Yelp Project

Shane McIntyre

12/8/2020

```
yelp_test = read.csv('/Users/shanemcintyre/Desktop/Fall Semester 2019/Stats
333/Yelp_test.csv')
yelp = read.csv('/Users/shanemcintyre/Desktop/Fall Semester 2019/Stats
333/Yelp_train.csv')
yelp_val = read.csv('/Users/shanemcintyre/Desktop/Fall Semester 2019/Stats
333/Yelp_validate.csv')
attach(yelp_test)
attach(yelp)

## The following objects are masked from yelp_test:
##
##     adorable, affordable, amazing, amazingly, apologize, apology,
##     awesome, awful, beautifully, burned, categories, charged, city,
##     complained, contact, cool, date, delectable, delicious,
##     deliciously, deliciousness, delightful, delish, die, dirty,
##     disappoint, disgusting, divine, downhill, dream, edible, everyday,
##     excellent, fabulous, fantastic, favorite, favorites, flag,
##     flavorless, funny, garbage, gem, gross, heaven, heavenly, hidden,
##     highly, horrible, Id, ignored, inattentive, incredible, inedible,
##     knowledgable, knowledgeable, lukewarm, management, manager,
##     mcdonald.s, mediocre, messed, name, nasty, nchar, nope, notch,
##     nword, outstanding, paired, perfect, perfection, perfectly,
##     phenomenal, poor, poorly, proceeded, receipt, refund, refused,
##     relaxed, response, rotating, rubbery, rude, scones, screw,
##     scrumptious, secret, sentiment, skeptical, sourced, stale, subpar,
##     sucks, superb, supposed, tasteless, terrace, terrible, text,
##     undercooked, unpleasant, upset, useful, waste, watering, wonderful,
##     wonderfully, worse, worst, X, yum, zero

attach(yelp_val)

## The following objects are masked from yelp:
##
##     adorable, affordable, amazing, amazingly, apologize, apology,
##     awesome, awful, beautifully, burned, categories, charged, city,
##     complained, contact, cool, date, delectable, delicious,
##     deliciously, deliciousness, delightful, delish, die, dirty,
##     disappoint, disgusting, divine, downhill, dream, edible, everyday,
##     excellent, fabulous, fantastic, favorite, favorites, flag,
##     flavorless, funny, garbage, gem, gross, heaven, heavenly, hidden,
##     highly, horrible, Id, ignored, inattentive, incredible, inedible,
##     knowledgable, knowledgeable, lukewarm, management, manager,
```

```
##     mcdonald.s, mediocre, messed, name, nasty, nchar, nope, notch,
##     nword, outstanding, paired, perfect, perfection, perfectly,
##     phenomenal, poor, poorly, proceeded, receipt, refund, refused,
##     relaxed, response, rotating, rubbery, rude, scones, screw,
##     scrumptious, secret, sentiment, skeptical, sourced, stale, subpar,
##     sucks, superb, supposed, tasteless, terrace, terrible, text,
##     undercooked, unpleasant, upset, useful, waste, watering, wonderful,
##     wonderfully, worse, worst, X, yum, zero
##
## The following objects are masked from yelp_test:
##
##     adorable, affordable, amazing, amazingly, apologize, apology,
##     awesome, awful, beautifully, burned, categories, charged, city,
##     complained, contact, cool, date, delectable, delicious,
##     deliciously, deliciousness, delightful, delish, die, dirty,
##     disappoint, disgusting, divine, downhill, dream, edible, everyday,
##     excellent, fabulous, fantastic, favorite, favorites, flag,
##     flavorless, funny, garbage, gem, gross, heaven, heavenly, hidden,
##     highly, horrible, Id, ignored, inattentive, incredible, inedible,
##     knowledgable, knowledgeable, lukewarm, management, manager,
##     mcdonald.s, mediocre, messed, name, nasty, nchar, nope, notch,
##     nword, outstanding, paired, perfect, perfection, perfectly,
##     phenomenal, poor, poorly, proceeded, receipt, refund, refused,
##     relaxed, response, rotating, rubbery, rude, scones, screw,
##     scrumptious, secret, sentiment, skeptical, sourced, stale, subpar,
##     sucks, superb, supposed, tasteless, terrace, terrible, text,
##     undercooked, unpleasant, upset, useful, waste, watering, wonderful,
##     wonderfully, worse, worst, X, yum, zero

yelp_out = rbind(yelp_test,yelp_val)
str(yelp)

## 'data.frame':    55342 obs. of  114 variables:
##  $ X            : int  2677 4349 6109 6302 7224 13664 13743 16193 17312
19712 ...
##  $ Id           : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ stars        : int  4 4 5 5 5 4 2 1 2 3 ...
##  $ name         : chr  "1847 At the Stamm House" "1847 At the Stamm House"
"1847 At the Stamm House" "1847 At the Stamm House" ...
##  $ text         : chr  "Went early on a Friday night (5:45pm-ish) and got
right in.  They seated us in the upstairs dining room, which "| __truncated__
"Great food, and atmosphere but I wouldn't recommend for large groups bc
otherwise the time it takes to get the "| __truncated__ "Bottom Line (at the
top): We were pleasantly surprised with our visit to 1847 tonight.  I've
never left a Yelp "| __truncated__ "We spent a lovely evening here with
dinner on the outside patio. Our meal started with cocktails, Moscow Mules,"|
__truncated__ ...
##  $ date         : chr  "2016-11-30 22:58:58" "2018-05-18 18:53:09" "2015-
10-31 05:22:56" "2017-05-30 01:18:21" ...
##  $ useful       : int  0 0 5 2 0 0 0 0 2 0 ...
```

```
##  $ funny        : int  0 0 0 0 0 0 0 0 0 1 ...
##  $ cool         : int  0 1 1 0 0 0 0 0 0 0 ...
##  $ city         : chr  "Middleton" "Middleton" "Middleton" "Middleton" ...
##  $ nchar        : int  604 132 1293 805 152 124 2764 768 468 2447 ...
##  $ nword        : int  113 26 238 142 31 25 497 143 92 473 ...
##  $ sentiment    : num  2.8 2 1.8 2 2.5 ...
##  $ categories   : chr  "Supper Clubs, French, Restaurants, Gastropubs,
American (New), American (Traditional)" "Supper Clubs, French, Restaurants,
Gastropubs, American (New), American (Traditional)" "Supper Clubs, French,
Restaurants, Gastropubs, American (New), American (Traditional)" "Supper
Clubs, French, Restaurants, Gastropubs, American (New), American
(Traditional)" ...
