Wisdom of Crowds at OpenIDEO

Juho Salminen

22 Apr 2015

# Munging, merging and data transformations

# Summary of the dataset

## E-waste challenge data consists of 106 observations of 14 variables.  
## Unemployment challenge data consists of 149 observations of 14 variables.  
## Business celebration challenge data consists of 95 observations of 14 variables.

Names of variables and a sample of data:

## [1] "Order" "Concept" "Views"   
## [4] "Comments" "Applause" "Shortlist"   
## [7] "Shortlist.views" "Shortlist.comments" "Shortlist.applause"  
## [10] "Winner" "ViewsStd" "CommentsStd"   
## [13] "ApplauseStd" "OrderStd" "Challenge"

## Order Concept Views Comments Applause Shortlist Shortlist.views  
## 1 1 concept\_73 596 19 11 Rejected NA  
## 2 2 concept\_60 696 26 25 Rejected NA  
## 3 3 concept\_58 225 3 7 Rejected NA  
## 4 4 concept\_35 394 28 36 Shortlisted 581  
## 5 5 concept\_86 309 15 15 Shortlisted 471  
## 6 6 concept\_106 242 15 11 Rejected NA  
## Shortlist.comments Shortlist.applause Winner ViewsStd CommentsStd  
## 1 NA NA 0 2.00178813 1.5837195  
## 2 NA NA 0 2.55338828 2.5223137  
## 3 NA NA 0 -0.04464839 -0.5616386  
## 4 33 45 0 0.88755585 2.7904834  
## 5 17 21 1 0.41869573 1.0473800  
## 6 NA NA 0 0.04912364 1.0473800  
## ApplauseStd OrderStd Challenge  
## 1 0.65016559 -1.707675 ewaste  
## 2 2.81349683 -1.675147 ewaste  
## 3 0.03207095 -1.642620 ewaste  
## 4 4.51325709 -1.610093 ewaste  
## 5 1.26826023 -1.577566 ewaste  
## 6 0.65016559 -1.545039 ewaste

## 'data.frame': 350 obs. of 15 variables:  
## $ Order : int 1 2 3 4 5 6 7 8 9 10 ...  
## $ Concept : Factor w/ 149 levels "concept\_1","concept\_10",..: 78 64 61 36 92 9 102 75 26 17 ...  
## $ Views : int 596 696 225 394 309 242 540 454 244 217 ...  
## $ Comments : int 19 26 3 28 15 15 15 16 9 10 ...  
## $ Applause : int 11 25 7 36 15 11 15 17 8 6 ...  
## $ Shortlist : Factor w/ 2 levels "Rejected","Shortlisted": 1 1 1 2 2 1 2 1 1 1 ...  
## $ Shortlist.views : int NA NA NA 581 471 NA 723 NA NA NA ...  
## $ Shortlist.comments: int NA NA NA 33 17 NA 19 NA NA NA ...  
## $ Shortlist.applause: int NA NA NA 45 21 NA 18 NA NA NA ...  
## $ Winner : int 0 0 0 0 1 0 0 0 0 0 ...  
## $ ViewsStd : num [1:350, 1] 2.0018 2.5534 -0.0446 0.8876 0.4187 ...  
## ..- attr(\*, "dimnames")=List of 2  
## .. ..$ : NULL  
## .. ..$ : NULL  
## $ CommentsStd : num [1:350, 1] 1.584 2.522 -0.562 2.79 1.047 ...  
## ..- attr(\*, "dimnames")=List of 2  
## .. ..$ : NULL  
## .. ..$ : NULL  
## $ ApplauseStd : num [1:350, 1] 0.6502 2.8135 0.0321 4.5133 1.2683 ...  
## ..- attr(\*, "dimnames")=List of 2  
## .. ..$ : NULL  
## .. ..$ : NULL  
## $ OrderStd : num [1:350, 1] -1.71 -1.68 -1.64 -1.61 -1.58 ...  
## ..- attr(\*, "dimnames")=List of 2  
## .. ..$ : NULL  
## .. ..$ : NULL  
## $ Challenge : chr "ewaste" "ewaste" "ewaste" "ewaste" ...

The most interesting thing to explore is the relationship between different variables and selection to shortlist, as this offers an opportunity to compare preferences of the crowd to expert decision.

