Northampton Survival Center Data Analysis Report

Ji Young Yun, Lucy Shen, Yue Kuang

May 30th, 2017

Objective

- Studying the relationship between weather and number of people coming each day and testing whether people will tend to come early on rainy days.
- Building models to analyze the demographic information for rainy and not rainy days and predicting weather based on the model.
- Analyzing and comparing the demographic information of people coming to survival center.

Data Cleaning

- *Duplicated and Missing Data:* The data of same people who came several times a day were combined and the missing data were deleted. The data file was combined with the demographics file based on ID.
- *Waiting Time:* The waiting time was calculated as the difference(in minutes) between arrival time and intake time. Since the arrival time was provided only for May, the waiting time was omitted in the data file attached here (See "NSC_combinedDataUpdated.csv"). However, an analysis concerning correlation between weather and waiting time would be possible only if more data were collected .
- *Weather and Temperature:* Since weather and temperature are factors of interest, historical data(rain/not rain; highest/lowest/average temperature) for each day were added to the data file[1]. In column L of attached data file, 0 indicates "no rain" and 1 indicates "rain".
- *Categorical Values:* Factors, such as number of people coming each day and level of monthly income per person, were converted into categorical factors. The level of income was divided into 5 groups with equal number of data: Low(<345), MedLow(>345 and <615), Medium(>615 and <889), MedHigh(>889 and <1051), and High(>1051).(The unit is dollar/month*person) Also, the number of people coming each day was divided into into three groups based on the average(36): Less(< 36), Average(= 36), and Above(> 36).

Data Analyzing

Relationship between weather and number of people coming and the time they waited

The slight positive slope implied that people are more likely to come to survival center during raining days. (Figure 1) However, the p-value calculated by Tableau was 0.6358, larger than 0.05 significance level, so it was not reasonable to conclude that the difference of number of people coming each day was due to the weather. Similarly, the regression analysis of the time people waited and weather yielded a p-value of 0.9147, larger than 0.05.(Figure 2) So the difference of the waiting time in raining and not raining days might also due to chance.

[1] https://www.wunderground.com/us/ma/northampton

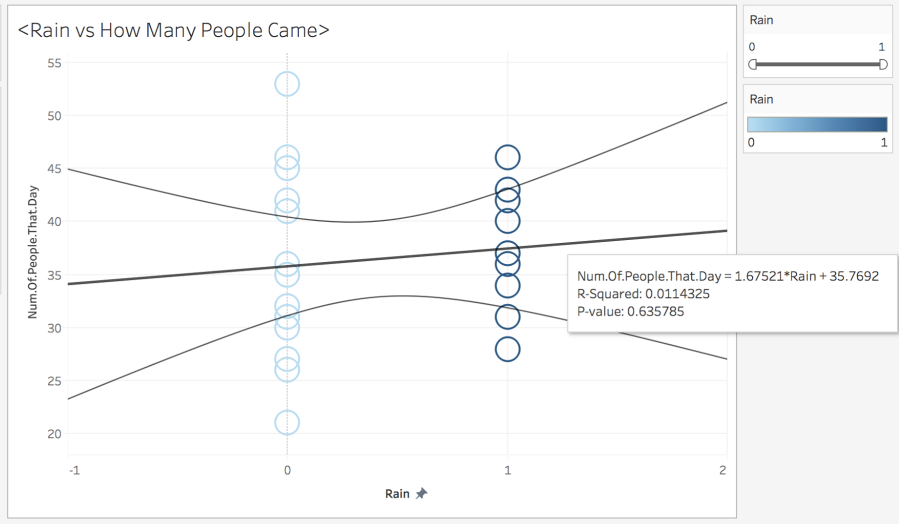

Figure 1. Scatterplot of Number of People Coming Each Day vs. Weather(rain or not rain)

```
Call:
lm(formula = survival_may$wait_open ~ survival_may$rain)

Residuals:
     Min       1Q   Median       3Q      Max
-175.948  -55.464   -6.948   70.021  134.052

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)        40.948      5.317   7.702 1.89e-13 ***
survival_may$rain   1.030      9.608   0.107    0.915
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 77.59 on 305 degrees of freedom
Multiple R-squared:  3.77e-05,  Adjusted R-squared:  -0.003241
F-statistic: 0.0115 on 1 and 305 DF,  p-value: 0.9147
```

Figure 2. Result of Linear Regression of Weather and Time People Waited

Rain and Demographical Characteristics

As the regression model was not applicable for all demographic variables, quadratic discriminant analysis(QDA) was chosen to analyze what were the demographical characteristics of people who were more likely to come on rainy days and those of people who were more likely to come on non-rainy days.