##  $ gem          : int  0 0 1 0 0 0 1 0 0 0 ...
##  $ incredible   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ perfection   : int  0 0 0 1 0 0 0 0 0 0 ...
##  $ heaven       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ phenomenal   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ divine       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ die          : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ deliciously  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ highly       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ heavenly     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ superb       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ amazing      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ favorites    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ delectable   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ perfect      : int  0 0 1 1 0 0 0 0 0 0 ...
##  $ sourced      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ knowledgeable: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ wonderfully  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ deliciousness: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ knowledgable : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ fantastic    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ adorable     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ wonderful    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ fabulous     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ scrumptious  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ hidden       : int  0 0 1 0 0 0 0 0 0 0 ...
##  $ notch        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ favorite     : int  0 0 0 0 0 0 0 0 0 1 ...
##  $ disappoint   : int  0 0 0 0 0 0 1 1 0 0 ...
##  $ watering     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ delightful   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ yum          : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ outstanding  : int  1 0 0 0 0 0 0 0 0 0 ...
##  $ awesome      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ everyday     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ relaxed      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ dream        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ amazingly    : int  0 0 0 0 0 0 0 0 0 0 ...
```

```
##  $ delicious    : int  0 0 1 1 0 0 0 0 0 0 ...
##  $ affordable   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ scones       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ excellent    : int  1 0 0 0 0 0 0 0 0 0 ...
##  $ rotating     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ paired       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ secret       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ beautifully  : int  0 0 0 0 0 0 1 1 0 0 ...
##  $ terrace      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ skeptical    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ perfectly    : int  0 0 1 0 0 0 0 0 0 0 ...
##  $ delish       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ supposed     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ mcdonald.s   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ undercooked  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ management   : int  0 0 0 0 0 0 1 0 0 0 ...
##  $ sucks        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ flag         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ messed       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ unpleasant   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ dirty        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ mediocre     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ manager      : int  0 0 0 1 0 0 0 0 0 0 ...
##  $ burned       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ nope         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ upset        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ zero         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ stale        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ screw        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ subpar       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ inattentive  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ edible       : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ complained   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ rubbery      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ lukewarm     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ garbage      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ contact      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ flavorless   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ poor         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ apologize    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ gross        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ charged      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ receipt      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ worse        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ ignored      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ response     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ poorly       : int  0 0 0 0 0 0 0 0 0 0 ...
##   [list output truncated]
```

```r
yelp = yelp[,-1]
yelp_test = yelp_test[,-1]
yelp_val = yelp_val[,-1]
yelp_out = yelp_out[,-1]

# convert text into actual strings
yelp$text = as.character(yelp$text)
yelp_out$text = as.character(yelp_out$text)
yelp$categories = as.character(yelp$categories)
yelp_out$categories = as.character(yelp_out$categories)

# Refactorize yelp_out city and restaurant names after binding
yelp_out$name = as.character(yelp_out$name)
yelp_out$city = as.character(yelp_out$city)
yelp_out$city = factor(yelp_out$city)

# Fix date variable into actual dates
yelp$date = as.Date(yelp$date)
yelp_out$date = as.Date(yelp_out$date)

#benchmark
dat = yelp[,-c(1,3:5,9,13)]
benchmark = lm(stars~., data=dat)
sqrt(sum(benchmark$residuals^2)/benchmark$df.residual) #RMSE check
```

```
## [1] 0.9313123
```

```r
summary(benchmark)
```

```
##
## Call:
## lm(formula = stars ~ ., data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.8846 -0.5763  0.0710  0.6563  3.8330
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.0891864  0.0096830 319.033  < 2e-16 ***