# Descriptive statistics

## Views Comments Applause  
## nbr.val 350.000 350.000 350.000  
## nbr.null 0.000 35.000 1.000  
## nbr.na 0.000 0.000 0.000  
## min 13.000 0.000 0.000  
## max 1448.000 49.000 43.000  
## range 1435.000 49.000 43.000  
## sum 83524.000 2216.000 2313.000  
## median 185.000 4.000 5.000  
## mean 238.640 6.331 6.609  
## SE.mean 10.161 0.354 0.323  
## CI.mean.0.95 19.985 0.696 0.635  
## var 36138.334 43.827 36.514  
## std.dev 190.101 6.620 6.043  
## coef.var 0.797 1.046 0.914  
## skewness 1.587 2.275 2.302  
## skew.2SE 6.086 8.726 8.828  
## kurtosis 4.537 8.506 7.253  
## kurt.2SE 8.724 16.355 13.947  
## normtest.W 0.867 0.797 0.773  
## normtest.p 0.000 0.000 0.000

##   
## Rejected Shortlisted   
## 290 60

## 17 % of concepts have been selected on the shortlist and 4 % of concepts are winners.

Concepts have much more views than comments or applause, which makes sense as it is easier just to view a concept than do something about it. Interestingly statistics on comments and applause are very similar to each other. They have similar ranges, means, medians, standard deviations, sums and even skewness. 20 concepts per challenge have been selected on the shortlist. Acceptance rate is higher than on either Threadless or Quirky.

### Descriptive statistics by challenge

## openideo$Challenge: celebrate  
## Views Comments Applause Shortlist Shortlist.views  
## median 9.100000e+01 3.0000000 3.0000000 NA 0  
## mean 1.559684e+02 4.9157895 4.9789474 NA 0  
## SE.mean 1.572530e+01 0.5307273 0.4295832 NA 0  
## CI.mean.0.95 3.122296e+01 1.0537715 0.8529476 NA 0  
## var 2.349209e+04 26.7587906 17.5314670 NA 0  
## std.dev 1.532713e+02 5.1728900 4.1870595 NA 0  
## coef.var 9.827073e-01 1.0523010 0.8409527 NA NaN  
## Shortlist.comments Shortlist.applause Winner  
## median 0 0 0  
## mean 0 0 0  
## SE.mean 0 0 0  
## CI.mean.0.95 0 0 0  
## var 0 0 0  
## std.dev 0 0 0  
## coef.var NaN NaN NaN  
## --------------------------------------------------------   
## openideo$Challenge: ewaste  
## Views Comments Applause Shortlist Shortlist.views  
## median 166.500000 6.0000000 5.0000000 NA 6.105000e+02  
## mean 233.094340 7.1886792 6.7924528 NA 6.170000e+02  
## SE.mean 17.608514 0.7243813 0.6285677 NA 4.985474e+01  
## CI.mean.0.95 34.914430 1.4363144 1.2463337 NA 1.043472e+02  
## var 32866.333872 55.6212040 41.8803235 NA 4.970989e+04  
## std.dev 181.290744 7.4579625 6.4715009 NA 2.229572e+02  
## coef.var 0.777757 1.0374593 0.9527487 NA 3.613568e-01  
## Shortlist.comments Shortlist.applause Winner  
## median 18.0000000 16.5000000 0.00000000  
## mean 22.9000000 18.9000000 0.08490566  
## SE.mean 2.7046743 2.2229662 0.02720236  
## CI.mean.0.95 5.6609484 4.6527218 0.05393726  
## var 146.3052632 98.8315789 0.07843666  
## std.dev 12.0956713 9.9414073 0.28006545  
## coef.var 0.5281953 0.5260004 3.29854866  
## --------------------------------------------------------   
## openideo$Challenge: unemployment  
## Views Comments Applause Shortlist Shortlist.views  
## median 235.000000 5.0000000 5.0000000 NA 9.330000e+02  
## mean 295.295302 6.6241611 7.5167785 NA 1.194250e+03  
## SE.mean 16.248801 0.5495143 0.5353124 NA 1.549219e+02  
## CI.mean.0.95 32.109621 1.0859075 1.0578430 NA 3.242553e+02  
## var 39339.506802 44.9929258 42.6973517 NA 4.800159e+05  
## std.dev 198.341894 6.7076766 6.5343211 NA 6.928318e+02  
## coef.var 0.671673 1.0126077 0.8692981 NA 5.801397e-01  
## Shortlist.comments Shortlist.applause Winner  
## median 30.5000000 25.0000000 0.00000000  
## mean 30.9000000 28.3000000 0.04054054  
## SE.mean 4.2425787 3.1764760 0.01626670  
## CI.mean.0.95 8.8798192 6.6484407 0.03214679  
## var 359.9894737 201.8000000 0.03916161  
## std.dev 18.9733886 14.2056327 0.19789293  
## coef.var 0.6140255 0.5019658 4.88135887

