STSCI 5080 Probability Models and Inference

Final Project

Instructor: Maria Caterina Bramati

Member: Ben Liu
Gunjan Sood
Han Fang
Shu Yang

Contents

1 Introduction

3-1-1 is a special telephone number supported in many communities in Canada and the United States. The number provides access to non-emergency municipal services, such as needle clean-up,graffiti removal, pothole repair and street cleaning.

In this project, the 311 service requests that took place in the city of Boston was researched into. Section 2 provided probabilistic models of service requests flow to measure the expected number of service requests for a certain time interval. In section 3-5, focus was drawn on four factors of service requests, respectively source, time, location and status, to better qualify the citizens' preference on service request. In section 6, several potential significant factors that affect the amount of service requests were explored. Through those quantitative analysis of the data, the ultimate goal is to help the Mayor of Boston to restructure the service in a cost-effective way.

Our study was based on the 311 service request dataset[1]. The original data set contains around 200 thousands records with 34 variables, ranging from January 1 to November 11 in 2016. In order to achieve high performance, only variables that fit into our interests were selected. Moreover, the time interval was refined for more detailed analysis(see Table 1).

Table 1: Variables In Revised 311 Service Request Dataset.

Variable Name	Label	Type
Open_DT	Case Open date	Date
Date.hour	Hour Part of case open date, (mmddhh)	Integer
Date	Date Part of case open date,(mmdd)	Integer
Timezone	Time section of open date, 1 is 00:00 am-5:59 am.	Integer
Weekday	Weekday of open date, 1 is Monday.	Integer
Weeknum	Number of the week in the year	Integer
Holiday	Whether open date is a school holiday[1].	Boolean
Case_Status	Status of a case	Character
Department	Department assigned to a case	Character
Neighborhood	Neighborhood case is within	Character
Source	Source of case	Character
Type	Individual case type	Character

2 Model of Service Requests Flow

The 311 service request data were generated by date. Therefore, a general model that captures the pattern of service requests flows is necessary. Figure 1 shows the general model,

[1] School Holidays include national and state holidays, recess and breaks between school sessions. We used Boston University academic calendar as our reference.

which counts the number of requests in each day. There are two peaks displayed in this model, which indicates a possible bimodal distributed model. The mean and the standard deviation are estimated by applying the EM algorithm. However, the fitted curve is not well-approximated as shown in Figure 1.

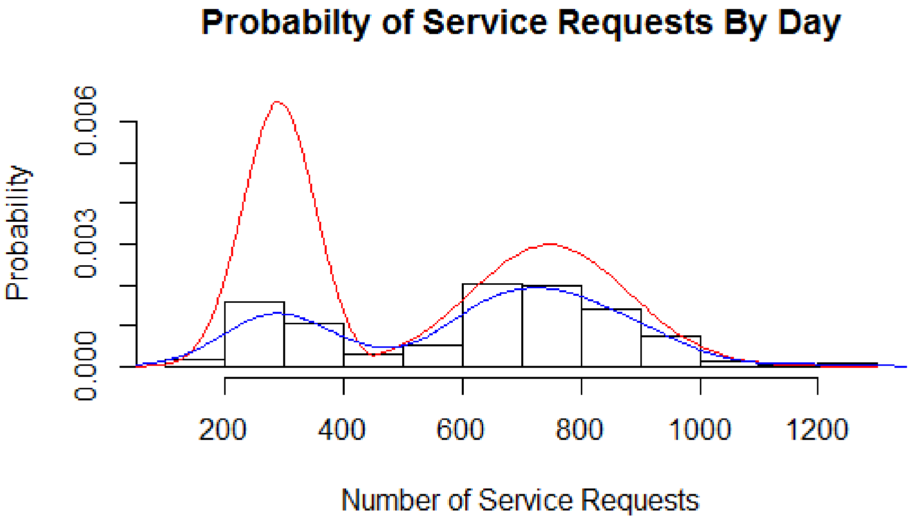

Figure 1: Probability of Service Request by Day

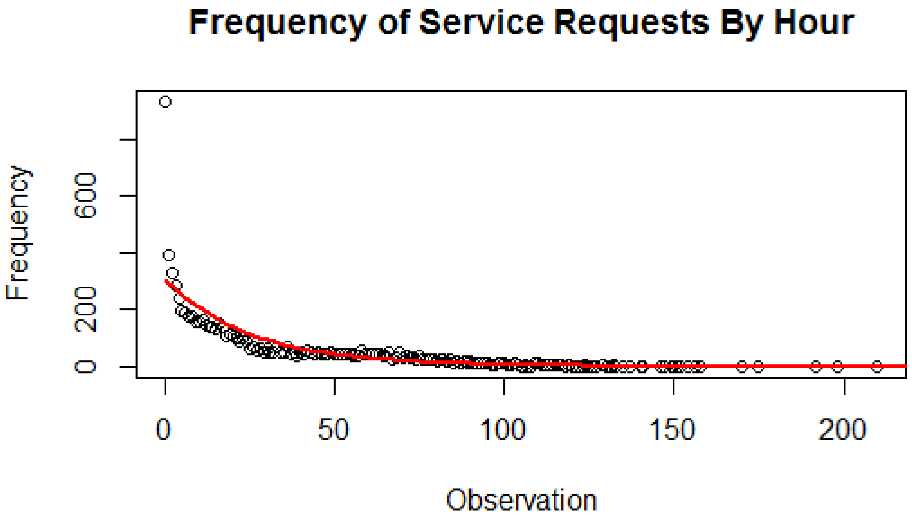

Figure 2: Number of Service Requests by Hour

Consequently, a model that counts the number of service requests by hour was considered. The pattern in Figure 2 suggests an exponential distribution of the sample. The probability density function is

