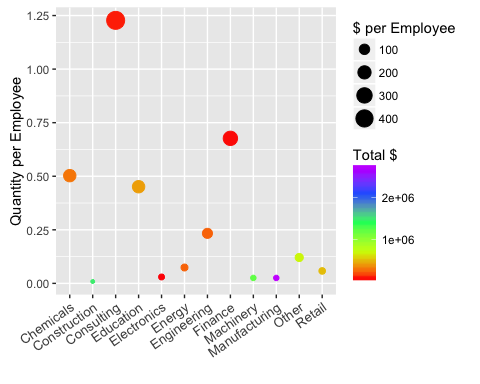
The following plot describes the overall relationship per industry of the median order size (Quantity and $) to size of the customer (# of Employees).

Along the x-axis are the industries, along the y-axis is the mean quantity per median employee count. The mean quantity is used here because the median across all industries is 1 and the range of all values is from .25 - 17, so the magnitude of the outliers is far smaller and should make the mean an acceptible central tendancy value to use in this place. This is not true for the employee count.

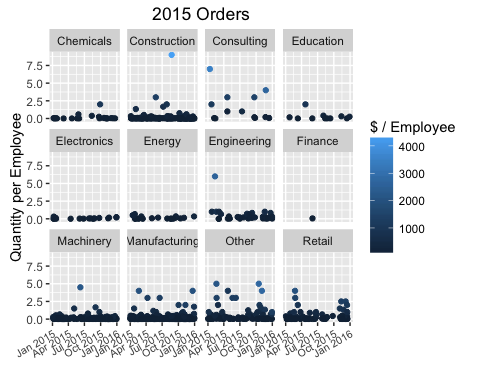
The size of the point shows the dollars spent per employee and the color indicates the total dollars spent by that industry. Cooler colors translates to higher amounts.

Cause for concern would be a large and "cool" point that is high on the y-axis. One point is relatively high on the y-axis but it is "hot" owing to the fact that the Consulting industry is significantly less represented in the client base.

The larger industries - Construction, Machinery, Manufacturing - are very low and "cool."



The next plot displays all orders in 2015 in terms of the dollar and quantity per employee. A large number of floating points on these plots would be concerning but does not appear to be the case.



Looking at these plots it appears that none of the Industries represented have a large group of purchases that are above the overall median of 1. This should be good news because that would suggest that the majority of clients have not overbought.

