EMORY GOIZUETA BUSINESS SCHOOL

Master of Science in Business Analytics MSBA

ISOM 670 Business Analytics

Sea Watch

Group 5: Carl Xi, Stella Li, Ivy Zhou, Jake Arendsen

Addressing Key Concerns

- We think a model using highly correlated variables including potentially college graduates (ρ=.81), visit number (ρ=.47), per capita income (ρ=.39), and political leaning (ρ≅.57, depending on a particular party) would be the best way for Len to convince the Board to proceed with his plans.. If Len can prove that his model somewhat accurately predicts gross receipts through both past results and field testing, the board will likely trust in his model and projections made from it highlighting worthy places to expand towards. We would like to point out that the factors we chose are good predictors for demographics of any background, and should be suitable for predicting similar North American cities like the one our dataset is based off of. However, significant cultural deviations (e.g. the ones in Barcelona) could cause problems, as a city there who may have similar values as a city in boston in terms of our chosen variables above may have drastically different gross receipt numbers.

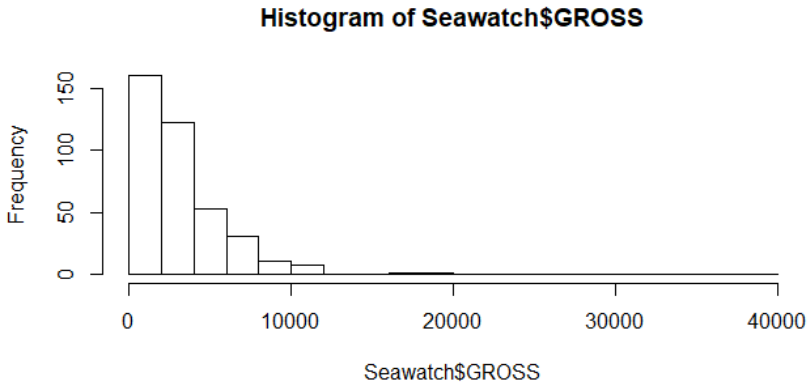
- Regarding repeat visits, we do see a somewhat positive trend between visit number and gross receipts, which by itself would provide some reason to doubt whether towns are truly “burned out”. That being said, common sense tells that it is highly unlikely everybody will give more money on repeat visits, so these areas can likely be viewed as prime areas to increase efficiency. For example, if the towns are driven by few, large scale donators, these households could be hit quickly once a year and moved on from.

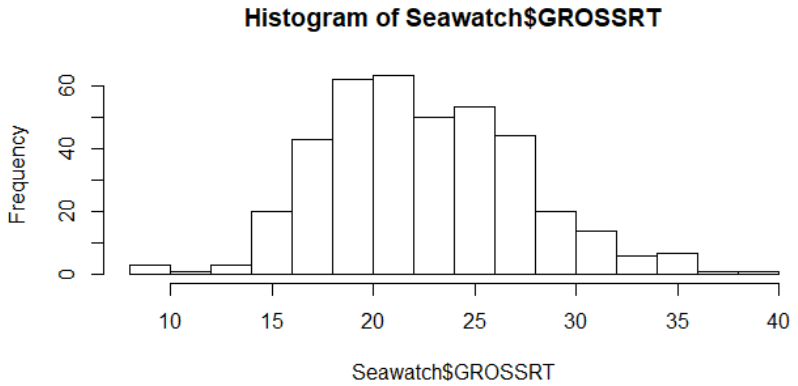
- From our inflation adjusted analysis, we continue to see a strong growth even when adjusted for inflation. It’s likely therefore that a minimal amount of SeaWatch’s growth is due to inflation, which makes sense for a firm that likely relies on many small scale donations. People are unlikely to increase their donations from $10 to $11 to keep up with inflation.

- The ethical/moral dilemma behind the location of the new branch is likely the hardest problem to solve with data alone. Data models are great as long as the status quo during the data collection period is maintained through execution, but the location of this branch could have large social repercussions. Could locating in Greenwich be seen as a purely financial move, harming reputation and decreasing gross receipts? Likewise, could Bridgeport be seen as super environmentally conscious, therefore boosting reputation and gross receipts? Neither of these questions can be answered with any model we generate using Massachusetts data.

- As we are mostly relying on past data from Boston for our prediction of New Your operations, we think a big criteria for the selection of a specific New York site is for it to have variables that strongly resonate with the key factors we highlighted above. With this being said, developing a model will quickly help us filter out the key candidates and narrow it down from there based on other operational and logistical factors like rent, commute distance, etc. If we were members of Sea Watch’s Board, we would want Len to produce a projected Gross Receipt distribution range for New York expansion where the mean outperforms the expected gross receipt from not expanding in order to justify for the risk associated with this expansion. The 95% confidence interval should also be within reasonable limits.