In our analysis, the data file was divided into two parts: training data set and test data set. The test data set was needed because it would be used to test the accuracy of the model. Furthermore, to make data in each factor less variant, certain categories in multiple factors were combined. For instance, for the factor "usual transportation", data were categorized into "car" and "other transportation methods".

```
Call:
qda(rain ~ WIC + Fuel.Assist + Masshealth + level.of.income +
    household_size + car + Northampton + English.speaker, data = train)

Prior probabilities of groups:
        0         1
0.4556962 0.5443038

Group means:
     WICTRUE Fuel.AssistTRUE MasshealthTRUE level.of.incomeLow level.of.incomeLowMed level.of.incomeMed
0 0.04166667       0.2731481      0.7685185          0.1759259             0.1574074          0.2175926
1 0.05813953       0.2558140      0.7906977          0.1976744             0.2170543          0.1899225
  level.of.incomeMedHigh household_size   carTRUE NorthamptonTRUE English.speakerTRUE
0              0.2407407       2.074074 0.8101852       0.5601852           0.8935185
1              0.1860465       2.081395 0.8410853       0.5310078           0.8953488
> predictions_qd %>%
+ summarize(score = mean(class == rain))
     score
1 0.442348
```

Figure 3. Result of QDA Analysis

The model we have build was able to predict not even more than half(44.23%) of the test data correctly. To conclude, relationship between whether people will come on raining days and their demographical characteristics was not significant.

Exploring Demographical Features

Chi-square test was used to test for the correlation between factors of interest: WIC, fuel assistance, health insurance, level of monthly income per person, language, transportation, and living area. 14 pairs of factors were significantly dependent to each other and the 8 most significant pairs were plotted in Tableau.

The first graph(Figure 4) shows the difference in the percent of population using car and other vehicles to transport to the survival center. According to the graph, Northampton residents are more likely to go to the survival center using other vehicles rather than cars, such as PVTA bus, walking, etc. For most non-Northampton residents(93.43%), car is the most common way to transport to the survival center. Such result is reasonable because people who live far away from the survival center would possibly have more demand of cars, the most convenient way to go to the center; while people who live in Northampton have more flexibility of choosing other vehicles.

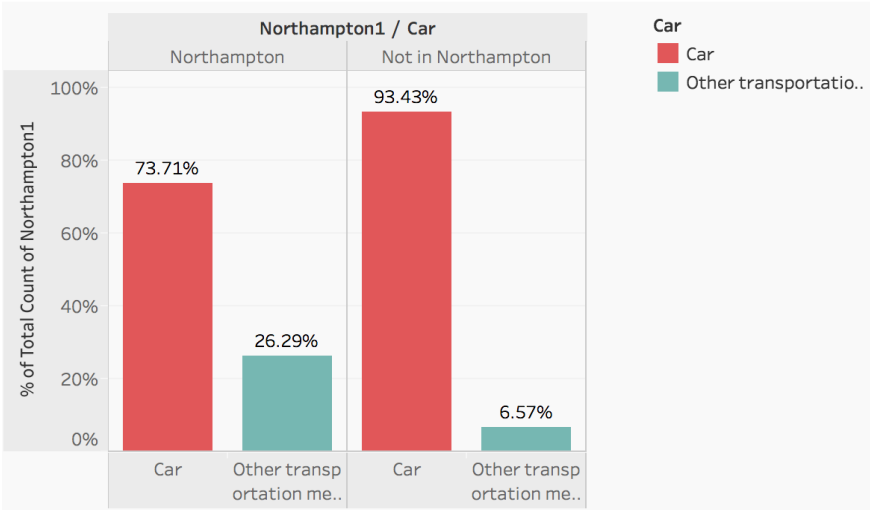

Figure 4. Bar Plot of Transportation vs. Living Location

The stacked bar plot shows that the percentage of the population with health insurance are largest in medium-high and medium-low income groups but is lowest in the high income group.(Figure 5) A possible reason might be that people with high income would get health

examination more frequently than people from other groups, so high-income people would buy less health insurance.