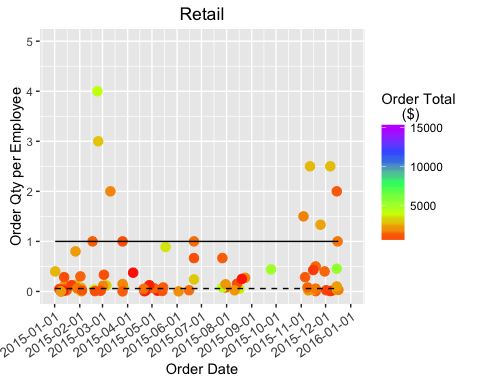
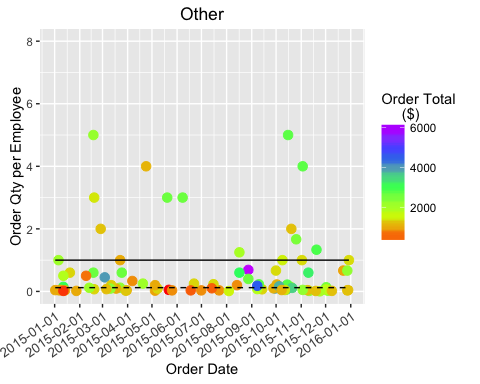
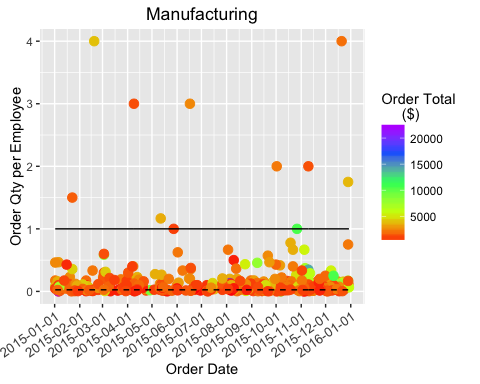
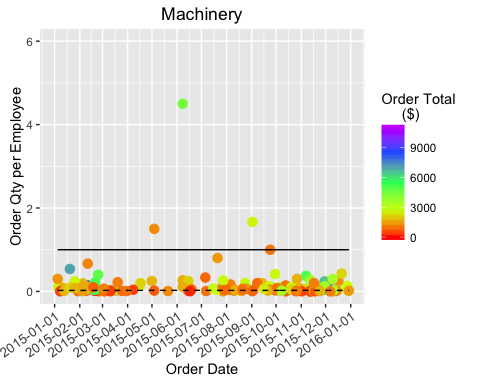
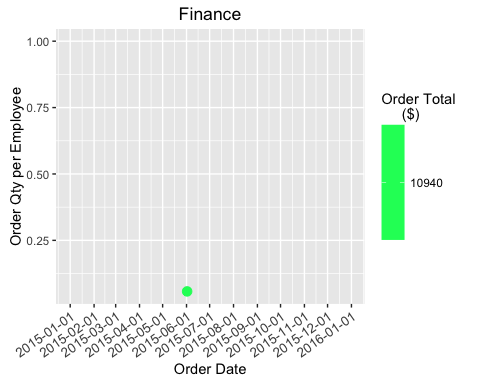
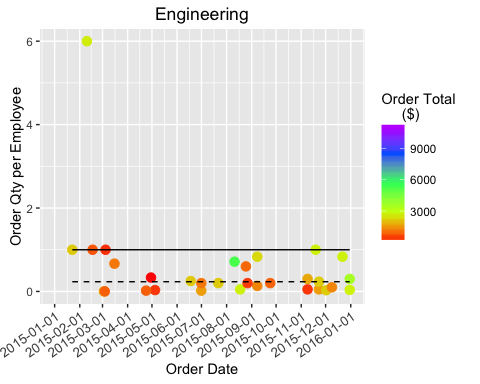
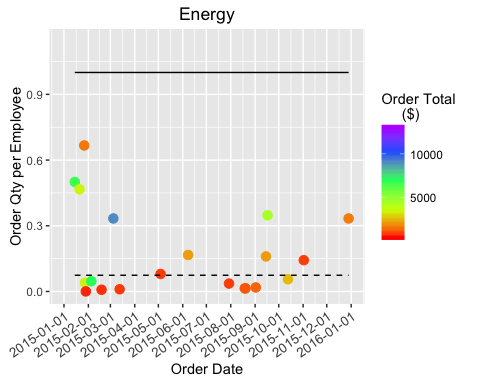
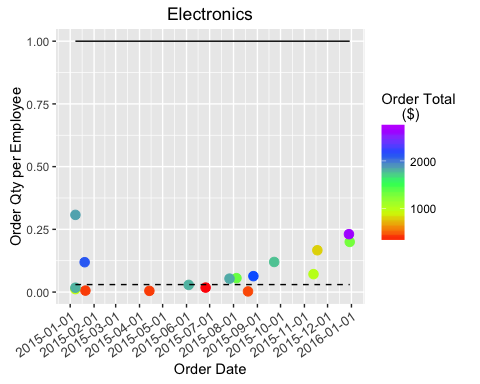
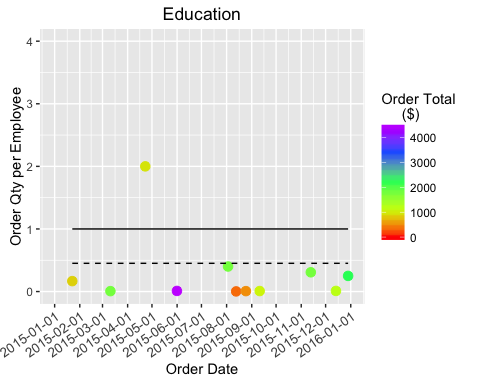
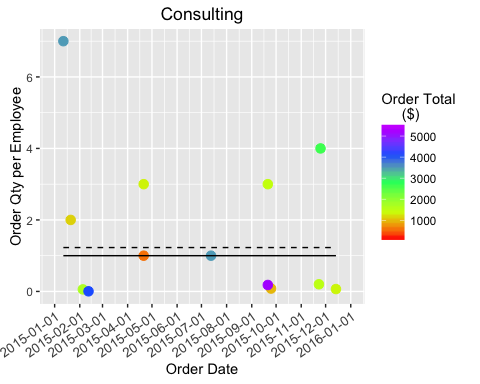
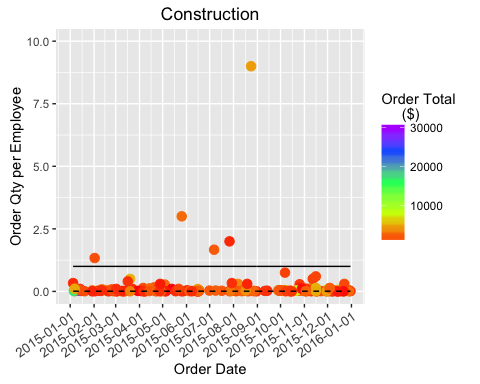
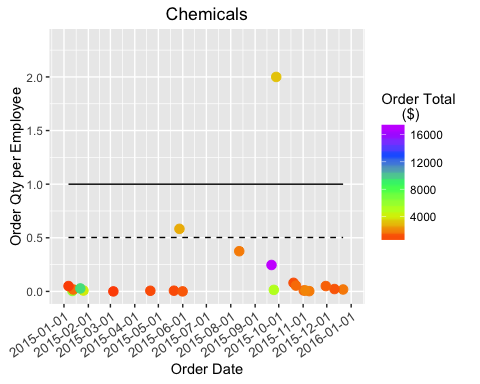
However, the display is fairly crowded and doesn't include the baseline value per industry so I'll show a plot for each industry separately.