## useful        -0.0577548  0.0028668 -20.146  < 2e-16 ***
## funny         -0.0829936  0.0042098 -19.714  < 2e-16 ***
## cool           0.1852480  0.0045599  40.625  < 2e-16 ***
## nchar          0.0010083  0.0001165   8.658  < 2e-16 ***
## nword         -0.0063289  0.0006130 -10.324  < 2e-16 ***
## sentiment      0.3998109  0.0041778  95.699  < 2e-16 ***
## gem            0.2799987  0.0334504   8.371  < 2e-16 ***
## incredible     0.8036893  0.0709230  11.332  < 2e-16 ***
## perfection     0.0775190  0.0409644   1.892 0.058449 .
## heaven         0.3746516  0.0474424   7.897 2.91e-15 ***
```

```
## phenomenal     0.3243697  0.0408534   7.940 2.06e-15 ***
## divine         0.1855574  0.0656607   2.826 0.004715 **
## die            0.1665433  0.0127144  13.099  < 2e-16 ***
## deliciously    0.0173388  0.0801952   0.216 0.828827
## highly         0.3329703  0.0202531  16.440  < 2e-16 ***
## heavenly      -0.0836687  0.0867021  -0.965 0.334542
## superb        -0.0003135  0.0455265  -0.007 0.994506
## amazing        0.2136531  0.0108947  19.611  < 2e-16 ***
## favorites      0.0306767  0.0358525   0.856 0.392203
## delectable     0.2381609  0.0803750   2.963 0.003047 **
## perfect        0.2199324  0.0142467  15.437  < 2e-16 ***
## sourced        0.1594288  0.0616291   2.587 0.009687 **
## knowledgeable  0.2195770  0.0398752   5.507 3.67e-08 ***
## wonderfully    0.1546324  0.0710143   2.177 0.029449 *
## deliciousness -0.0243057  0.0761661  -0.319 0.749641
## knowledgable   0.3196812  0.0850073   3.761 0.000170 ***
## fantastic      0.1866539  0.0163602  11.409  < 2e-16 ***
## adorable       0.2012420  0.0711092   2.830 0.004656 **
## wonderful      0.1487266  0.0177296   8.389  < 2e-16 ***
## fabulous       0.1179691  0.0359097   3.285 0.001020 **
## scrumptious    0.2657841  0.0804797   3.302 0.000959 ***
## hidden         0.1921794  0.0508022   3.783 0.000155 ***
## notch          0.3632581  0.0372226   9.759  < 2e-16 ***
## favorite       0.3175478  0.0121941  26.041  < 2e-16 ***
## disappoint    -0.1088343  0.0125365  -8.681  < 2e-16 ***
## watering       0.2469797  0.0677308   3.646 0.000266 ***
## delightful     0.2408598  0.0440981   5.462 4.73e-08 ***
## yum            0.1591653  0.0155897  10.210  < 2e-16 ***
## outstanding    0.1031219  0.0267578   3.854 0.000116 ***
## awesome        0.1777371  0.0142045  12.513  < 2e-16 ***
## everyday       0.2923938  0.0736289   3.971 7.16e-05 ***
## relaxed        0.2940164  0.0537608   5.469 4.55e-08 ***
## dream          0.4193816  0.0493936   8.491  < 2e-16 ***
## amazingly      0.0119979  0.0668639   0.179 0.857595
## delicious      0.3025384  0.0087442  34.599  < 2e-16 ***
## affordable     0.2206039  0.0407279   5.417 6.10e-08 ***
## scones         0.1493058  0.0433957   3.441 0.000581 ***
## excellent      0.2093327  0.0119949  17.452  < 2e-16 ***
## rotating       0.1736878  0.0737394   2.355 0.018505 *
## paired         0.0129436  0.0539736   0.240 0.810476
## secret         0.2705495  0.0576278   4.695 2.68e-06 ***
## beautifully   -0.0192309  0.0580524  -0.331 0.740443
## terrace        0.2789547  0.0539091   5.175 2.29e-07 ***
## skeptical      0.4071186  0.0765952   5.315 1.07e-07 ***
## perfectly     -0.0978024  0.0245952  -3.976 7.00e-05 ***
## delish         0.2499442  0.0467518   5.346 9.02e-08 ***
## supposed      -0.2958733  0.0333197  -8.880  < 2e-16 ***
## mcdonald.s    -0.2505454  0.0543776  -4.608 4.08e-06 ***
## undercooked   -0.4705585  0.0545825  -8.621  < 2e-16 ***
## management    -0.4977050  0.0489716 -10.163  < 2e-16 ***
```

```
## sucks         -0.3381900  0.0639893  -5.285 1.26e-07 ***
## flag          -0.2420702  0.0546143  -4.432 9.34e-06 ***
## messed        -0.1268703  0.0731721  -1.734 0.082948 .
## unpleasant    -0.4572210  0.0740836  -6.172 6.80e-10 ***
## dirty         -0.2087756  0.0281203  -7.424 1.15e-13 ***
## mediocre      -0.5955654  0.0286939 -20.756  < 2e-16 ***
## manager       -0.1785734  0.0174063 -10.259  < 2e-16 ***
## burned        -0.4695676  0.0684943  -6.856 7.18e-12 ***
## nope          -0.2933806  0.0623935  -4.702 2.58e-06 ***
## upset         -0.1358210  0.0608534  -2.232 0.025623 *
## zero          -0.2718498  0.0496903  -5.471 4.50e-08 ***
## stale         -0.4502562  0.0503892  -8.936  < 2e-16 ***
## screw         -0.3502364  0.0514256  -6.811 9.82e-12 ***
## subpar        -0.8263314  0.0826360 -10.000  < 2e-16 ***
## inattentive   -0.4397579  0.0837088  -5.253 1.50e-07 ***
## edible        -0.4493075  0.0645477  -6.961 3.42e-12 ***
## complained    -0.0973106  0.0669722  -1.453 0.146230
## rubbery       -0.5104571  0.0722095  -7.069 1.58e-12 ***
## lukewarm      -0.6072710  0.0600648 -10.110  < 2e-16 ***
## garbage       -0.2395133  0.0837916  -2.858 0.004259 **
## contact       -0.2901692  0.0538260  -5.391 7.04e-08 ***
## flavorless    -0.5430040  0.0423707 -12.816  < 2e-16 ***
## poor          -0.3622912  0.0275626 -13.144  < 2e-16 ***
## apologize     -0.0026555  0.0391204  -0.068 0.945881
## gross         -0.4270127  0.0394664 -10.820  < 