Project 1

Jianqiao Wang

2019-10-07

Introduction

- Urban Ministries of Durham (UMD) is a program that helps homeless people by providing neighbors with emergency shelter and case management to help them overcome barriers such as unemployment, medical and mental health problems.
- Data provided by UMD recorded different kinds of support that UMD provided for homeless people from 1931.
- Data has more than ten variables. Our analysis is mainly based on three variables:
  - Date: Time
  - Food.Pounds: Food Pounds UMD provided one time
  - Food.Provided.for: Number of People Receiving Food one time

Purpose of Analysis

We want to answer the following three questions:

- Does the total number of people receiving food every day increase?
- Does the total food pounds UMD provided every day increase?
- What is the average food pounds per person? Is there a difference among different families and people?

Data Cleaning

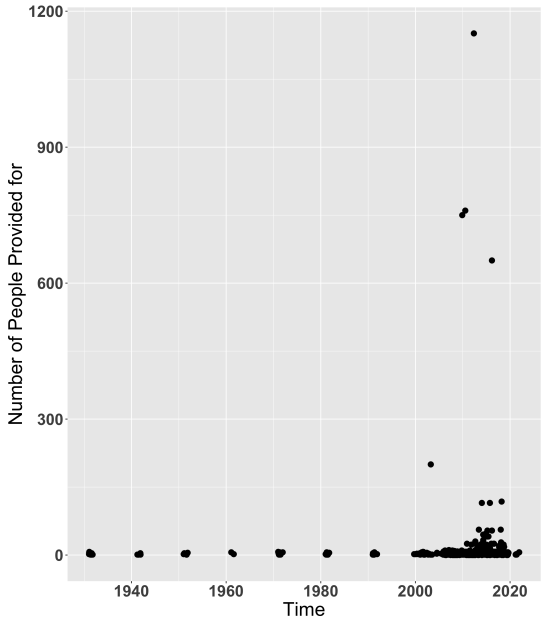

Figure 1: Number of People Every Day over Time

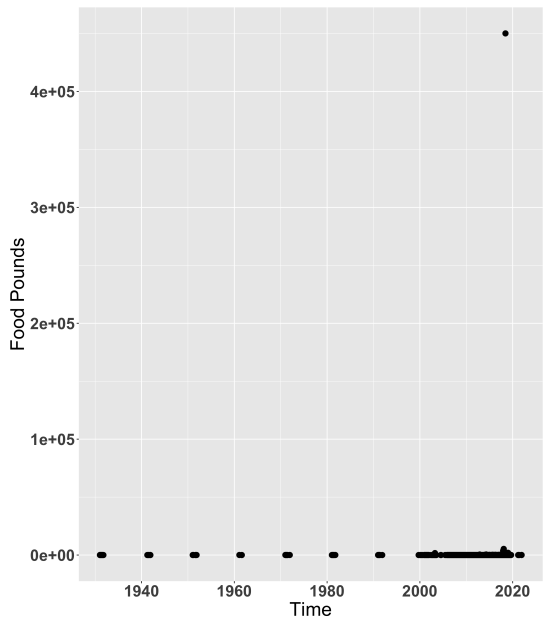

Figure 2: Total Food Pounds Every Day over Time

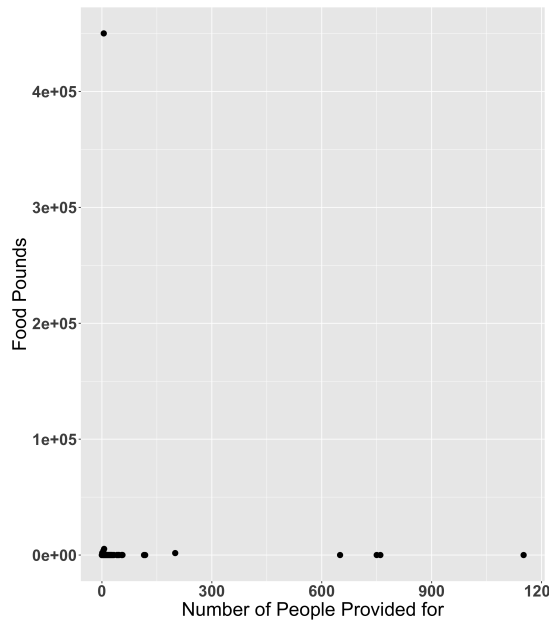

Figure 3: Food Pounds by Number of People

Question 1&2

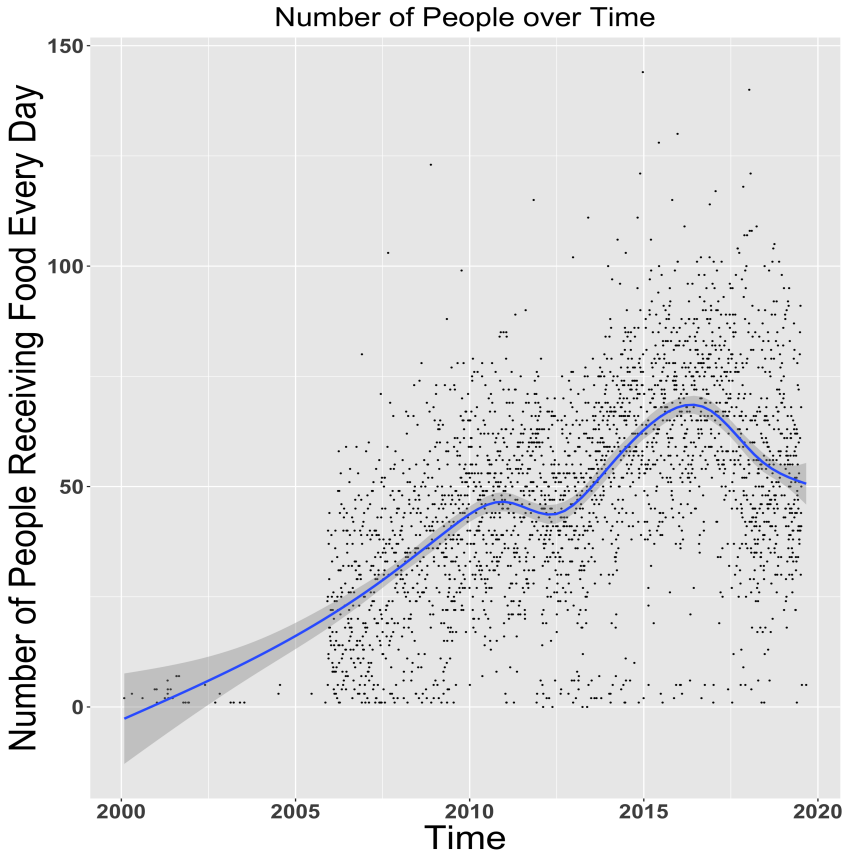

Figure 4: Number of People Every Day over Time

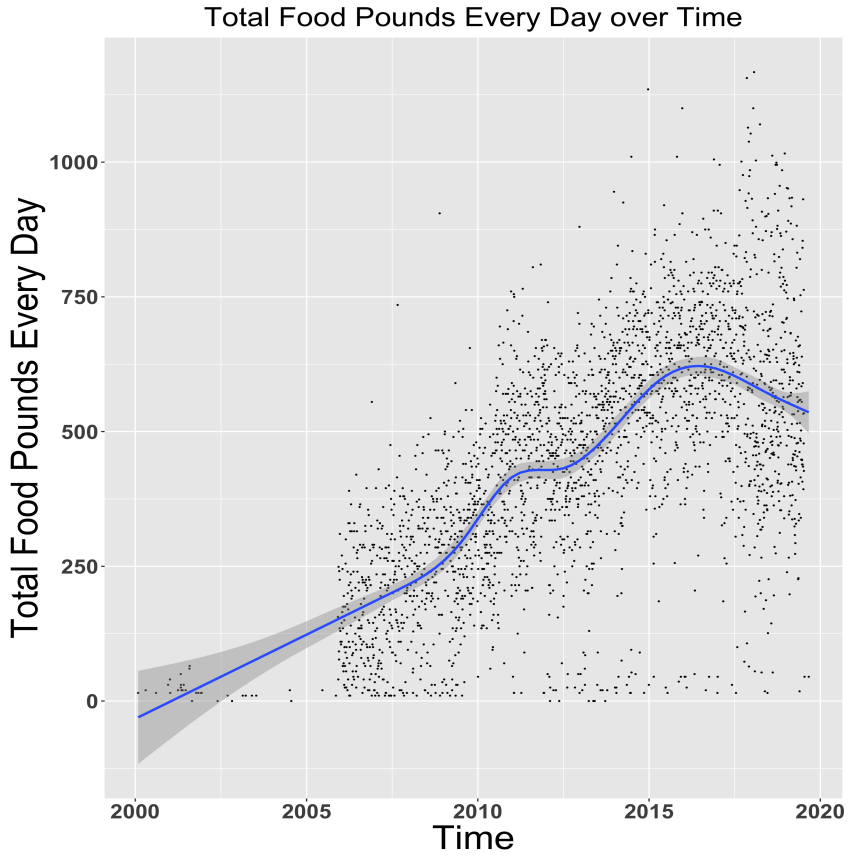

Figure 5: Total Food Pounds Every Day over Time

Question 1&2

- Total food pounds provided every day and number of people receiving food every day have the same trend over time.