$$f(x) = \frac{1}{\mu} e^{-\frac{1}{\mu}x}, \quad x \geq 0$$

where μ is the average number of requests happening in an hour. The maximum likelihood estimation (MLE) was applied to estimate the parameter for its asymptotically efficient, consistent and normally distributed. Let X denotes the number of requests in one hour. The maximum likelihood estimator of the exponential distribution is

$$\hat{\mu}^{MLE} = \frac{\sum_{i=1}^{n} x_i}{n} \approx 25.56,$$

which is just the sample mean. Then, the fitted exponential curve was added on the plot and the Kolmogorov-Smirnov test was conducted. However, the result was still not that satisfactory. This is mainly because the exponential curve does not fit smaller values of X very well. In fact, to fit a perfect model into a sample dataset is really difficult in real life scenario. Since the experimental model captured the trend and fitted X well except for first three values, the model with exponential distribution is the "second best" choice to represent the service request flow. The expected number of service requests is $\hat{\mu} = 25.56$ in an hour and 614 in a day.

3 Type of Request Source

The number of requests by each type of source is listed in Table 2. It is straightforward to see that the most popular way to request a service was Constituent Call, with 85,356 total requests.

Table 2: Source of Requests

Source	Number
Constituent Call	85,356
Citizens Connect App	69,801
Employee Generated	14,021
Self Service	9,923
Twitter	798
Maximo Integration	9

Furthermore, it is worth to investigate whether at night time [2] the Citizens Connect App was more often used than the traditional call. The original dataset was split into one containing *Source* with only "Constituent Call" or "Citizens Connect App". A hypothesis test was

[2] Night time is defined from 0:00 to 5:59 a.m.

conducted based on one sample proportion test. The null and the alternative hypothesis are stated as below:

H_0 : The Citizens Connect App is equally used as the traditional call at night time.
H_1 : The Citizens Connect App is more often used than the traditional call at night time.

In symbols:

$$H_0 : p_{App} = p_{Call}$$
$$H_1 : p_{App} > p_{Call}$$

where p_{App} is the proportion of requests by Citizens Connect App and p_{Call} is the proportion of requests by traditional call. The z-score was calculated as following:

$$z = \frac{p_{app} - p_{call} - 0}{\sqrt{\frac{0.5\times0.5}{n}}} \approx 121.2$$

The null hypothesis is rejected at the 1% significance level, which means the Citizens Connect App was more often used than the traditional call at night time.
Moreover, the top 10 types of requests associated with a Twitter is displayed in Table 3. "Parking Enforcement " is most often associated with Twitter.

Table 3: Types of Requests by Twitter

Type	Total Amount
Parking Enforcement	47
Ground Maintenance	41
Traffic Signal Inspection	40
Contractor Complaints	39
Space Savers	39
Unshoveled Sidewalk	37
Improper Storage of Trash (Barrels)	36
Missed Trash/Recycling/Yard Waste/Bulk Item	31
Street Light Outages	31

4 Service Requests by Time and Location

4.1 Considering Time Periods

This section mainly focuses on the relationship of service requests with time and location. The amount of request among different time periods was first studied. In this case, one day was divided into four time periods, respectively 6:00 a.m.-11:59 a.m. as morning, 12:00

p.m.-17:59 p.m. as noon, 18:00 p.m.-23:59 p.m. as night, and 0:00 a.m.-5:59 a.m as early morning.
With simple categorization and analysis, the most recurrent type of service for the four time periods is concluded in Table 4.

Table 4: Most Recurrent Type of Service by Time Period

Time Period	Request Type
1	Street Light Outages
2	Parking Enforcement
3	Parking Enforcement
4	Parking Enforcement

Furthermore, the top three request types for the four time periods was concluded as Table 5.

Table 5: Top Three Types of Service by Time Period

Time	1st Request Type	2nd Request Type	3rd Request Type
1	Street Light Outages	Parking Enforcement	Requests for Street Cleaning
2	Parking Enforcement	Requests for Street Cleaning	Schedule a Bulk Item Pickup
3	Parking Enforcement	Schedule a Bulk Item Pickup	Requests for Street Cleaning
4	Parking Enforcement	Requests for Street Cleaning	Street Light Outages

There are four types of requests in Table 5. Respectively, they are “Street Light Outages”, “Parking Enforcement”, “Requests for Street Cleaning” and “Schedule a Bulk Item Pickup”. This raised the question whether four request types were of equal possibility in the four time periods considered. To solve this, ANOVA test was conducted separately on the data of this four request types with the following hypothesis.

H_0 : The service requests in mean are the same in the four time periods.
H_1 : At least for one time period, the service request in mean is different.

In symbols:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4$$
$$H_1 : \exists i, j \in \{1,2,3,4\} \text{ s.t. } \mu_i \neq \mu_j.$$

where 1-4 are integers that represent time periods. And the null hypothesis is the same for the four request types, which respectively, “Street Light Outages”, “Parking Enforcement”, “Requests for Street Cleaning” and “Schedule a Bulk Item Pickup”. The results were shown in Table 6- 9.

