Project 1: Urban Ministries of Durham

Yangjianchen Xu

9/24/2019

```
library(tidyverse)
library(readr)
library(lubridate)
```

Background

The data is provided by Urban Ministries of Durham, of which the people in this organization connect with the community to end homelessness and fight poverty by offering food, shelter and a future to neighbors in need. The data includes 79838 observations and 18 variables.

Data import

```
UMD=read_tsv("~/Documents/GitHub/bios611-projects-fall-2019-Poutine1025/project_1/data/UMD_Services_Pro
metadata=read_tsv("~/Documents/GitHub/bios611-projects-fall-2019-Poutine1025/project_1/data/UMD_Service
```

Questions of interest

The basic questions I am interested in are

- What is the relationship among Food Pounds, Clothing Items and Number of people in the family for which food was provided?
- How does time influence the amount of food and clothing items?

By the analysis of the above questions, I will try to answer some further questions like

- What is the amount of food need to be provided in 2019?

Data cleaning

```
head(UMD)
```

```
## # A tibble: 6 x 18
##   Date  `Client File Nu~ `Client File Me~ `Bus Tickets (N~ `Notes of Servi~
##   <chr>            <dbl>            <dbl>            <dbl> <chr>
## 1 1/22~              212                0               NA <NA>
## 2 1/29~              738                0               NA <NA>
## 3 1/20~             3455                0               NA <NA>
## 4 11/2~             1804            21804               NA <NA>
## 5 12/2~             1806            21806               NA <NA>
## 6 10/1~             1614            21614               NA financial refer~
## # ... with 13 more variables: `Food Provided for` <dbl>, `Food
## #   Pounds` <dbl>, `Clothing Items` <dbl>, Diapers <dbl>, `School
## #   Kits` <dbl>, `Hygiene Kits` <dbl>, Referrals <chr>, `Financial
## #   Support` <dbl>, `Type of Bill Paid` <chr>, `Payer of Support` <chr>,
## #   Field1 <lgl>, Field2 <lgl>, Field3 <lgl>
```

The data consists of 79838 observations with 18 variables such as Date, Client File Number, Food Pounds, Clothing Items and Number of people in the family for which food was provided. I will foucus on these 5

variables to answer the questions of interest and dicard observations with NA. For convenience, I will simplify the variable names. Hence, clean the data as follows:

```
UMD_selected=UMD %>%
  select(`Date`, `Client File Number`, `Food Pounds`, `Clothing Items`, `Food Provided for`) %>%
  rename(CFN=`Client File Number`, food=`Food Pounds`, clothing=`Clothing Items`, number=`Food Provided
  drop_na()
head(UMD_selected)
```

```
## # A tibble: 6 x 5
##   Date        CFN  food clothing number
##   <chr>     <dbl> <dbl>    <dbl>  <dbl>
## 1 1/22/2009   212    20        5      3
## 2 1/29/2009   738    25       26      4
## 3 1/20/2009  3455    40       39      6
## 4 2/19/2008  1814     0        0      0
## 5 5/1/1931      4    15       10      1
## 6 1/26/2006     4    10        2      1
```

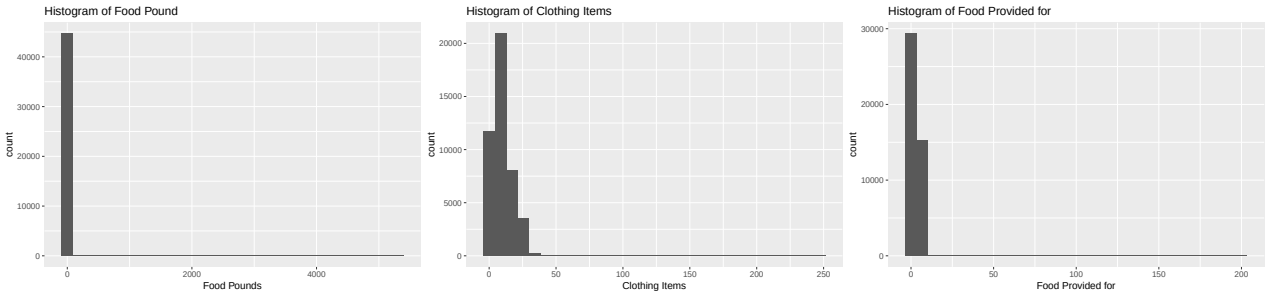
Here, "CFN" stands for "Client File Number", "food" stands for "Food Pounds", "clothing" stands for "Clothing Items", and "number" stands for "Number of people in the family for which food was provided".