2e-16 ***
## charged       -0.2787452  0.0364785  -7.641 2.18e-14 ***
## receipt       -0.2412892  0.0569810  -4.235 2.29e-05 ***
## worse         -0.2173959  0.0402611  -5.400 6.70e-08 ***
## ignored       -0.2947661  0.0602022  -4.896 9.79e-07 ***
## response      -0.4095200  0.0603077  -6.791 1.13e-11 ***
## poorly        -0.1095021  0.0688542  -1.590 0.111763
## nasty         -0.2653322  0.0613285  -4.326 1.52e-05 ***
## proceeded     -0.3021905  0.0706331  -4.278 1.89e-05 ***
## terrible      -0.3734419  0.0242651 -15.390  < 2e-16 ***
## waste         -0.4374465  0.0406594 -10.759  < 2e-16 ***
## tasteless     -0.5760351  0.0452082 -12.742  < 2e-16 ***
## inedible      -0.0694323  0.0876383  -0.792 0.428213
## downhill      -0.7373187  0.0831679  -8.865  < 2e-16 ***
## awful         -0.3307123  0.0316455 -10.451  < 2e-16 ***
## rude          -0.4684396  0.0230964 -20.282  < 2e-16 ***
## horrible      -0.4368221  0.0303162 -14.409  < 2e-16 ***
## refused       -0.2368852  0.0756638  -3.131 0.001744 **
## apology       -0.3306422  0.0594842  -5.558 2.73e-08 ***
## disgusting    -0.3345741  0.0459934  -7.274 3.53e-13 ***
## worst         -0.5122428  0.0259276 -19.757  < 2e-16 ***
## refund        -0.3205240  0.0530401  -6.043 1.52e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9313 on 55235 degrees of freedom
```

```
## Multiple R-squared:  0.4913, Adjusted R-squared:  0.4903
## F-statistic: 503.3 on 106 and 55235 DF,  p-value: < 2.2e-16
```

```
library(stringr)
new_words <- c("5 stars", "4 stars", "3 stars", "2 stars", "1 star", "huge
plus", "loud", "too loud", "weird", "bummer", "lifeless", "impressed",
"disappointing", "lovely", "super", "unique", "enjoy", "enjoyed", "so much",
"definitely", "flavorful", "keep coming back", "recommended", "highly
recommended", "good service", "expensive",
"overpriced","recommend","down","ice cream machine","prefer","complain","try
again","extremely rude","convenient","unsatisfied","satisfied","best
location","best pizza","pizza","chinese","mexican","indian","best
food","worst food","wine","beer","curds","burgers","downtown","short
wait","no wait","terrible service","bad service","friendly staff","great
service","long wait","frozen","moist","dry","cold","hot","new","second",
"wait", "ecstatic","reseated", "brilliant", "time", "view","not
good","tip","love","good","service","great","one","two","three","four","five"
,"negative","happy","tasty","savory","best","hate","friendly","price")
new_X <- matrix(0, nrow(yelp), length(new_words))
colnames(new_X) <- new_words
for (i in 1:length(new_words)){
  new_X[,i] <- str_count(yelp$text, regex(new_words[i], ignore_case=T))
}
dat1 = cbind(dat,new_X)
dat1$nchar = log(dat1$nchar)
dat1$nword = log(dat1$nword)
dat1$sentiment = log(dat1$sentiment+6)
dat1$useful = log(dat1$useful+1)
dat1$funny = log(dat1$funny+1)
dat1$cool = log(dat1$cool+1)
#mymodel
model1 = lm(stars~., data=dat1)
summary(model1)
```

```
##
## Call:
## lm(formula = stars ~ ., data = dat1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.3154 -0.5448  0.0715  0.6144  4.2858
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -0.9071111  0.1267382  -7.157 8.33e-13 ***
## useful              -0.1233693  0.0077068 -16.008  < 2e-16 ***
## funny               -0.2095485  0.0113456 -18.470  < 2e-16 ***
## cool                 0.4664465  0.0103242  45.180  < 2e-16 ***
## nchar                0.5106227  0.0628787   8.121 4.73e-16 ***
## nword               -0.7195924  0.0619747 -11.611  < 2e-16 ***
```

```
## sentiment             2.2120026  0.0285968  77.352  < 2e-16 ***
## gem                   0.2577987  0.0320253   8.050 8.45e-16 ***
## incredible            0.6782776  0.0678537   9.996  < 2e-16 ***
## perfection            0.0725965  0.0391672   1.854 0.063816 .
## heaven                0.3450130  0.0453750   7.604 2.93e-14 ***
## phenomenal            0.2530341  0.0391154   6.469 9.95e-11 ***
## divine                0.1652109  0.0629537   2.624 0.008684 **
## die                   0.1317101  0.0120995  10.886  < 2e-16 ***
## deliciously           0.0487893  0.0766683   0.636 0.524539
## highly                0.2337192  0.0234748   9.956  < 2e-16 ***
## heavenly             -0.0468426  0.0828808  -0.565 0.571953
## superb                0.0459242  0.0453596   1.012 0.311329
## amazing               0.2095062  0.0104382  20.071  < 2e-16 ***
## favorites             0.0264084  0.0342728   0.771 0.440986
## delectable            0.2151437  0.0768066   2.801 0.005094 **
## perfect               0.1931393  0.0136408  14.159  < 2e-16 ***
## sourced               0.1495254  0.0589150   2.538 0.011152 *
## knowledgeable         0.1776980  0.0382407   4.647 3.38e-06 ***
## wonderfully           0.0895042  0.0678971   1.318 0.187431
## deliciousness         0.0099997  0.0728003   0.137 0.890748
## knowledgable          0.2735260  0.0813653   3.362 0.000775 ***
## fantastic             0.1877051  0.0156496  11.994  < 2e-16 ***
## adorable              0.1609949  0.0680184   2.367 0.017940 *
## wonderful             