BIOS 611 PROJECT 1

OCT 7TH, 2019

BY WONKYUNG JANG

Background & Data source

The Urban Ministries of Durham (UMD) Project aims for connecting with the community to diminish homelessness and fighting poverty through providing food, shelter and a future to neighbors with special needs (UMD, 2019).

The dataset is offered by the Urban Ministries of Durham (UMD) Project Team (http://www.umdurham. org/), which includes the dataset with 79838 observations from 1990's to 2019.

URBAN MINISTRIES OF DURHAM

Variables

1. Client File Number (Identifier)
2. Bus Tickets: Service discontinued
3. Food: # of people in the family for which food was provided
4. Food Pounds: # of pounds of food that each individual or family received when shopping the food pantry
5. Clothing Items: # of clothing items that each individual or family received in the clothing closet
6. Diapers: # of packs of diapers received (individuals/families are given 2 packs of diapers per child, and packs contain 22 diapers on average)
7. School Kits
8. Hygiene Kits: # of kits received per individual or family. Kits contain soap, shampoo, conditioner, lotion, deodorant, a toothbrush, toothpaste, a washcloth, a disposable razor, and a bottle of shaving cream.
9. Financial Support: Service discontinued

Rename the variables

```
data <- rawdata %>%
  rename(ClientID = Client.File.Number, Bus = Bus.Tickets..Number.of., Note = Notes.of.Service,
         Food = Food.Provided.for, Clothing = Clothing.Items) %>%
  select(Date, ClientID, Bus, Food, Food.Pounds, Clothing, Diapers, School.Kits, Hygiene.Kits,
         Referrals, Note, Financial.Support)
```

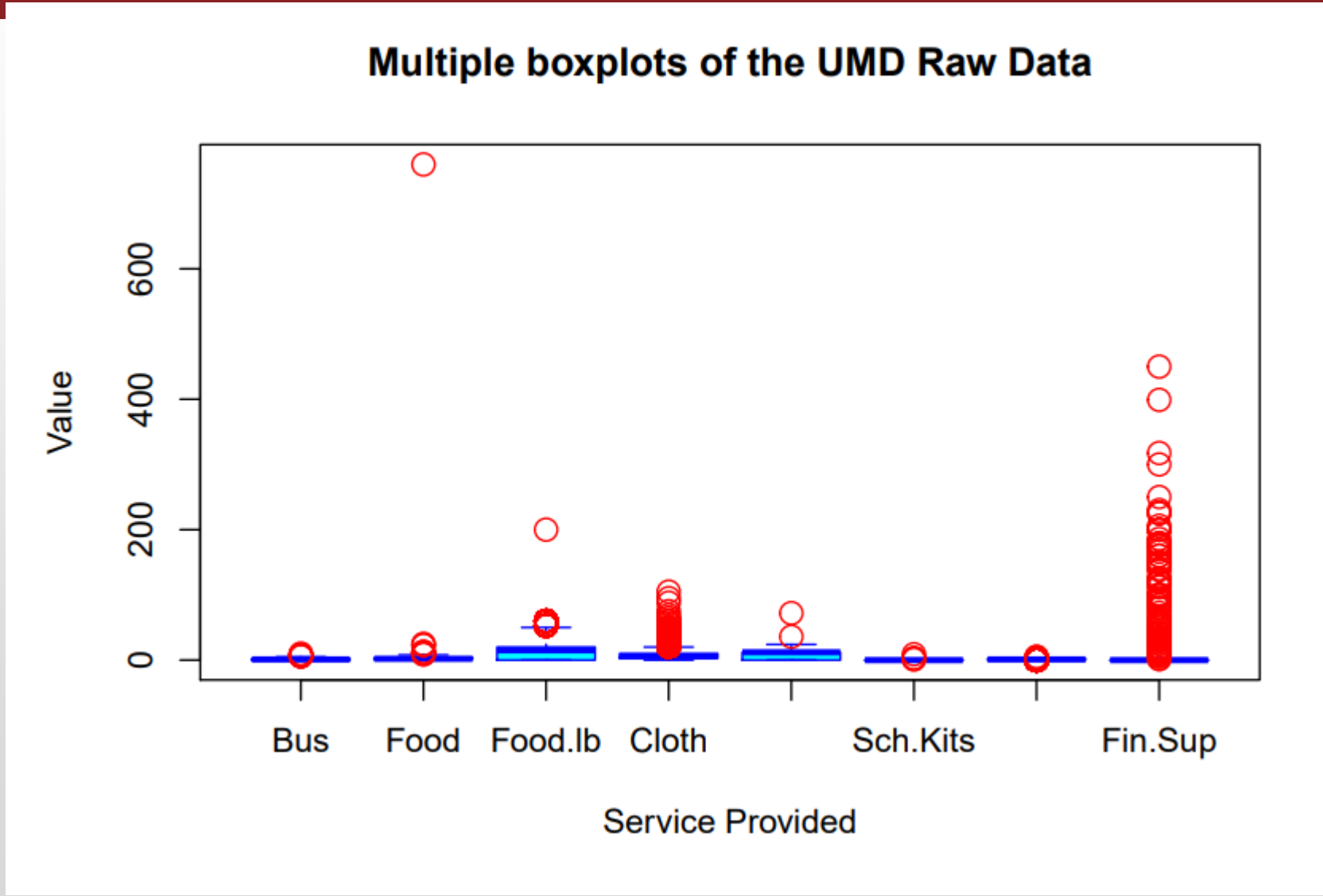
Filtering

```
data$Date <- as.Date(data$Date, format = "%m/%d/%Y")
data = data %>%
  filter(Date >= "1983-01-01" & Date <= "2019-10-01")
```

Given that the UMD was established in 1983, we should remove some rows with dates before 1983 or after 2019. Here, the Date data was converted into DATE format.