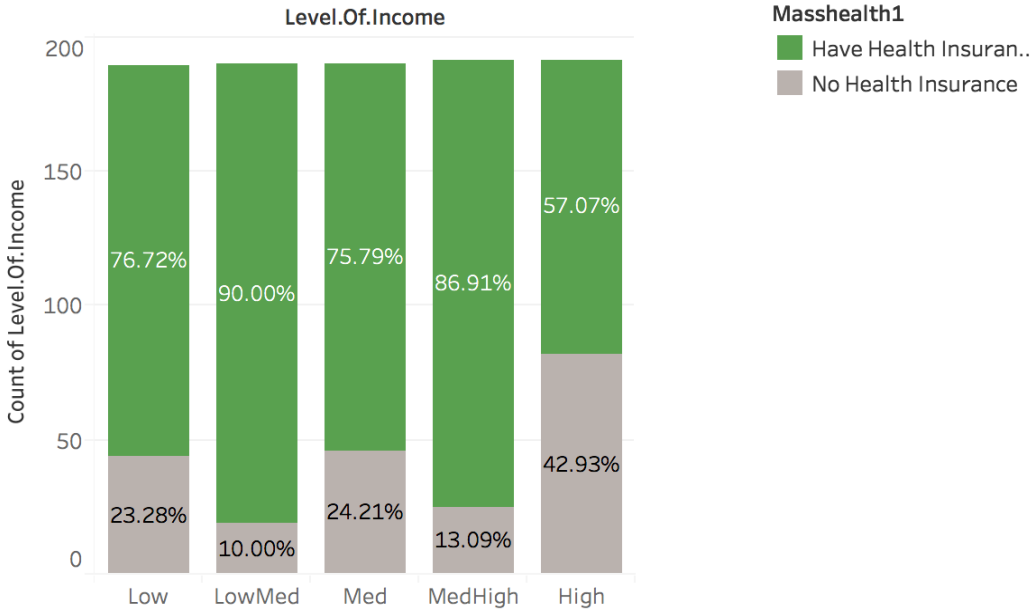

Figure 5. Stacked Bar Plot of Income Levels and Health Insurance

Figure 6 shows the comparison of the distribution of different income levels(income per capita) between English speaking population (people who speak English at home) and non-English speaking population. Obviously, individual whose language spoken at home is not English tends to earn less money than an English speaking individual. Among non-English speaking population, more than half(57.42%) of the population has income level less than medium level. Since people not speaking English at home possibly are immigrants, higher threshold for them to find a job caused by deficiency in English might account for their overall lower income level.

