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,

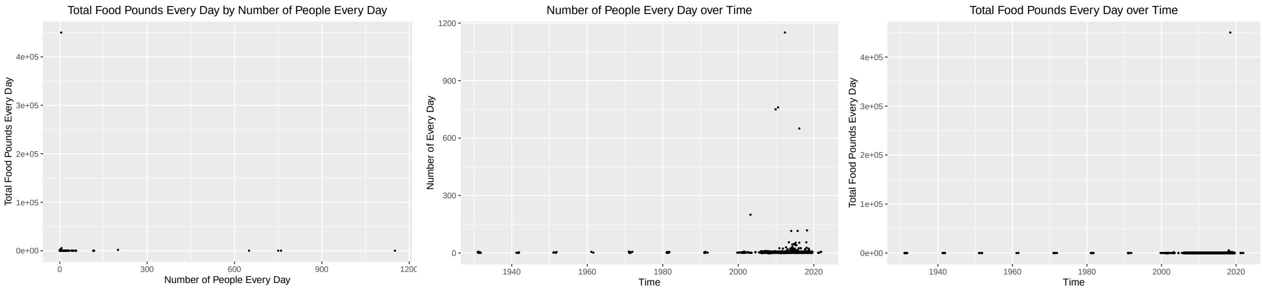
- Does the total number of people UMD provided food for every day increase?
- Does the total food pounds UMD provided every day increase?
- What is the average food pounds per person? Is there a difference among different families and people?

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

```r
umd_df = select(umd_df, c(Date, Food.Provided.for, Food.Pounds))
```

Data Cleaning

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



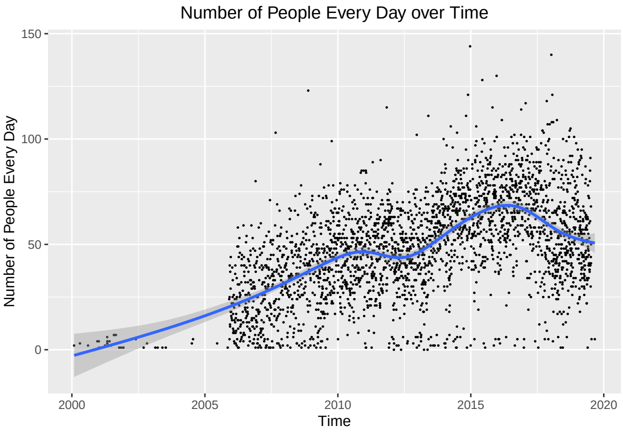
As we can see in the first plot, there are some outliers in the data that do not make sense in real life. In the second and third plot, we notice that UMD even contains data after 2020, which is impossible. Therefore, we have to remove those data before analysis. Moreover, most data lies between year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

```r
umd_df = umd_df %>%
  filter(Food.Pounds < 100, Food.Provided.for < 60) %>%
  filter(Date < as.Date('2019-09-24'), Date > as.Date('2000-01-01'))
```