Table 6: ANOVA Test for Street Light Outages

	Degree of Freedom	Sum of Square	Mean Sum of Square	F-value	P-value
Time Period	3	389.6	129.871	4.8257	0.002417
Residuals	1135	30545.3	26.912		

Table 7: ANOVA Test for Parking Enforcement

	Degree of Freedom	Sum of Square	Mean Sum of Square	F-value	P-value
Time Period	3	89141	29714	471.5	<2.2e-16
Residuals	1250	78774	63		

Table 8: ANOVA Test for Requests for Street Cleaning

	Degree of Freedom	Sum of Square	Mean Sum of Square	F-value	P-value
Time Period	3	61839	20612.9	323.17	<2.2e-16
Residuals	1201	76604	63.8		

Table 9: ANOVA test for Schedule a Bulk Item Pickup

	Degree of Freedom	Sum of Square	Mean Sum of Square	F-value	P-value
Time Period	3	41005	13668.5	131.17	<2.2e-16
Residuals	873	90973	104.2		

Under significance level of 5%, the conclusion can be drawn that all the four requests were not of equal probable in the four time periods.

4.2 Considering Neighborhoods

The main neighborhoods were used as indications of different locations. Similarly, with simple categorization, the top ten neighborhoods with the highest number of requests are listed in Table 10.
With the data of time and location, the dependency of time and location can be checked. This is done by conducting chi-square test. The hypothesis yields:

$$H_0 : \text{Time and Location are independent.}$$
$$H_1 : \text{Time and Location are not independent}$$

After the test was conducted, the p-value of the test is 2.2×10^{-16}, indicating that the null hypothesis is rejected and as a result, the time and location were not independent.

Table 10: Top Ten Neighborhoods with the Highest Number of Requests

rank	Neighborhood
1	Dorchester
2	Roxbury
3	South Boston / South Boston Waterfront
4	Downtown / Financial District
5	Allston / Brighton
6	Jamaica Plain
7	Back Bay
8	South End
9	East Boston
10	Greater Mattapan

4.3 Considering Departments

The variable *Department* was considered and the top ten departments that were more often called to respond to request are listed as in Table 11.

Table 11: Top Ten Departments Solicited to Respond to Service Request

Rank	Departments
1	PWDx
2	BTDT
3	ISD
4	PARK
5	INFO
6	PROP
7	HS_O
8	GEN
9	BWSC
10	DISB

4.4 Considering The Day of Week

Instead of time periods, the number of service requests for each weekday was explored. Interestingly, the most recurrent type of service request was "Parking Enforcement" for every weekday. Furthermore, the top three request types are listed in Table 12.

With the basic information, it is natural to detect whether there were more service requests in a certain day of week or they were evenly distributed across the week. To achieve this, the ANOVA test is carried with following hypothesis(see Table 13).

Table 12: Top 3 Type of Service By the Day of Week

Time	1st Request Type	2nd Request Type	3rd Request Type
1	Parking Enforcement	Requests for Street Cleaning	Improper Storage of Trash (Barrels)
2	Parking Enforcement	Schedule a Bulk Item Pickup	Requests for Street Cleaning
3	Parking Enforcement	Requests for Street Cleaning	Schedule a Bulk Item Pickup
4	Parking Enforcement	Requests for Street Cleaning	Schedule a Bulk Item Pickup
5	Parking Enforcement	Requests for Street Cleaning	Schedule a Bulk Item Pickup
6	Parking Enforcement	Requests for Street Cleaning	Missed Trash/Recycling/Yard Waste/Bulk Item
7	Parking Enforcement	Requests for Street Cleaning	Improper Storage of Trash (Barrels)

H_0 : The service requests in mean are the same in each day of the week.

H_1 : At least for one day of the week, the service requests in mean is different.

In symbols:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu_6 = \mu_7$$
$$H_1 : \exists i, j \in \{1, 2 \cdots 7\} \text{ s.t. } \mu_i \neq \mu_j.$$

Table 13: ANOVA Test for Number of Requests Across the Week

	Degree of Freedom	Sum of Square	Mean Sum of Square	F-value	P-value
Day of week	6	13238854	2206476	134.89	<2.2E-16
Residuals	318	5201781	16358		

Where the integer 1-7 corresponding to each weekday from Monday to Sunday. Under the significance level of 5%, the null hypothesis that the number of requests are the same across the week was rejected. Furthermore, Tukey test was conducted to check specifically in which weekday that the service requests in mean was different(See Figure 3).

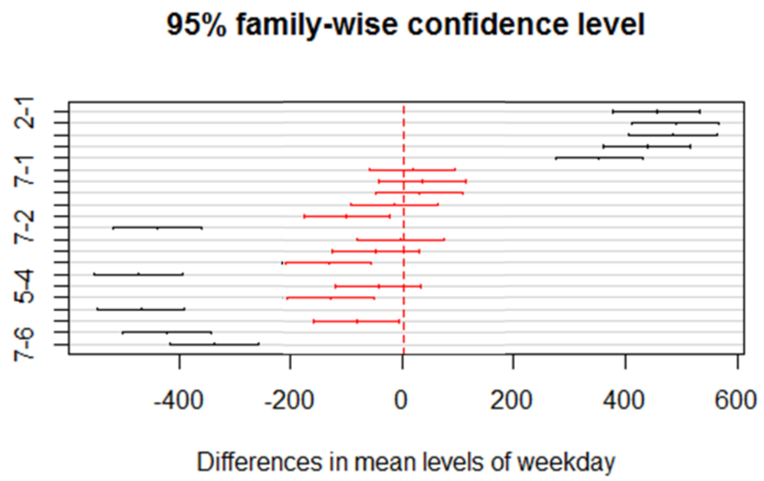

Figure 3: The Tukey Multiple Comparison of Means.

