Final Capstone Project

Moira Lennox

November 22nd, 2015

Title & Introduction

The data set for this capstone project was supplied by Yelp and is a subset of data for the years 2004 – 2015 and contains a number of counties. The yelp data supplied was made up of 5 files namely: reviews, business, user, tips, checkin. My work will focus on only two files namely the review and business files. My objective for this report is to analyze the "yelp" data set by exploring and researching the relationship between sets of variables and the business attribute that identifies as "dogallowed" to see if I could answer the question(s) below.

Dogs can be a major part of the family. We love taking our dog everywhere with us, however she is not always welcome. So here is my question: Can the dataset tell me if there was a trend in welcoming pets into various businesses? I plan to narrow my focus down to two major categories hotels and restaurants since these are the areas I care about the most. I am going to look at the review data from 2008-2014. I will compare business that allow dogs to business that do not and then try to correlate those counts to dog references I find from data mining the review text for those references. I will look at the country trends to see if locations makes a difference.

Data processing and Transformation

I read in the json files for Business and Review, flatten them created a few new columns namely: main category, country and review year. I aggregated the review data up to the business level. The final step was mergeing the dataset into a single dataset.

```r
str(dt_merge_pet)
```

```
## Classes 'data.table' and 'data.frame':   10125 obs. of  17 variables:
##  $ business_id    : chr  "--pOlFxITWnhzc7SHSIP0A" "--pOlFxITWnhzc7SHSIP0A" "-3Qu8aYgOleRw-OThNPovA"
##  $ categories     : chr  "AMERICAN (NEW), RESTAURANTS" "AMERICAN (NEW), RESTAURANTS" "ITALIAN, PIZZA
##  $ city           : chr  "Charlotte" "Charlotte" "Karlsruhe" "Karlsruhe" ...
##  $ state          : chr  "NC" "NC" "BW" "BW" ...
##  $ bsnreviewcount : int  109 109 7 7 16 16 16 16 16 487 ...
##  $ bsnstars       : num  4 4 4 4 4 4 4 4 4 4.5 ...
##  $ dogsallowed    : Factor w/ 2 levels "FALSE","TRUE": 1 1 1 1 1 1 1 1 1 2 ...
##  $ maincat        : Factor w/ 2 levels "HOTEL","RESTAURANT": 2 2 2 2 2 2 2 2 2 2 ...
##  $ country        : Factor w/ 4 levels "CAN","DEU","UK",..: 4 4 2 2 1 1 1 1 1 4 ...
##  $ usercount      : int  60 45 1 1 1 3 1 5 5 3 ...
##  $ revcount       : int  60 45 1 1 1 3 1 5 5 3 ...
##  $ votefuncount   : int  35 7 0 0 0 0 0 3 3 10 ...
##  $ voteusefulcount: int  96 20 0 0 0 1 0 7 4 18 ...
##  $ votecoolcount  : int  50 13 0 0 0 0 0 3 2 16 ...
##  $ starsavg       : num  3.93 4.2 5 3 5 ...
##  $ petcount       : int  14 8 0 1 0 0 0 1 3 1 ...
##  $ revyear        : Factor w/ 7 levels "2008","2009",..: 6 7 5 6 3 4 5 6 7 1 ...
##  - attr(*, "sorted")= chr "business_id"
##  - attr(*, ".internal.selfref")=<externalptr>
```

Exploratory Data Analysis

A quick histogram revealed that the pet review count data are heavily Skewed. This means we will need to transform the data when we apply the various models.