Outliers

Trimming: detect, visualize and test for outliers



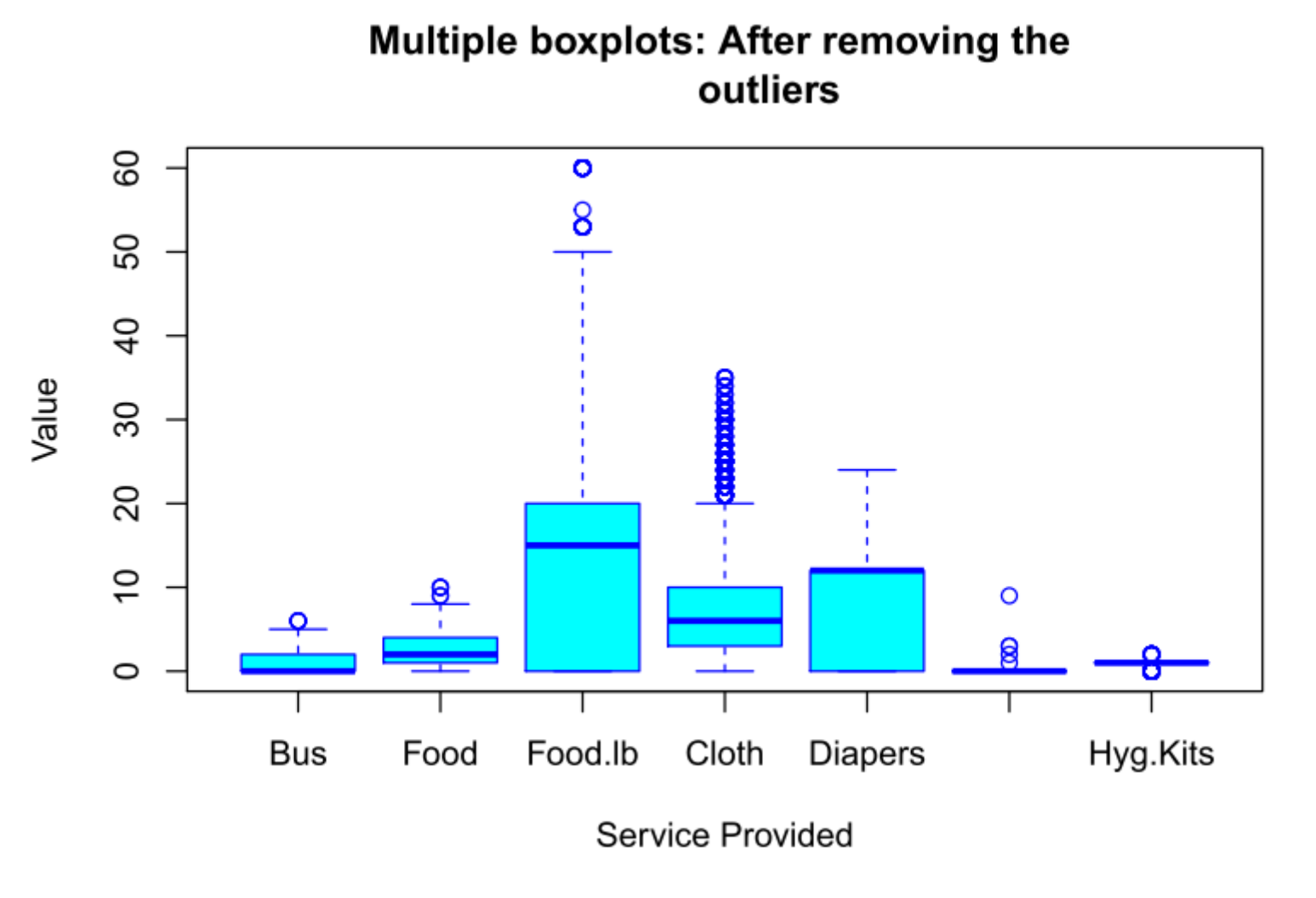
Rosner test

$$R_{i+1} = \frac{|x^{(i)} - \bar{x}^{(i)}|}{s^{(i)}}$$

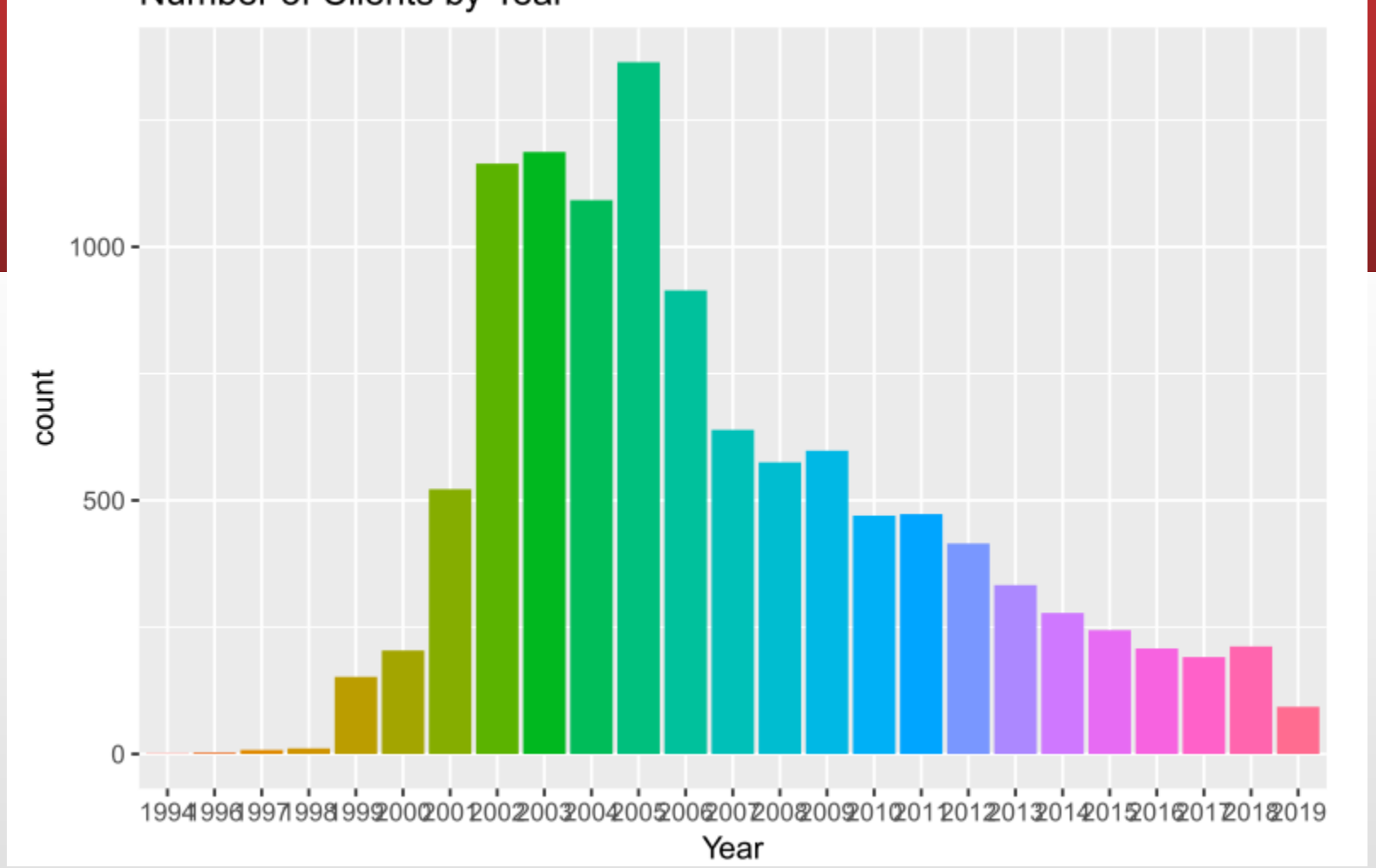
```
rosnerTest(data$Bus, k = 10, warn = F)
```

```
## Alternative Hypothesis:           Up to 10 observations are not
##                                   from the same Distribution.
##
## Type I Error:                     5%
##
## Number of Outliers Detected:      6
##
##      i    Mean.i      SD.i Value Obs.Num    R.i+1 lambda.i+1 Outlier
## 1    0 1.348837 1.842647    10      10 4.694964   3.680486    TRUE
## 2    1 1.315175 1.764970     8      11 3.787500   3.679364    TRUE
## 3    2 1.289062 1.717967     8      12 3.906326   3.678238    TRUE
## 4    3 1.262745 1.668839     8      13 4.037091   3.677106    TRUE
## 5    4 1.236220 1.617379     8      14 4.181938   3.675969    TRUE
## 6    5 1.209486 1.563341     7      15 3.703934   3.674828    TRUE
## 7    6 1.186508 1.523042     6      16 3.160445   3.673681   FALSE
## 8    7 1.167331 1.495289     6      17 3.231931   3.672528   FALSE
## 9    8 1.148000 1.466526     6      18 3.308499   3.671371   FALSE
## 10   9 1.128514 1.436685     6      19 3.390783   3.670208   FALSE
```

