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)
```

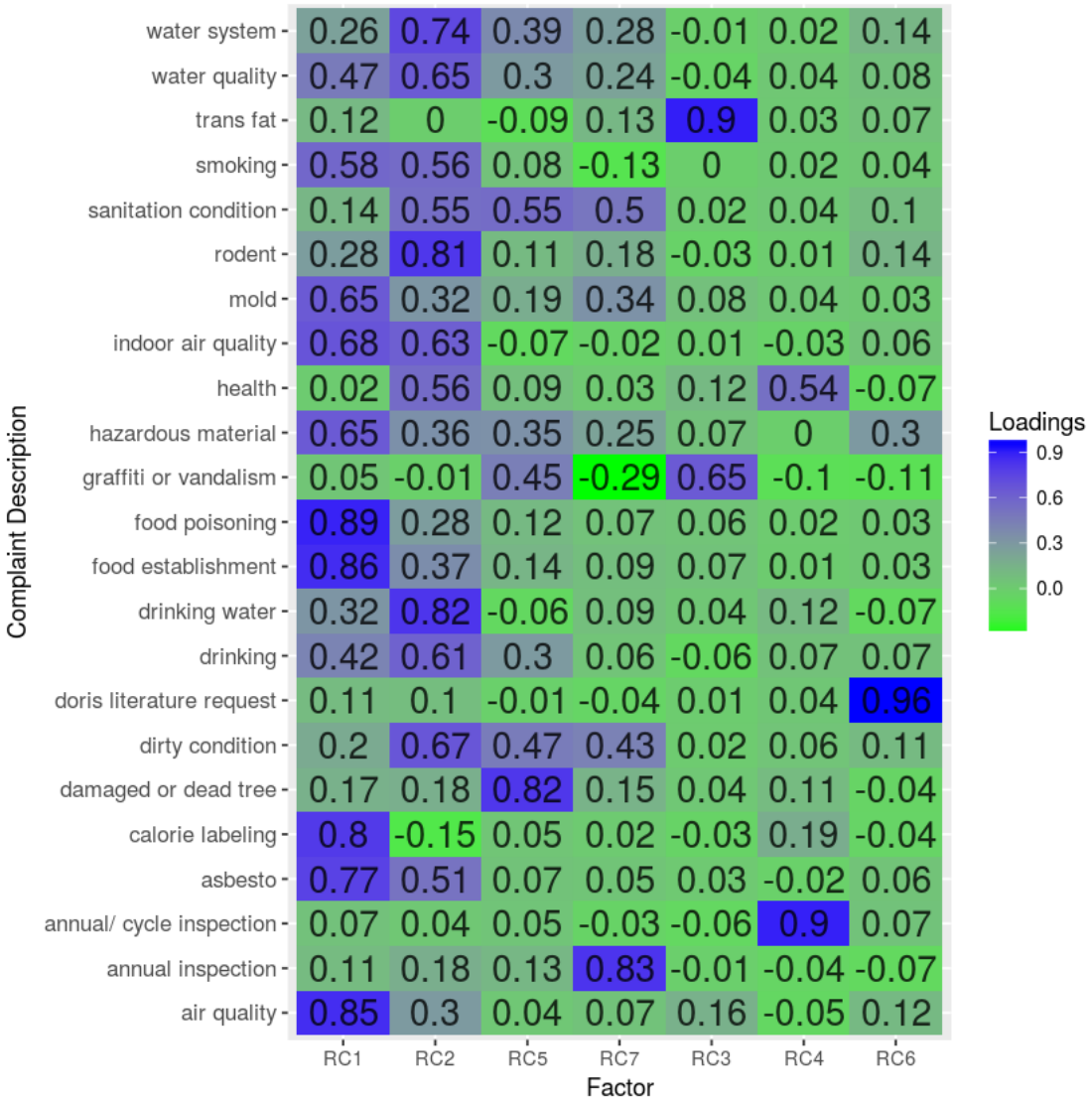
In [14]:
```
library(tidyr) #prepare data for pca
raw = spread(NYC311byZip, complaint_type, n)
raw[is.na(raw)] = 0
counts = which(colSums(raw[,-1]) < 10)
zipcodes = raw[,1]
raw = raw[,-1]
raw = raw[,-counts]
processed = scale(raw, center=T, scale=T)

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))
```



In [27]:
```
#Cluster data
set.seed(400)
cluster=kmeans(processed, 7)

#Visualize cluster results
library(scatterplot3d)
#library(rgl)
NYCPCs = pca$scores
scatterplot3d(NYCPCs[,3], NYCPCs[,1], NYCPCs[,2], color=cluster$cluster, xlab='', ylab=
```

```
                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)
```