EMORY
GOIZUETA BUSINESS SCHOOL

Master of Science in Business Analytics MSBA

ISOM 670 Business Analytics Sea Watch Pt. 2

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

The Variables We Will Be Utilizing - Simple Statistics

GROSS	VISIT	POP80	POVPR	COLLPR	CART	REAG
Min. : 43	Min. :1.000	Min. : 688	Min. : 0.400	Min. : 8.20	Min. : 97	Min. : 84
1st Qu.: 1070	1st Qu.:1.000	1st Qu.: 6345	1st Qu.: 3.500	1st Qu.:18.65	1st Qu.: 1089	1st Qu.: 1442
Median : 2420	Median :2.000	Median : 13212	Median : 5.100	Median :25.90	Median : 2158	Median : 3067
Mean : 3441	Mean :2.003	Mean : 19179	Mean : 6.204	Mean :28.72	Mean : 3624	Mean : 3882
3rd Qu.: 4479	3rd Qu.:3.000	3rd Qu.: 24100	3rd Qu.: 8.050	3rd Qu.:38.30	3rd Qu.: 4091	3rd Qu.: 5496
Max. :38256	Max. :5.000	Max. :161799	Max. :26.100	Max. :61.70	Max. :31225	Max. :23339
NA's :2		NA's :9	NA's :9	NA's :25	NA's :19	NA's :16

^ Massachusetts Data

CT/NY/NJ Data -->

POP80	POVPR	COLLPR	CART	REAG
Min. : 19	Min. : 0.600	Min. : 4.81	Min. : 1	Min. : 15
1st Qu.: 5978	1st Qu.: 3.000	1st Qu.:14.30	1st Qu.: 853	1st Qu.: 1629
Median : 12166	Median : 4.200	Median :21.48	Median : 1771	Median : 3222
Mean : 24051	Mean : 5.494	Mean :23.58	Mean : 6911	Mean : 7239
3rd Qu.: 24142	3rd Qu.: 6.700	3rd Qu.:30.75	3rd Qu.: 4278	3rd Qu.: 6110
Max. :738497	Max. :32.800	Max. :64.00	Max. :288893	Max. :251333
NA's :6	NA's :15	NA's :14	NA's :50	NA's :50

**All numbers within the highlighted red boxes were flagged for being 'interesting' and further investigated*

Just to recap, we decided to focus on college graduates population, poverty population and political leaning between the two major parties (excluding independent) based on intuition and statistics. Statistically speaking, these variables were among the highest correlated variables with gross, and are definitely usable with some simple transformation. Intuitively speaking, college graduates population should indicate **environmental awareness**, poverty population should indicate **ability to donate**, and political leaning should indicate **ideology standpoint** for whether the environment is worth/in need of/or should be saved.

We first **replaced all obvious errors data from our database with NaN**. (The most notable of this being **Longmeadow's number of Carter votes***). This is to ensure that both our analysis and the final model will be as accurate as reasonable. We then conducted simple statistics for both the Massachusetts data (top row) and CT/NY/NJ data (bottom row). Across all our percentage statistics, we see little reason to worry. The distributions, skew, and size of mins and maxs are **relatively similar** across both sets. The one concern we do have is with **population**. The new data set features one town **(Hempstead, NY) much larger than any of the towns in our Massachusetts data set**. As such, we should treat any prediction for this town cautiously. This concern isn't extended to our political stats because we use them as a **percentage** in our model.

**Longmeadow's number of Carter votes exceeded the population of Longmeadow.*

Variables For Our Model

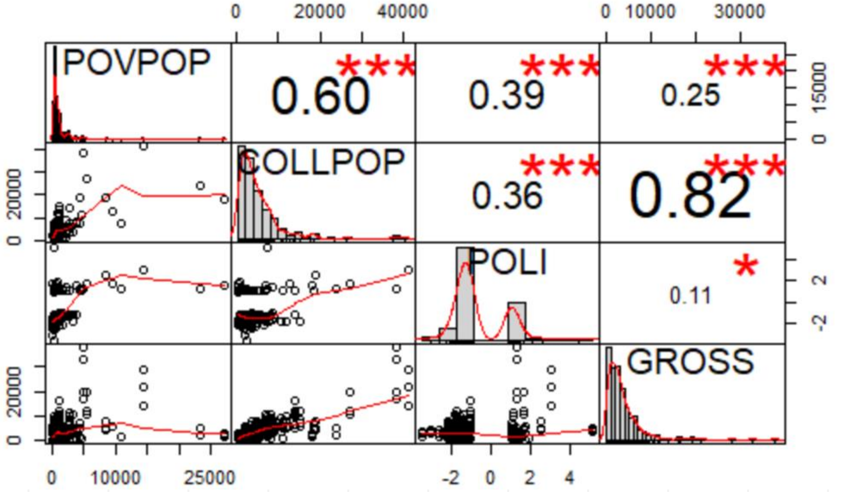
Massachusetts Data (Transformed)

```
    POVPOP            COLLPOP             POLI
Min.   :   16.2   Min.   :  373.4   Min.   :-3.4589
1st Qu.:  299.2   1st Qu.: 1637.0   1st Qu.:-1.6824
Median :  618.8   Median : 3657.1   Median :-1.2921
Mean   : 1456.2   Mean   : 5566.2   Mean   :-0.7254
3rd Qu.: 1061.2   3rd Qu.: 6891.3   3rd Qu.: 1.0277
Max.   :27112.8   Max.   :41083.8   Max.   : 5.1505
NA's   :9         NA's   :25        NA's   :16
```