Question 1

The first question of interest is what the relationship is among Food Pounds, Clothing Items and Number of people in the family for which food was provided.

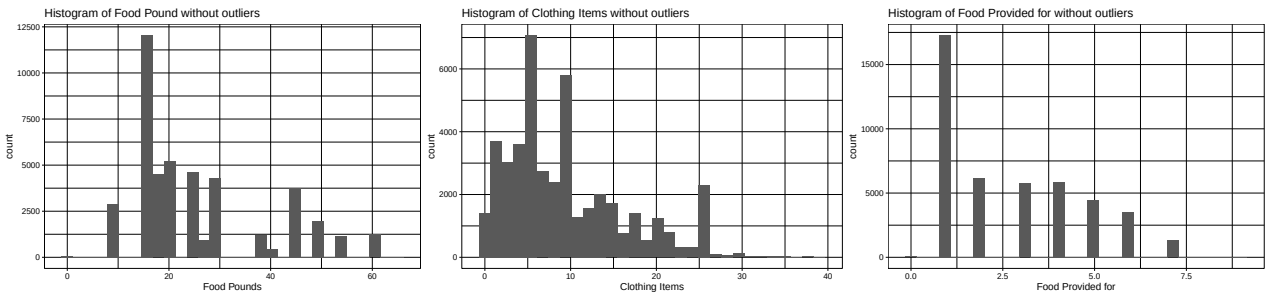
Data preview

The histograms of Food Pounds, Clothings Items and Food Provided for are as follows:



As we can see, there are some extreme values among these 3 variables. After examining the variables, I decided to discard observations with Food Pounds larger than 75 or Clothing Items larger than 40 or Food Provided for larger than 10. Thus, I clean the data and plot histograms again as follows:

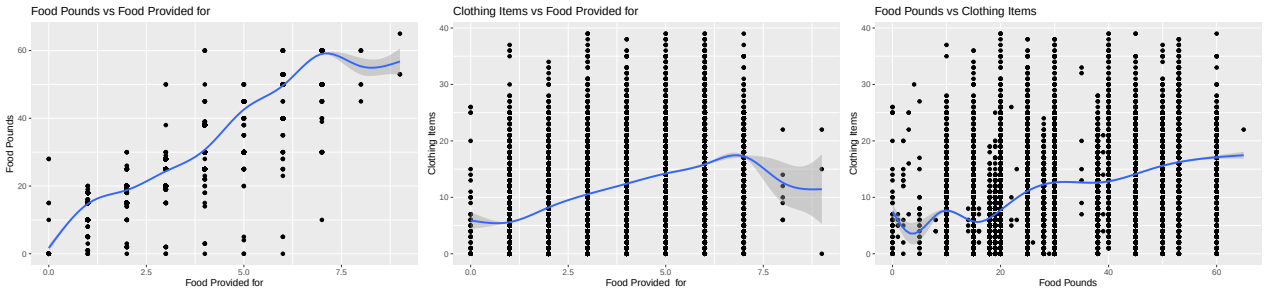
```
UMD_selected=UMD_selected %>% filter((food<75) & (clothing<40) & (number<10))
```



We can see after removing the outliers the distributions of these 3 variables become more normal, that is, they in a reasonable range.

Correlation of variables

In order to explore the relationship among them, I plotted the scatterplots for each pair of the variables.