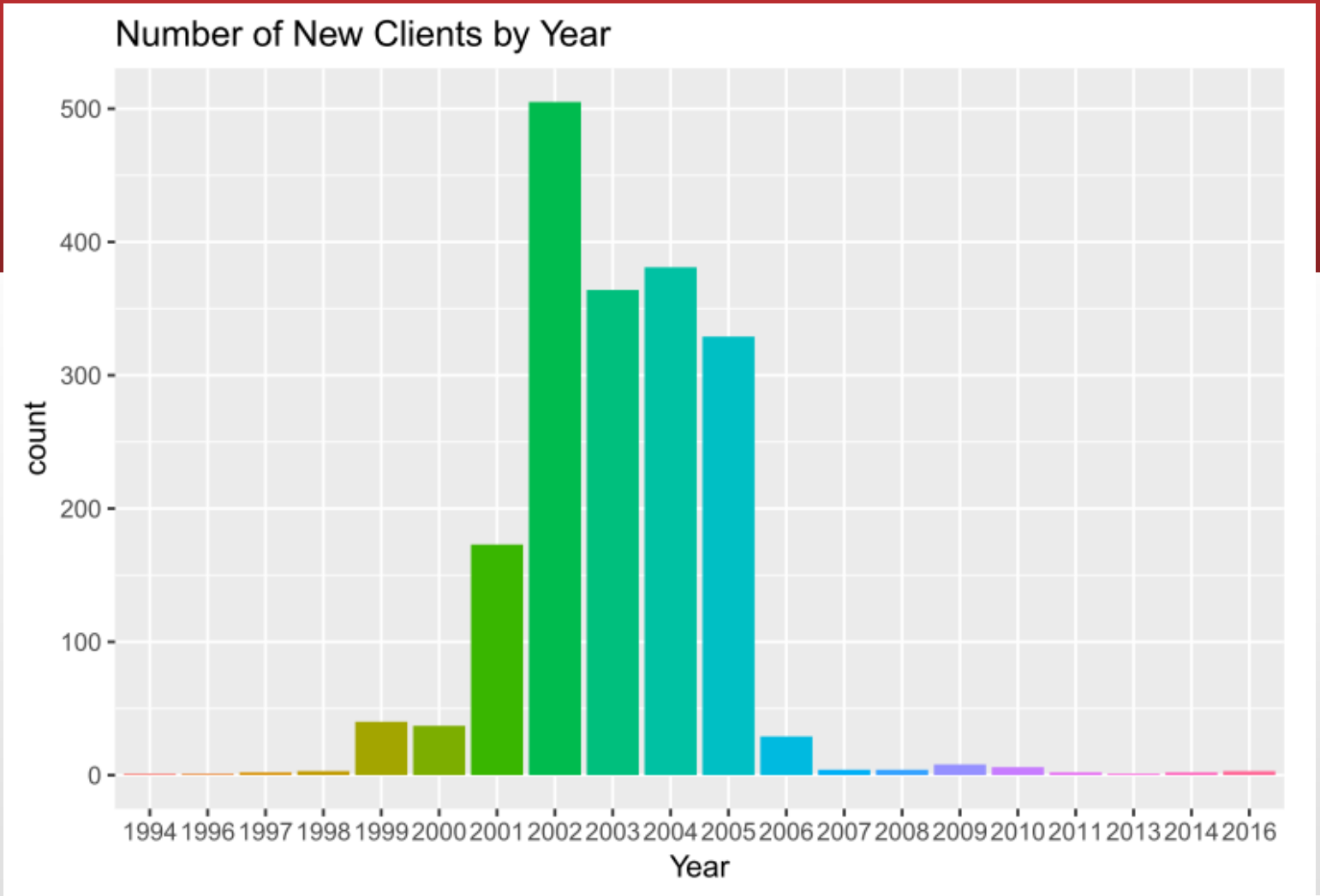
Rosner test


Multiple boxplots: After removing the outliers
Value
0
10
20
30
40
50
60
Bus
Food
Food.lb
Cloth
Diapers
Hyg.Kits
Service Provided

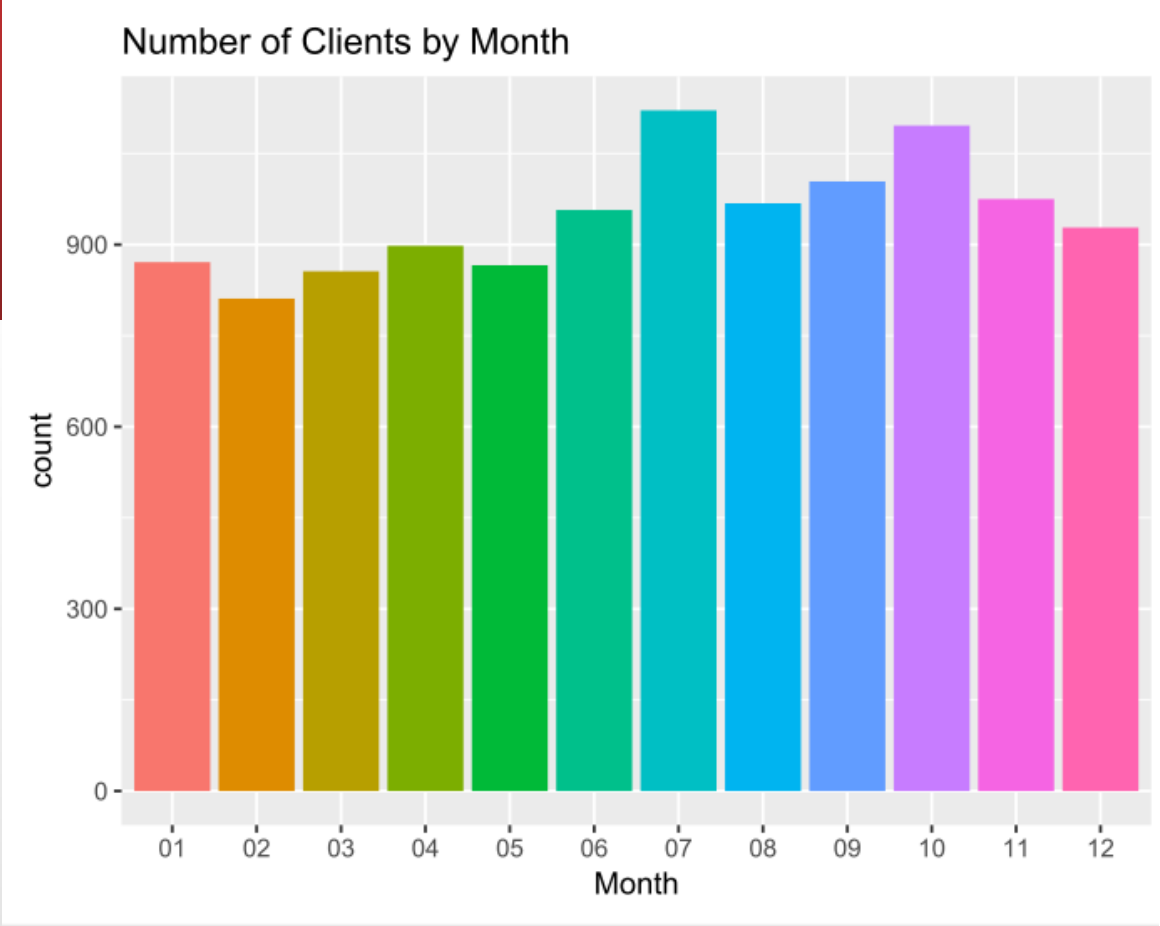
Number of Clients by Year and Month & Duration of Assistance