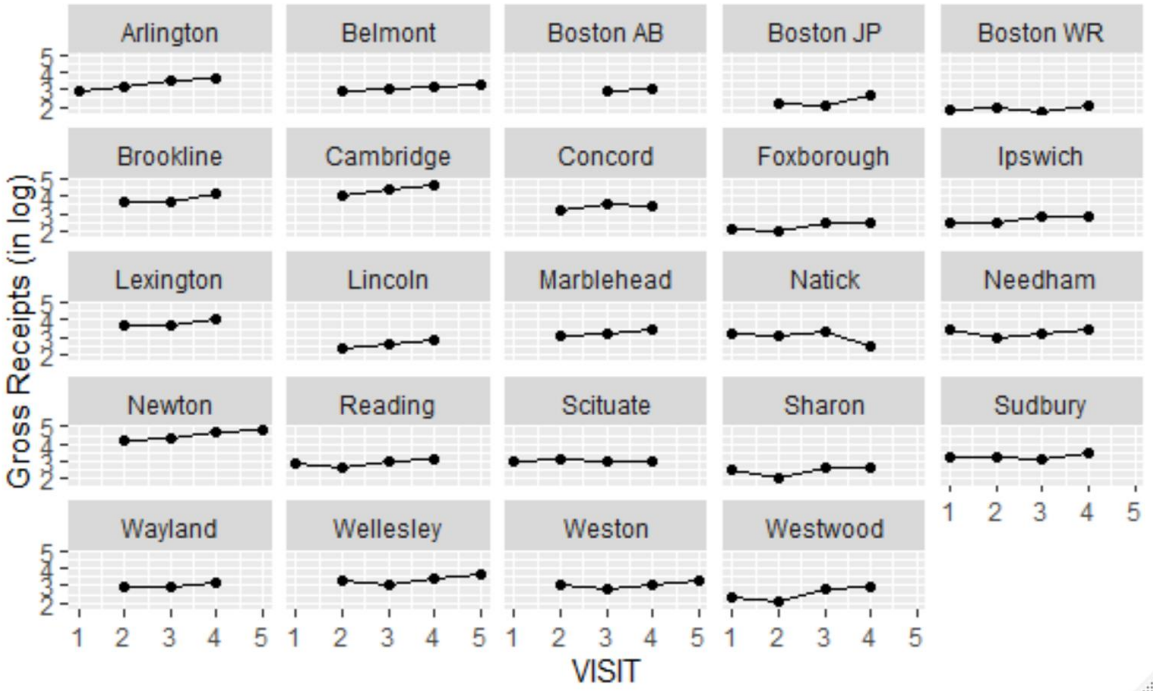
Gross Receipt Histogram Insight





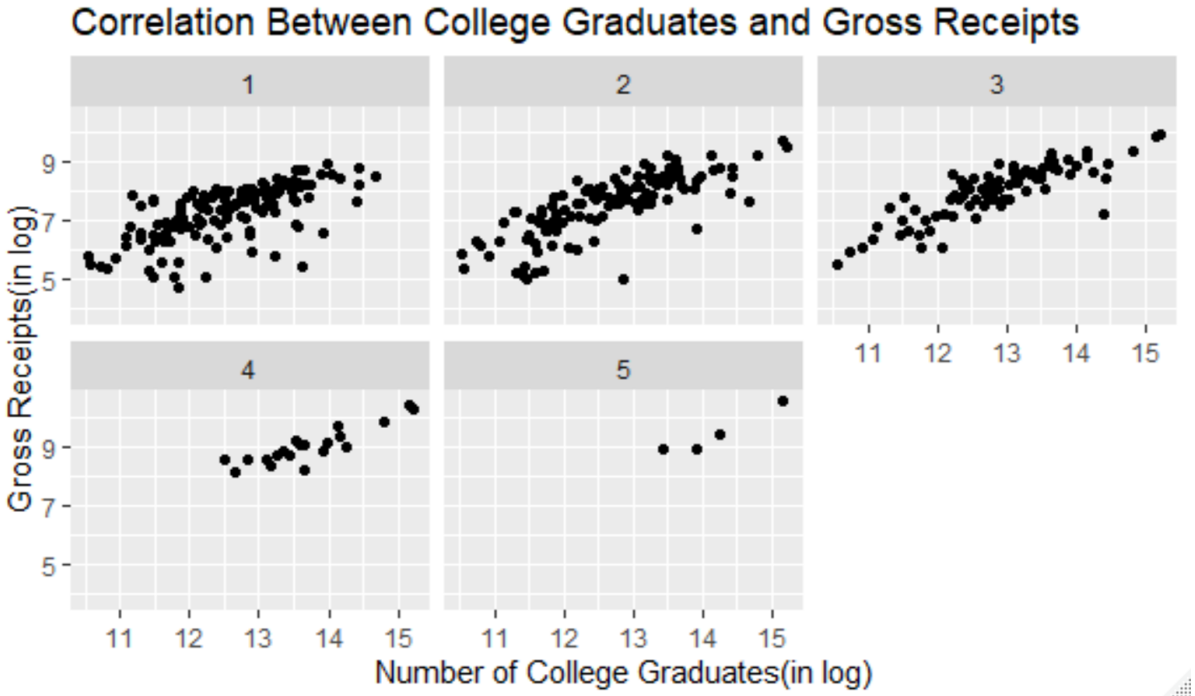
We noted some interesting discoveries involving our two likely response variables, Gross Receipts and Gross Receipts per Canvas Hour. Gross Receipts itself its very heavily skewed right, with a mean of 3341.30 compared to a mean of 2419.50, with standard deviation 4056.82. One key insights from this discovery is the outlier with well above $30,000 in Gross Receipts: Newton. It would be wise to thoroughly examine this point to determine any factors that differentiate it. At a preliminary observation level, no single factor about the city stands out, though it favorably compares to the mean for many seemingly telling statistics, like population, college graduation rate, and per capita income. Secondly, the shape of our data implies that a log function, rather than a linear function, would be much more useful for predicting Gross Receipts. Gross Receipts per Canvas Hour, however, has an approximately normal shape. This leads us to believe that a linear model could be a good fit for this variable and it is a much wiser variable for comparison in field studies or experiments than Gross Receipts itself.