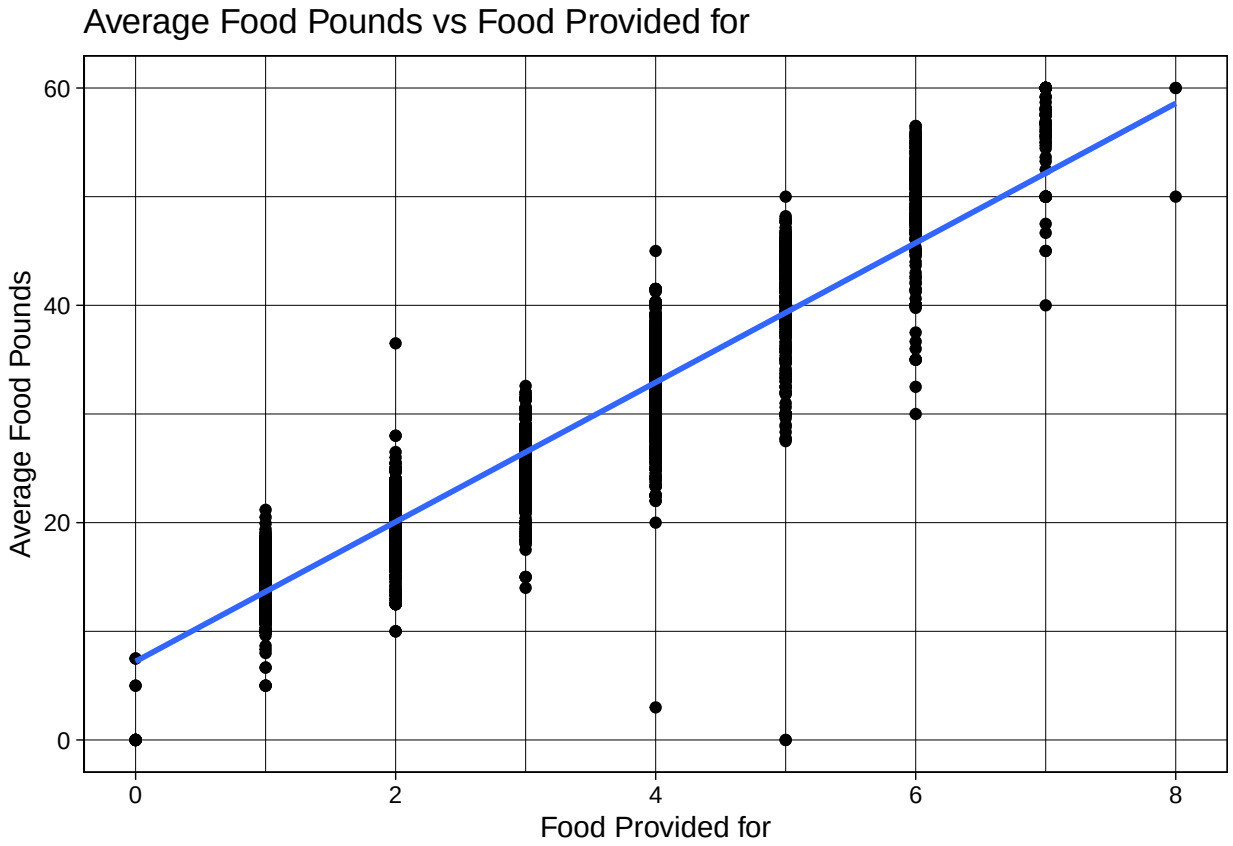


The scatterplots show that Food Pounds is positive correlated with Food Provided for and there is no apparent relationship in the other two pairs of variables. Since many records share the same Client File Number, it is necessary to group the data by Client File Number and see what happens. The following code groups the data by CFN and creats 2 new variables - Average Food Pounds and Average Clothing Items.

```
#subset1
UMD_subset1=UMD_selected %>%
  group_by(CFN) %>%
  summarize(number=round(mean(number)), food=sum(food), clothing=sum(clothing), freq=n()) %>%
  mutate(food_mean=food/freq, clothing_mean=clothing/freq)
```

