STAA57 W20 Final Project Report

Group #17L Girvan Tse, Edmond Lee, Brandom Lam

## Introduction

### Description of Report

  This report wants to glean insights into the emergency shelter situation in Toronto using data. At its core we will be asking two main questions: “How operational are emergency shelters?” and ‘What influences emergency shelter usage?’.

## Data

### 211 Caller Data

Usage:   Geographical, Caller Reason, Caller demographic, and Caller Gender data for emergency shelters  
#### Variables and Observations:  
  Level3Name   Category variable that helps us slice 211 caller data for emergency shelter  
  PostalArea   Postal Codes of Toronto  
  DateOfCall   Date of call  
  NeedWasMet   Boolean specifying if the needs of the caller were met Bias   The PostalArea has a non-response bias. That is there is a large amount of calls that contain NULL in PostalArea

### Daily Shelter Occupancy [<https://open.toronto.ca/dataset/daily-shelter-occupancy/>]

Usage:   Geographical Data on the location of shelters as well as occupancy to calculate demographics and usage  
Variables and Observations:  
  OCCUPANCY\_DATE   Date of shelter operation  
  SHELTER\_POSTAL\_CODE   postal code of shelter  
  SECTOR   Categorisation variable for different parts of any shelter  
  OCCUPANCY   The count of occupants in any shelter sector  
  CAPACITY   The count of available spaces in any shelter

See Appendix for Postal Codes reference

## Main Analysis

To conserve space the following abbrevations are used: ES = Emergency Shelter, CD = 211 Caller Data (callerData), ESD = Emergency Shelter Data (shelterData), OC = Overcapacity, OCR = Overcapacity Rate

### How functional are ESs?

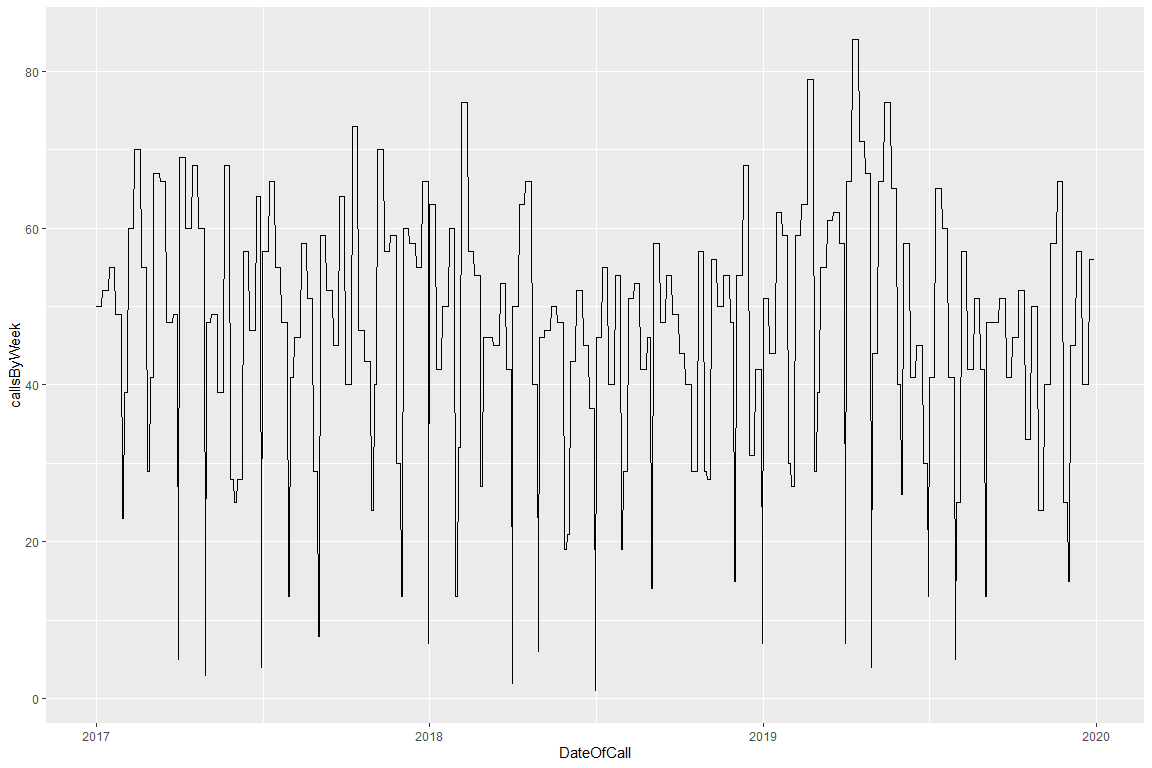
#### 1.0 : What are the ES trends?