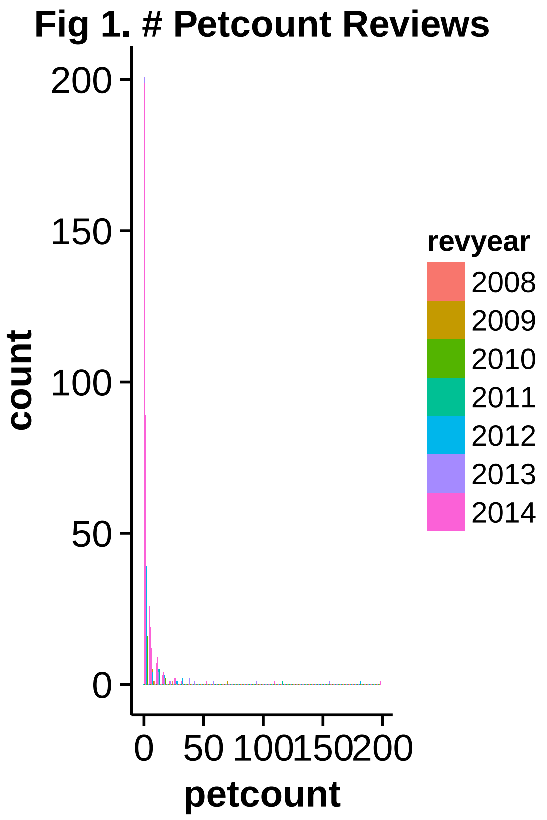


Fig 1. # Petcount Reviews

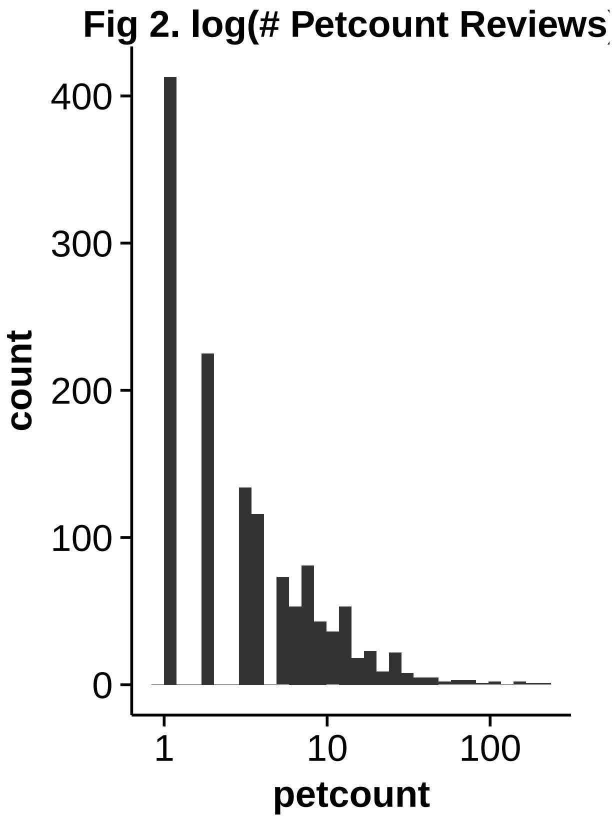


Fig 2. log(# Petcount Reviews)

I quick look at the summary data for the petcount shows and consistent increase year over year. So we could make the initial assumption that pets are increasing being mentioned for pet friendly businesses and maybe this is implying that there might be an increase in pet friendly places.

```
##   revyear   N     mean        sd        se
## 1    2008 160 0.943750  2.775107 0.2193915
## 2    2009 205 1.931707  5.749273 0.4015466
## 3    2010 287 2.439024  6.446304 0.3805133
## 4    2011 343 3.093294  8.227222 0.4442282
## 5    2012 416 3.637019 11.277250 0.5529127
## 6    2013 507 3.859961 11.848939 0.5262298
## 7    2014 539 4.294991 11.717722 0.5047180
```

I did a boxplot and scatterplot matrix to see if there might be linear correlation between multiple variables. I used this is will help me pinpointing specific variables that might have similar correlations to my dogsalloed petcount data. I looked at the correlation and it is not very strong but revcount shows the best correlation with voteusefulcount a close second.

Methods and Data

I tested three differnt models from the GLM family and show my outcomes. As a result of my initial data analysis I have selected to use a Zero-inflated negative binomial model for modeling the data. This model works well for count variables with excessive zeros. There is 45% response of zero for petcount variable.

```
# Look at three types of models
fit1 <- glm(petcount~revcount+revyear+country+votefuncount+voteusefulcount+
                +starsavg, data=dt_merge_pet2,family=poisson(link=log))

fit2 <- zeroinfl(petcount~revcount+revyear+votefuncount+voteusefulcount+
                +starsavg+country|revcount+revyear+votefuncount+voteusefulcount+
                +starsavg+country,data=dt_merge_pet2)

fit3 <- zeroinfl(petcount~revcount+revyear+votefuncount+voteusefulcount+
                +starsavg+country|revcount+revyear+votefuncount+voteusefulcount+
                +starsavg+country,data = dt_merge_pet2, dist = "negbin", EM = TRUE)
```

The Vuong test compares the three models and we can see that the test statistic is significant, for model three which indicates that the zero-inflated model is best model for the job.

