Project 1: Urban Ministries Durham

Jianqiao Wang

Backgroud and 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, past criminal convictions and addiction. The data provided by UMD recorded different kinds of support that UMD provided for homeless people from 1931. It has more than 10 variables including date, family identifiers, finicial support, etc.

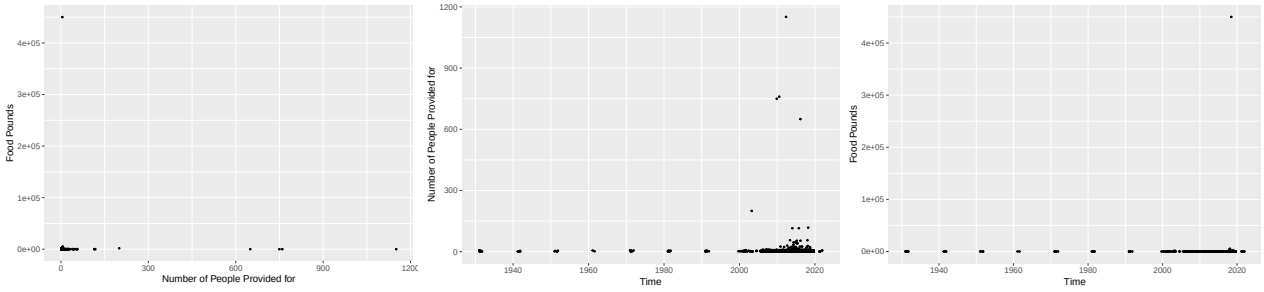
Since food plays an important role in helping homeless people, it is very useful to extract information about food from data. Therefore, analysis in this report will mostly focus on food that UMD provides. Specifically,

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

The analysis will mostly based on variables Date, Food.Pounds and Food.Provided.for.

Data Cleaning

First of all, we summarize the data into three plots to see what kind of data should be removed in our analysis.

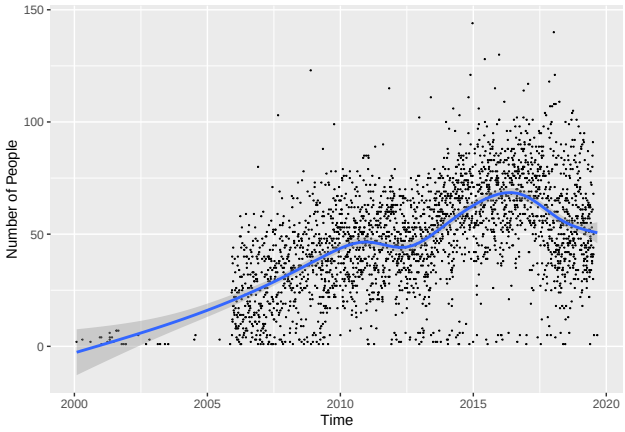


As we can see in the first plot, there are some outliers in the data that do not make sense in the real life. In the second and third plot, we notice that UMD even contains data after 2020, which is impossible. Therefore, we have to remove those data before analysis. Moreover, most data lies between year 2000 and year 2020. Therefore, we will only focus on data during that time.

Number of people that food is provided for and time

We first find the relationship between time and number of people UMD provided food for. The plot is as follows:

```
ggplot(total_food_provided_for, aes(x=Date, y=Food.Provided.for)) +
  geom_point(size=0.1) +
  labs(x='Time', y='Number of People') +
  geom_smooth()
```

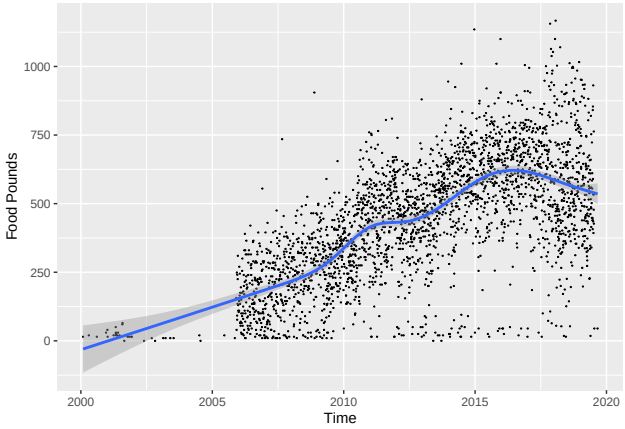


Generally, number of people that UMD helped in one day slowly increases during 2005 and 2019. Notice that there are two local maximum in the smoothed line. (Why?)

