205_project

Thong Bui

11/18/2016

Figuring out the correlations from Amazon's overall and NLTK values

This section is using a sample file generated from parse_json.py and find out which variable has the best correlation. Then, we will try to combine the variables to figure out the outliers, anomalies to hopefully find out "weird" items

```
library(car)
setwd("~/Desktop/MIDS/205_storage_retrieval/github/w205-project/code/analysis")
sa = read.csv("sample.csv")
summary(sa)
```

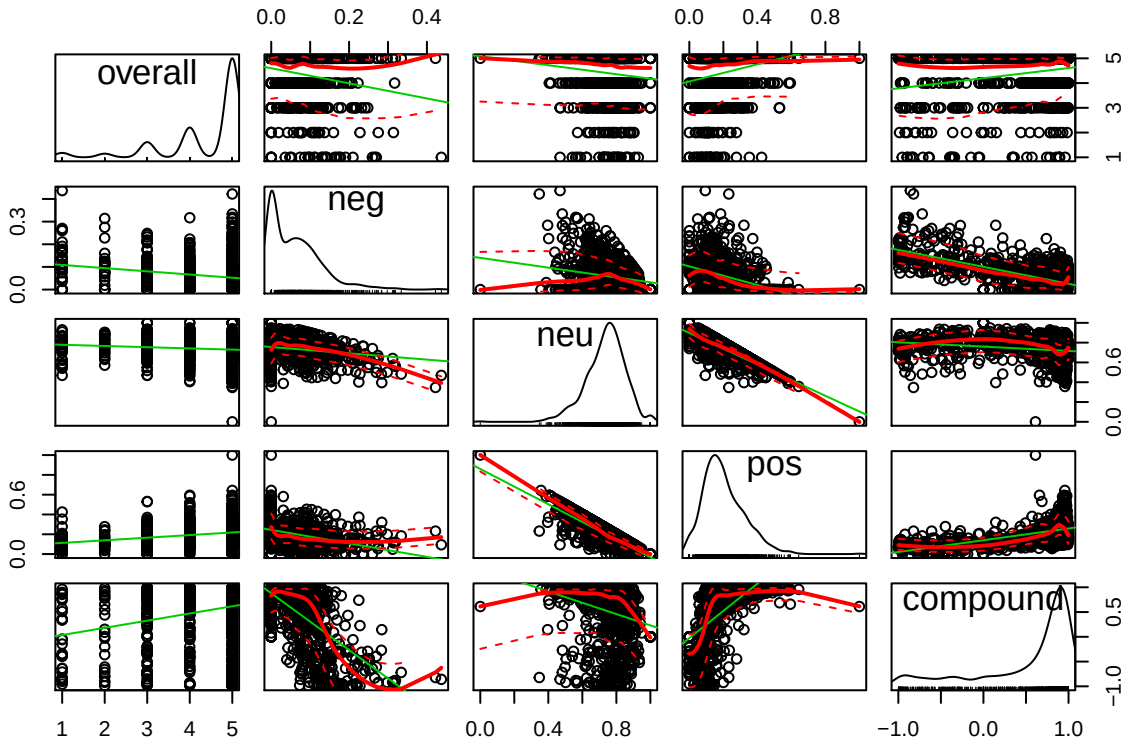
```
##     overall           neg               neu              pos        
##  Min.   :1.000   Min.   :0.00000   Min.   :0.0000   Min.   :0.0000  
##  1st Qu.:4.000   1st Qu.:0.00000   1st Qu.:0.6715   1st Qu.:0.1190  
##  Median :5.000   Median :0.04900   Median :0.7475   Median :0.1825  
##  Mean   :4.416   Mean   :0.06053   Mean   :0.7364   Mean   :0.2030  
##  3rd Qu.:5.000   3rd Qu.:0.09725   3rd Qu.:0.8100   3rd Qu.:0.2732  
##  Max.   :5.000   Max.   :0.43700   Max.   :1.0000   Max.   :1.0000  
##     compound       
##  Min.   :-0.9996  
##  1st Qu.: 0.3400  
##  Median : 0.8074  
##  Mean   : 0.5324  
##  3rd Qu.: 0.9326  
##  Max.   : 0.9993
```

```
nrow(sa)
```

```
## [1] 1000
```

```
scatterplotMatrix(~ overall + neg + neu + pos + compound, data=sa)
```