Inflation-Adjusted Gross Receipts on Repeated Visits



- From the charts to the left, we can see how gross receipts change between visits for all counties that were visited more than 3 times. As mentioned on our previous slide, we decided to use a log transformation on the gross receipts variable to visualize the pattern.
- We adjusted the gross receipts using the Consumer Price Index (CPI) for the same month of the visit.
- For most counties, repeated visits leads to an increase in gross receipts with the exceptions of Concord and Natick.
- We believe collecting more data on county population trend will help us separate out the population growth effect as well.

Gross Receipts vs. College Graduates



- We calculated the number of college graduates using the census population and percentage of college graduates. We then plotted gross receipts with the number of college graduates while performing log transformations on both to visualize the pattern, separated by visits.

- We notice a strong positive correlation between the log transformations of gross receipts and number of college graduates. The correlation between the log transformations of both is 0.7761. The correlation between college graduates and gross receipts is 0.8151. We think by mitigating heteroscedasticity in our data the slight drop in correlation is completely reasonable.

POLITICAL Factors Influencing Total Gross Receipts

```
> cor(GROSS,REAG,use="complete.obs")
[1] 0.5133656
> cor(GROSS,CART,use="complete.obs")
[1] 0.5036925
> cor(GROSS,ANDR,use="complete.obs")
[1] 0.6837335
```

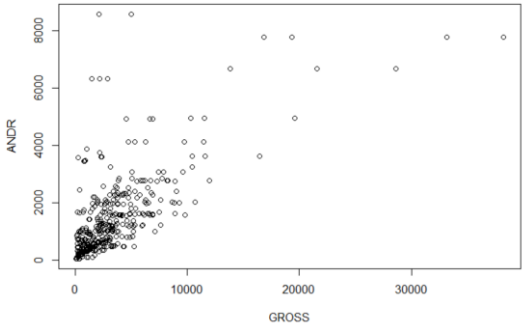
- We originally wanted to see if there is a correlation between number of Carter supporters and gross receipts for each county, as democrats are usually more likely to be pro-environmentalism. Despite being an independent candidate, we know that John Anderson was Republican and subsequently had a very Republican voter base. On the other hand, we also know that even many Republican voters thought Reagan's anti-environmentalist stance was too much, and subsequently seeked out to donate to Sea Watch. All this is to say that there are many forces at play and that we should exercise caution when exploring the idea of whether political alignment truly plays a significant role in gross receipt numbers.

- We tried multiplying our population with the % of the population that voted (sum of votes for all three parties divided by population census) and % of voters who voted for Carter (number of carter votes divided by total number of votes), but quickly realized that we were foolish and that after canceling out top and bottoms the answer is the same as the number of Carter Votes. What was interesting was that of all three parties, the independent (Anderson) had the strongest correlation with total gross receipts at a correlation of 0.6837335.

- Plotting the scatterplots of the three voter bases immediately showed us heteroscedasticity, which is why we decided to continue our explorations using logs of both total gross receipts and number of votes for each party. Doing so revealed a decent linear correlation between the two variables. We were definitely surprised by the fact that Carter's voter base, despite being democratic, had the least correlation to gross receipts amongst the three voter pools. Further examination and modeling may be needed to examine why this may be the case.

- While examining the data, we also noticed that in some cases the combined number of votes for the three parties exceeded the total population. Logically, this makes no sense. A closer examination at the worst offender (Longmeadow county) revealed a staggering 23888 votes for Carter alone despite having a recorded population of only 16301. This suggests significant error in either of the two variables.

Select example of a graph of the correlation between Total Gross Receipts and Number of Votes for John Anderson (before we logged everything) (Note the heteroscedasticity shape)




```
> cor(log(GROSS),log(ANDR),use="complete.obs")
[1] 0.7080688
```

```
> cor(log(GROSS),log(REAG),use="complete.obs")
[1] 0.6486222
```

```
> cor(log(GROSS),log(CART),use="complete.obs")
[1] 0.5590689
```

Log of Gross Receipts as a Function of Log of Number of Votes for Ronald Reagan, John Anderson and Jimmy Carter (Charts are color coded respectively)