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 countries. The Yelp data supplied was made up of 5 files: reviews, business, user, tips, checkin. My work will focus on two files, the review and business files. My objective is to analyze the data by exploring and researching the relationship between sets of variables and the business attribute that identifies as "dogsallowed" 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 data show there is an increase in welcoming pets into various businesses? I narrowed my focus to two major categories, hotels and restaurants, since these are the areas I most care about. I looked at the review years from 2008-2014, compared businesses that allow dogs to businesses that do not and then correlated those counts to dog references I found from data mining the review text. I looked at the country trends to see if location makes a difference.

Data Processing and Transformation

I read in the business and review json files, flattened them, and created a few new columns: main category, country and review year. I aggregated the review data up to the business level. The final step was merging the data set into a single data set for analysis.

```r
str(dt_merge_pet)
```

```
## Classes 'data.table' and 'data.frame':   10125 obs. of  17 variables:
##  $ business_id    : chr  "--pOlFxITWnhzc7SHSIPOA" "--pOlFxITWnhzc7SHSIPOA" "-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 the “petcount”" data is heavily skewed. This inferred I needed to transform the data to test various models.

Fig 1. # Petcount Reviews

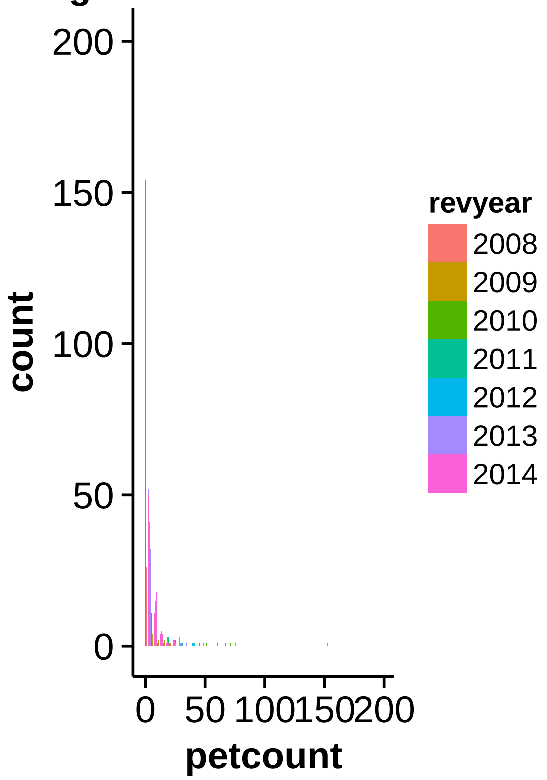
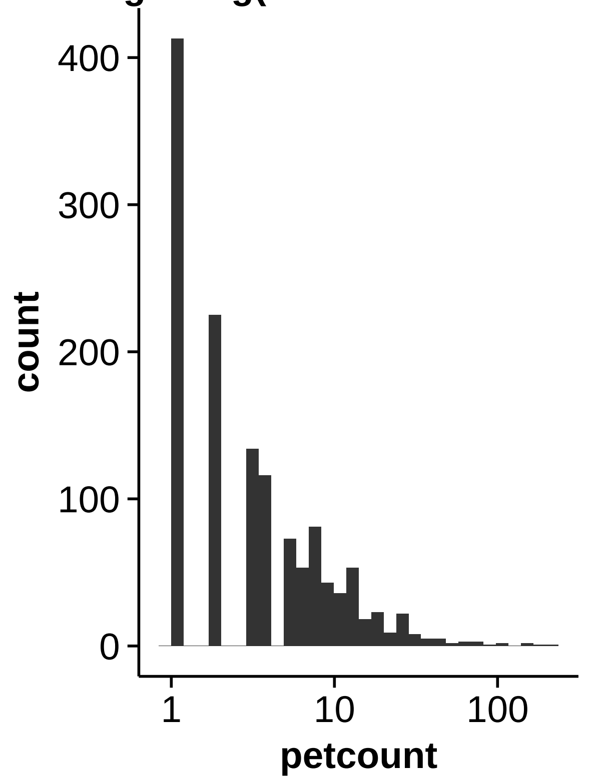


Fig 2. log(# Petcount Reviews)



A quick look at the summary data for “petcount”" showed a consistent increase year over year. An initial assumption could be made that the mention of pets increased for pet-friendly businesses and maybe this is implying 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
```

A box plot and scatter plot matrix were done to see if there was a linear correlation between multiple variables. I used this to help me pinpoint specific variables that might have similar correlations to my dogsallowed petcount data. The correlations are not very strong, but “revcount” shows the best correlation while “voteusefulcount” is a close second.