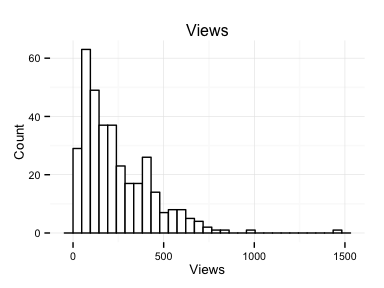
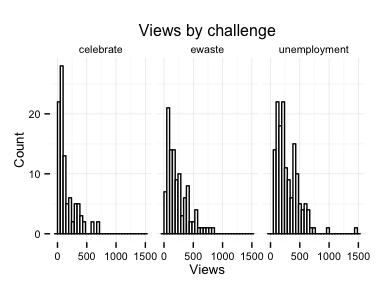
There is some variability between challenges. Celebrating innovative businesses challenge appears to have been the least popular of the three. That challenge is also missing data on what happened after the shortlist selection. This is not a problem, though, as here the focus is on events before the shortlist selection.

### Descriptive statistics by shortlist status

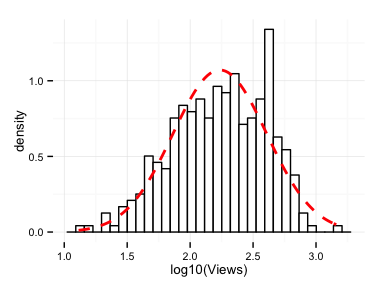
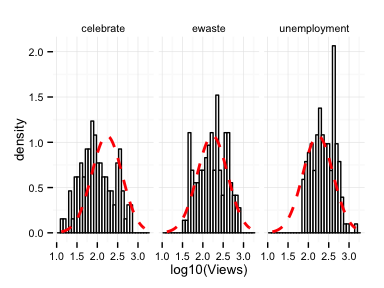
## openideo$Shortlist: Rejected  
## Views Comments Applause Shortlist Shortlist.views  
## median 151.500000 4.0000000 4.0000000 NA 0  
## mean 207.862069 5.2758621 5.6034483 NA 0  
## SE.mean 9.931698 0.3204059 0.2960060 NA 0  
## CI.mean.0.95 19.547631 0.6306250 0.5826010 NA 0  
## var 28605.198902 29.7713877 25.4096766 NA 0  
## std.dev 169.130715 5.4563163 5.0408012 NA 0  
## coef.var 0.813668 1.0342037 0.8995891 NA NaN  
## Shortlist.comments Shortlist.applause Winner  
## median 0 0 0  
## mean 0 0 0  
## SE.mean 0 0 0  
## CI.mean.0.95 0 0 0  
## var 0 0 0  
## std.dev 0 0 0  
## coef.var NaN NaN NaN  
## --------------------------------------------------------   
## openideo$Shortlist: Shortlisted  
## Views Comments Applause Shortlist Shortlist.views  
## median 3.710000e+02 10.5000000 9.000000 NA 5.800000e+02  
## mean 3.874000e+02 11.4333333 11.466667 NA 6.037500e+02  
## SE.mean 2.783574e+01 1.1652615 1.021096 NA 8.290691e+01  
## CI.mean.0.95 5.569918e+01 2.3316828 2.043208 NA 1.658963e+02  
## var 4.648970e+04 81.4700565 62.558192 NA 4.124133e+05  
## std.dev 2.156147e+02 9.0260765 7.909374 NA 6.421942e+02  
## coef.var 5.565687e-01 0.7894528 0.689771 NA 1.063676e+00  
## Shortlist.comments Shortlist.applause Winner  
## median 16.500000 13.5000000 0.00000000  
## mean 17.933333 15.7333333 0.25000000  
## SE.mean 2.371424 1.9902115 0.05637345  
## CI.mean.0.95 4.745209 3.9824039 0.11280302  
## var 337.419209 237.6564972 0.19067797  
## std.dev 18.368974 15.4161116 0.43666688  
## coef.var 1.024292 0.9798376 1.74666753

On average the shortlisted designs gather about the double the amount of views, comments and applause compared to rejected designs. There is also more variance in these statistics among the shortlisted designs than among the rejected designs.