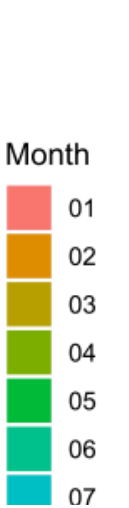


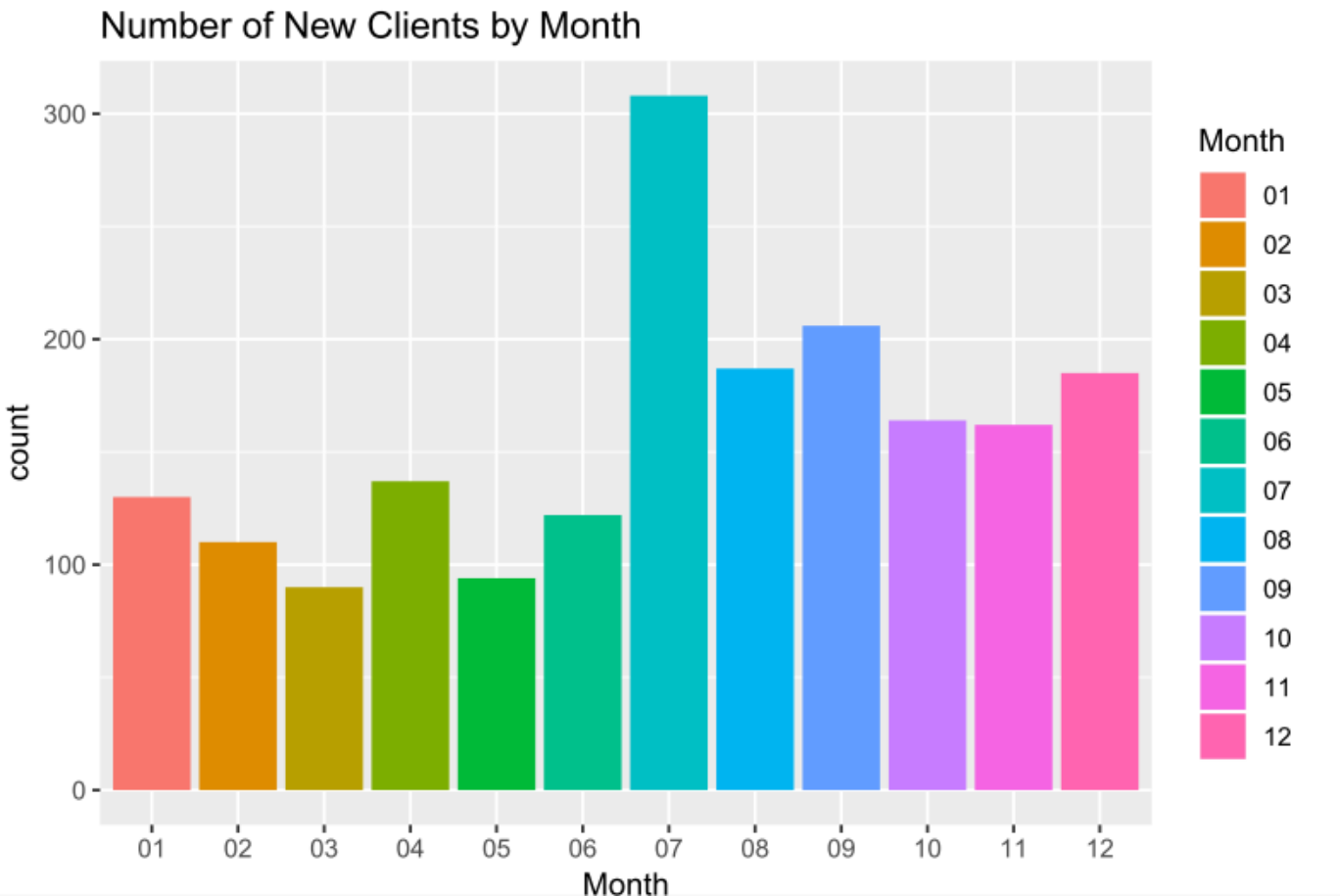

```r
client.data = data %>%
  select(Date, ClientID) %>%
  drop_na() %>%
  separate(Date, sep = "-", into = c("Year", "Month", "Day"))
ggplot(data = client.data) +
  geom_bar(mapping = aes(x = Year, fill = Year)) +
  labs(title = "Number of Clients by Year") +
  theme(legend.position = "none")
```