##### Method

We’ll use our CD see if the frequency of calls increases over time using weekly view.

##### Analysis

##### 1.0a CD

Here we’ll perform a simple summation of the data to see how it trend weekly.  See Appendix for 1.0b ESD

##### Conclusion

In the graph we can see that the data trends towards a centre line with not upward or downward preference.

#### 1.1 Is every caller able to reach a nearby ES?

##### Method

Here we’ll take the CD and do a check to see if it matches a postal code of any shelter, if it is then we can consider that the user is able to reach an ES (regardless of if it is full/OC or not)

##### Analysis

Pull all FSA codes of shelters in Toronto and filter all FSA codes of callers not in the shelters FSA. Summarise any calls that aren’t in a shelters FSA.

## # A tibble: 3 x 2  
## PostalArea calls  
## <chr> <int>  
## 1 M4J 13  
## 2 M6N 8  
## 3 M4L 7

Proprotion of callers calculation in Appenedix 1.1

##### Conclusion

41.28% of callers are unable to reach shelters are in the EY region if we do not consider CD nonresponce bias.

#### 1.2 Which regions are ESs in high demand?

##### Method

Here we can do an inner join on postal code; this will yield us the caller’s borough that we do a simple summarise on to yield the total number of calls over the years for ESs.

##### Analysis

We’ll retrieve CD with geographical data and join with postalBorough. Then perform a summation to get the number of calls by borough.

## # A tibble: 3 x 2  
## Borough Calls  
## <chr> <int>  
## 1 Downtown Toronto 193  
## 2 Scarborough 112  
## 3 Etobicoke 97

##### Conclusion

We can see that lots of callers are from the DT, Scarborough, North York and Etobicoke boroughs.

### What influences ES usage?

#### 2.1: Does borough influence number of calls in a day?

##### Method

We are testing if the number of calls vary from borough to borough, or if it is standard across all boroughs. The assumption that its standard means we are assuming the need for ESs is the same across all boroughs while the alternate proposes the opposite.

##### Analysis

Here we join ES with postalBorough to get boroughs. We’ll test number of calls in a day and borough to see if they’re related. See Appendix 2.1 for details on the independence test

##   
## Approximative General Independence Test  
##   
## data: Calls by  
## Borough (Central Toronto, Downtown Toronto, East Toronto, East York, Etobicoke, North York, Queen's Park, Scarborough, West Toronto, York)  
## maxT = 6.67, p-value < 1e-04  
## alternative hypothesis: two.sided

##### Conclusion

Given that our p-value resulted in 0.0004 we can conclude that borough does influence our calls in a day.

See Appendix for 2.2: Does sector of ES user influence OCR

#### 2.3: Is ES usage seasonal (monthly) ?

##### Method

This question will help us determine if shelter usage is affected by the month.

##### Analysis

We can use shelter data like previously to get the OCR. The difference is now we’ll use that in tandem with the factor SECTOR and an independence test. See Appendix 2.3 for details on the indepedence test.

##   
## Approximative General Independence Test  
##   
## data: MONTH by LOAD  
## Z = 7.9429, p-value < 1e-04  
## alternative hypothesis: two.sided

##### Conclusion

Given that our p-value resulted in 0.0004 we can conclude that the month does affect ES load. It seems that June, July, August, September, and October have higher loads (See Appendix 2.3).

#### 2.4: Can we classify shelters as regularly OC based solely on their load?

##### Method

Here we will use threshold classification to find a classifer that will tell us if a shelter is OC based solely on its load.

##### Analysis

Here we’ll graph our shelter data to see what a good initial classifer would be. We can determine that 0.895 is a pretty good divide between OC and high load (See Appendix 2.4). To confirm this we can make a crosstab with our data comparing our guess and the real state of the ESs

## OC\_REAL  
## OC\_GUESS TRUE FALSE Sum  
## TRUE 8180 1287 9467  
## FALSE 541 621 1162  
## Sum 8721 1908 10629

Doing a small calculation for accuracy yields

## [1] 0.8280177

See Appendix 2.4 for Precision/Recall Plot

##### Conclusion

The classifer can’t achive above a 75% precision/75% recall. For this given context of ESs I would say this isn’t accurate enough even at 83%.

