Untitled2

May 12, 2018

```
In [28]: library(ggplot2) # Data visualization
         library(readr) # CSV file I/O, e.g. the read_csv function
         library(plyr)
         library(dplyr)
         library(reshape2)
In [29]: NYC311 = read.csv('cleaned_data.csv', header=T)
         NYC311$complaint_type = tolower(NYC311$complaint_type)
In [5]: NYC311$complaint_type = gsub('s$', '', NYC311$complaint_type)
        NYC311$incident_zip = gsub('-[[:digit:]]{4}$', '', NYC311$incident_zip)
        NYC311$complaint_type = gsub('paint - plaster', 'paint/plaster', NYC311$complaint_type)
        NYC311$complaint_type = gsub('general construction', 'construction', NYC311$complaint_ty
        NYC311$complaint_type = gsub('nonconst', 'construction', NYC311$complaint_type)
        NYC311$complaint_type = gsub('street sign - [[:alpha:]]+', 'street sign', NYC311$complai
        NYC311$complaint_type = gsub('fire alarm - .+','fire alarm', NYC311$complaint_type)
        idx = grepl('[[:digit:]]{5}', NYC311$incident_zip)
        NYC311clean = NYC311[idx,]
In [6]: NYC311byZip = ddply(NYC311clean, .(incident_zip, complaint_type), count)
In [14]: library(tidyr) #prepare data for pca
         raw = spread(NYC311byZip, complaint_type, n)
         raw[is.na(raw)] = 0
         counts = which(colSums(raw[,-1]) < 10)</pre>
         zipcodes = raw[,1]
         raw = raw[,-1]
         raw = raw[,-counts]
         processed = scale(raw, center=T, scale=T)
         library(psych)
         pca = principal(processed, nfactor=8, covar=F)
In [18]: loadings = as.data.frame(pca$loadings[,1:7])
         loadings$complaint.type = rownames(loadings)
         loadings_m = melt(loadings, id='complaint.type')
         ggplot(loadings_m, aes(x=variable, y=complaint.type, label = round(value,2), fill=value
```

```
geom_tile()+xlab('Factor')+ylab('Complaint Description')+geom_text(size=5.75, alp
scale_fill_continuous(low='green', high='blue', name='Loadings')+
theme(axis.text.y = element_text(size=10))

water system-
variety-
0.26 0.74 0.39 0.28 -0.01 0.02 0.14
water quality-
0.47 0.65 0.3 0.24 -0.04 0.04 0.08
```

0.47 0.65 0.3 0.24 -0.04 0.04 0.08 0.12 -0.09 0.13 **0.9** 0.03 0.07 trans fat smoking -0.58 0.56 0.08 -0.13 0 0.02 0.04 0.14 0.55 0.55 0.5 0.02 0.04 0.1 sanitation condition -0.28 0.81 0.11 0.18 -0.03 0.01 0.14 **0.65** 0.32 0.19 0.34 0.08 0.04 0.03 **0.68 0.63** -0.07 -0.02 0.01 -0.03 0.06 indoor air quality -0.02 0.56 0.09 0.03 0.12 0.54 -0.07 0.65 0.36 0.35 0.25 0.07 hazardous material -0.05 -0.01 0.45 -0.29 0.65 -0.1 graffiti or vandalism -**0.89** 0.28 0.12 0.07 0.06 0.02 0.03 food poisoning -**0.86** 0.37 0.14 0.09 0.07 0.01 0.03 food establishment -0.32 0.82 -0.06 0.09 0.04 0.12 -0.07 drinking water -0.42 0.61 0.3 0.06 -0.06 0.07 0.07 drinking -0.1 -0.01 -0.04 0.01 0.04 0.96 doris literature request -0.67 0.47 0.43 0.02 0.06 0.11 0.2 dirty condition -0.17 0.18 0.82 0.15 0.04 0.11 -0.04 damaged or dead tree --0.15 0.05 0.02 -0.03 0.19 -0.04 calorie labeling -**0.77 0.51 0.07 0.05 0.03 -0.02 0.06** asbesto -0.07 0.04 0.05 -0.03 -0.06 0.9 0.07 annual/ cycle inspection -annual inspection -0.3 0.04 0.07 0.16 -0.05 0.12 air quality -0.85 RC1 RC2 RC5 RC7 RC3 RC4 RC6 Factor

Complaint Description

```
Loadings
- 0.9
- 0.6
- 0.3
- 0.0
```

```
tick.marks=FALSE, main='Cluster Assignments')
table(cluster$cluster)
```

Error in library(scatterplot3d): there is no package called scatterplot3d Traceback:

- 1. library(scatterplot3d)
- 2. stop(txt, domain = NA)