
ISOM-670 Business Statistics

Group Assignment 3

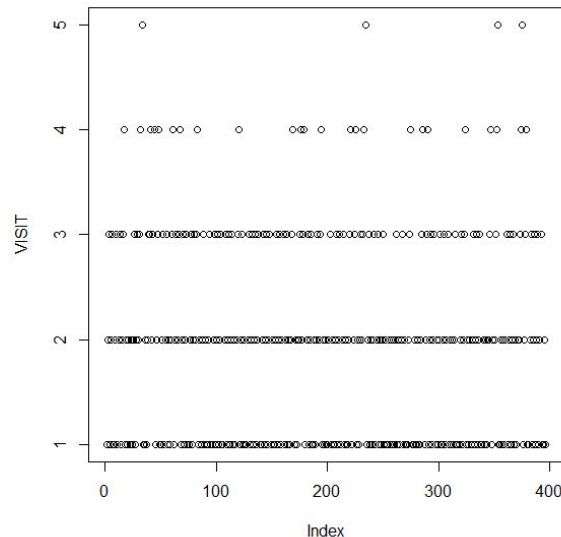
— Team 2: Yifei, Jie, Eleanor & Susie —

Concern of the Board:

What's the Trend?

How does frequency of visits impact GROSS?

- ❑ For the 5th visit, we have only 4 data points. The sample size is not big enough to determine the actual effect of 5th VISIT on GROSS. Thus, we decide to analyze situations when VISIT is in the range of 1 to 4.
- ❑ The average GROSS grows from 2049 to 9451, when VISIT increases from 1 to 4. This indicates that more visits actually lead to higher GROSS per observation.
- ❑ Selection bias: It's highly possible and logical for Sea Watch to pay more visits to one city if the previous visit brought considerable GROSS (high-potential city). In other words, the positive correlation between GROSS and VISIT may not be the case for cities with unsuccessful previous visit (low-potential city).



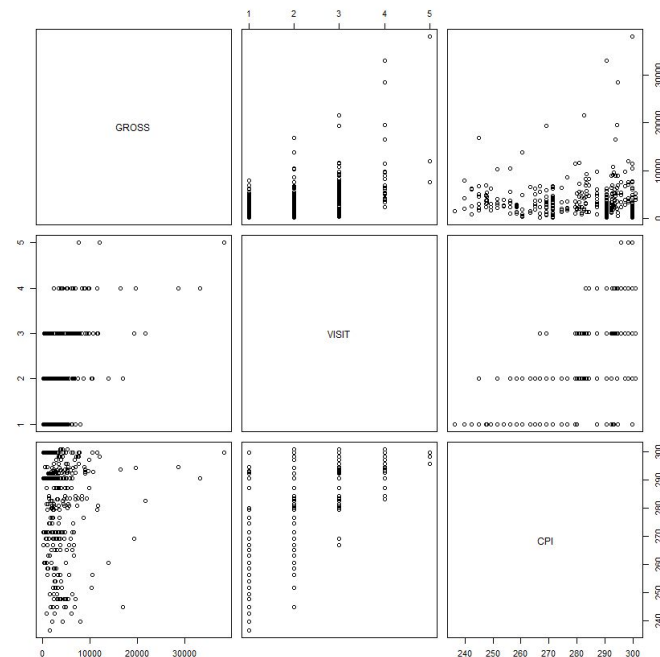
```
> describe(GROSS[VISIT==1])
vars  n   mean    sd median trimmed   mad min  max range skew kurtosis   se
x1    1 146 2048.92 1552.65  1904 1875.32 1679.04  92 7860  7768 1.03    1.08 128.5
> describe(GROSS[VISIT==2])
vars  n   mean    sd median trimmed   mad min  max range skew kurtosis   se
x1    1 131 2936.33 2736.56  2229 2523.9 2087.5  43 16833 16790 2.02    5.88 239.09
> describe(GROSS[VISIT==3])
vars  n   mean    sd median trimmed   mad min  max range skew kurtosis   se
x1    1  89 4268.27 3707.35  3336 3745.89 2937.03 257 21568 21311 2.07    6.23 392.98
> describe(GROSS[VISIT==4])
vars  n   mean    sd median trimmed   mad min  max range skew kurtosis   se
x1    1  24 9451.21 7723.9 6865.5 7968.75 3482.63 2376 33176 30800 1.84    2.53 1576.63
> cor(GROSS,VISIT,use="complete.obs")
[1] 0.4674409
```

Concern of the Board: What's the Trend?