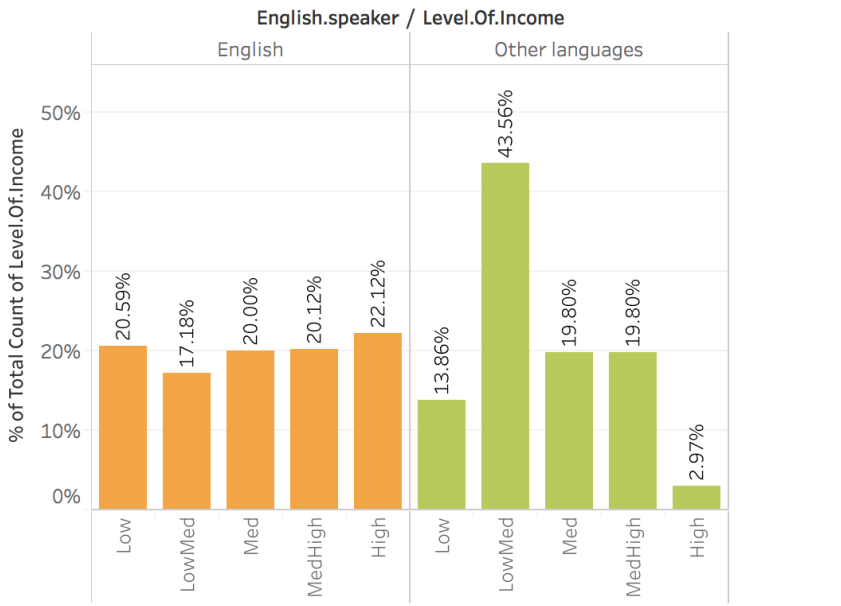

Figure 6. Bar Plot of Language among Income Levels

Figure 7 shows the distribution of different income levels(income per capita) between families with WIC grant (grants of foods, health care, and nutrition education for low-income pregnant, infants and children) and families with no WIC grant. According to the plot, families with no WIC grant have relatively uniform distribution in all income levels; while most families(80.35%) with WIC grant tend to have less than medium income level and only 5.36% have income level above the average. Because families which are qualified to the grant are those with low-income pregnant, infant and children, normally families with WIC grant should have lower personal income.

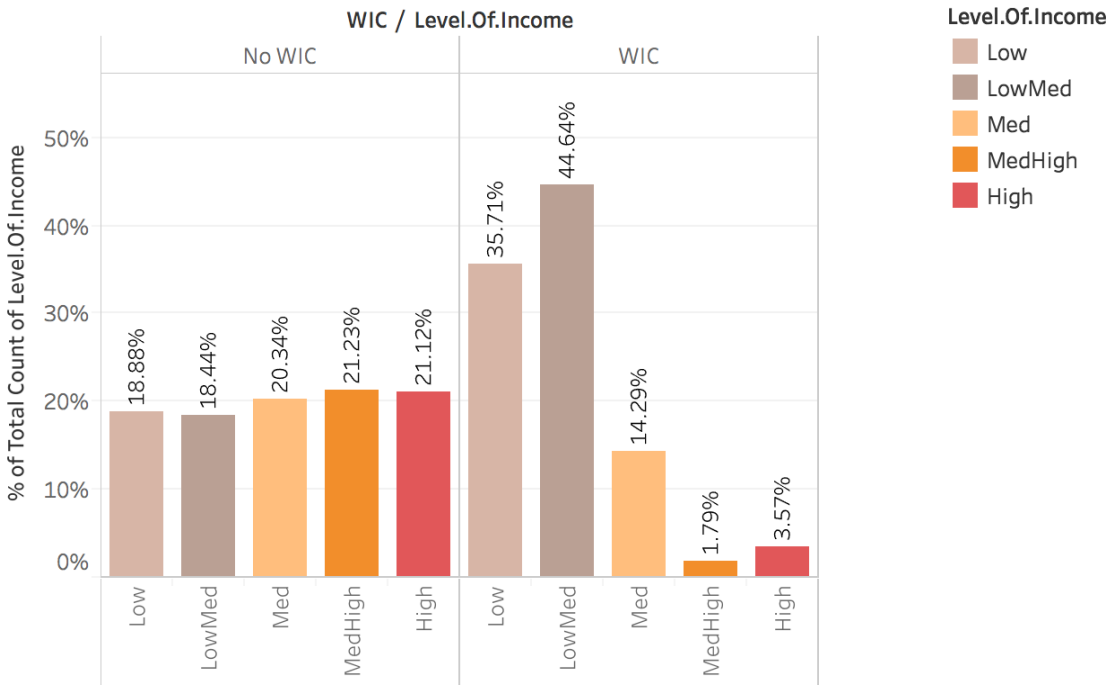

Figure 7. Bar Plot of WIC Grant among Income Levels

Figure 7. Bar

Figure 8 shows that people who have fuel assist are more likely to have health insurance(MassHealth), comparing to people who don't have fuel assist. Certain criteria for eligibility of both of these two welfares might account for the correlation between them. However, there is no enough background information to validate our hypothesis.