# Distributions of variables

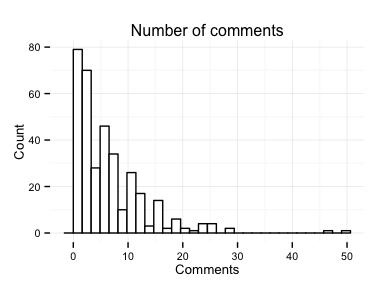
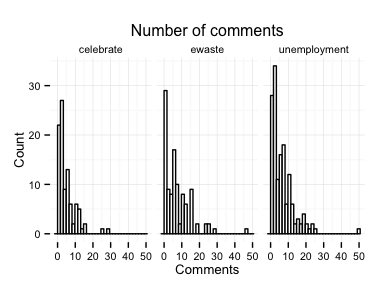
 

The distribution of views is right-skewed due to natural limit at zero views and no limit at the other end of the scale. The situation is the same both at the aggregate level and within individual challenges.

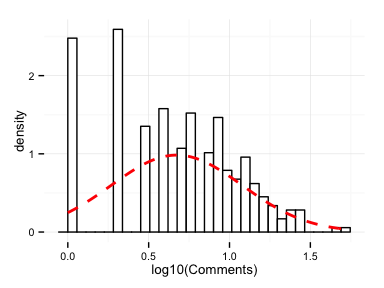
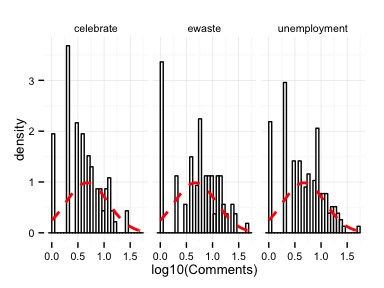
 

Log-transforming the number of views results in a near-normal distribution for the combined data set, but not for the individual challenges, perhaps due to smaller amount of data.

## Comments

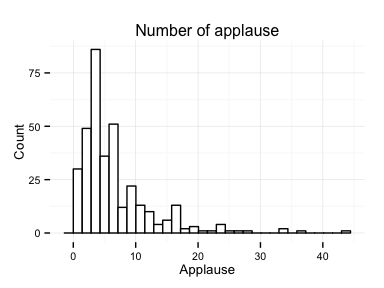
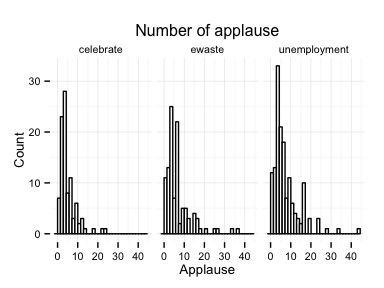
 

Situation is the same with the distribution of number of comments as with the number of views.

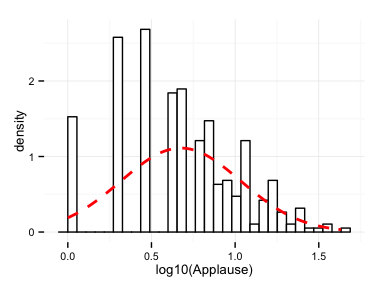
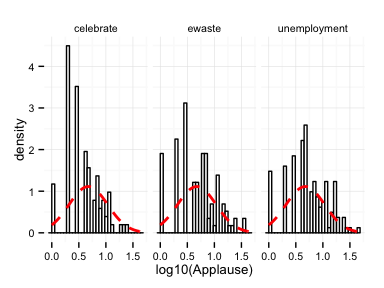
 

Log-transformation does not result in normal-looking distribution. This might be problematic in further analysis.