## Summary

Here we’ll restate and answer all the questions using insight gleaned from the data. I’ll add a short answer to the subsections then answer the main question.

### How functional are ESs?

Using the subsections we can see that ESs in Toronto are functional. We can see that (vocal minority) callers aren’t able to shelters in their FSA (mainly East York) but they are a very small fraction of callers when compared to the most in demands areas (13~ vs 193~).

### What influences ES usage?

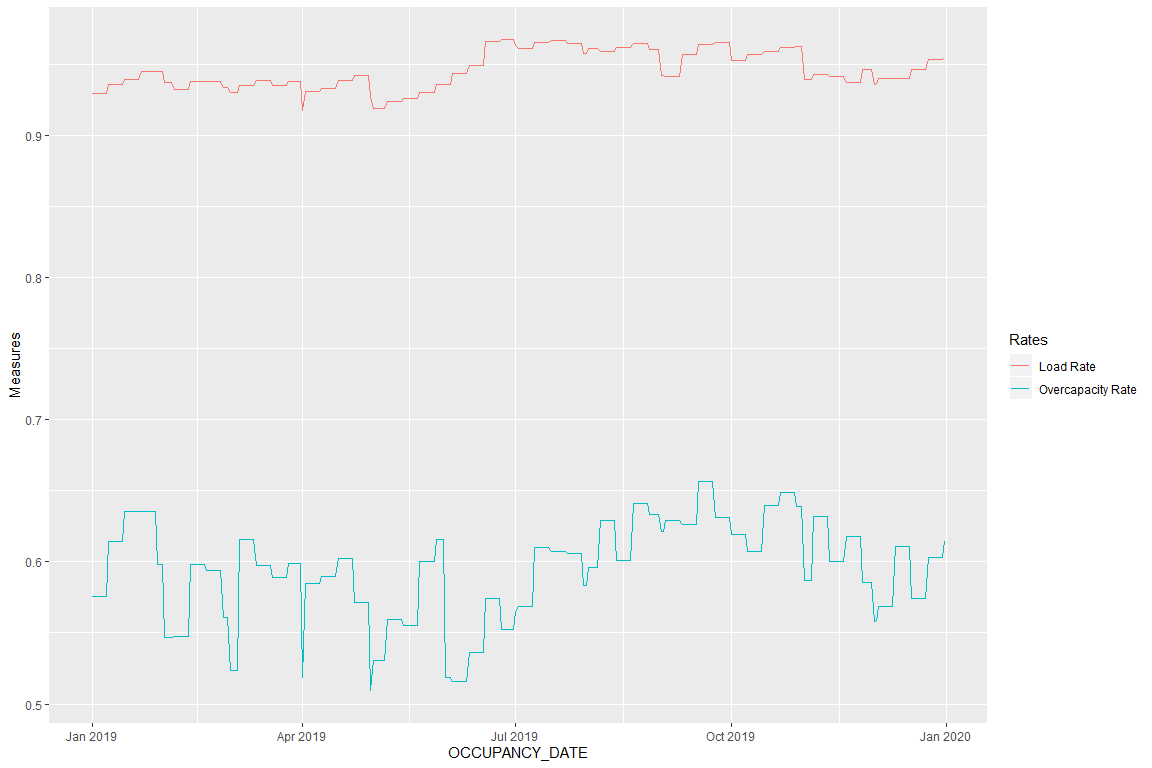
Here we can confirm a few factors (namely borough, month) that influence usage. Another factor (namely demographic) that influences OCR. A linear classifier can be made but is not accurate enough.

## Appendix

### Toronto Postal Codes

Link: <https://www.canadapost.ca/cpo/mc/personal/postalcode/fpc.jsf> Usage:   Converting Postcodes to City of Toronto Boroughs  
#### Variables and Observations   Postcode   Postal Codes of Toronto  
  Borough   Boroughs of Toronto  
  Neighbourhoods   Neighbourhoods of Toronto

##### 1.0b ESD



#### 1.1 Proprotion

Using the sum of callers without a shelter in their FSA we can get a proportion

## [1] 0.4127907

That is 41.28% of callers (with geographical data) do not have a shelter in their FSA

#### 2.1 Indepedence Test Details

H0 : The borough and number of calls are not related  
Halt : The borough and number of calls are related  
We’ll work with a 95% confidence interval (and thus require p <= 0.05 to reject the null)

#### 2.2: Does sector of ES user influence OCR?

##### Method

Here we need to get the OCR of ESs, afterwards we’ll need to do an independence test to see if sectors and OCR are related. This specifically will point out if some sectors are often OC, more so than others.