The lines ranged around 0 were pairs of weekdays with evenly distributed requests under the significance level of 5%. Those pairs are Monday and Sunday, Tuesday and Wednesday, Tuesday and Thursday, Tuesday and Friday, Wednesday and Thursday, Wednesday and Friday, Thursday and Friday. For the rest of weekday pairs, the null hypothesis that the number of service requests is evenly distributed was rejected.

4.5 Considering School Holiday

After checking the calendar for Boston University, preprocessing of the data was done and the original time was categorized into two groups, one was school holiday and one was not school holiday. Similarly as above, the test of whether there were more service requests in school holidays than in regular days was performed and this time the two sample t test was conducted with following hypothesis:

H_0 : The proportion of the requests in school holiday is the same as that of regular days.

H_1 : The proportion of the requests in school holiday is different from that of regular days.

In symbols:

$$H_0 : p_h = p_r$$
$$H_1 : p_h \neq p_r$$

Where h represents school holidays, and r represents regular days.
The result of the test showed that the p-value was smaller than 2.2×10^{-16}, which is much smaller than the significance level 5%. In conclusion, the null hypothesis that the proportion of requests in school holiday and in regular days were the same is rejected. In fact, there were more service requests on regular days than in school holidays.

5 Status of the Request

Distributions of requests of two case status are separately displayed in Figure 4 and Figure 5. The distributions distinct from each other greatly. If only considering the number of requests greater than 3,000, the distribution of closed requests roughly formed normal distribution. The percentage of requests with open statue is 11.7%.

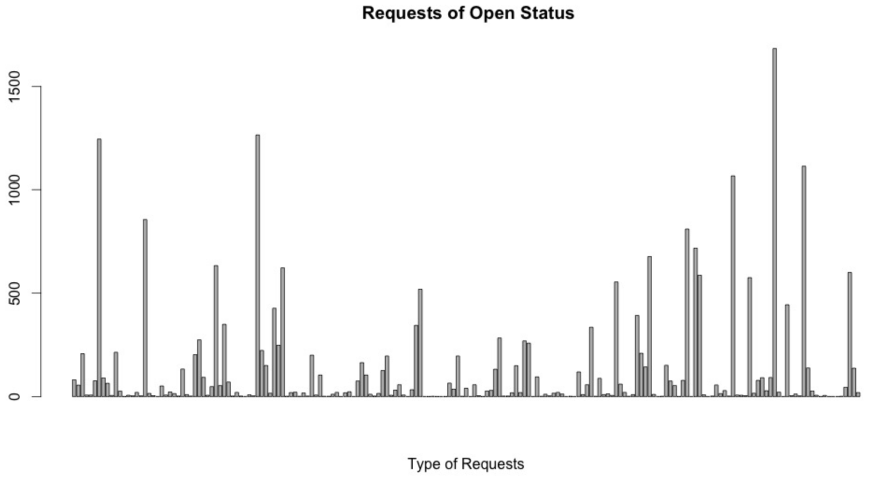

Figure 4: Distribution of Open Requests

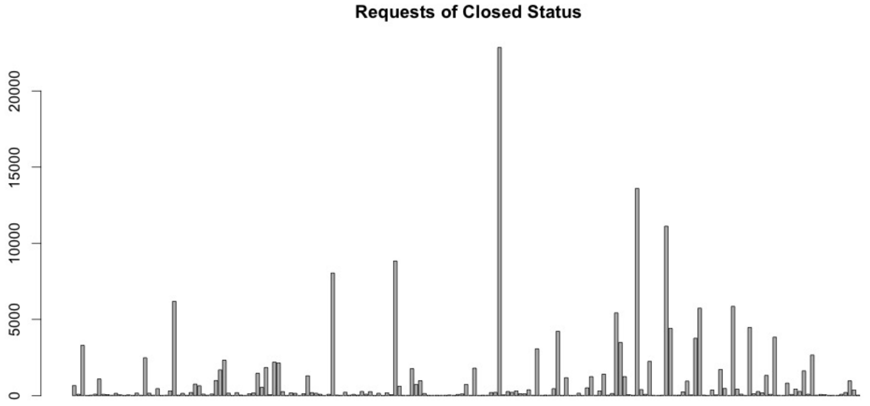

Figure 5: Distribution of Closed Requests

Furthermore, the modal request type of open status is “Tree Maintenance. The most frequently requested service by each neighborhood is displayed in Table 14. “Parking Enforcement" requests was listed 13 times, while none of the neighborhood had the "Abandoned Bicycle" as their top requested service.

Since none of the neighborhood required "Abandoned Bicycle" the most, which neighborhood requested "Abandoned Bicycle" the most compared to others is investigated. The top five neighborhoods are shown in Table 15. It can be seen that the financial district in downtown requested "Abandoned Bicycle" most compared to others.