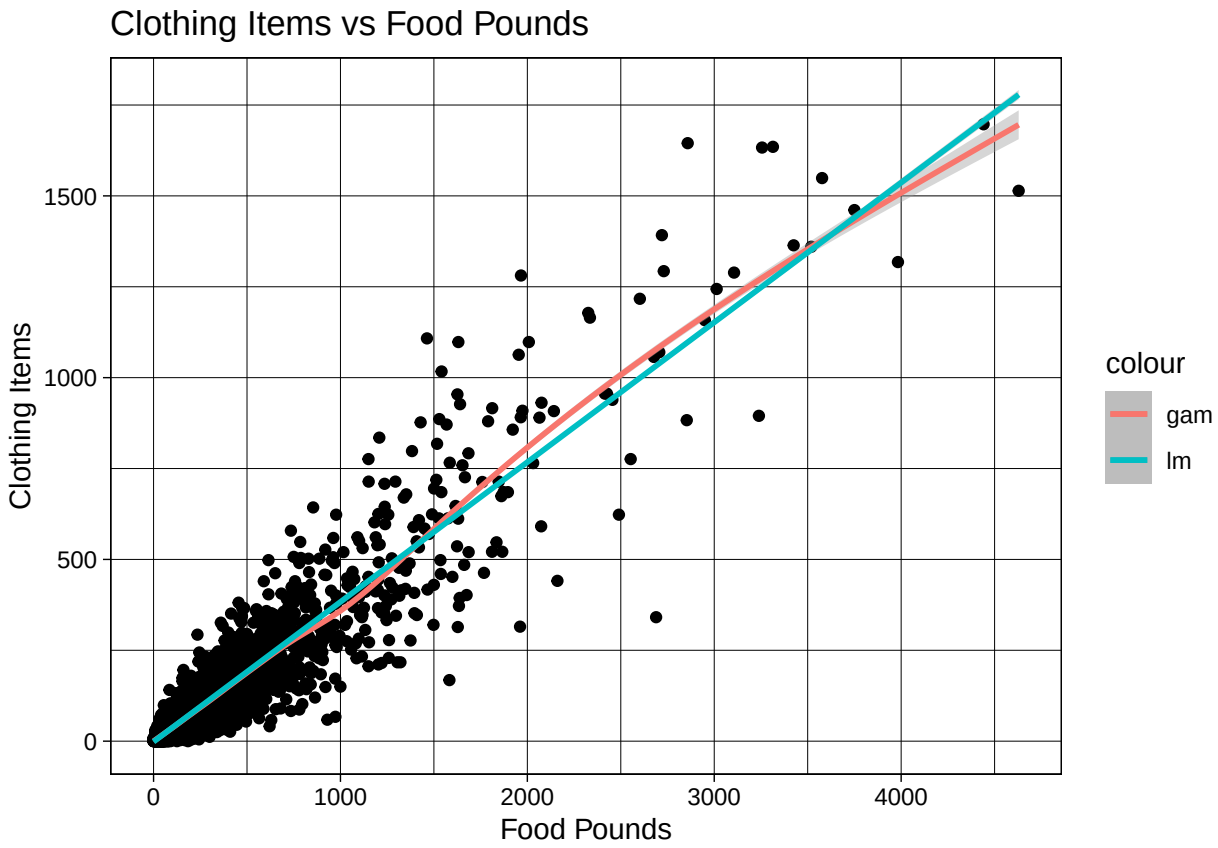
After grouping, I plotted the scatterplots for 2 pairs of the variables and found some correlation.

```
#number vs food_mean
ggplot(data=UMD_subset1,aes(x=number, y=food_mean)) +
  theme_linedraw() +
  geom_point() +
  geom_smooth(method = "lm", se = TRUE) +
  labs(x="Food Provided for",
       y="Average Food Pounds",
       title = "Average Food Pounds vs Food Provided for")
```



We can see this figure shows a stronger positive correlation between Average Food Pounds and Food Provided for.

```
#food vs clothing
ggplot(data=UMD_subset1,aes(x=food, y=clothing)) +
  theme_linedraw() +
  geom_point() +
  geom_smooth(aes(colour="gam")) +
  geom_smooth(method = "lm", aes(colour="lm")) +
  labs(x="Food Pounds",
       y="Clothing Items",
       title="Clothing Items vs Food Pounds")
```



After grouping the data, Clothing Items and Food Pounds show a much stronger positive correlation. Thus, we can fit a linear model to the data and predict one variable by the other to some degree.