##### Analysis

We can use shelter data like previously to get the OCR. The difference is now we’ll use that in tandem with the factor SECTOR and an independence test.  
H0 : All sectors are equally likely to have OC issues  
Halt : Some sectors are more often OC than others  
We’ll work with a 95% confidence interval (and thus require p <= 0.05 to reject the null)

##   
## Approximative General Independence Test  
##   
## data: SECTOR by OC\_RATE  
## maxT = 33.866, p-value < 1e-04  
## alternative hypothesis: two.sided

##### Conclusion

Given that our p-value resulted in 0.0004 we can conclude that ES sector does influence OCR. Succinctly, some sectors are more often OC than other sectors. This means we should seek to expand ES housing for some sectors more than others.

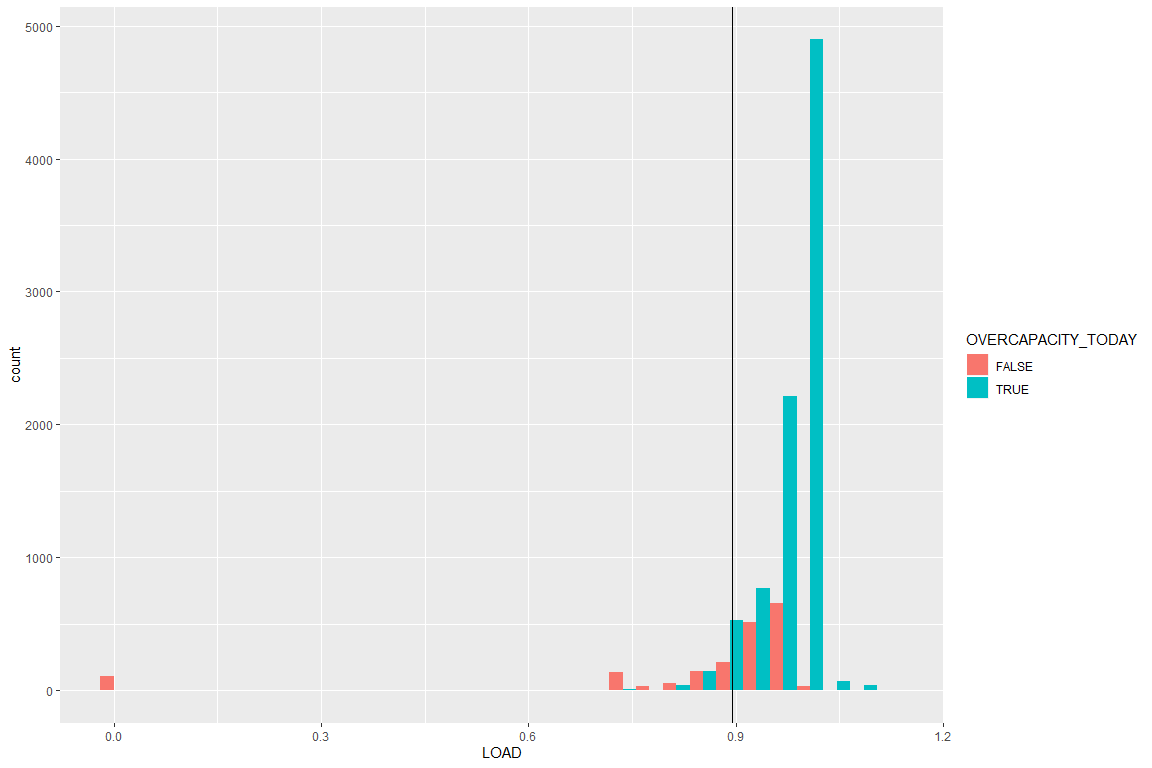
#### 2.3 Independence Test

H0 : The month doesn’t affect the ES load  
Halt : The month does affect the ES load  
We’ll work with a 95% confidence interval (and thus require p <= 0.05 to reject the null