Methods and Data

I ran and tested three models from the GLM family. The Vuong test compared the three models and found the test statistic significant for model three which indicates the Zero-inflated negative binomial model is the best model for the job. This model was also chosen because it works well for count variables with excessive zeros and petcount variable reflected this with a 45% zero count.

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

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

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

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 demonstrate this, we can compare the current model to a null model without predictors using the 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.' 2.703384e-295 (df=27)
```

Below are the results for final model.

```
##
## Call:
## zeroinfl(formula = petcount ~ revcount + revyear + country + voteusefulcount +
##     starsavg | revcount + revyear + country + voteusefulcount +
##     starsavg, data = dt_merge_pet2, dist = "negbin", EM = TRUE)
##
## Pearson residuals:
##      Min       1Q   Median       3Q      Max
## -0.91402 -0.52276 -0.32211  0.08093 27.84684
##
## Count model coefficients (negbin with log link):
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -0.788999   0.370777  -2.128 0.033340 *
## revcount         0.008250   0.001485   5.555 2.78e-08 ***
## revyear2009      0.170656   0.209016   0.816 0.414228
## revyear2010      0.506014   0.192604   2.627 0.008608 **
## revyear2011      0.540951   0.187564   2.884 0.003925 **
## revyear2012      0.684963   0.183830   3.726 0.000194 ***
## revyear2013      0.633806   0.182299   3.477 0.000508 ***
## revyear2014      0.564143   0.184072   3.065 0.002178 **
## countryDEU      -1.889546   0.743668  -2.541 0.011058 *
## countryUK        0.778208   0.325642   2.390 0.016859 *
## countryUSA       0.410526   0.253558   1.619 0.105434
## voteusefulcount  0.009142   0.001228   7.446 9.63e-14 ***
## starsavg         0.126024   0.050196   2.511 0.012051 *
## Log(theta)      -0.133032   0.051128  -2.602 0.009270 **
##
## Zero-inflation model coefficients (binomial with logit link):
```

```
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)       0.88971    0.77478   1.148   0.2508
## revcount         -0.27879    0.05164  -5.399 6.72e-08 ***
## revyear2009      -0.21835    0.48952  -0.446   0.6556
## revyear2010      -0.09840    0.47030  -0.209   0.8343
## revyear2011       0.30019    0.47112   0.637   0.5240
## revyear2012       0.21384    0.45745   0.467   0.6402
## revyear2013      -0.16863    0.46061  -0.366   0.7143
## revyear2014      -0.24727    0.47658  -0.519   0.6039
## countryDEU       -1.55439    1.53183  -1.015   0.3102
## countryUK         0.36154    0.54630   0.662   0.5081
## countryUSA       -0.17856    0.46177  -0.387   0.6990
## voteusefulcount -0.07203    0.03535  -2.037   0.0416 *
## starsavg          0.24002    0.11004   2.181   0.0292 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Theta = 0.8754
## Number of iterations in BFGS optimization: 1
## Log-likelihood: -4426 on 27 Df
```