CT/NY/NJ Data (Transformed)

```
    POVPOP              COLLPOP              Poli
Min.   :     21.05   Min.   :    130.2   Min.   :-15.000
1st Qu.:    216.38   1st Qu.:   1279.4   1st Qu.: -2.064
Median :    451.02   Median :   2663.6   Median : -1.674
Mean   :   2072.05   Mean   :   5199.3   Mean   : -1.432
3rd Qu.:   1215.95   3rd Qu.:   5693.1   3rd Qu.: -1.242
Max.   :107993.34    Max.   :154345.9    Max.   :  5.977
NA's   :15           NA's   :14          NA's   :50
```

Correlation Matrix for SeaWatch C Data




```
> describe(SeaWatchC_Data)
        vars   n    mean      sd  median trimmed     mad    min      max    range skew kurtosis     se
POVPOP     1 387 1456.20 3318.10  618.85  776.63  532.66  16.20 27112.78 27096.59 5.70    36.67 168.67
COLLPOP    2 371 5566.18 6606.57 3657.06 4203.40 3363.40 373.45 41083.78 40710.33 3.15    12.09 343.00
POLI       3 380   -0.73    1.46   -1.29   -0.86    0.61  -3.46     5.15     8.61 1.12     0.85   0.08
> describe(SeaWatchD_Data)
        vars   n    mean       sd  median trimmed     mad    min       max     range  skew kurtosis     se
POVPOP     1 390 2072.05  7287.75  451.02  738.24  425.36  21.05 107993.34 107972.29  9.63   120.06 369.03
COLLPOP    2 391 5199.32 10627.21 2663.55 3449.52 2598.58 130.17 154345.87 154215.70  8.94   106.76 537.44
POLI       3 355   -1.43     1.52   -1.67   -1.56    0.61 -15.00      5.98     20.98 -0.93    18.87   0.08
```

We know that variables may behave very different after undergoing transformation, so we decided to check our chosen variables after transformation. Other than the obvious outliers that carried over from our raw data, everything else looked to be in order. We do want to note that Teterboro, NJ only has a population of 19 with 15 REAG voters and 1 CART voter. With a POLI coefficient of -15, this town may obtain an extremely inaccurate prediction. That being said, a town of 19 people is not likely a priority anyways, so this seems to be a non-issue.

Our Model

Our Model = lm(GROSS ~ COLLPOP + factor(VISIT) + POVPOP + POLI)

COLLPOP = The number of college graduates of each town

COLLPOP = COLLPR * POP80/100

VISIT = The number visit for a particular town (i.e. 1st visit, 2nd visit, etc.)

POV= The poverty population of each town

POV = POVPR * POP80/100

POLI= The ratio of number of votes the winning party won by

POLI = (# OF VOTES FOR WINNING PARTY)/(# OF VOTES FOR WINNING PARTY) * P

P = 1 for Carter & P = -1 for Reagan

- If Carter won, Poli >= 1
- If Reagan won, Poli =< -1

We have discussed in detail why

**Please see final slide for further thoughts on further improvements for our model*

Model Regression Output

We chose college population (COLLPOP) as one of the variable because it has the highest correlation with Gross among all the possible variables.

We chose factor(visit) because we can see a clear difference in Gross with every increase in visit.