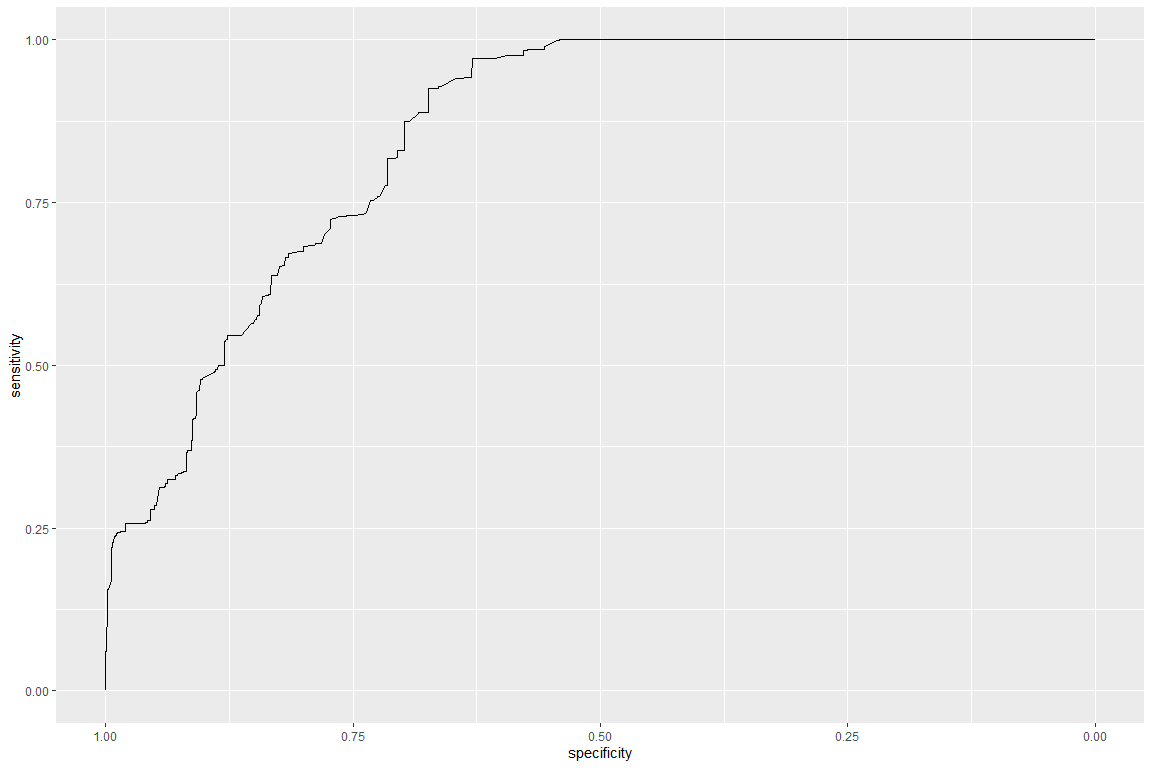
#### 2.3 Monthly Distribution of Load

## # A tibble: 12 x 2  
## MONTH LOAD  
## <dbl> <dbl>  
## 1 1 0.938  
## 2 2 0.936  
## 3 3 0.936  
## 4 4 0.935  
## 5 5 0.926  
## 6 6 0.954  
## 7 7 0.964  
## 8 8 0.961  
## 9 9 0.956  
## 10 10 0.958  
## 11 11 0.941  
## 12 12 0.945

#### 2.4 Classifier Graph



#### 2.4 ROC curve at best a balanced 75% precision/75% recall



## Area under the curve: 0.8546

### Footnote for Professor

Girvan Here  
I tried using ggmaps but found that in 2018 google changed the requirement to use the API to a private key from google cloud. If I wanted to put a ggmap into the report, I’d need to also publicly put in my private key for my google account in the R code to allow readers to run the report. Also tried using the maps and sp package but they have their own problems (maps is doesn’t contain up to date Canadian political borders; sp uses GADM which takes ages to render).

Also I tried using PCCF but found it was too detailed and unwieldly (it seems to be more compatible for SAS) for an FSA -> Borough conversion. I do understand however that PCCF is a more academic source (and frankly better than just pulling postal codes by hand.)

Also I think 5 pages was too little only because the R code outputs take up so much space so a lot of main report stuff was cut and put into the appendix.