Table 14: Most Frequently Service by Neighborhood

Neighborhood	Type	Number
Allston	Ground Maintenance	24
Brighton	Parking Enforcement	1,895
Back Bay	Parking Enforcement	1,318
Beacon Hill	Improper Storage of Trash (Barrels)	1,015
Boston	Parking Enforcement	520
Brighton	Parking Enforcement	45
Charlestown	Parking Enforcement	481
Chestnut	Mechanical	2
Dorchester	Schedule a Bulk Item Pickup	2746
Downtown / Financial District	Parking Enforcement	2,940
East Boston	Parking Enforcement	1,077
Fenway	Parking Enforcement	563
Greater Mattapan	Schedule a Bulk Item Pickup	1,077
Hyde Park	Schedule a Bulk Item Pickup	980
Jamaica Plain	Parking Enforcement	1,298
Mattapan	Unsatisfactory Living Conditions	28
Mission Hill	Parking Enforcement	371
Roslindale	Schedule a Bulk Item Pickup	709
Roxbury	Requests for Street Cleaning	1,154
South Boston	Parking Enforcement	267
South Boston Waterfront	Parking Enforcement	5,013
South End	Parking Enforcement	1,533
West Roxbury	Schedule a Bulk Item Pickup	939

Table 15: "Abandoned Bicycle" Requests by Neighborhood

Neighborhood	Number
Downtown / Financial District	116
South End	102
Allston / Brighton	90
Beacon Hill	79
Back Bay	70

6 Key Factors

The most influential factor that affecting the number of service requests was timezone. In Figure 6, it is clear that the second and the third timezone had more requests than the other two timezones.

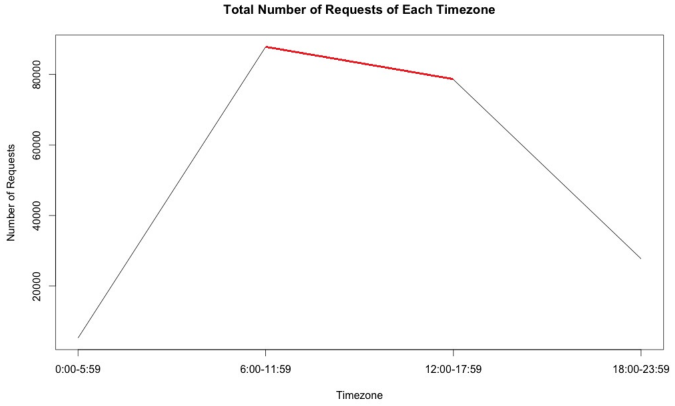

Figure 6: Number of Requests by Timezone

Furthermore, in section 4.1 whether top four frequent requests were of equal probable in the four time periods was investigated and the null hypothesis was rejected at the 1% significance level, which indicates that the top four requests were not of equal probable in the four time periods.
School holiday was another potential significant factors based on former research. Recall that the null hypothesis of whether there were more service requests in school holidays than those in regular days was rejected at the 1% significance level (see Section 4.5).

7 Conclusions and Recommendations

From Figure 6, it is clear to see that during time period from 6:00 a.m. to 17:59 p.m., numbers of called requests were considerably larger compared to those of other two periods. Therefore, this time period deserves further exploration. “Scheduling a Bulk Item Pickup” and “Missed Item reports” were top two request types during these two timezones . Therefore, these two issues should be put on the first priority to avoid call congestion during the time period from 6:00 a.m. to 17:59 p.m. Furthermore, by investigation and summary of the whole 311 dataset, it is easy to find that the most popular way for requesting these two service was “Constituent Call”(See Table 16 and Table 17).

Table 16: “Schedule a Bulk Item Pickup” Requests by Source

Source	Number
Constituent Call	10,321
Employee Generated	950
Twitter	3
Citizens Connect App	0
City Worker App	0

Table 17: “Missed Item” Requests by Source

Source	Number
Constituent Call	7,022
Self Service	1,453
Employee Generated	357
Twitter	31
Citizens Connect App	0
City Worker App	0

Meanwhile, Twitter is less popular and the number of requests sent by Citizens Connect App and City Worker App was 0 . This result suggests that the on-line sources for service requests were rarely used and hence should be improved. For instance, the software developer can add related functions about “Scheduling Bulks Pickup” and “Missed Item Reports” on those Apps. Also, the government can attract more app users by using some promotions. For example, requests sent from these on-line platforms will be dealt ahead of those traditional phone call requests. With those improvements, citizens may prefer using on line sources to pose service requests, which will share the bare with traditional phone calls and therefore reduce the call flow.

In general, the amount of total requests by source is listed in Table 18. Since Citizens Connect App, City Worker App and Twitter are all Internet products, they might be connected together by plug-ins to share same functions so that citizens have more options to request for services.

Moreover, there were four main issues should be further considered and dealt with - “Street Light Outages”, “Parking Enforcement, “Requests for Street Cleaning and “Schedule a Bulk Item Pickup”. The top one problem was “Parking Enforcement” since its amount was always the largest among all days. From Table 19, the Transportation Department (BTDT) is the main department dealing with the “Parking Enforcement” issue. Consequently, more budget and staffs should be allocated to BTDT to deal with “Parking Enforcement” more efficiently.

Table 18: Number of Requests by Source

Source	Number
Constituent Call	85,356
Citizens Connect App	69,804
City Worker App	19,494
Employee Generated	14,021
Self Service	9,923
Twitter	798
Maximo Integration	9

Table 19: Top Two Departments Handling "Parking Enforcement"

Department	Number
BTDT	22,730
PWDx	152

"Street cleaning" is the second frequently requested service type in every day. The number of requests for street cleaning greater than 10,000 by neighborhood are listed in Table 20. Therefore, the Mayor might consider to increase the number of cleaners in those regions, especially in Dorchester.