Total food pounds provided one day and time

The plot of food pounds UMD provided in one day and time is as follows:

```
ggplot(total_food_pound, aes(x=Date, y=Food.Pounds)) +
  geom_point(size=0.1) +
  labs(x='Time', y='Food Pounds') +
  geom_smooth()
```

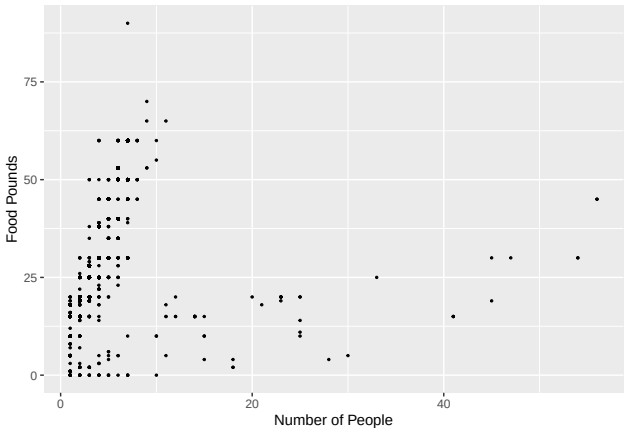


Similarly with last section, food pounds that UMD provided in one day increases during 2005 and 2020.

Average food pounds per person

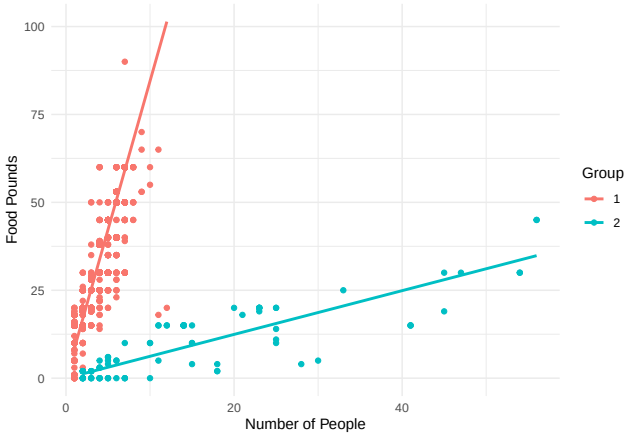
Finally, we give the plot of food pounds and number of people that UMD provided food for.

```
ggplot(umd_df, aes(Food.Provided.for, Food.Pounds)) +
  geom_point(size=0.5) +
  labs(x='Number of People', y='Food Pounds')
```



It is very interesting that there are two 'group' in this plot. And there is very likely a difference in average food pounds per person UMD provided between the two group. In order to see that difference, we need to model the two groups separately. To justify which group that these data points belongs to, we may use EM clustering method. The plot of data points by group is as follows:

```
ggplot(umd_df, aes(x=Food.Provided.for, y=Food.Pounds, color=cluster, group=cluster)) +
  geom_point() +
  geom_smooth(method='lm', se=FALSE, formula=y~x-1) +
  labs(x='Number of People', y='Food Pounds') +
  theme_minimal() +
  scale_color_discrete('Group')
```



It is clear that data points in each group seems to fit well by simple linear model. We assume that UMD does not provide food if there is no people. Therefore, the intercept for linear model should be zero. For group 1, the R squared statistic is

```
## [1] 0.9631234
```

The R squared statistic shows that the linear model fits very well for group 1. The estimated coefficient of group 1 is

```
## Food.Provided.for
##          8.449425
```

That is, UMD provides 8.45 pounds of food per person for group 1. For group 2, the R squared statistic is

```
## [1] 0.8551064
```

Therefore, linear model fits well for group 2. The estimated coefficient is

```
## Food.Provided.for
##         0.6222627
```

Then, UMD provides 0.62 pounds of food per person for group 2. Therefore, there is a big difference between these two groups. Most likely, the difference is caused by wealth difference between two groups. That is, people in group 2 may be richer than people in group 1.