## Applause

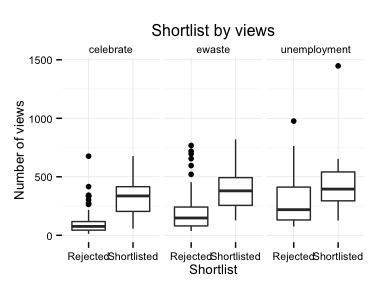
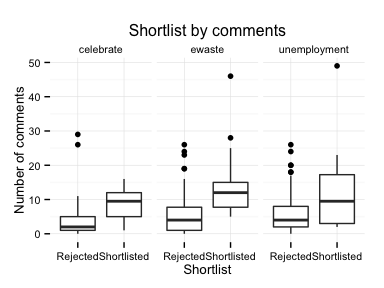
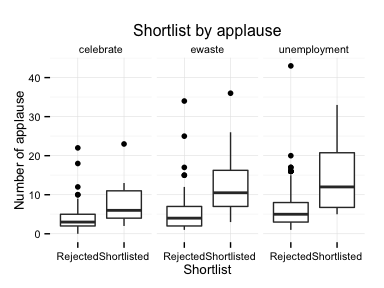
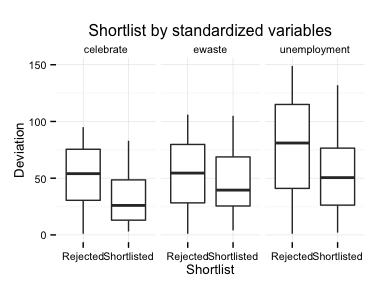
 

Distribution of number of applause repeats the familiar shape. In contrast to capped scores used at Threadless the measurements used at OpenIDEO are not limited, which leads to less helpful distributions.

Log-transformation does not result in normal distribution. This might be challenging in further analysis.

# Relationships between variables

In all challenges the shortlisted designs tend to have more views, comments and applause. There also appears to be a small tendency for shortlisted designs to have been submitted earlier in the challenge. Unfortunately data on exact submission dates is not available in the dataset.

# Logistic regression analysis

##   
## Call:  
## glm(formula = Shortlist ~ ApplauseStd + CommentsStd + ViewsStd +   
## OrderStd, family = "binomial", data = train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.6876 -0.5061 -0.3413 -0.2688 2.4829   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.9707060 0.2268870 -8.686 < 2e-16 \*\*\*  
## ApplauseStd 0.2982571 0.2309490 1.291 0.196551   
## CommentsStd 0.2860971 0.2352681 1.216 0.223967   
## ViewsStd 0.8870787 0.2344347 3.784 0.000154 \*\*\*  
## OrderStd -0.0004701 0.2238484 -0.002 0.998324   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 220.94 on 243 degrees of freedom  
## Residual deviance: 168.28 on 239 degrees of freedom  
## AIC: 178.28  
##   
## Number of Fisher Scoring iterations: 5

When trying to predict the shortlist status based on the standardized number of views, comments, applause and submission order, only the number of views is statistically significant. Therefore, a model containing only the number of views is used.

Correlations between variables:

## Order ViewsStd CommentsStd ApplauseStd  
## Order 1.0000000 -0.3576367 -0.3445924 -0.3486417  
## ViewsStd -0.3576367 1.0000000 0.6655894 0.6288069  
## CommentsStd -0.3445924 0.6655894 1.0000000 0.6653346  
## ApplauseStd -0.3486417 0.6288069 0.6653346 1.0000000

There is moderate correlation between variables, further suggesting the most of the information available is already contained in the most significant variable.

##   
## Call:  
## glm(formula = Shortlist ~ ViewsStd, family = "binomial", data = train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.1091 -0.4970 -0.3658 -0.2930 2.4415   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.9205 0.2165 -8.871 < 2e-16 \*\*\*  
## ViewsStd 1.1879 0.2032 5.846 5.04e-09 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 220.94 on 243 degrees of freedom  
## Residual deviance: 173.79 on 242 degrees of freedom  
## AIC: 177.79  
##   
## Number of Fisher Scoring iterations: 5

In the simpler model the number of views is statistically significant predictor of shortlist status.

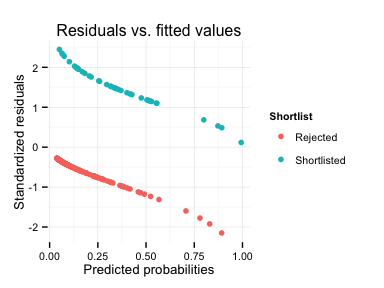
## Analysis of Deviance Table  
##   
## Model 1: Shortlist ~ ApplauseStd + CommentsStd + ViewsStd + OrderStd  
## Model 2: Shortlist ~ ViewsStd  
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)  
## 1 239 168.28   
## 2 242 173.79 -3 -5.5142 0.1378

The more complex model has slightly smaller residual deviance than the simple model, but because other variables were far from being statistically significant, there is a good change that the more complex model is just fitting the noise. The performance of models is not very different overall.