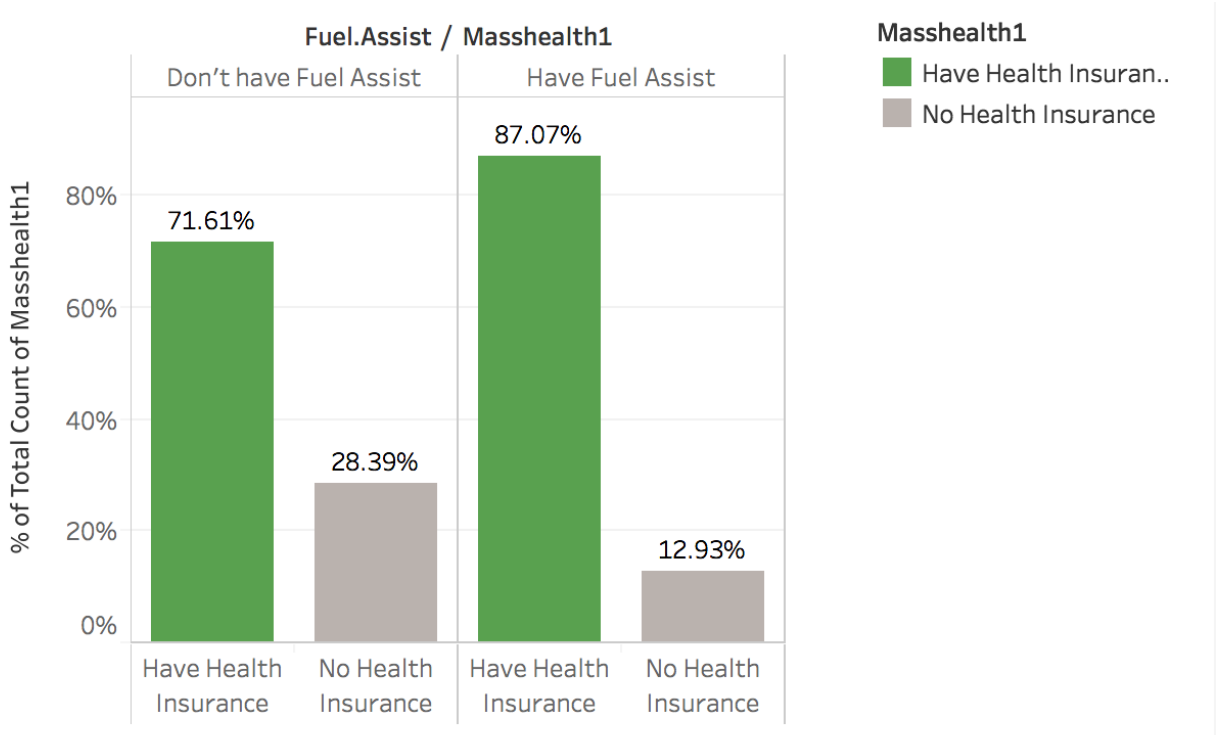

Figure 8. Bar Plot of Fuel Assistance in Population with/without Health Insurance

Figure 9 shows the fuel assistance(have or don't have) between English-speaking and non-English speaking populations. To study the net effect of language on fuel assistance, the transportation factor(having cars or not) was filtered out. Obviously, non-English speaking individuals are more likely to have fuel assistance---about 20% higher than that of the English speaking individuals. Since non-English speaking individuals are more likely to have lower or medium low income(Figure 6), people who are non-English speakers might need more help with the fuel expense.