0.1447034  0.0169786   8.523  < 2e-16 ***
## fabulous              0.1251953  0.0343481   3.645 0.000268 ***
## scrumptious           0.2325097  0.0769342   3.022 0.002511 **
## hidden                0.1545066  0.0485865   3.180 0.001473 **
## notch                 0.3310243  0.0356277   9.291  < 2e-16 ***
## favorite              0.2693566  0.0117371  22.949  < 2e-16 ***
## disappoint           -0.0328001  0.0137068  -2.393 0.016715 *
## watering              0.2298732  0.0647698   3.549 0.000387 ***
## delightful            0.2141739  0.0421441   5.082 3.75e-07 ***
## yum                   0.1396027  0.0149632   9.330  < 2e-16 ***
## outstanding           0.1347801  0.0255974   5.265 1.40e-07 ***
## awesome               0.1707692  0.0135958  12.560  < 2e-16 ***
## everyday              0.2666685  0.0703795   3.789 0.000151 ***
## relaxed               0.1919059  0.0514486   3.730 0.000192 ***
## dream                 0.3734808  0.0472287   7.908 2.67e-15 ***
## amazingly             0.0077650  0.0639174   0.121 0.903307
## delicious             0.2802306  0.0084192  33.285  < 2e-16 ***
## affordable            0.1717740  0.0390106   4.403 1.07e-05 ***
## scones                0.1635816  0.0418823   3.906 9.40e-05 ***
## excellent             0.2174476  0.0115212  18.874  < 2e-16 ***
## rotating              0.1039672  0.0706456   1.472 0.141115
## paired                0.0038578  0.0517015   0.075 0.940519
## secret                0.2195571  0.0551367   3.982 6.84e-05 ***
## beautifully           0.0311828  0.0554423   0.562 0.573821
## terrace               0.2737575  0.0516388   5.301 1.15e-07 ***
## skeptical             0.3317839  0.0732496   4.529 5.92e-06 ***
## perfectly            -0.0824173  0.0235229  -3.504 0.000459 ***
```

```
## delish               0.1928137  0.0447159   4.312 1.62e-05 ***
## supposed            -0.2460117  0.0318028  -7.736 1.05e-14 ***
## mcdonald.s          -0.2329309  0.0519855  -4.481 7.46e-06 ***
## undercooked         -0.3605445  0.0522414  -6.902 5.20e-12 ***
## management          -0.4658019  0.0469206  -9.927  < 2e-16 ***
## sucks               -0.3551441  0.0612088  -5.802 6.58e-09 ***
## flag                -0.1316094  0.0530677  -2.480 0.013140 *
## messed              -0.1106739  0.0700144  -1.581 0.113946
## unpleasant          -0.3725043  0.0708653  -5.257 1.47e-07 ***
## dirty               -0.1875333  0.0268921  -6.974 3.13e-12 ***
## mediocre            -0.5288307  0.0274908 -19.237  < 2e-16 ***
## manager             -0.1627904  0.0167025  -9.746  < 2e-16 ***
## burned              -0.3918762  0.0654843  -5.984 2.19e-09 ***
## nope                -0.3318390  0.0596398  -5.564 2.65e-08 ***
## upset               -0.0922282  0.0582417  -1.584 0.113304
## zero                -0.2333827  0.0475281  -4.910 9.11e-07 ***
## stale               -0.3653603  0.0482381  -7.574 3.67e-14 ***
## screw               -0.3283262  0.0492062  -6.672 2.54e-11 ***
## subpar              -0.7340451  0.0790435  -9.287  < 2e-16 ***
## inattentive         -0.3616953  0.0800892  -4.516 6.31e-06 ***
## edible              -0.3790545  0.0617289  -6.141 8.28e-10 ***
## complained          -0.1997786  0.0675618  -2.957 0.003108 **
## rubbery             -0.3904926  0.0690701  -5.654 1.58e-08 ***
## lukewarm            -0.5412576  0.0574784  -9.417  < 2e-16 ***
## garbage             -0.1960001  0.0801091  -2.447 0.014422 *
## contact             -0.2728169  0.0515130  -5.296 1.19e-07 ***
## flavorless          -0.4559626  0.0405497 -11.245  < 2e-16 ***
## poor                -0.2968087  0.0265204 -11.192  < 2e-16 ***
## apologize            0.0743691  0.0376260   1.977 0.048099 *
## gross               -0.3913563  0.0377489 -10.367  < 2e-16 ***
## charged             -0.2226575  0.0349344  -6.374 1.86e-10 ***
## receipt             -0.2173699  0.0546160  -3.980 6.90e-05 ***
## worse               -0.1759508  0.0385139  -4.569 4.92e-06 ***
## ignored             -0.2568872  0.0576383  -4.457 8.33e-06 ***
## response            -0.3498350  0.0577349  -6.059 1.38e-09 ***
## poorly              -0.1497731  0.0659650  -2.270 0.023182 *
## nasty               -0.2482634  0.0586422  -4.234 2.30e-05 ***
## proceeded           -0.2583547  0.0676426  -3.819 0.000134 ***
## terrible            -0.3142364  0.0245904 -12.779  < 2e-16 ***
## waste               -0.4305263  0.0389267 -11.060  < 2e-16 ***
## tasteless           -0.4324381  0.0433660  -9.972  < 2e-16 ***
## inedible            -0.0787750  0.0838247  -0.940 0.347345
## downhill            -0.6348410  0.0802717  -7.909 2.65e-15 ***
## awful               -0.3068575  0.0302987 -10.128  < 2e-16 ***
## rude                -0.4217239  0.0229860 -18.347  < 2e-16 ***
## horrible            -0.3810758  0.0291338 -13.080  < 2e-16 ***
## refused             -0.2300977  0.0725042  -3.174 0.001507 **
## apology             -0.2733953  0.0569892  -4.797 1.61e-06 ***
## disgusting          -0.3196073  0.0439865  -7.266 3.75e-13 ***
## worst               -0.4493704  0.0252204 -17.818  < 2e-16 ***
```