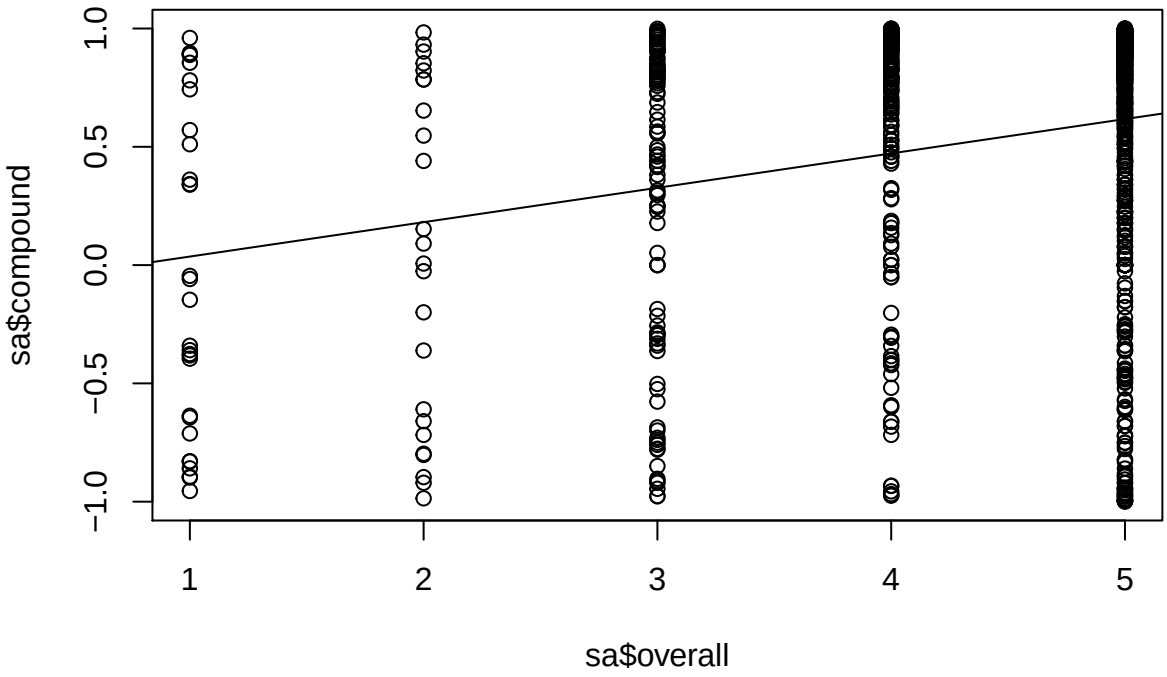
```
## Warning in smoother(x, y, col = col[2], log.x = FALSE, log.y = FALSE,
## spread = spread, : could not fit smooth

## Warning in smoother(x, y, col = col[2], log.x = FALSE, log.y = FALSE,
## spread = spread, : could not fit smooth

## Warning in smoother(x, y, col = col[2], log.x = FALSE, log.y = FALSE,
## spread = spread, : could not fit smooth

## Warning in smoother(x, y, col = col[2], log.x = FALSE, log.y = FALSE,
## spread = spread, : could not fit smooth
```




```
(m1=lm(compound ~ overall, data=sa))
```

```
##
## Call:
## lm(formula = compound ~ overall, data = sa)
##
## Coefficients:
## (Intercept)      overall
##     -0.1093       0.1453
```

```
plot(sa$overall, sa$compound)
abline(m1)
```




```
cor(sa$overall, sa$compound)
```

```
## [1] 0.2440726
```

```
cor(sa$overall, sa$pos)
```

```
## [1] 0.2141229
```

```
cor(sa$overall, sa$neg)
```

```
## [1] -0.2077541
```

```
cor(sa$overall, sa$neu)
```

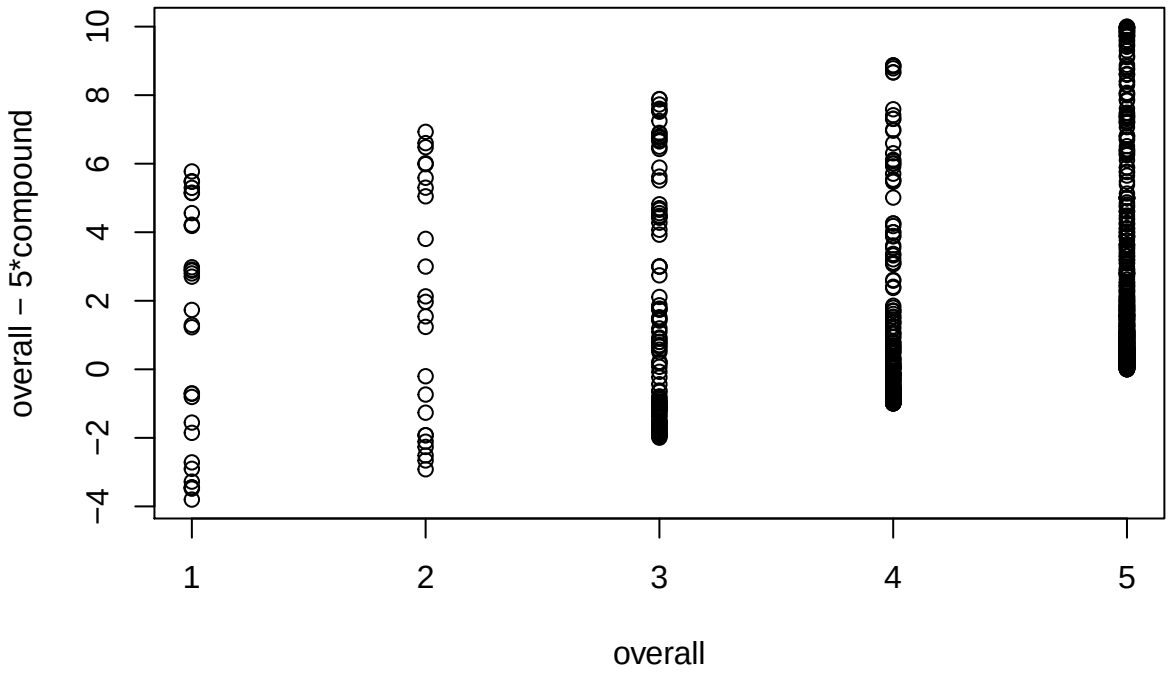
```
## [1] -0.1055931
```

Because compound values is in the range of (-1, 1) and overall is in the range [0,5], I thought using this function:

overall - 5 * compound

will help us figure out the outliers

```
plot(sa$overall, -(5*sa$compound) +sa$overall, xlab = "overall", ylab="overall - 5*compound")
```



As you can see, the outliers lay when overall - 5 * compound $<$ -3 or $>$ 8