```
##                  Count_Model Zero_Model
## Intercept          0.4542993  2.4344335
## revcount           1.0082838  0.7566983
## revyear2009        1.1860830  0.8038433
## revyear2010        1.6586672  0.9062828
## revyear2011        1.7176402  1.3501169
## revyear2012        1.9836988  1.2384252
## revyear2013        1.8847697  0.8448232
## revyear2014        1.7579403  0.7809307
## countryDEU         0.1511404  0.2113188
## countryUK          2.1775662  1.4355353
## countryUSA         1.5076102  0.8364774
## voteusefulcount    1.0091842  0.9305021
## starsavg           1.1343097  1.2712746
```

Results

The average petcount is 4.6 and a one unit increase in revcount increased the average petcount by 1.01 times. The revyear(s) show a general increase year over year. A one unit increase in the UK increased the average petcount by 2.25 times, followed by the USA by 1.51 times.

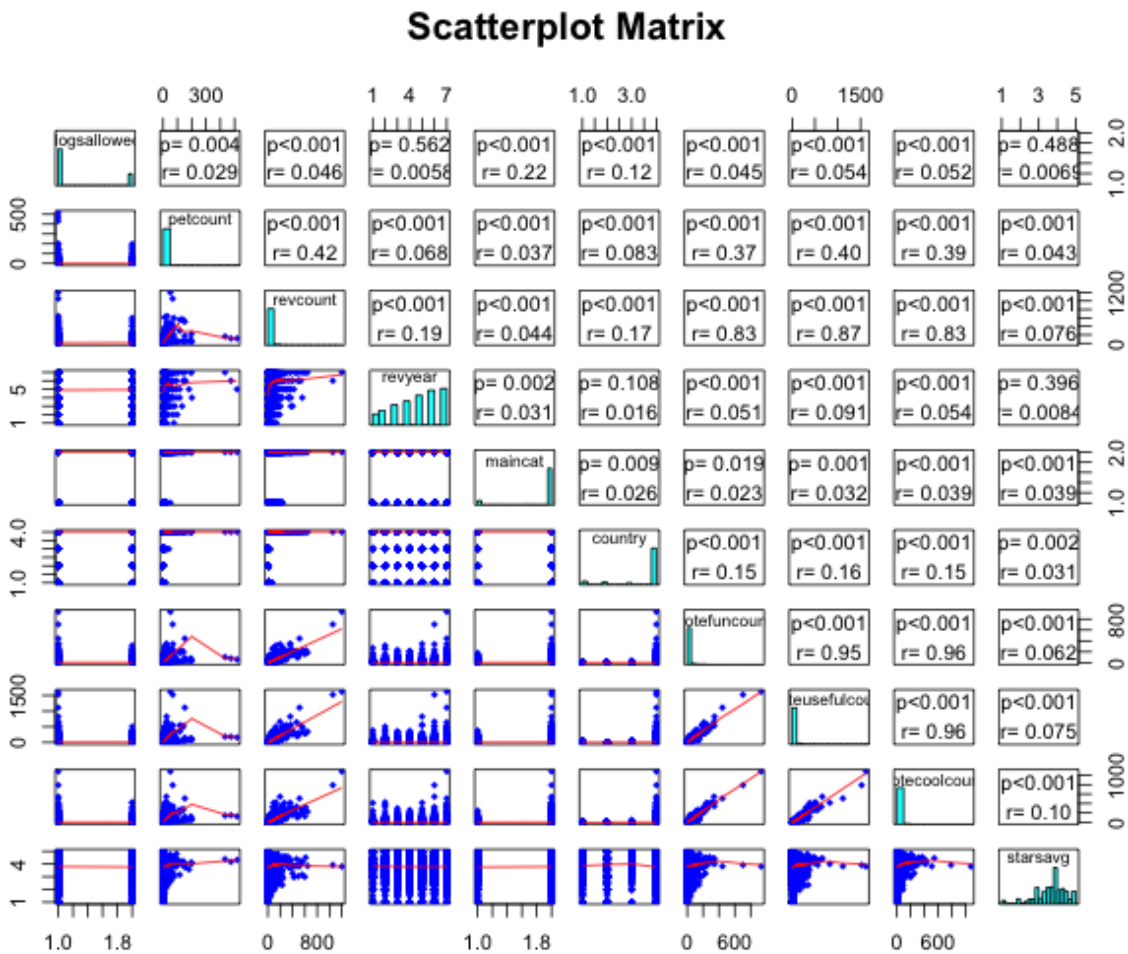
Discussion & Conclusion

I took a very simplistic approach and extracted words like "pet" and "dog" to create a new variable from the review text data so it could be researched. That said I do believe having better designed data elements would almost always guarantee better results. 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 commentary around pets in hotel and restaurant businesses. This could lead us to believe that this indicates an increase in pet-friendly places.

For country location, the UK showed a stronger correlation than the USA and led all other countries.

Appendix

Scatter plot 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))
```