How can we separate out the growth to different variables?

- ❑ Static variables (1980 census) are not helpful in trend analyses. We omit these variables and focus on time-sensitive variables: VISIT, CPI, and MOY. †
- ❑ **Scale** is measured by the number of VISIT.
- ❑ **Inflation** is measured by CPI. CPI and GROSS are negatively correlated, showing that people tend to donate less when the inflation is high. Although its correlation with GROSS is relatively low, it still could be statistically significant when used in building a model.
- ❑ **Awareness** increases as Sea Watch launches more successful campaigns and becomes a bigger organization. It could be partially indicated by the change in the voting stats of the three presidential candidates (some candidates focus more on environmental issues), but we need these data in different months to do further analyses.
- ❑ **Other possible factors** include MOY and omitted static variables. MOY currently has low correlation and needs to be broken into dummy variables in regression model to recognize potential seasonality trends.

† CNVHRS is not a good indicator of scale here, due to its ultra-high correlation with GROSS. Since VISIT measures how many times Sea Watch has been to a city, it is a better option.



	GROSS
GROSS	1.0000000
VISIT	0.46422947
CPI	-0.02880838
CNVHRS	0.96973762
MOY	-0.01571797

Concern of the Board:

Satellite Location Selection Criteria

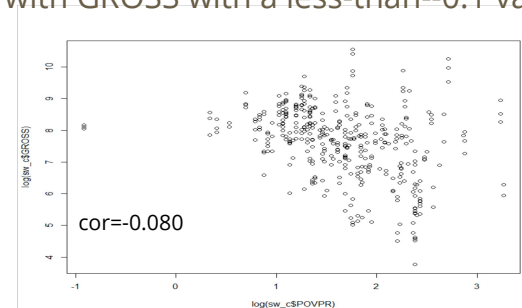
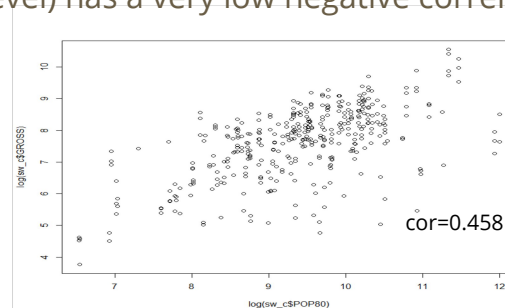
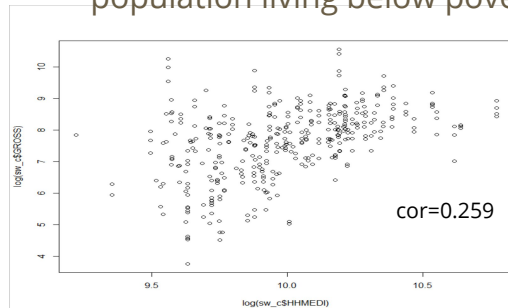
We summarized the following main criteria for choosing the site - **potential donation market, social awareness of environmental issues, severity of existing environmental problems & accessibility/convenience of travel.**