All of the predictors in both the count and inflation portions of the model are statistically significant. This model fits the data significantly better than the null model, i.e., the intercept-only mode. To show that this is the case, we can compare with the current model to a null model without predictors using chi-squared test on the difference of log likelihoods, see below.

```
mnull <- update(fit3, . ~ 1)
pchisq(2 * (logLik(fit3) - logLik(mnull)), df = 6, lower.tail = FALSE)
```

```
## 'log Lik.' 7.756167e-296 (df=29)
```

Review the results from model three which was selected.

```
##
## Call:
## zeroinfl(formula = petcount ~ revcount + revyear + votefuncount +
##     voteusefulcount + +starsavg + country | revcount + revyear +
##     votefuncount + voteusefulcount + +starsavg + country, data = dt_merge_pet2,
##     dist = "negbin", EM = TRUE)
##
## Pearson residuals:
##      Min       1Q   Median       3Q      Max
## -0.91167 -0.52265 -0.32047  0.07697 29.23391
##
## Count model coefficients (negbin with log link):
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -0.768113   0.370249  -2.075 0.038025 *
## revcount         0.008150   0.001472   5.537 3.08e-08 ***
## revyear2009      0.182975   0.207974   0.880 0.378969
## revyear2010      0.481230   0.192619   2.498 0.012477 *
## revyear2011      0.498801   0.188982   2.639 0.008305 **
## revyear2012      0.644507   0.185346   3.477 0.000506 ***
## revyear2013      0.601601   0.182844   3.290 0.001001 **
## revyear2014      0.529825   0.184897   2.866 0.004163 **
## votefuncount    -0.005017   0.003210  -1.563 0.118051
## voteusefulcount  0.011472   0.001952   5.877 4.17e-09 ***
## starsavg         0.125727   0.050102   2.509 0.012093 *
## countryDEU      -1.863254   0.745514  -2.499 0.012444 *
## countryUK        0.812524   0.326560   2.488 0.012842 *
## countryUSA       0.415808   0.253822   1.638 0.101383
## Log(theta)      -0.131506   0.051112  -2.573 0.010085 *
##
## Zero-inflation model coefficients (binomial with logit link):
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)      0.95626    0.77821   1.229   0.2192
## revcount        -0.27938    0.05159  -5.415 6.13e-08 ***
## revyear2009     -0.19743    0.48708  -0.405   0.6852
## revyear2010     -0.13751    0.47172  -0.291   0.7707
## revyear2011      0.24688    0.47327   0.522   0.6019
## revyear2012      0.17146    0.45829   0.374   0.7083
## revyear2013     -0.22286    0.46274  -0.482   0.6301
## revyear2014     -0.29526    0.47815  -0.618   0.5369
## votefuncount    -0.03513    0.07085  -0.496   0.6200
## voteusefulcount -0.06145    0.04023  -1.528   0.1266
## starsavg         0.23491    0.11054   2.125   0.0336 *
## countryDEU      -1.52015    1.50354  -1.011   0.3120
## countryUK        0.38820    0.54613   0.711   0.4772
## countryUSA      -0.18530    0.46179  -0.401   0.6882
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Theta = 0.8768
## Number of iterations in BFGS optimization: 1
## Log-likelihood: -4425 on 29 Df
```

```
##                 Count_Model Zero_Model
## Intercept         0.4638875  2.6019411
## revcount          1.0081829  0.7562545
## revyear2009       1.2007848  0.8208335
## revyear2010       1.6180627  0.8715293
## revyear2011       1.6467458  1.2800290
## revyear2012       1.9050475  1.1870340
## revyear2013       1.8250382  0.8002251
## revyear2014       1.6986351  0.7443374
## votefuncount      0.9949953  0.9654765
## voteusefulcount   1.0115381  0.9404010
## starsavg          1.1339723  1.2647957
## countryDEU        0.1551668  0.2186788
## countryUK         2.2535882  1.4743230
## countryUSA        1.5155954  0.8308566
```