- UMD is helping more and more people!
  - Both of them increase during 2005 and 2017.
- UMD is ending homelessness!
  - Growth slowed down during 2012 and 2013.
  - Both of them start to decrease after 2017.

Question 3

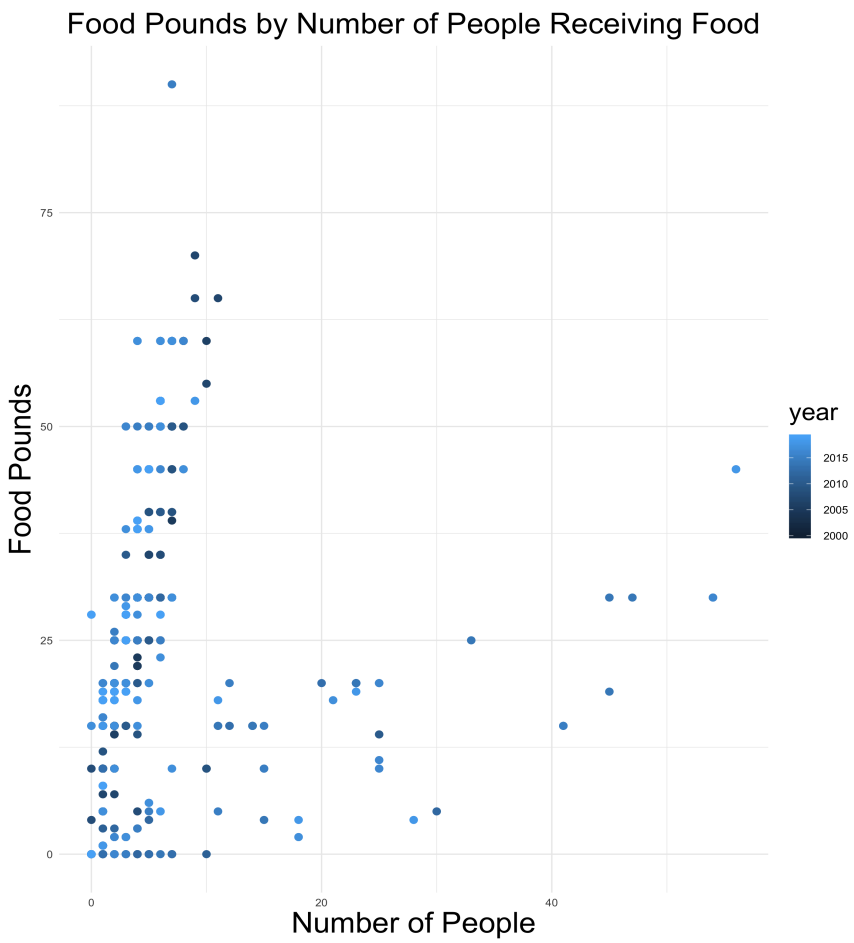

Figure 6: Food Pounds by Number of People Receiving Food

- Clearly, there are two groups of data points in this plot: one with big derivative and the other with small derivative.
- Derivative can be seen as estimated Average Food Pounds per person.
- We could use EM clustering method to justify which group that these data points belong to.
- EM algorithm finds the cluster of data points by iteratively maximizing marginal log likelihood of observed data.