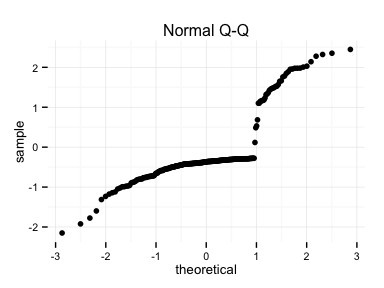
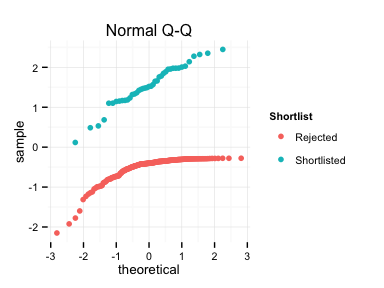
## modelChi 47.15083   
## chidf 1   
## chisq.prob 6.572853e-12   
## Pseudo R^2 for logistic regression  
## Hosmer and Lemeshow R^2 0.213   
## Cox and Snell R^2 0.176   
## Nagelkerke R^2 0.295   
## Odds ratios:  
## (Intercept) ViewsStd   
## 0.147 3.280   
## Confidence intervals:  
## 0.5 % 99.5 %  
## (Intercept) 0.080 0.246  
## ViewsStd 2.017 5.792

The model is clearly better than random baseline model, but the effect size is only moderate, as estimated by the pseudo R^2 statistics. On average, a concept gaining one standard deviation more views in a challenge increases the odds of that concept being selected on the shortlist by a factor of about 3.

# Model diagnostics



The model makes largest mistakes by missing the concepts that get on the shortlist, similarly to problems with Threadless model. Here the effect is smaller though, presumably due to larger ratio of submissions being selected.

QQ-plots show the same issue. Because selected concepts are relatively rare, the model tends to predict that nothing is selected, but suffers only small punishments for the mistakes it makes. The predictive performance might be almost trivial.

## 8 observations have residuals larger than 2 standard deviations.  
## 23 observations have leverage more than 2 times larger than the average.

Observations with large residuals:

##   
## Rejected Shortlisted  
## FALSE 202 34  
## TRUE 1 7

Most of the observations with large residuals are shortlisted concepts.

Observations with large leverage:

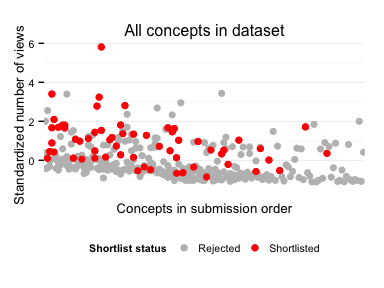
##   
## Rejected Shortlisted  
## FALSE 193 28  
## TRUE 10 13

With large leverage the pattern is less clear. The shortlisted concepts are still over represented, but not as clearly as with the large residuals.

## train$Shortlist[train$large.leverage]: Rejected  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.335 1.538 1.770 2.085 2.600 3.393   
## --------------------------------------------------------   
## train$Shortlist[train$large.leverage]: Shortlisted  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.338 1.632 1.693 1.971 1.803 3.399

## train$Shortlist[!train$large.leverage]: Rejected  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -1.11100 -0.83490 -0.50870 -0.35810 -0.04465 1.28400   
## --------------------------------------------------------   
## train$Shortlist[!train$large.leverage]: Shortlisted  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -0.84850 0.09187 0.41950 0.58540 0.99190 5.81200

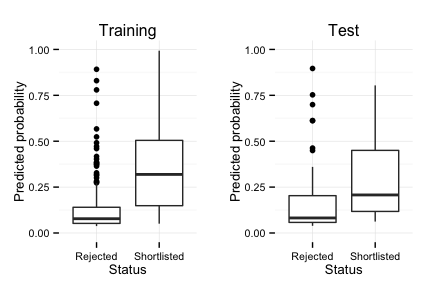
It is not obvious what causes some observations to have a large leverage.



The concepts with least views are rarely selected on the shortlist, but after that the pattern is not clear. Concepts with more than average number of views seem to have quite equal changes of getting on the shortlist. This figure also shows the small trend of older concepts having slightly more views and being selected more often on the shortlist.

# Validation

To validate the model the predicted probabilities between training and test sets are compared.



The shortlisted concepts in the test set tend to have higher predicted probabilities than rejected designs, but the model is not accurate enough to discriminate the concepts.