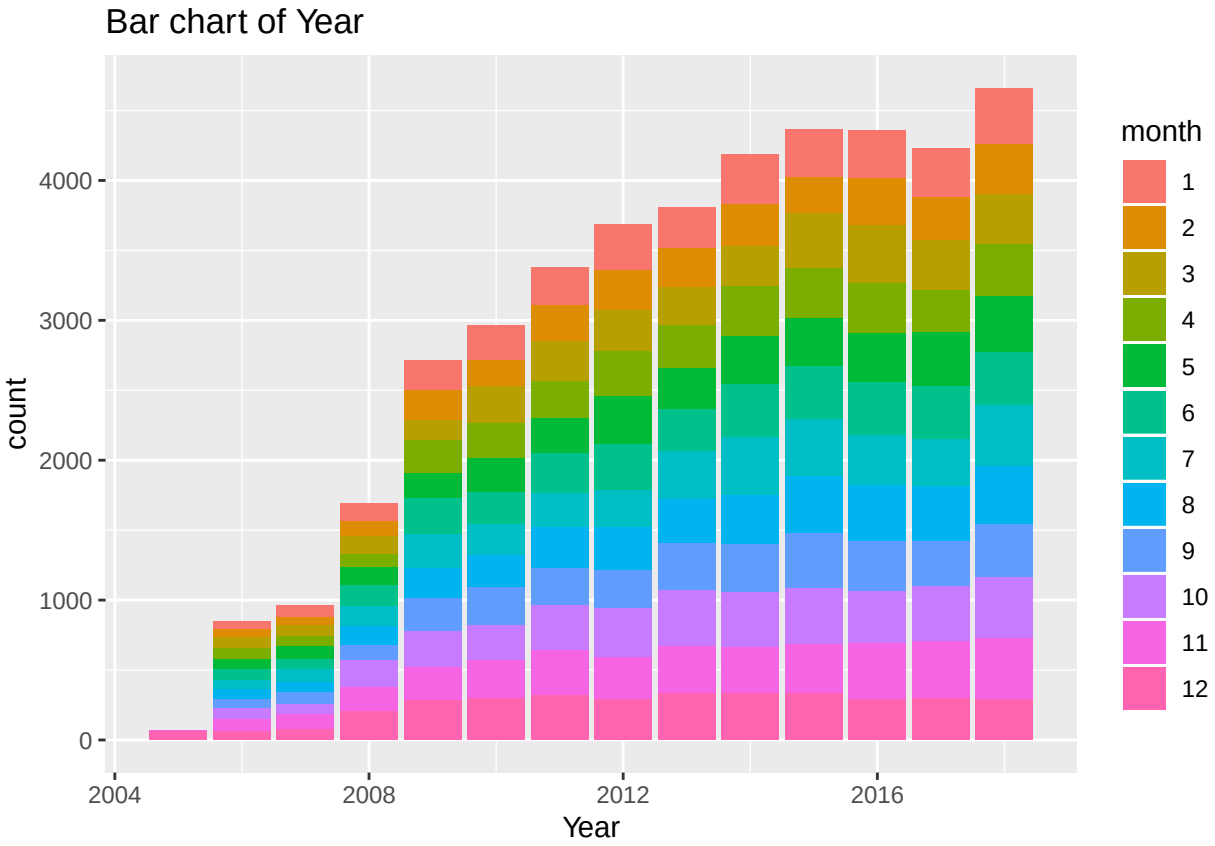
Question 2

The second question of interest is hows time influences the amount of food and clothing items. I first extracted the information from variable Date and remove observations with too small samples size. The records since 2019 are not complete so I remove them as well.