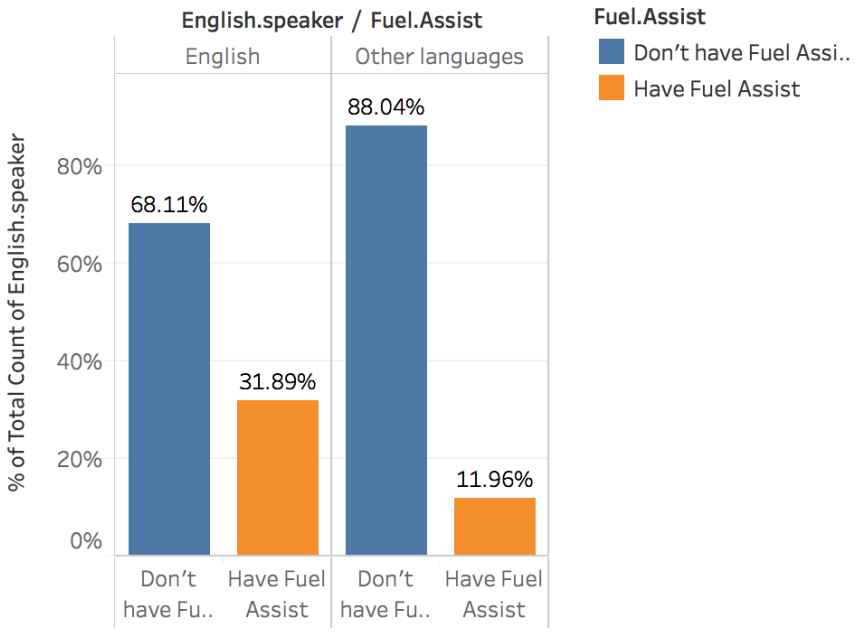

Figure 9. Bar Plot of Language vs. Fuel Assistance

Figure 9. Bar Plot of La

The following graph(Figure 10) suggests that people who don't speak English at home are much more likely to live in Northampton than people who speak English at home. This result might be due to different racial makeup of different locations, or to better information accessibility of Northampton residents who don't speak English at home comparing to those in other locations.

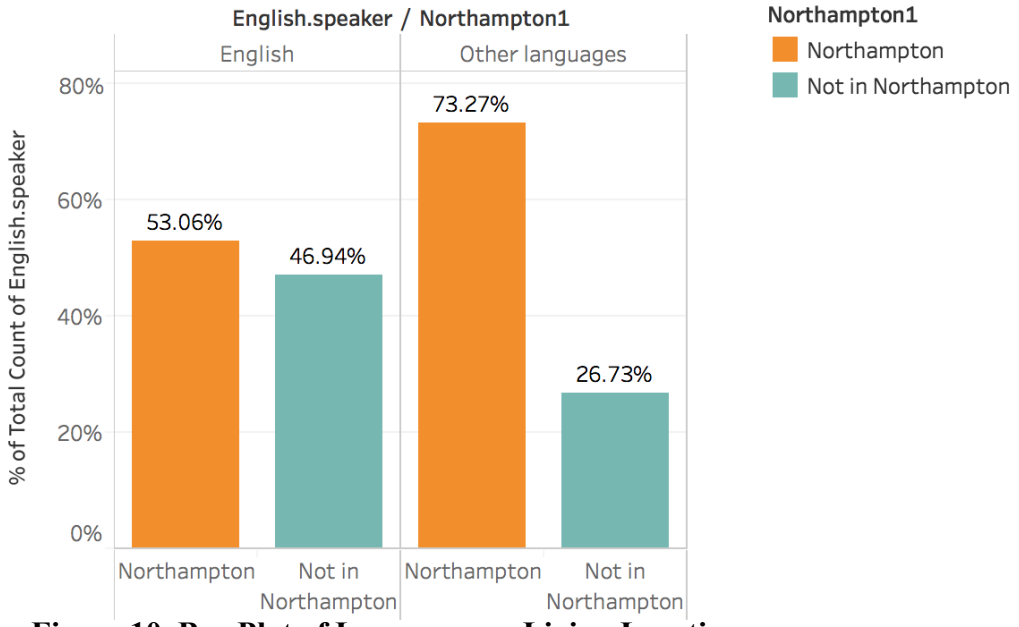

Figure 10. Bar Plot of Language vs. Living Location

Figure 11 shows that most of families in WIC program are also in the MassHealth insurance program, comparing to families not in WIC program. The result is not surprising since both WIC program and MassHealth program are under the rubric of Massachusetts government.