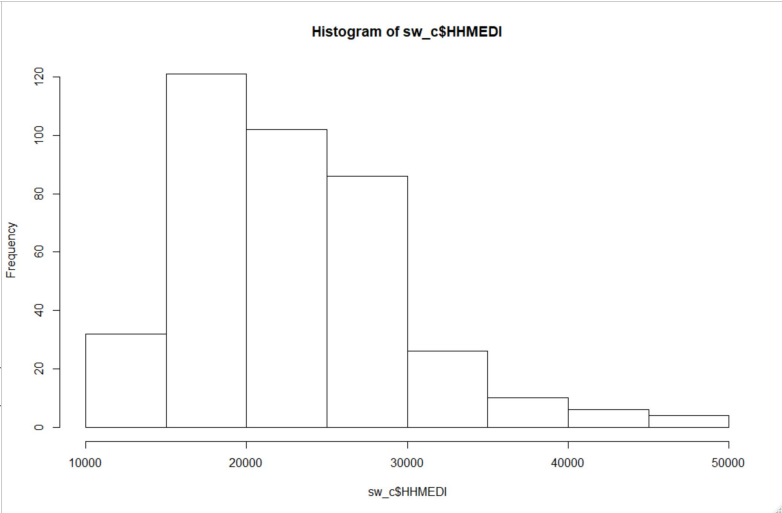
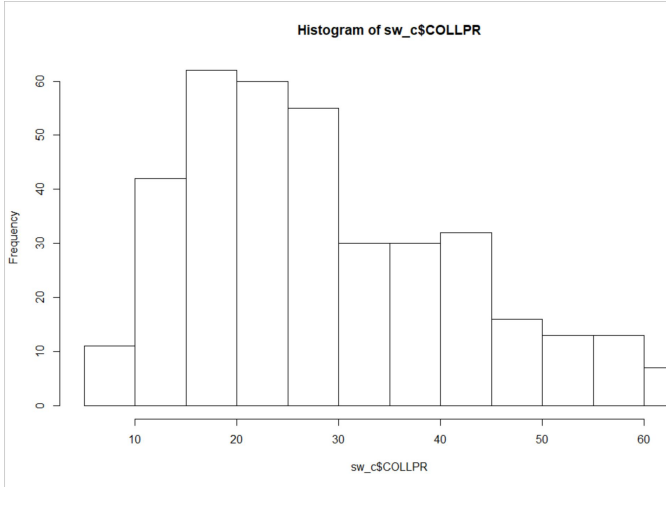
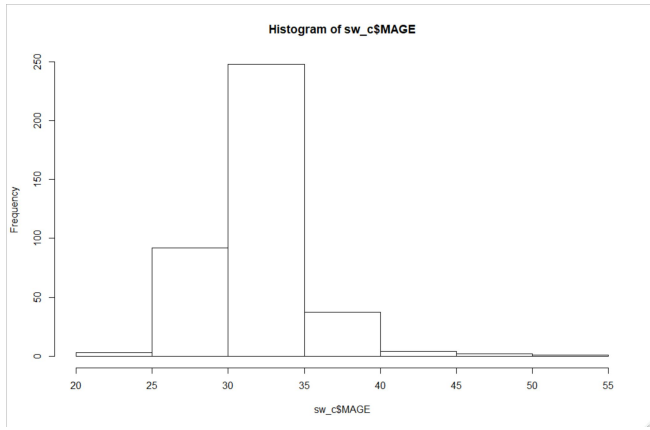
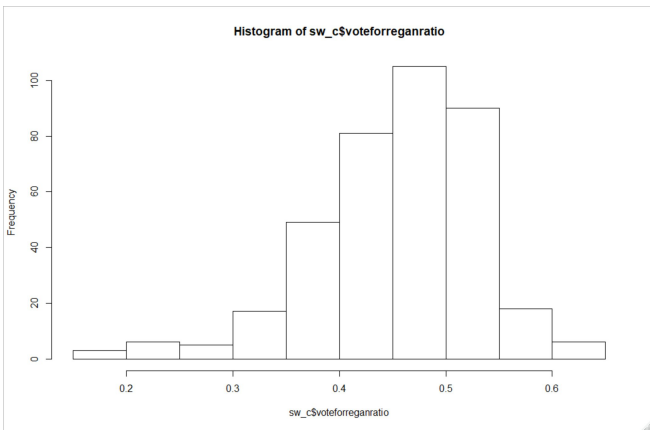
Based on the available variables, possibly appropriate key indicators we considered to represent the criteria:

- potential donation market: HHMEDI (household median income) + POP80 (population) + POVPR (below poverty)
- social awareness: COLLPR (4+ year education) + MAGE (median age) + vote for REAG % (Ronald Reagan)
- severity of environmental issues: MFGPR (percentage employed in manufacturing jobs)

Criteria - Potential Donation Market

- First, we glimpse at histograms of GROSS, HHMEDI, POP80 and POVPR, and notice that all these distributions are skewed → take logs on both sides to try handling the issue
- From the scatter plots below, we could observe that HHMEDI (Household Median Income) is slightly positively correlated to GROSS (Total Gross Receipts for one obs.). Aligned with our intuition, GROSS is also positively correlated with POP80 (Total population in 1980) with a stronger relationship. However, POVPR (percentage population living below poverty level) has a very low negative correlation with GROSS with a less-than--0.1 value.