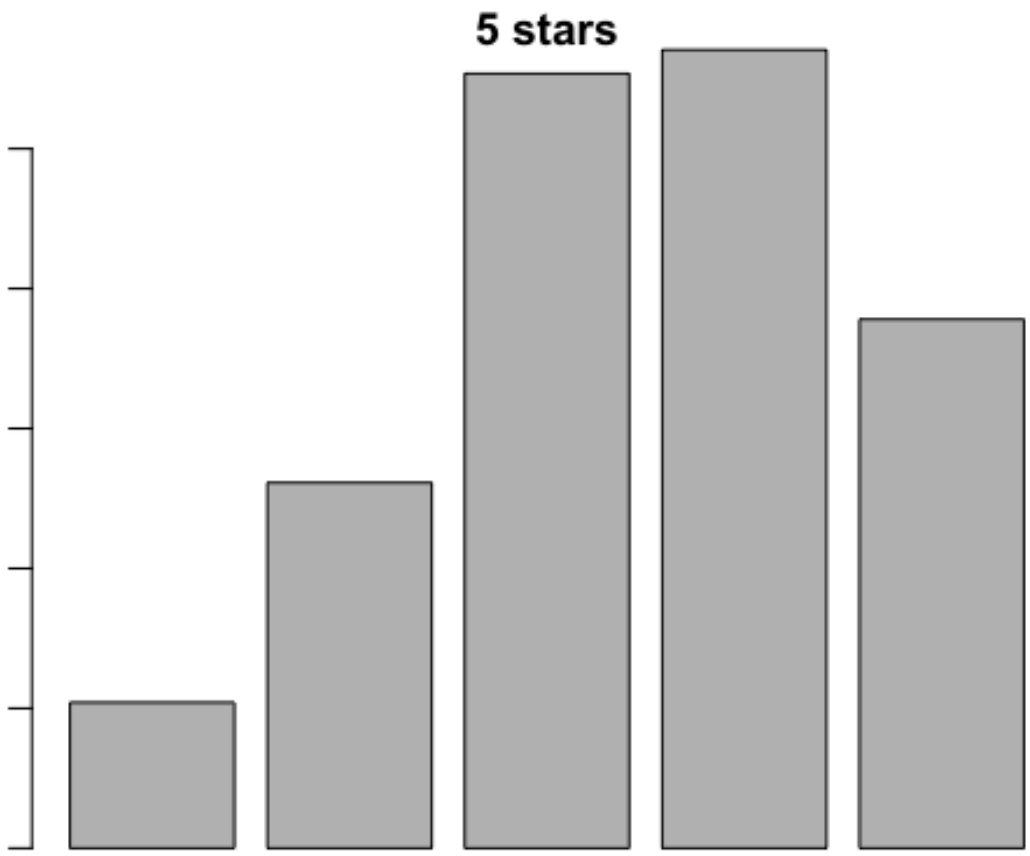
```
## refund                -0.3094324  0.0507241  -6.100 1.07e-09 ***
## `5 stars`              0.1418044  0.0244052   5.810 6.27e-09 ***
## `4 stars`             -0.0080864  0.0357183  -0.226 0.820896
## `3 stars`             -0.0645679  0.0423722  -1.524 0.127558
## `2 stars`             -0.1838395  0.0511407  -3.595 0.000325 ***
## `1 star`              -0.2858103  0.0554315  -5.156 2.53e-07 ***
## `huge plus`            0.1623726  0.1314726   1.235 0.216825
## loud                  -0.0889242  0.0229195  -3.880 0.000105 ***
## `too loud`             0.0830699  0.0760168   1.093 0.274493
## weird                 -0.0202021  0.0284197  -0.711 0.477180
## bummer                -0.0381402  0.0670382  -0.569 0.569404
## lifeless              -0.2240435  0.2309264  -0.970 0.331954
## impressed             -0.1435393  0.0205638  -6.980 2.98e-12 ***
## disappointing         -0.3429819  0.0286212 -11.984  < 2e-16 ***
## lovely                -0.0421107  0.0295302  -1.426 0.153868
## super                 -0.0051655  0.0130487  -0.396 0.692208
## unique                 0.1064225  0.0214222   4.968 6.79e-07 ***
## enjoy                  0.0890829  0.0141953   6.276 3.51e-10 ***
## enjoyed               -0.0007555  0.0209062  -0.036 0.971173
## `so much`             -0.0131795  0.0241385  -0.546 0.585073
## definitely             0.1167034  0.0099707  11.705  < 2e-16 ***
## flavorful              0.0480212  0.0189203   2.538 0.011149 *
## `keep coming back`     0.1182157  0.0855700   1.382 0.167128
## recommended           -0.0352016  0.0355352  -0.991 0.321880
## `highly recommended`   0.0657032  0.0654074   1.005 0.315131
## `good service`         0.0474309  0.0365413   1.298 0.194291
## expensive             -0.1063425  0.0207835  -5.117 3.12e-07 ***
## overpriced            -0.5392164  0.0315826 -17.073  < 2e-16 ***
## recommend              0.0637536  0.0131808   4.837 1.32e-06 ***
## down                   0.0094163  0.0110512   0.852 0.394182
## `ice cream machine`    0.2823238  0.3367806   0.838 0.401865
## prefer                 0.0363887  0.0248212   1.466 0.142646
## complain               0.1316657  0.0204568   6.436 1.23e-10 ***
## `try again`           -0.3377835  0.0720279  -4.690 2.74e-06 ***
## `extremely rude`      -0.1940639  0.1150135  -1.687 0.091549 .
## convenient             0.0195723  0.0386438   0.506 0.612523
## unsatisfied           -0.2953746  0.1340559  -2.203 0.027573 *
## satisfied              0.1844884  0.0351331   5.251 1.52e-07 ***
## `best location`       -0.1854331  0.2575379  -0.720 0.471514
## `best pizza`           0.3190210  0.0494948   6.446 1.16e-10 ***
## pizza                 -0.0060171  0.0048580  -1.239 0.215506
## chinese                0.0393639  0.0130042   3.027 0.002471 **
## mexican                0.0036757  0.0142341   0.258 0.796227
## indian                 0.0484445  0.0164999   2.936 0.003326 **
## `best food`            0.1251960  0.0719248   1.741 0.081750 .
## `worst food`          -0.3420074  0.1551330  -2.205 0.027486 *
## wine                   0.0042818  0.0121713   0.352 0.724993
## beer                   0.0095013  0.0061991   1.533 0.125354
## curds                 -0.0128691  0.0119227  -1.079 0.280425
## burgers               -0.0241596  0.0145882  -1.656 0.097705 .
```

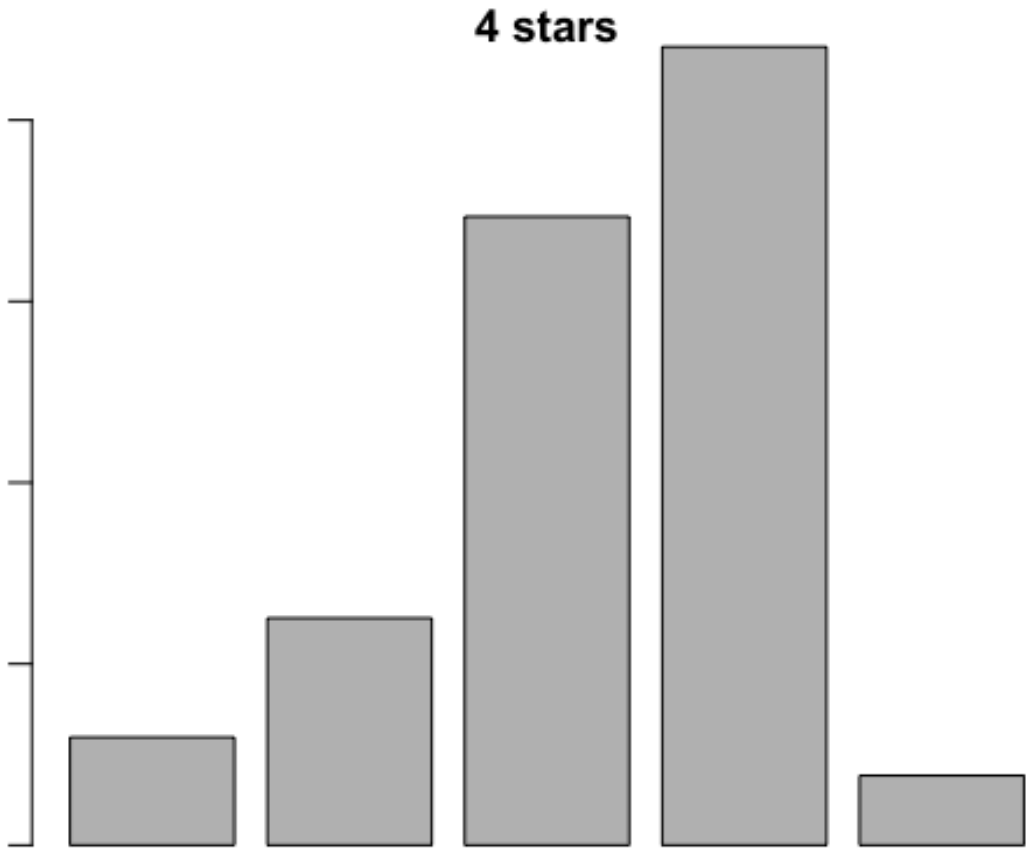
```