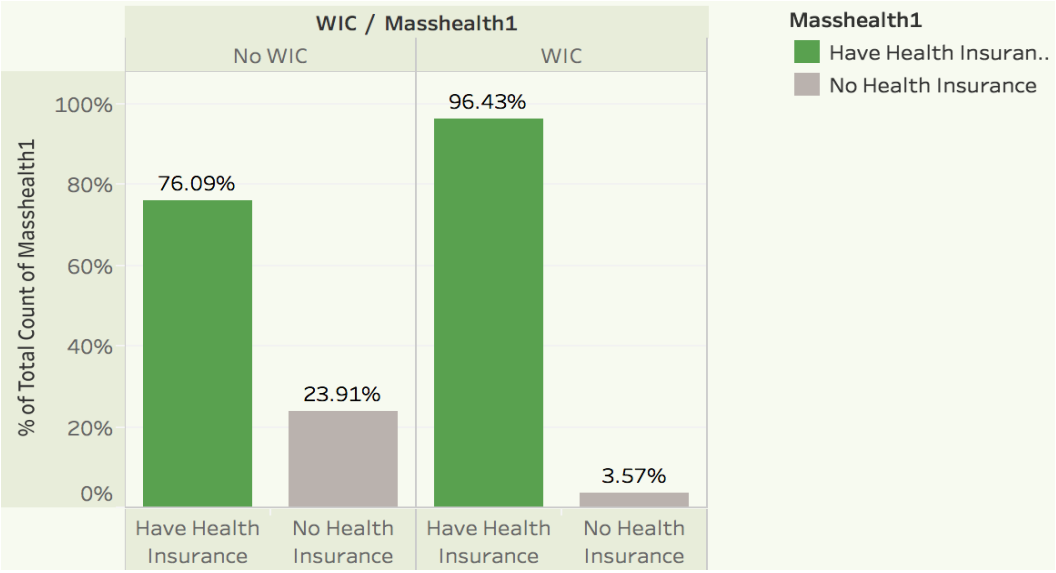

Figure 11. Bar Plot of WIC Grant vs. Health Insurance

More people coming toward the end of months

Figure 12 shows that for both April and May (note: there are only 12 days data available for May), there is a trend that more people tend to come when the date is approaching the end of a month. Although an explanation is still lacked to explain the trend, to avoid the survival center becoming crowded on certain days, one could consider about how to convince people to come more often at the beginning of each month.

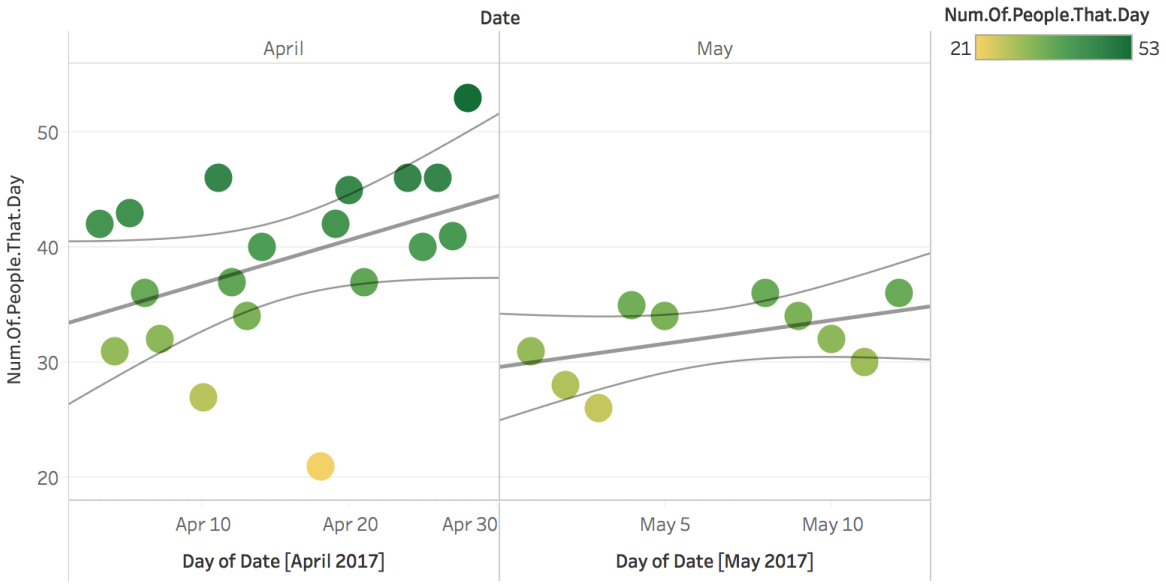

Figure 11. Scatter Plot of Number of People Coming Daily vs. Date

Summary

- More data is needed to find the effect of weather(rain or not rain) on number of people coming each day and people's waiting time. For the data we have and the test we have done so far, we cannot conclude that there is a relationship between weather and number of people coming each day and people's waiting time.
- There are noticeable correlation between demographical features, such as the relationship between language spoken at home and income per capita.
- More people tend to come when the date is approaching the end of a month. The center could work on convincing people to come at different dates and therefore avoid too many people coming on certain days. However, since only two months of data are available, we cannot conclude that the increasing trend will occur in every single month.