Question 3

- Formally, let X be observed data, Z be the latent variable, which is the estimated cluster in our problem, and θ be unknown parameters along with a likelihood function $L(\theta; X, Z) = p(X, Z|\theta)$.
- EM algorithm finds clusters for each data points by iteratively applying *Expectation* step and *Maximization* step.
- *Expectation* step:
  - $Q(\theta|\theta^{(t)}) = \mathbb{E}_{Z^{(t)}|X,\theta^{(t)}}[\log(L(\theta; X, Z^{(t)}))]$
- *Maximization* step:
  - $\theta^{(t+1)} = \arg\max_\theta Q(\theta|\theta^{(t)})$
  - $Z^{(t+1)}|X = \arg\max_Z \log(L(\theta^{(t+1)}; X, Z))$

Question 3

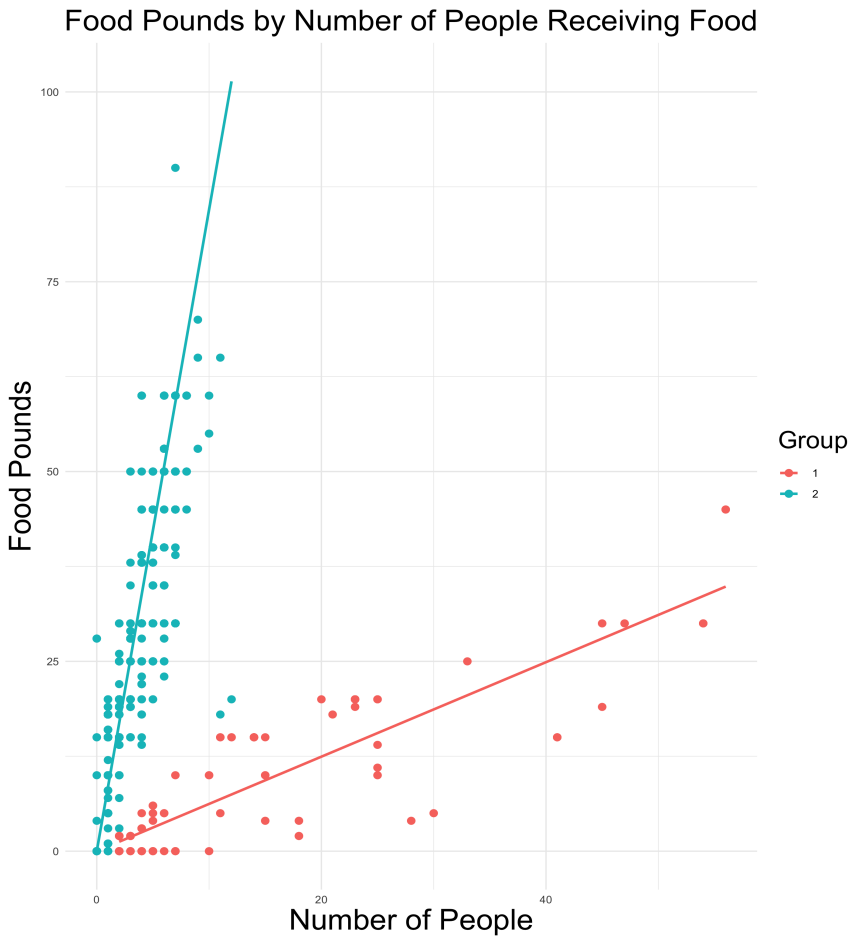

Figure 7: Food Pounds by Number of People Receiving Food

- EM algorithm appropriately divides the data into two groups as expected.
- Intuitively, data points in each group seems to be fitted well by simple linear regression.
- Since UMD does not need to provide food if there is no people, intercepts should be zero .
- $\text{Food.Pounds} = \beta \cdot \text{Food.Provided.for} + \epsilon$

Question 3

- The coefficient of Food.Provided.for (Number of People Receiving Food) is the estimated Average Food Pounds per person.
- For group 1, UMD provides 0.62 pounds of food for each person.
- For group 2, UMD provided 8.45 pounds of food for each person.
- A big difference in average food pounds per person exists between these two groups!

Conclusion

- Total Food Pounds provided every day and Number of People Receiving Food every day increase before 2017 and start to decrease after during 2017 and 2019.
- UMD is ending homelessness!
- People that UMD provided food for can be divided into two groups by Average Food Pounds per person.
- There is a huge difference in Average Food Pounds per person between two groups.

Future Analysis

- Future analysis may focus on the reason why differences exist between two groups. More variables should be added into analysis such as financial support, clothing items and identity numbers.