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, financial 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 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?

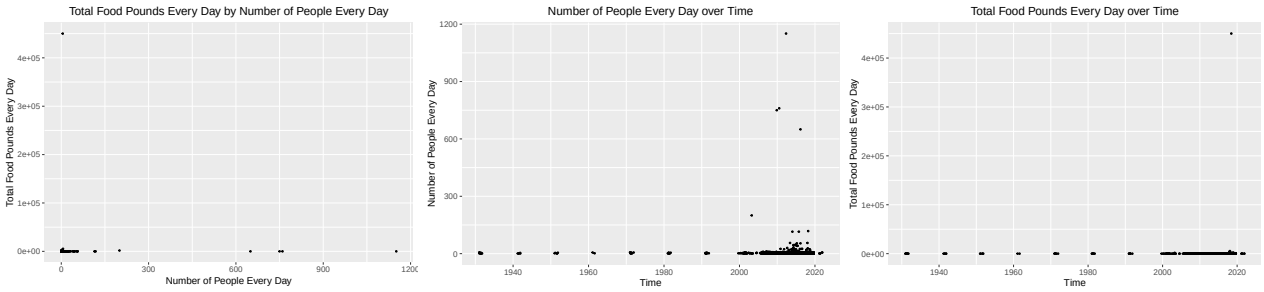
The analysis will mostly based on variables Date, Food.Pounds and Food.Provided.for, where

- Date: Time.
- Food.Pounds: Food Pounds UMD provided one time.
- Food.Provided.for: Number of People Receiving Food one time.

```
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.



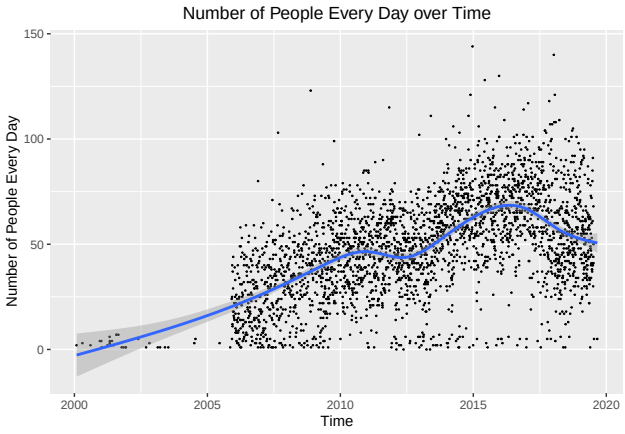
As we can see in the first plot, there are some outliers (Number of People Every Day > 75 or Total Food Pounds Every Day $>$ 4e+05) in the data that do not make sense in real life. In the second and third plot, we notice that UMD even contains data after 2020, which must be a typo. We will only consider data before the day that we begin our analysis, that is 2019-09-24. Moreover, the data before 2000 is too sparse and discrete so that the data may not provide sufficient and reliable information. Therefore, we will only focus on data during 2000 and 2019, which will provide most information in recent years.

```
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'))
```

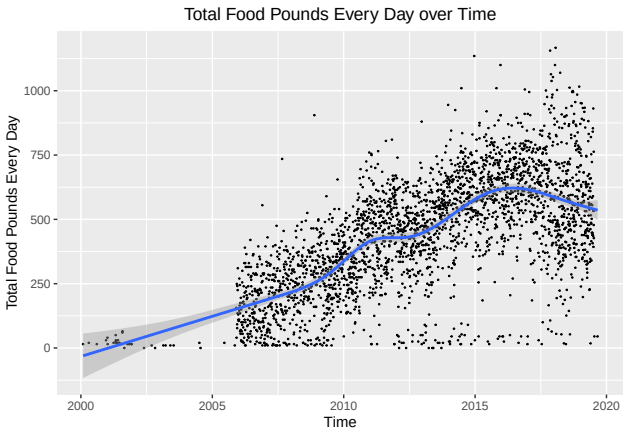
Food Plots over Time

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:

```
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()
```

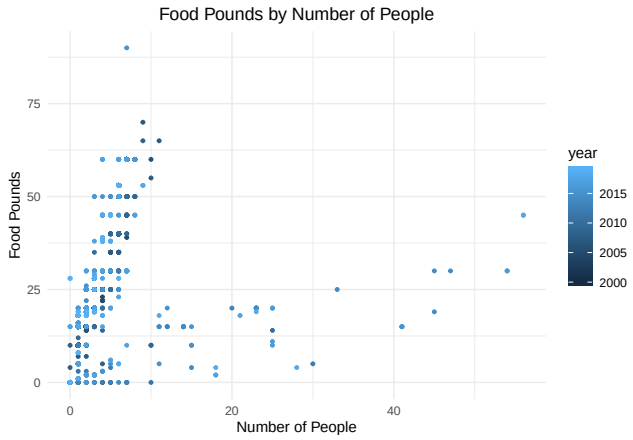


Generally, total food pounds every day and number of people that UMD helped every day have the same trend over time. Both of them increase during 2005 and 2017, which may be attributable to UMD's great work. However, the growth slowed down during 2012 and 2013, which indicates that UMD may end some people's homelessness. Moreover, total food pounds every day and number of people every data start to decrease after 2017, which indicates that UMD has ended many people's homelessness since 2017.

Average Food Pounds per Person

Next, we try to figure out average food pounds per person UMD provided. Intuitively, UMD should treat people in the same way and provide the same amount of food per person. To see if there is a difference in it, we give the plot of food pounds for one time by corresponding number of people. If there is significant divide of all data points, the difference may exist.

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



It is very interesting to note that there are two significant 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 belong to, we could use EM clustering method. EM algorithm finds the cluster of data points by iteratively maximizing marginal log likelihood of observed data. 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 $L(\theta; X, Z) = p(X, Z|\theta)$. Then the EM algorithm finds clusters for each data points by iteratively applying following steps:

- *Expectation Step*:
  - $Q(\theta|\theta^{(t)}) = \mathbb{E}_{Z^{(t)}|X,\theta^{(t)}}[\log(L(\theta; X, Z^{(t)}))]$
- *Maximization Step*:
  - $\theta^{(t+1)} = \arg\max_\theta Q(\theta|\theta^{(t)})$
  - $Z^{(t+1)}|X = \arg\max_Z \log(L(\theta^{(t+1)}; X, Z))$.

The plot of data points by estimated group through EM algorithm 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='Food Pounds by Number of People') +
  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 be fitted 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] 0.8551064
```

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
##         0.6222627
```

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

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

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

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

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.

Conclusion

In this report, we analysis the trend of food provided and number of people UMD provided food for overtime. Generally, both of them increase before 2017. They began to decrease since 2017, which may be resulted from great work of UMD. UMD is ending homelessness! We also find out that people that UMD provided food for can be divided into two groups. UMD provided different average food pounds per person for two groups. We calculate average food pounds per person for each group, which is the coefficient of number of people.

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.