The next plots combine serveral bits of information. First, each point is an individual order and is plotted by order date along the x-axis and the quantity of the order normalized by employee count along the y-axis. The total size of the order in dollars governs the color of the point. Larger and "cooler" dots should be cause for concern if they fall above the overall median order quantity.

The dashed horizontal line represents the ratio of the mean order quantity to the median employee count. The solid horizontal line represents the overall median order quantity across all clients.

The expectation is that half of the points would fall below the dashed line and that *most* of the points would fall below the solid line - indicating that most purchases are within expectations if clients are purchasing only what they need.

The data points above the line in all the critical industries are relatively small in number indicating a relatively low risk for those clients about which we have demographic data.



It looks as though for most industries the majority of purchases would be characterized as "correct."  
Consulting is unique probably due to the fact that consultants probably have a higher ratio of employees that use the software compared to other industries.

### How can we quantify the risk?

Specifically, how much of a problem could there be if the accounts that may be over purchased were to normalize their purchasing? How much of a risk is represented by the accounts that were excluded from the plots due to missing demographics?

Calculating the sum of the orders that lie above the overall median Order quantity per employee for the most recent full year should give a decent idea of the overall risk.

Assuming that the rate of "overbuy" will be the same in the whole population of accounts as it is in the accounts for which we know demographics - we can approximate the risk in dollars from last year.

The overbuy rate will be calculated from the sample set with employee data as:

overbuy\_rate = Total order amounts above the overall median / Total orders

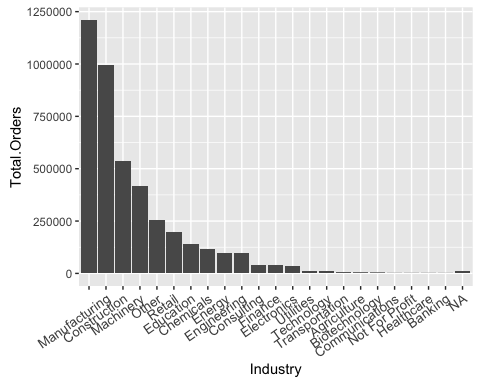
The total amount of dollars at risk across all orders for 2015 will be calculated as the product of the overbuy\_rate and the total orders for 2015.

dollars\_at\_risk = overbuy\_rate \* Total 2015 Orders

### The rate of overbuy in the subset with demographics: 0.0471

### Total dollars at risk from 2015: $199079

# Relative sizes of the industries in dollars



Note here that the empty Industry category is very large but not larger than the sum of the rest of the Industries.

#### Data clean up

* Removed "kugo2p" and "c" strings from field names in the source data to make joins easier.
* Removed "Competitor", "Inspector", "Prospect" and "Reseller" account types from Account table
* Removed extraneous columns from Account & Order & Order Product Line tables
* Removed test items from order table
* Cleaned up address fields in Account table

Much of the clean up was performed using OpenRefine - those operations are available for inspection / replication as a json.