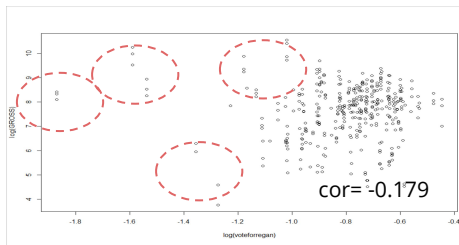
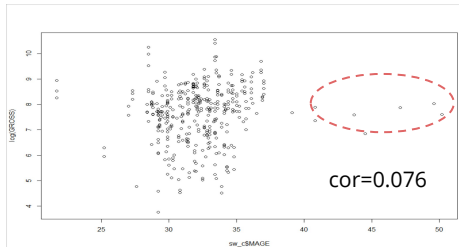
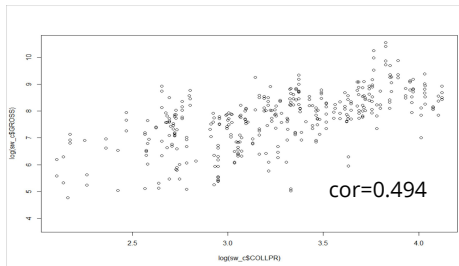




Concern of the Board:

Satellite Location Selection Criteria

Criteria - Social Awareness



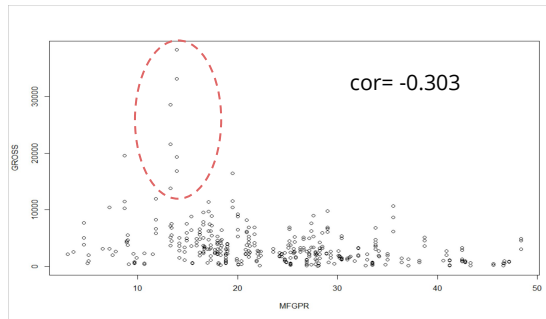
From the plots on the left, we could observe positive correlation between COLLPR(% population with 4+ yrs of college education) and GROSS, signaling that people with higher education usually have more willingness to contribute to environmental issues indeed.

The middle plot clearly shows most donators' ages lie between 28 and 37 years old, but have weaker relationship with GROSS.

We calculate the percentage of people voting for President Ronald over total voters, and found a modest negative relationship.

However, we note that there tends to be many extreme values that might considerably distort the results..

Criteria - Environmental Problem



We consider MFGPR(% of population employed in manufacturing job) to partly indicate which locations rely on manufacturing more heavily, thus risk having more environmental issues currently. We assume that a higher percentage indicates a more urgent need for addressing environmental severity. We find a negative relationship between GROSS and MFGPR. This might be due to the lower public awareness or interest in environment when manufacturing prospers.

More Thoughts about Data Limitation & Improvements

In this case, we consider the possibility of including both more data observations and more variables to have more accurate analysis and estimates.

For more variables, for instance, if we could gather the number of employees involved in each observation of canvass, we might get a better idea of the canvassing staff's productivity/effectiveness at different locations for each time previously, which is critical for recruiting and allocating staffing resources when setting up the new office, especially given the expected constraints during the first year of establishment.

In addition, more historical data entries on some aspects might give the board a more holistic view of the trends besides cross-sectional data. In the data related to demographics, economics and politics, we only have sequential records of the CPI variable, with the remaining limited to the census data of 1980. The access to time series data of these variables as well might help separate out different sources of growth better.

Thank you!

Case C - Analytical Concerns

- Recommended Indicators to look into by the board
 - Commercial real estate markets
 - Travel times on area highways
 - Local labor markets
- Growth <- scale, inflation, awareness, frequency of visits
 - frequency - The effect of repeated visits (ppl seems to be unwilling to write check after they have already donated.) (check towns' donation that doesn't grow are towns which have been visited 3+yrs)
 - Inflation - Apparent YoY growth might simply be shrinking values of dollars
 - Awareness <- not only canvass visits but also presence in media

Case C - Satellite location selection

- Len : chase after money – **Greenwich**
- David : help the community – **Bridgeport**
(middle class + with local environ. Issue)
- Gwen: NYC <- Len: too expensive, safety
- Gwen: NJ <- Len: too far from BA and NYC