Table 20: "Street Cleaning" Requests by Neighborhood

Neighborhood	Number
Dorchester	28,119
South Boston Waterfront	15,744
Downtown / Financial District	15,186
Allston / Brighton	13,782
Downtown / Financial District	15,186
Jamaica Plain	12,376
Back Bay	10,537
South End	10,349

"Street Light Outages" was another main issue. Table 21 shows the top five areas that should improve their street light quality. Overall, the top three request types for the top ten neighborhoods is covered in Table 22. With such information, the government can balance their budget and labor forces to target on most influential issues among most troublesome regions.

Table 21: Number of "Street Light Outage" Requests by Neighborhood

Source	Number
Dorchester	698
Downtown / Financial District	644
Back Bay	637
Beacon Hill	515
Roxbury	403

Table 22: Top Three Request Types for the Top Ten Neighborhoods

Neighborhood	Service Request Type	Number
Dorchester	Schedule a Bulk Item Pickup	2,746
	Parking Enforcement	2,413
	Requests for Street Cleaning	2,227
Roxbury	Requests for Street Cleaning	1,154
	Parking Enforcement	1,113
	Schedule a Bulk Item Pickup	1,056
South Boston (Waterfront)	Parking Enforcement	5,013
	Requests for Street Cleaning	934
	Improper Storage of Trash (Barrels)	666
Downtown / Financial District	Parking Enforcement	2,940
	Requests for Street Cleaning	1,542
	CE Collection	1038
Allston / Brighton	Parking Enforcement	1,895
	Schedule a Bulk Item Pickup	815
	Requests for Street Cleaning	772
Jamaica Plain	Parking Enforcement	1298
	Missed Trash/Recycling/Yard Waste/Bulk Item	783
	Requests for Street Cleaning	708
Back Bay	Parking Enforcement	1,318
	Improper Storage of Trash (Barrels)	1,157
	Requests for Street Cleaning	895
South End	Parking Enforcement	1,533
	Improper Storage of Trash (Barrels)	997
	Requests for Street Cleaning	917
East Boston	Parking Enforcement	1,077
	Requests for Street Cleaning	819
	Schedule a Bulk Item Pickup	512
Greater Mattapan	Schedule a Bulk Item Pickup	1,077
	Requests for Street Cleaning	626
	Missed Trash/Recycling/Yard Waste/Bulk Item	576