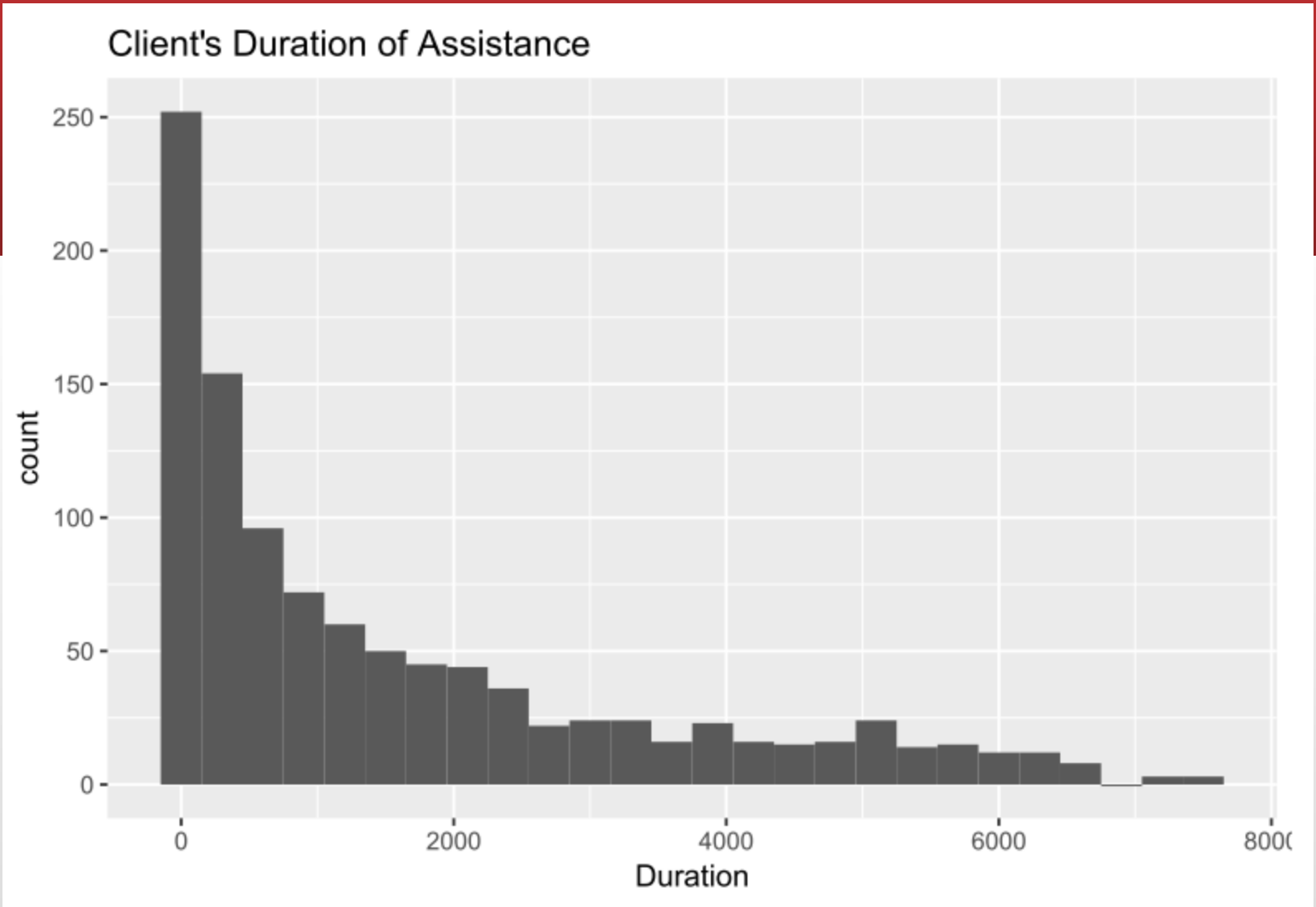



```
client.data = data %>%
  select(Date, ClientID) %>%
  drop_na() %>%
  separate(Date, sep = "-", into = c("Year", "Month", "Day")) %>%
  distinct(ClientID, .keep_all = TRUE)
ggplot(data = client.data) +
  geom_bar(mapping = aes(x = Year, fill = Year)) +
  labs(title = "Number of New Clients by Year") +
  theme(legend.position = "none")
```


Number of Clients by Month
count
900
600
300
0
01
02
03
04
05
06
07
08
09
10
11
12
Month


Month
01
02
03
04
05
06
07


Number of New Clients by Month
count
300
200
100
0
01
02
03
04
05
06
07
08
09
10
11
12
Month
Month
01
02
03
04
05
06
07
08
09
10
11
12

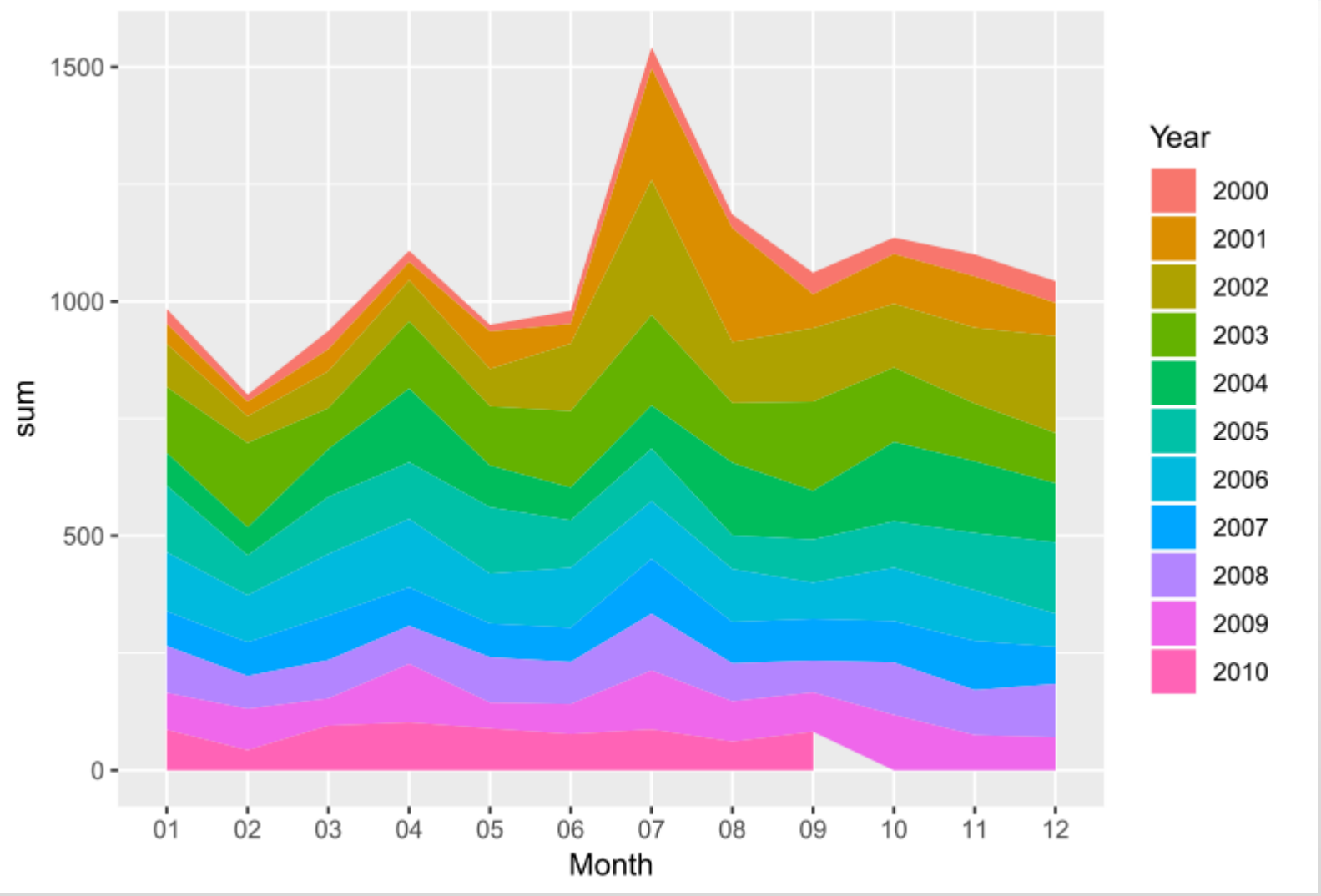



```
client.data.duration <- data %>%
  group_by(ClientID) %>%
  summarize(Duration = difftime(max(Date),min(Date))) %>%
  filter(Duration > 0 & Duration < 10000)
ggplot(data = client.data.duration) +
  geom_histogram(mapping = aes(x = Duration, fill = Duration), binwidth = 300) +
  labs(title = "Client's Duration of Assistance")
```