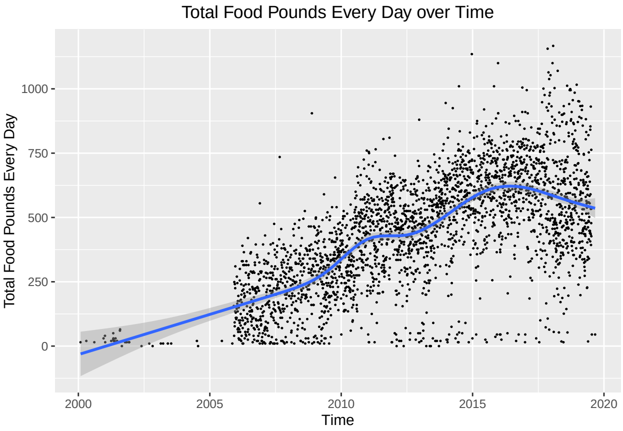
Food Plots over Time

First of all, we can find the trend of total food pounds every day and number of people every day over time. The plots are as follows:

```r
ggplot(number_of_people_every_day, aes(x=Date, y=Food.Provided.for)) +
  geom_point(size=0.2) +
  labs(x='Time', y='Number of People Every Day', title='Number of People Every Day over Time') +
  theme(plot.title = element_text(hjust = 0.5)) +
  geom_smooth()
```



```
ggplot(total_food_pound_every_day, aes(x=Date, y=Food.Pounds)) +
  geom_point(size=0.2) +
  labs(x='Time', y='Total Food Pounds Every Day', title='Total Food Pounds Every Day over Time') +
  theme(plot.title = element_text(hjust = 0.5)) +
  geom_smooth()
```

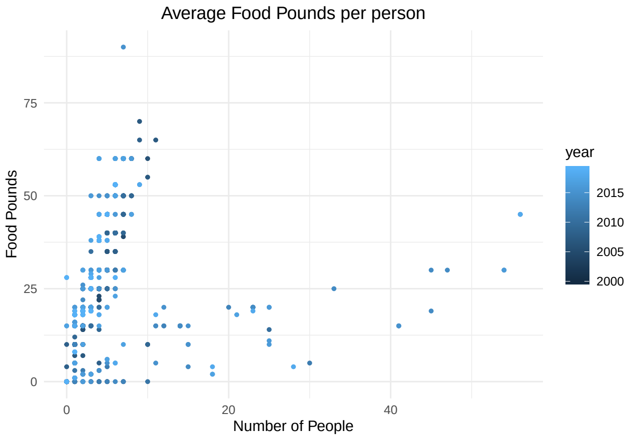


Generally, total food pounds every day and number of people that UMD helped every day have the same trend over time. Both of them increase during 2005 and 2017. However, the growth slowed down during 2012 and 2013. Moreover, total food pounds every day and number of people every data start to decreases after 2017.

Average Food Pounds per person

Next, we try to figure out average food pounds per person UMD provided. Intuitively, UMD may treat people in the same way and provide the same amount of food per person. We first give the plot of total food pounds every day and number of people every day.

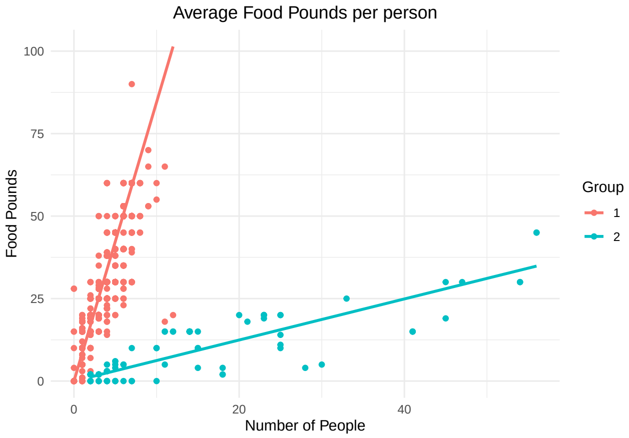
```
ggplot(umd_df, aes(Food.Provided.for, Food.Pounds, color=year)) +
  geom_point(size=1) +
  labs(x='Number of People', y='Food Pounds', title='Average Food Pounds per person') +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5))
```



It is very interesting to note that there are two groups in this plot and they are not relevant to time. Therefore, there is very likely a difference in average food pounds per person UMD provided. 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 could 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+0) +
  labs(x='Number of People', y='Food Pounds', title='Average Food Pounds per person') +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5)) +
  scale_color_discrete('Group')
```



We can see that EM algorithm fits the data well. Intuitively, data points in each group seem to fit well by simple linear regression and the intercepts should be zero since UMD does not need to provide food if there is no people. The fitted linear model has been plotted in the above figure. The specific model is

$$\text{Food.Pounds} = \beta \cdot \text{Food.Provided.for} + \epsilon$$

For group 1, the R squared statistic is

```
## [1] 0.9630706
```

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, the 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. There is a big difference in average food pounds per person between these two groups, which is counter-intuitive.

Future Analysis

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