Results

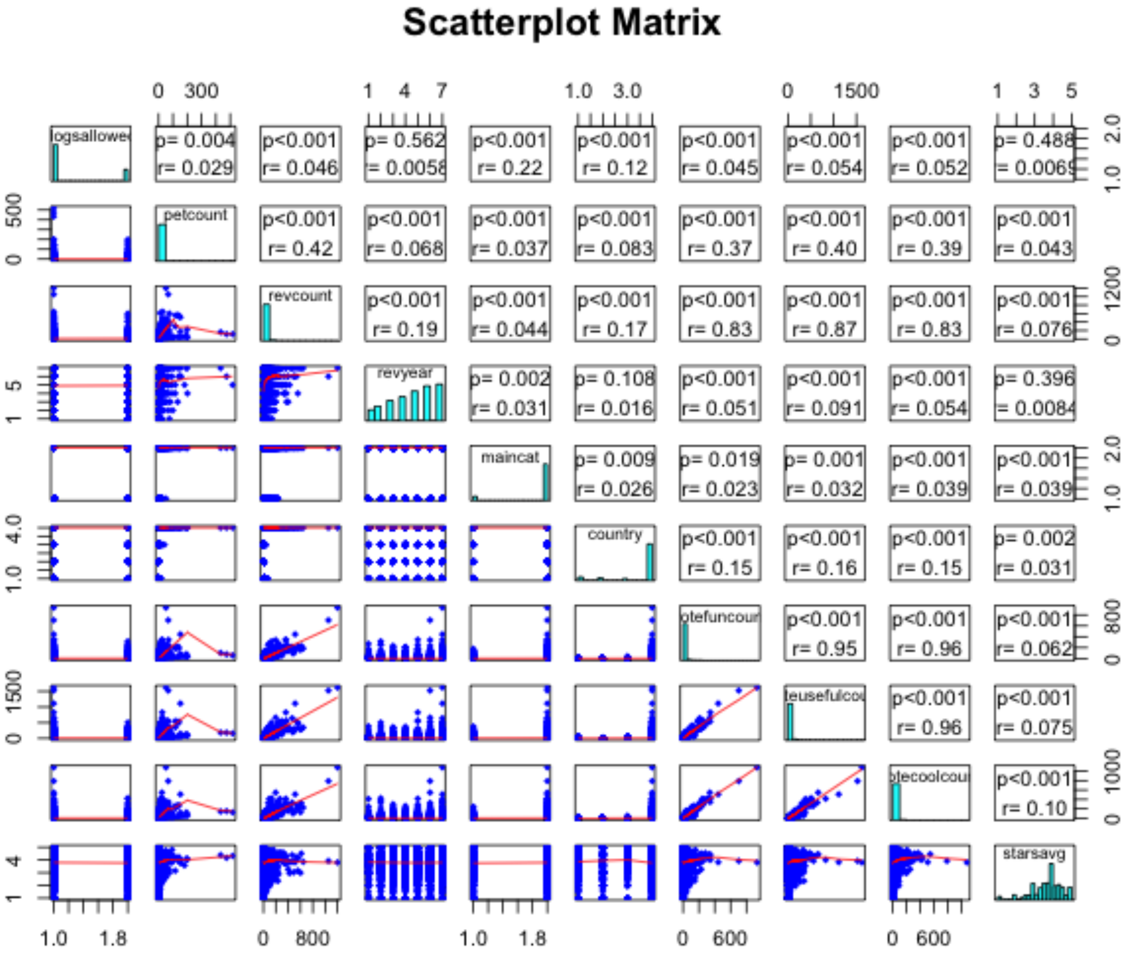
The average pet count is 4.6. One unit increase in revcount increases the average pet count increase by 1.01 times. The revyear(s) show a general increase year over year. One unit increase in UK increases the average pet count increase by 2.25 times, followed by USA by 1.51

Discussion & Conclusion

I took a very simplistic approach and extracted words like “pet”and “dog” to create a new variable from the review text data so that it could be explored. That said I do believe having better designed features will almost always guarantee better result. Even thou there was not a very strong correlation between the petcount and the other predictors there seems to be some indication that there is an increase in commentaty around pets and dogs which might indicate that there might be an increase on dog friendly places. As for country location the UK seemed to lead all the other locations for show an increase in the correlation.

Appendix

Scatterplot Matrix used for exploratory data analysis.



The Vuong test compares the three models to show which is the best model for the job.

```
# Model 3 is the best model
vuong(fit1, fit2)
vuong(fit2, fit3)
vuong(fit1, fit3)
```

```
# Correlations
cor1 <- with(dt_merge_pet2, cor.test(petcount, revcount))
cor2 <- with(dt_merge_pet2, cor.test(petcount, voteusefulcount))
```