We chose the poverty population (POP) because of its high correlation with Gross and stability as the population changes. We can see that the coefficient is negative, which makes sense in reality.

```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          8.06820  177.74592   0.045    0.964
SeaWatchC$COLLPOP    0.58659    0.01888  31.074  < 2e-16 ***
factor(VISIT)2     287.88642  215.61011   1.335    0.183
factor(VISIT)3    1160.57264  242.84662   4.779 2.59e-06 ***
factor(VISIT)4    3804.47520  434.28099   8.760  < 2e-16 ***
factor(VISIT)5    7902.16614 1044.48281   7.566 3.37e-13 ***
SeaWatchC$POP       -0.35344    0.03522 -10.036  < 2e-16 ***
SeaWatchC$Poli    -274.46806   69.29888  -3.961 9.04e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1717 on 354 degrees of freedom
  (34 observations deleted due to missingness)
Multiple R-squared:  0.8333,    Adjusted R-squared:  0.83
F-statistic: 252.8 on 7 and 354 DF,  p-value: < 2.2e-16
```

Most of the coefficients are significant except for the coefficient for visit 2. This means that there is no significant difference in gross receipts between the first visit and the second visit.

Adjusted R-squared of 0.83 means that this model is able to explain 83% of the variability in gross receipts for the Massachusetts data.

We can see here that apart from the few outliers, our POLI variable does a great job of predicting

We bring down the standard error from 4507(the original standard deviation of Gross) to 1717, a decrease of 62%.

The VIF data of our model

Here we use the vif function to calculate the variance-inflation and generalized variance-inflation factors of our model.

Notice in our model, we used both the <u>college population(COLLPOP)</u> and the <u>poverty population(POP)</u> for each town. So we checked the vif to see if there is any variance-inflation. It turned out that although the GVIF^(½*Df)) for College population and poverty population is slightly higher than the other two variable, they are way smaller than 2. So we can say the multicollinearity is not significant, and the influence on our model is not significant.

```
> vif(model)
                      GVIF Df GVIF^(1/(2*Df))
SeaWatchC$COLLPOP 1.932228  1        1.390046
factor(VISIT)     1.196298  4        1.022657
SeaWatchC$POP     1.771912  1        1.331132
SeaWatchC$Poli    1.244029  1        1.115361
```

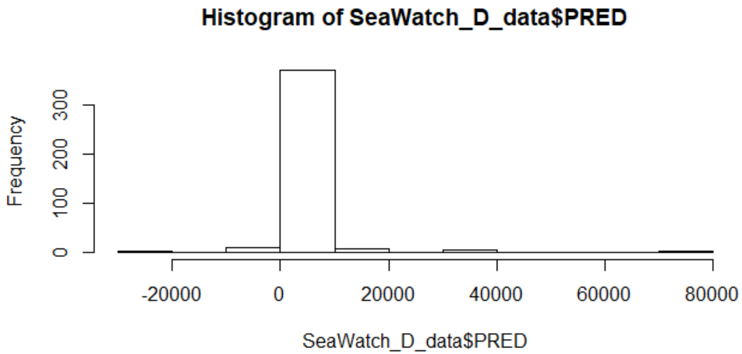
All variables have GVIF^(½*Df)) <2,

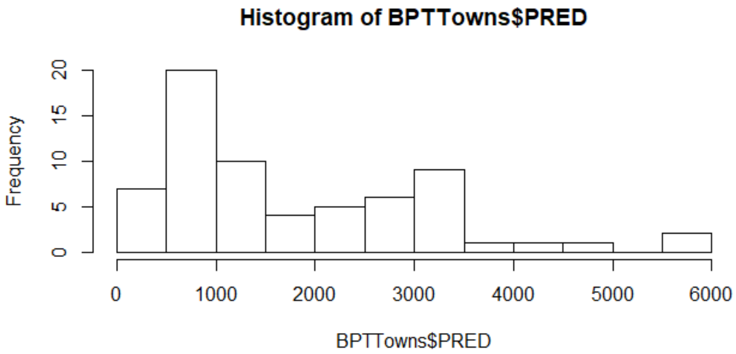
Comparison of Greenwich vs. Bridgeport

Towns	Sum of Predicted Gross Receipts ($)	Number of Towns (#)	Median Predicted Gross Receipts ($)
Within 40 Miles of Greenwich	511,480.30	100	2,563.86
Within 40 Miles of Bridgeport	337,859.40	116	2,120.19
Particularly closer to Greenwich (GRN<BPT)	368,524.90	52	3,256.40
Particularly closer to Bridgeport (BPT < GRN)	118,696.90	66	1,234.60