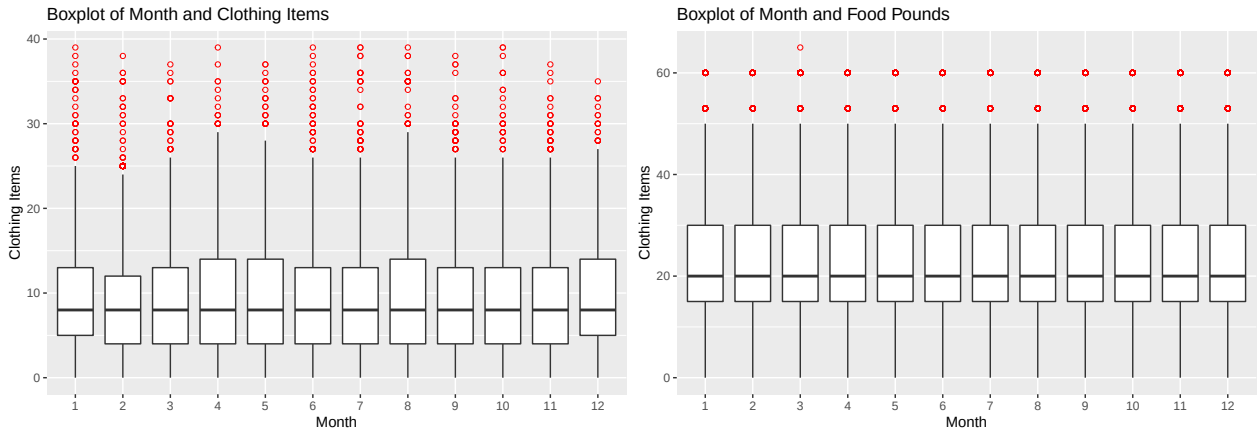
```
#subset2
UMD_subset2=UMD_selected %>%
  mutate(Date=as.Date(Date, format="%m/%d/%Y")) %>%
  mutate(year=year(Date), month=month(Date), day=day(Date)) %>%
  mutate(month=as.factor(month), day=as.factor(day)) %>%
  filter((year>2004) & (year<2019))
```

Then we can plot the bar chart of Year and examine the distribution across months.

```
#bar chart of year
ggplot(data=UMD_subset2,aes(year)) +
  geom_bar(aes(fill=month)) +
  labs(x="Year",
       title = "Bar chart of Year")
```



We can see that the amount of records is increasing by year and they are distributed uniformly across months. The following two boxplots show the distributions of Food Pounds and Clothing Items across months.