8 Appendix-R code

```
boston=read.csv("e:/5080/Project/311_SHU1.CSV",header=T)

boston$OPEN_DT=as.character(boston$OPEN_DT)

###Question 2

##Day
library(dplyr)
Q21=as.data.frame(count_(boston,c("date"),sort = T))
hist(Q21$n,breaks = 10,probability =T,xlab="Number of Service Requests",ylab="Probability",main = "Probabilty of Service Requests By Day",ylim = c(0,0.0065))
lines(density(Q21$n),col="blue")

library(mixtools)
out=normalmixEM(Q21$n,k=2,fast = FALSE)
summary(out)
curve(dnorm(x,747.165732,133.566575),450,1300,add = T,col="red")
curve(dnorm(x,291.532169,61.412108),0,450,add = T,col="red")

##Hour
Q22=as.data.frame(count_(boston,c("date.hour"),sort = T))
q2.add0=matrix(NA,933,2)
q2.add0[,1]=c(1:933)
q2.add0[,2]=rep(0,933)
colnames(q2.add0)=c("date.hour","n")
Q22.0=rbind(Q22,q2.add0)
hist(Q22.0$n,breaks = 100)
mu=mean(Q22.0$n)
ks.test(Q22.0$n,"pexp",1/24)

Q23=as.data.frame(count_(Q22,c("n"),sort = T))
Q23=rbind(c(0,933),Q23)
plot(Q23$n,Q23$nn,xlab="Observation",ylab="Frequency",main="Frequency of Service Requests By Hour")
curve((1/mu)*exp(-(1/mu)*x)*7800,0,250,add = T,col="red",lwd=2)

mydata=read.csv("e:/5080/Project/311_SHU1.CSV",header=T)
# Question 3(a)
# Most Popular Sourse
table(mydata$Source)
```

```
## Question 3(a)
## The most popular way to request a service
##is Constituent Call. The number is 85,356.


# Question 3(b)
nightdata=subset(mydata,timezone==4)
nightdata.app=subset(nightdata,Source=="Citizens Connect App")
nightdata.call=subset(nightdata,Source=="Constituent Call")
night.callapp=rbind(nightdata.call,nightdata.app)
View(night.callapp)
table(night.callapp$Source)
n=nrow(night.callapp)
p.app=sum(night.callapp$Source=="Citizens Connect App")/n
p.call=sum(night.callapp$Source=="Constituent Call")/n
p.diff=p.app-p.call
z.sta=p.diff/sqrt(0.5*0.5/n)


# Question 3(c)
summary(mydata)
twitter=subset(mydata,Source=="Twitter")
table(twitter$TYPE)
## The type of service requests most often associated
##with a Twitter request source is parking enforcement. The number is 47.


#Q4ACount the max
library(dplyr)
Q41=as.data.frame(count_(boston,c("timezone"),sort = T))
Q4=as.data.frame(count_(boston,c("timezone","TYPE"),sort = T))
Q4.a=as.data.frame(Q4 %>% group_by(timezone) %>% filter(n == max(n)))
#For time zone 1 it’s street light outages, for time zone 2, 3 and 4,
#it’s Parking enforcement

#Q4b
boston$date=as.factor(boston$date)
boston$timezone=as.factor(boston$timezone)
Q4.b=as.data.frame(Q4 %>% group_by(timezone) %>% arrange(desc(n)) %>% slice(1:3))
Q4.light=boston[boston$TYPE=="Street Light Outages",]
Q4.b.light=as.data.frame(count_(Q4.light,c("date","timezone"),sort = T))
%>% arrange(desc(date))
```

```
anova(lm(n~timezone,Q4.b.light))
#Reject the null hypthesis that the probable for each timezone is the same
light=aov(n~timezone,data = Q4.b.light)
TukeyHSD(light)
##Under level 5%, 2&1 are different, 4 and 2 are differnet.


Q4.parking=boston[boston$TYPE=="Parking Enforcement",]
Q4.b.parking=as.data.frame(count_(Q4.parking,c("date","timezone"),sort = T))%>%
arrange(desc(date))
anova(lm(n~timezone,Q4.b.parking))
#Reject the null hypthesis that the probable for each timezone is the same
parking=aov(n~timezone,data = Q4.b.parking)
TukeyHSD(parking)
##under level 5%, 3 and 2 are the same.


Q4.cleaning=boston[boston$TYPE=="Requests for Street Cleaning",]
Q4.b.cleaning=as.data.frame(count_(Q4.cleaning,c("date","timezone"),sort = T))%>%
arrange(desc(date))
anova(lm(n~timezone,Q4.b.cleaning))
#Reject the null hypthesis that the probable for each timezone is the same
cleaning=aov(n~timezone,data = Q4.b.cleaning)
TukeyHSD(cleaning)
##under level 5%, all are different.



Q4.pickup=boston[boston$TYPE=="Schedule a Bulk Item Pickup",]
Q4.b.pickup=as.data.frame(count_(Q4.pickup,c("date","timezone"),sort = T))%>%
arrange(desc(date))
anova(lm(n~timezone,Q4.b.pickup))
#Reject the null hypthesis that the probable for each timezone is the same
pickup=aov(n~timezone,data = Q4.b.pickup)
TukeyHSD(pickup)
##under level 5%, 4 and 1, 3 and 2 are the same.




##Q4c
Q4.neighbor=as.data.frame(count_(boston,"neighborhood",sort = T))
```

```
Q4.neighbor1=as.data.frame(count_(boston,c("neighborhood","TYPE"),sort = T))
write.csv(Q4.neighbor1,"e:/5080/Project/neighbor.csv")

Q4.c1=as.data.frame(Q4.neighbor1 %>% group_by(TYPE) %>% arrange(desc(n))
 %>% slice(1:3))



Q4.c=as.data.frame(Q4.neighbor %>% arrange(desc(n)) %>% slice(1:10))
library(MASS)
Q4.independence=table(boston$neighborhood,boston$timezone)
chisq.test(Q4.independence)
#Reject the null hypothesis

##Q4d
Q4.department=as.data.frame(count_(boston,"Department",sort = T))
Q4.d=as.data.frame(Q4.department %>% arrange(desc(n)) %>% slice(1:10))
##Top ten departments

###Q4e
Q4.weekday=as.data.frame(count_(boston,c("weekday","weeknum"),sort = T))
Q4.weekday1=as.data.frame(count_(boston,c("weekday","TYPE"),sort = T))
#Most recurrent type
Q4.e=as.data.frame(Q4.weekday1 %>% group_by(weekday) %>% filter(n == max(n)))
#Top 3 request
Q4.e3=as.data.frame(Q4.weekday1 %>% group_by(weekday) %>% arrange(desc(n)) %>%
slice(1:3))
##Number of requests for weekday
Q4.num=as.data.frame(count_(boston,"weekday",sort = T))
#ANOVA TEST????
Q4.weekday$weekday=as.factor(Q4.weekday$weekday)
anova(lm(n~weekday,data = Q4.weekday))

weekday=aov(n~weekday,data = Q4.weekday)
TukeyHSD(weekday)
plot(TukeyHSD(weekday),col="red")
plot(TukeyHSD(weekday))
##reject the null hypothesis, not evenly distributed


##School holiday
Q4.holiday=as.data.frame(count_(boston,c("holiday","TYPE"),sort = T))
```

```
Q4.holiday1=as.data.frame(count_(boston,c("holiday"),sort = T))
x=Q4.holiday[Q4.holiday$holiday=="N",]