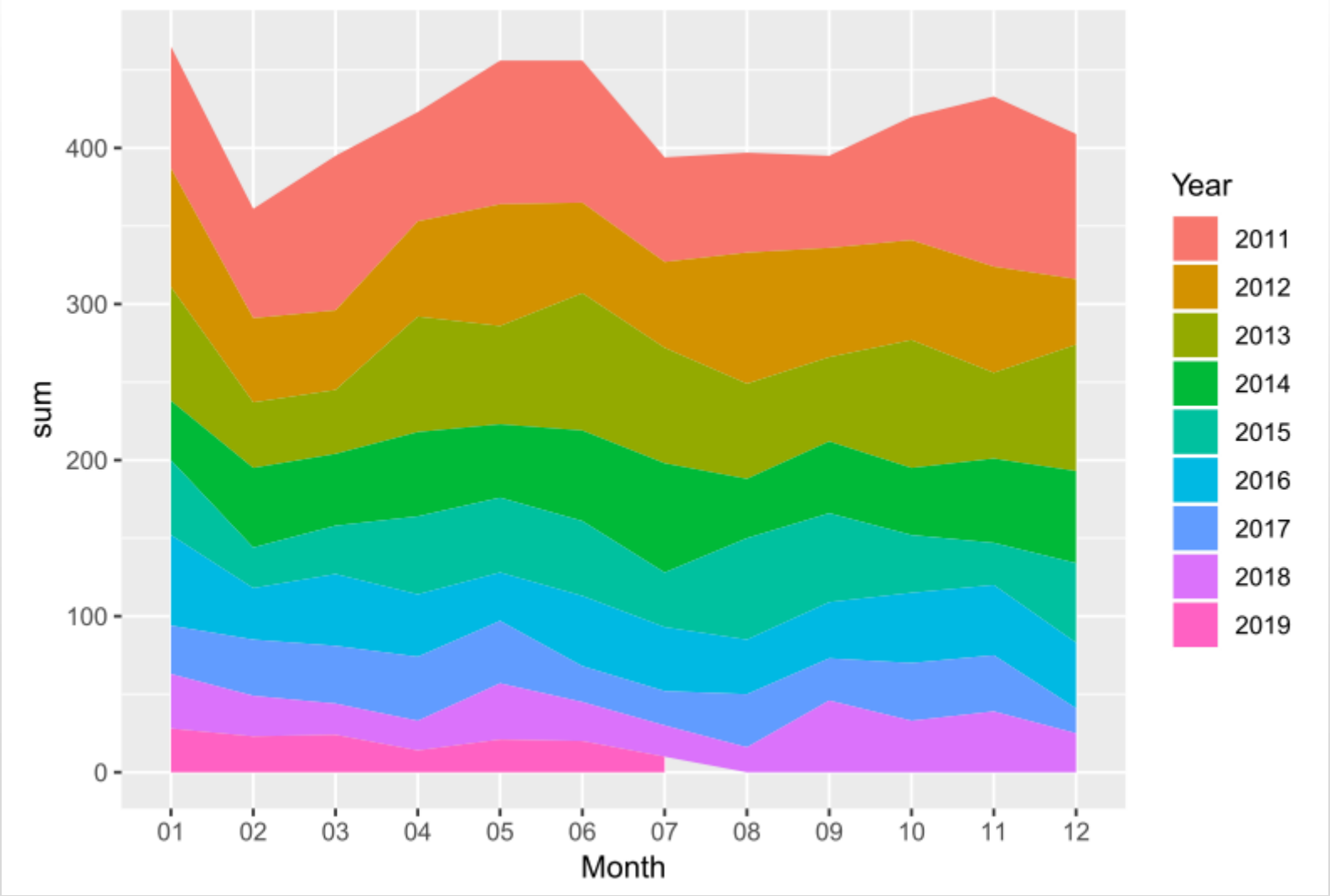
Seasonality

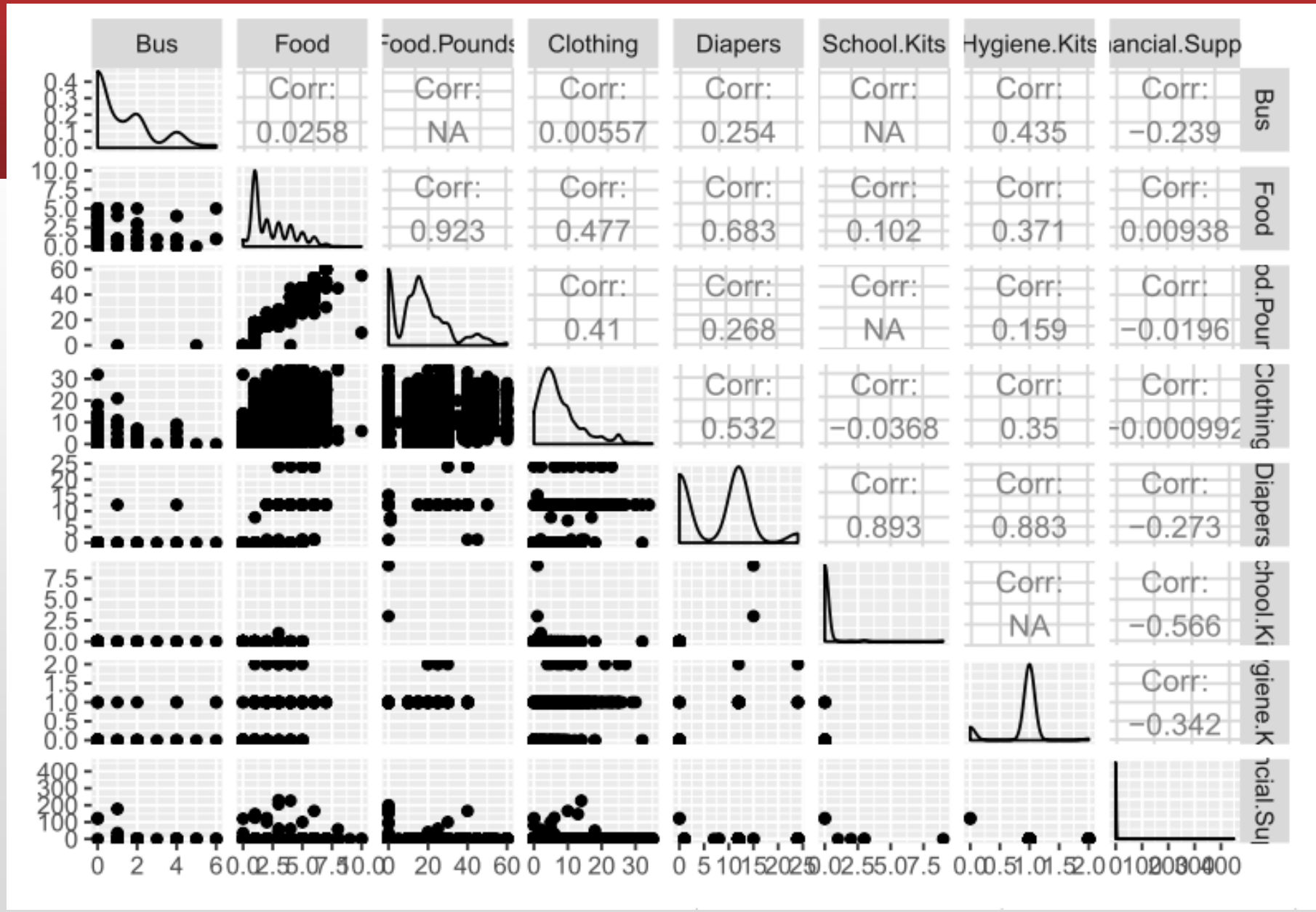
2000-2010
Food Sum by Year

```
food.data = data %>%
  select(Date, Food) %>%
  filter(Date >= "2000-01-01" & Date <= "2010-10-01") %>%
  drop_na() %>%
  separate(Date, sep = "-", into = c("Year", "Month", "Day"))%>%
  group_by(Year, Month) %>%
  summarise(sum = sum(Food))
ggplot(food.data, aes(x = Month, y = sum, group = Year)) +
  geom_area(aes(fill = Year), position = "stack")
```



2011-2019
Food Sum by Year




Bus
Food
Food.Pounds
Clothing
Diapers
School.Kits
Hygiene.Kits
ancial.Supp
Corr: 0.0258
Corr: NA
Corr: 0.00557
Corr: 0.254
Corr: NA
Corr: 0.435
Corr: −0.239
Corr: 0.923
Corr: 0.477
Corr: 0.683
Corr: 0.102
Corr: 0.371
Corr: 0.00938
Corr: 0.41
Corr: 0.268
Corr: NA
Corr: 0.159
Corr: −0.0196
Corr: 0.532
Corr: −0.0368
Corr: 0.35
Corr: 0.893
Corr: 0.883
Corr: −0.273
Corr: NA
Corr: −0.566
Corr: −0.342
Bus
Food
od.Pour
Clothing
Diapers
chool.Ki
giene.K
ncial.Su

Conclusion

1. The number of new clients served by UMD had been increasing until 2002 but has been decreasing up to now.
2. The number of clients served by UMD each month do differ. Generally, March is the month when the UMD serves the least amount of clients, and July is the month when the UMD serves the most amount of clients between 1994 and 2019.
3. The histogram of the duration of assistance was skewed righ with a tail going off to the right. The most amount of clients was served by the UMD for 0-500 days and there are people who have been served over 10-20 years.
4. Fourth, between 2000 to 2010, sum of the amount of foods tended to peak for July and then decline after summer. So time series of the food amount typically showed increasing pattern from January through July and declining pattern from July to December.
5. On the other hand, between 2011 to 2019, Sum of the amount of foods tend to peak for May and June, then decline after summer and rise again during winter. So time series of the food amount typically show increasing pattern from February through May , declining pattern from July to september and increasing pattern again from September to January.
6. Lastly, there isstatistically significant relationship between 1) Clothing and Food, 2) Diapers and Food, 3) Hygiene.Kits and Bus, 4) Hygiene.Kits and Food, 5) Hygiene.Kits and Diapers, and 6) Clothing and Food.pounds.

thank you!