## downtown               0.0555446  0.0262509   2.116 0.034357 *
## `short wait`           0.1045732  0.1062926   0.984 0.325206
## `no wait`              0.1368847  0.0673754   2.032 0.042192 *
## `terrible service`    -0.0291411  0.0816309  -0.357 0.721103
## `bad service`         -0.1545718  0.0639458  -2.417 0.015642 *
## `friendly staff`       0.0619793  0.0336917   1.840 0.065832 .
## `great service`        0.1879571  0.0290293   6.475 9.58e-11 ***
## `long wait`           -0.0007588  0.0440973  -0.017 0.986271
## frozen                -0.2017007  0.0280339  -7.195 6.33e-13 ***
## moist                  0.0719101  0.0374915   1.918 0.055111 .
## dry                   -0.2434161  0.0179732 -13.543  < 2e-16 ***
## cold                  -0.1503577  0.0141643 -10.615  < 2e-16 ***
## hot                    0.0131797  0.0085933   1.534 0.125103
## new                    0.0338445  0.0101389   3.338 0.000844 ***
## second                -0.0544522  0.0175983  -3.094 0.001975 **
## wait                  -0.0523051  0.0052261 -10.008  < 2e-16 ***
## ecstatic               0.0399206  0.1784309   0.224 0.822967
## reseated              -0.4980085  0.8947421  -0.557 0.577807
## brilliant              0.1284053  0.1026261   1.251 0.210869
## time                   0.0178175  0.0049546   3.596 0.000323 ***
## view                  -0.0224806  0.0098726  -2.277 0.022786 *
## `not good`            -0.6073052  0.0383935 -15.818  < 2e-16 ***
## tip                   -0.0057649  0.0145773  -0.395 0.692496
## love                   0.1226647  0.0069897  17.549  < 2e-16 ***
## good                  -0.0159607  0.0042715  -3.737 0.000187 ***
## service               -0.0774511  0.0070218 -11.030  < 2e-16 ***
## great                  0.1015140  0.0050336  20.167  < 2e-16 ***
## one                   -0.0096945  0.0039340  -2.464 0.013731 *
## two                   -0.0184990  0.0094470  -1.958 0.050214 .
## three                 -0.0569836  0.0154369  -3.691 0.000223 ***
## four                  -0.0334053  0.0214691  -1.556 0.119720
## five                   0.0828296  0.0224584   3.688 0.000226 ***
## negative               0.1666267  0.0394663   4.222 2.43e-05 ***
## happy                  0.0744211  0.0133688   5.567 2.61e-08 ***
## tasty                  0.0969595  0.0122965   7.885 3.20e-15 ***
## savory                 0.0366963  0.0314907   1.165 0.243900
## best                   0.2132035  0.0083691  25.475  < 2e-16 ***
## hate                   0.0633675  0.0221190   2.865 0.004174 **
## friendly               0.1771663  0.0106046  16.707  < 2e-16 ***
## price                 -0.0197830  0.0082909  -2.386 0.017030 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8892 on 55146 degrees of freedom
## Multiple R-squared:  0.537,  Adjusted R-squared:  0.5354
## F-statistic:   328 on 195 and 55146 DF,  p-value: < 2.2e-16

new_X2 <- matrix(0, nrow(yelp_out), length(new_words))
colnames(new_X2) <- new_words
for (i in 1:length(new_words)){
```

```
  new_X2[,i] <- str_count(yelp_out$text, regex(new_words[i], ignore_case=T))
# ignore the upper/lower case in the text
}
dat2 = cbind(yelp_out[,-c(1,2:4,8,12)],new_X2)
dat2$nchar = log(dat2$nchar)
dat2$nword = log(dat2$nword)
dat2$sentiment = log(dat2$sentiment+6)
dat2$useful = log(dat2$useful+1)
dat2$funny = log(dat2$funny+1)
dat2$cool = log(dat2$cool+1)

library(boot)

##
## Attaching package: 'boot'

## The following object is masked from 'yelp_val':
##
##     city

## The following object is masked from 'yelp':
##
##     city

## The following object is masked from 'yelp_test':
##
##     city

library(glmnet)

## Loading required package: Matrix

## Loaded glmnet 4.0-2

#ridge and lasso