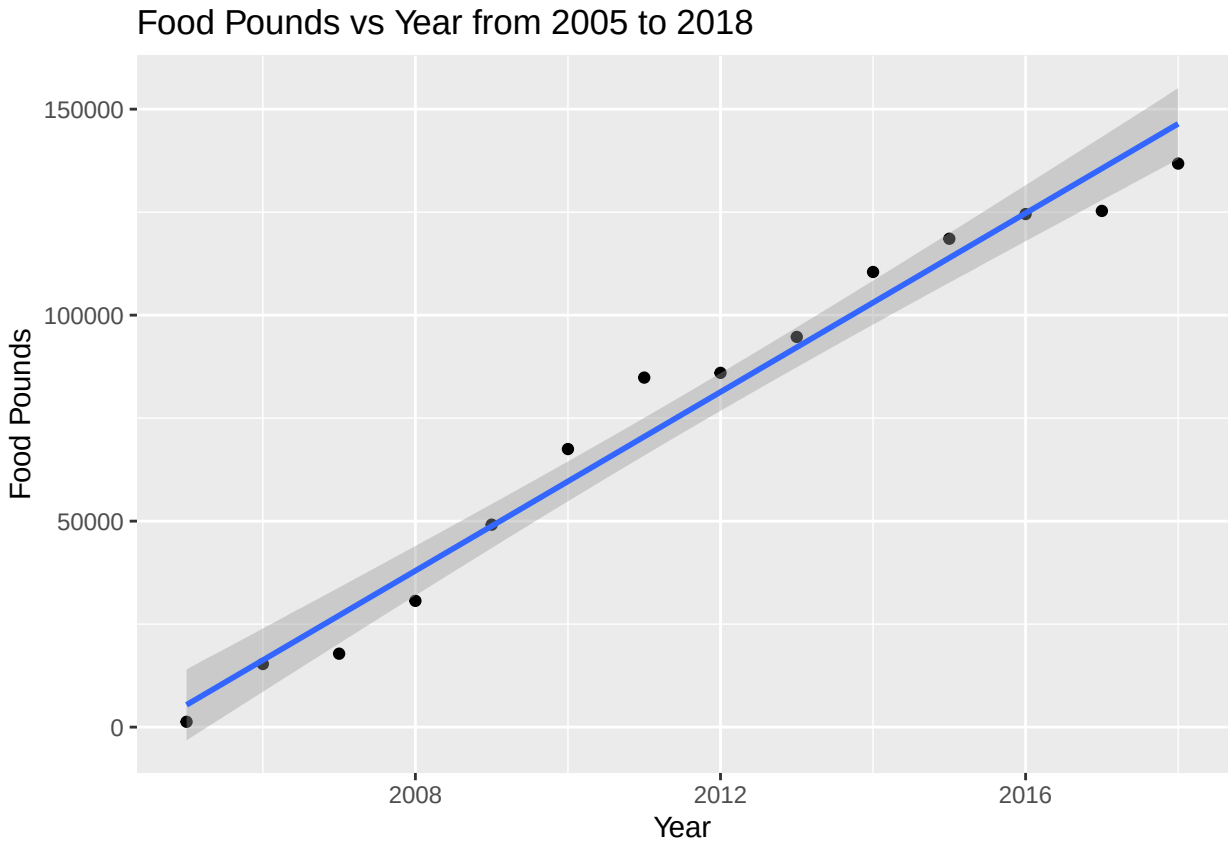
These two figures indicate the distributions of Food Pounds and Clothing Items are roughly the same across the months.

In order to predict the amount of Food Pounds of 2019, it is meaningful to plot the scatterplot of Food Pounds and Year.

```
#subset3
UMD_subset3=UMD_subset2 %>%
  group_by(year) %>%
  summarize(food=sum(food), clothing=sum(clothing), number=sum(number), freq=n())

ggplot(data = UMD_subset3, aes(x=year, y=food)) +
```

```
  geom_point() +
  geom_smooth(method = "lm") +
  labs(x="Year",
       y="Food Pounds",
       title="Food Pounds vs Year from 2005 to 2018")
```



It was found that there seemed to be a linear relationship between Food Pounds and Year. Thus, we can predict the amount of Food Pounds by fitting a linear model to the data.

Question 3

The last question is what the amount of food need to be provided is in 2019. I used the figure "Food Pounds vs Year from 2005 to 2018" to answer question 3. By fitting a linear model to the corresponding data, I found the predicted value.

```
model_FoodvsYear=lm(food~year, data = UMD_subset3)
Food2019=predict(model_FoodvsYear, newdata = data.frame(year=2019))
print(paste("The predicted value of Food Pounds of 2019 is", Food2019))
```

```
## [1] "The predicted value of Food Pounds of 2019 is 157291.472527474"
```

Conclusion

- There is a stronger positive correlation between Average Food Pounds and Food Provided for.
- There is a stronger positive correlation between Food Pounds and Clothing Items.
- The amount of records is increasing by year and they are distributed uniformly across months.
- The distributions of Food Pounds and Clothing Items are roughly the same across the months.
- There is a linear relationship between Food Pounds and Year.

- The predicted value of Food Pounds of 2019 is around 157291.