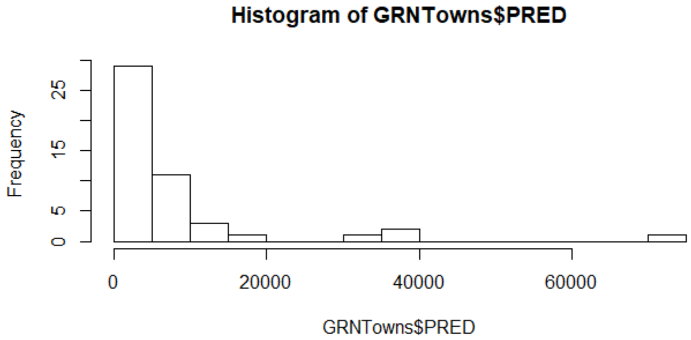
We compare towns that are within 40 miles of each city, and we predict that we will have more gross receipts from Greenwich for the first visit ($511,480 from Greenwich vs. $337,859 from Bridgeport). However, there are more towns in Bridgeport than in Greenwich, so we will need to look at total gross predicted from multiple visits to determine which town has a bigger long term value.

Deep Dive on Predicted Gross Receipts



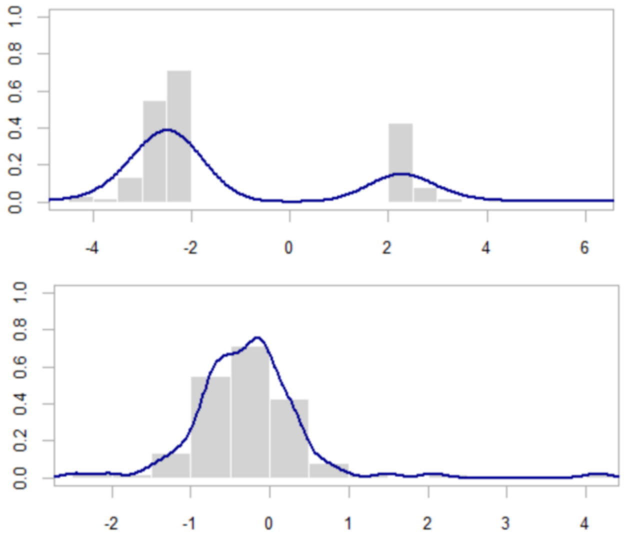


Across all data sets, we see a strong right skew in predicted gross. The negative values from the histogram of all towns can be ignored; they are simply dud towns not worth visiting. Greenwich towns have a number of high grossing towns compared to a more even dispersal from Bridgeport towns. That being said, Greenwich's valuable towns are way more valuable than Bridgeport's, making it the better choice to locate near. It should be noted that many of Greenwich's big towns are on Long Island, which could be impractical. Additionally, one of Greenwich's big towns is Hempstead, NY, which must be treated cautiously, as noted earlier. Even ignoring Long Island, Greenwich still sees more money coming from first visits, but could certainly tail off on future visits.



Future Improvement Concepts

Lastly, we wanted to recognize that there is always room for improvement. While the effect on our model and results will be pretty minimal, we can make our POLI factor more normally distributed by simply doing '-1' on all positive data points and '+1' on all negative points. This would center our data at 0 while still keeping the same outward shape in both directions. We can see in the regression output below that while doing so will change our coefficient for POLI, it will only reduce our residual standard error by 4 and improve our R-squared by around 0.01.



The top chart is the histogram of our POLI factor Before '-1' adjustment and the bottom is after. The blue line denotes the density distribution. Note how the bottom curve resembles a normal curve.

```
Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)              -7.73985  176.97097  -0.044    0.965
COLLPOP                   0.58595    0.01877  31.215  < 2e-16 ***
factor(VISIT)2          285.51378  215.03102   1.328    0.185
factor(VISIT)3         1156.26814  242.20816   4.774 2.65e-06 ***
factor(VISIT)4         3796.02352  433.14321   8.764  < 2e-16 ***
factor(VISIT)5         7942.92151 1040.93091   7.631 2.19e-13 ***
POVPOP                   -0.35028    0.03519  -9.953  < 2e-16 ***
SeaWatch_C_data$Poli   -183.14484   43.54068  -4.206 3.29e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1713 on 354 degrees of freedom
  (34 observations deleted due to missingness)
Multiple R-squared:  0.8342,    Adjusted R-squared:  0.8309
F-statistic: 254.5 on 7 and 354 DF,  p-value: < 2.2e-16
```

We may also want to collect more data on what characteristics lead to repeat visits, to help us decide if we should target towns that we did not go back to in Massachusetts. This will also help us determine the towns to have repeat visits in NY/CT/NJ and calculate a long term value for Greenwich and Bridgeport.