Xmat = as.matrix(dat1[,-c(1)])
Ymat = as.matrix(dat1[,1])
Xmat_test = as.matrix(dat2)
ridge.model = cv.glmnet(Xmat,Ymat, alpha=1, nfolds = 5, type.measure = "mse",
family="gaussian")
p = predict(ridge.model, s=ridge.model$lambda.1se, newx = Xmat_test)
#Checking Ridge and Lasso
penalty_regression = function(train, test){
  # new model based on all the new variable transformation
  dat = train[,-c(1,3:5, 9,13)]

  # let's try ridge regression
  xMat = as.matrix(dat[,c(-1)])
  yMat = dat[,1]
  dat.test = test[,-c(1,3:5, 9,13)]
  xMat.test = as.matrix(dat.test[,c(-1)])
```

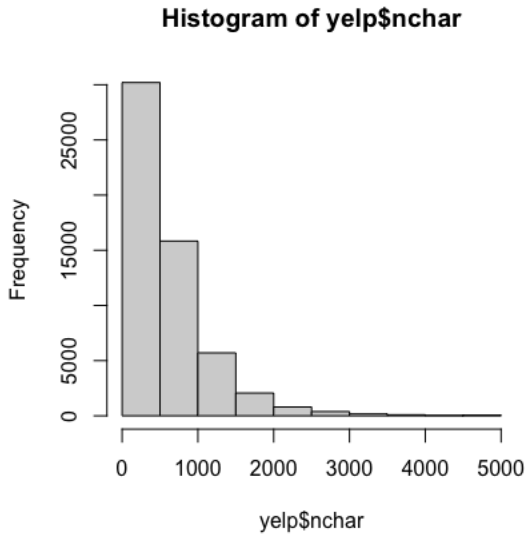
```
  model1_ridge = cv.glmnet(xMat, yMat, alpha = 0,nfold =5,type.measure="mse",
family="gaussian")
  p2 = predict(model1_ridge, s = model1_ridge$lambda.min, newx = xMat.test)
  r2 = rmse(p2)
  cat("RMSE for ridge with lambda min", round(r2,6),"\n")

  p2a = predict(model1_ridge, s = model1_ridge$lambda.1se, newx = xMat.test)
  r2a = rmse(p2a)
  cat("RMSE for ridge with lambda 1se ", round(r2a,6),"\n")

  # try lasso regression
  model2_lasso = cv.glmnet(xMat, yMat, alpha=1, nfold=5, type.measure="mse",
family="gaussian")
  p3 = predict(model2_lasso, s = model2_lasso$lambda.1se, newx = xMat.test)
  r3 = rmse(p3)
  cat("RMSE for Lasso with lambda 1se", round(r3,6),"\n")

  p3a = predict(model2_lasso, s = model2_lasso$lambda.min, newx = xMat.test)
  r3a = rmse(p3a)
  cat("RMSE for Lasso with lambda min", round(r3a,6),"\n")

  # try elastic net: i.e. combining both lasso and ridge
  model3_elastic = cv.glmnet(xMat, yMat, alpha = 0.5, nfold =5,
type.measure="mse", family = "gaussian")
  p4 = predict(model3_elastic, s=model3_elastic$lambda.min, newx = xMat.test)
  r4 = rmse(p4)
  cat("RMSE for Elastic Net with lambda min", round(r4, 6),"\n")
  p4a = predict(model3_elastic, s=model3_elastic$lambda.1se, newx =
xMat.test)
  r4a = rmse(p4a)
  cat("RMSE for Elastic Net with lambda 1se", round(r4a, 6) ,"\n")
}
rmse = function(predicted){
  n = length(y.test)
  return (sqrt(sum(((y.test - predicted)^2)  / n)))
}
# split the training data into 0.8 and 0.2 for training and testing
n = dim(yelp)[1]
train_row = sample(1:n, round(n*0.8))
train = yelp[train_row, ]
test = yelp[-train_row, ]


x.train = train[,-2]
y.train = train[,2]

x.test = test[,-2];
```

```
y.test = test[,2]
penalty_regression(train = train, test = test)

## RMSE for ridge with lambda min 0.931788
## RMSE for ridge with lambda 1se  0.932823
## RMSE for Lasso with lambda 1se 0.93232
## RMSE for Lasso with lambda min 0.930837
## RMSE for Elastic Net with lambda min 0.930837
## RMSE for Elastic Net with lambda 1se 0.93401

star_out = data.frame(Id=yelp_out$Id, Expected=predict(model1, newdata =
dat2))
res = star_out
len = dim(res)[1]
for(i in 1:len){
  if (res[i,2] > 5){
    res[i,2] = 5
  } else if (res[i,2] < 1 ){
    res[i,2] = 1
  }
}



write.csv(res, file="FinalResults", row.names=F)
# plotting the word count against star rating
plotWordStar <- function(stars, wordcount, wordname){
  meancount <- rep(0,5)
  names(meancount) <- 1:5
  for (i in 1:5)    meancount[i] <- mean(wordcount[stars==i])
  barplot(meancount, main=wordname, xlab="Stars", ylab="Average word count")
}
par(mar=c(1,1,1,1))
for (i in 1:2){
  plotWordStar(yelp$stars, new_X[,i], colnames(new_X)[i])
}
```

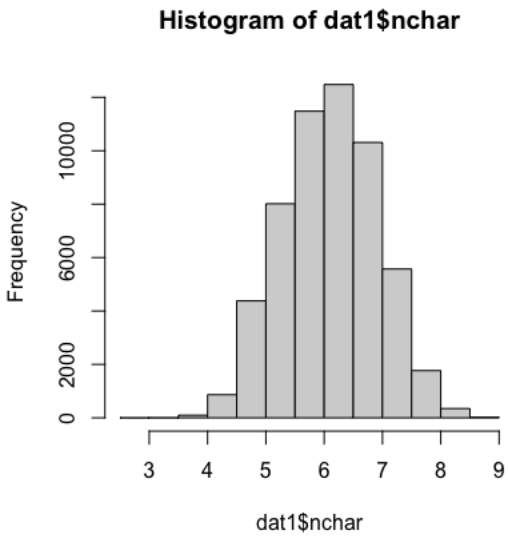

5 stars




```
#plotting word/phrase vs rating
hist(yelp$nchar,breaks = 10)
```




```
hist(dat1$nchar, breaks=10)
```


Histogram of dat1$nchar
Frequency
0
2000
6000
10000
3
4
5
6
7
8
9
dat1$nchar