y=Q4.holiday[Q4.holiday$holiday=="Y",]
t.test(x$n,y$n)
anova(lm(n~holiday,data = Q4.holiday))
x=t(as.matrix(Q4.holiday1$n))
prop.test(x)
##reject the null hypothesis, holiday and regular days are different.

## Question 5(a)
mydata.open=subset(mydata,CASE_STATUS=="Open")
mydata.closed=subset(mydata,CASE_STATUS=="Closed")
# Requests of Open Status
barplot(table(mydata.open$TYPE),main = " Requests of Open Status",
xlab="Type of Requests",xaxt="n" )
# Requests of Closed Status
barplot(table(mydata.closed$TYPE),xaxt="n",main = " Requests of Closed Status" )
# Percentage of Open Requests
100*sum(mydata$CASE_STATUS=="Open")/nrow(mydata)

# Question 5(b)
table(mydata.open$TYPE)>1500

# Question 5(c)
location0=subset(mydata,neighborhood=="Allston" )
summary(location0)
location1=subset(mydata,neighborhood=="Allston / Brighton" )
summary(location1)
location2=subset(mydata,neighborhood=="Back Bay" )
summary(location2)
location3=subset(mydata,neighborhood=="Beacon Hill")
summary(location3)
location4=subset(mydata,neighborhood=="Boston" )
summary(location4)
location5=subset(mydata,neighborhood=="Brighton" )
summary(location5)
location6=subset(mydata,neighborhood=="Charlestown" )
summary(location6)
location7=subset(mydata,neighborhood=="Chestnut Hill" )
summary(location7)
location8=subset(mydata,neighborhood=="Dorchester" )
summary(location8)
```

```
location9=subset(mydata,neighborhood=="Downtown / Financial District" )
summary(location9)
location10=subset(mydata,neighborhood=="East Boston" )
summary(location10)
location11=subset(mydata,neighborhood=="Fenway / Kenmore
/ Audubon Circle / Longwood" )
summary(location11)
location12=subset(mydata,neighborhood=="Greater Mattapan" )
summary(location12)
location13=subset(mydata,neighborhood=="Hyde Park" )
summary(location13)
location14=subset(mydata,neighborhood=="Jamaica Plain"  )
summary(location14)
location15=subset(mydata,neighborhood=="Mattapan"  )
summary(location15)
location16=subset(mydata,neighborhood=="Mission Hill"  )
summary(location16)
location17=subset(mydata,neighborhood=="Roslindale"  )
summary(location17)
location17=subset(mydata,neighborhood=="Roxbury"  )
summary(location17)
location18=subset(mydata,neighborhood=="South Boston"  )
summary(location18)
location19=subset(mydata,neighborhood=="South Boston / South Boston Waterfront"  )
summary(location19)
location20=subset(mydata,neighborhood=="South End"  )
summary(location20)
location21=subset(mydata,neighborhood=="West Roxbury"  )
summary(location21)


table(mydata$neighborhood_services_district,mydata$TYPE)[,1]
boston.bike=subset(mydata,TYPE=="Abandoned Bicycle")
bike.neighbour=table(boston.bike$neighborhood)
View(bike.neighbour)

# Question 6
tz1=subset(mydata,timezone==1)
tz2=subset(mydata,timezone==2)
tz3=subset(mydata,timezone==3)
tz4=subset(mydata,timezone==4)
x=c(1,2,3,4)
```

```
y=c(nrow(tz1),nrow(tz2),nrow(tz3),nrow(tz4))
plot(x,y,type="l",xlab="Timezone",ylab="Number of Requests",main="Number of
Requests in Each Timezone",xaxt="n")
axis(1, at=1:4, labels=c("0:00-5:59","6:00-11:59","12:00-17:59","18:00-23:59"))

# Question 7
call=subset(mydata,Source=="Constituent Call")
tz1.call=subset(call,timezone==1)
tz2.call=subset(call,timezone==2)
tz3.call=subset(call,timezone==3)
tz4.call=subset(call,timezone==4)
x=c(1,2,3,4)
y=c(nrow(tz1.call),nrow(tz2.call),nrow(tz3.call),nrow(tz4.call))
plot(x,y,type="l",xlab="Timezone",ylab="Number of Requests",main="Number of
Requests by Call",xaxt="n")
axis(1, at=1:4, labels=c("0:00-5:59","6:00-11:59","12:00-17:59","18:00-23:59"))
tz1.boston=subset(mydata,timezone==1)
tz2.boston=subset(mydata,timezone==2)
tz3.boston=subset(mydata,timezone==3)
tz4.boston=subset(mydata,timezone==4)
x=c(1,2,3,4)
y=c(nrow(tz1.boston),nrow(tz2.boston),nrow(tz3.boston),nrow(tz4.boston))
plot(x,y,type="l",xlab="Timezone",ylab="Number of Requests",main="Total Number of
Requests of Each Timezone",xaxt="n")
axis(1, at=1:4, labels=c("0:00-5:59","6:00-11:59","12:00-17:59","18:00-23:59"))
barplot(table(tz2.call$TYPE))
barplot(table(tz3.call$TYPE))
table(tz3.call$TYPE)>1300
which.max(table(tz2.call$TYPE))
which.max(table(tz3.call$TYPE))

boston.bulk=subset(mydata,TYPE=="Schedule a Bulk Item Pickup")
boston.missed=subset(mydata,TYPE=="Missed Trash/Recycling/Yard Waste/Bulk Item")
table(boston.bulk$Source)
table(boston.missed$Source)

# Question 7.2
# In boston, there are four main issues need to be considered and dealt with,
#"Street Light Outages", "Parking Enforcement", "Requests for Street Cleaning"
#and "Schedule a Bulk Item Pickup".
boston.parking=subset(mydata,TYPE=="Parking Enforcement")
temp=table(boston.parking$Department)
```

```
View(temp)

# The top one problem is Parking Enforcement since the number
#of which is always the largest in each day.
boston.street=subset(mydata,TYPE="Requests for Street Cleaning")
which.max(table(boston.street$neighborhood))


# The number of requests for street cleaning greater 10,000 by
#neighbourhood are listed in table"1". The Mayor might consider to
#increase the number of cleaners in these regions.  Meanwhile, the
#above 8 regions are also the top eight neighborhoods with the
#highest number of requests, which indicates these areas should be
#more concerned. Besides, the number of source is listed in table.
temp=table(mydata$Source)
View(temp)

# Street Light Outages by neighborhood
boston.outage=subset(mydata,TYPE=="Street Light Outages")
temp=table(boston.outage$neighborhood)
View(temp)
```

References

[1] City of Boston, Data Boston(2016). ”311,Service Requests”. Retrieved from https://data.cityofboston.gov/City-Services/311-Service-Requests/awu8-dc52.